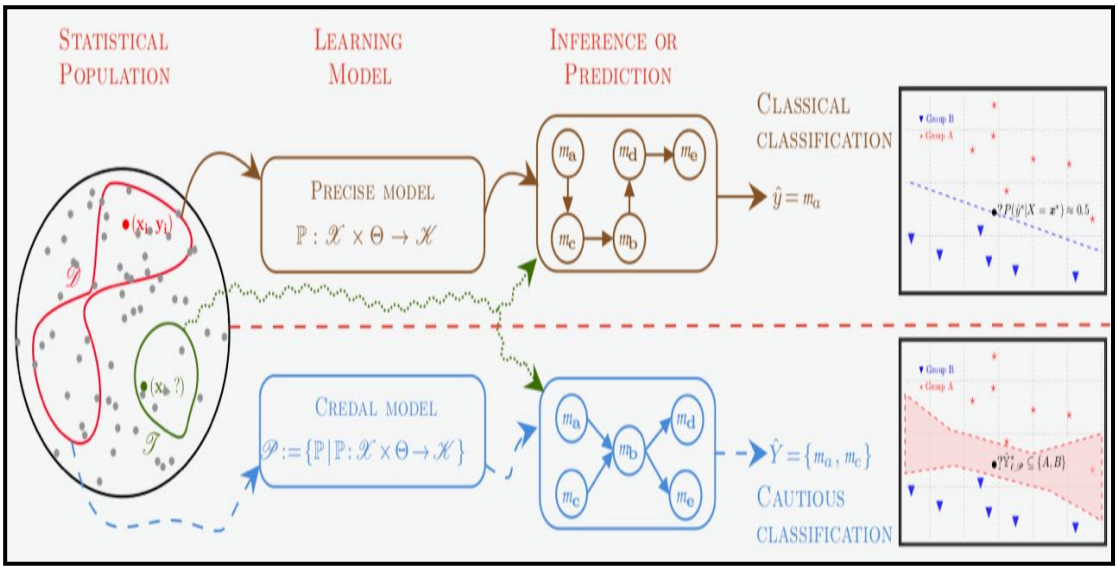
Distributionally robust, skeptical inferences in supervised classification using imprecise probabilities

Yonatan Carlos Carranza Alarcón

Par **Yonatan Carlos CARRANZA ALARCÓN**

Distributionally robust, skeptical inferences in supervised classification using imprecise probabilities

Thèse présentée
pour l'obtention du grade
de Docteur de l'UTC



Soutenue le 8 décembre 2020
Spécialité : Informatique : Unité de recherche Heudyasic (UMR-7253)

D2567

University of Technology of Compiègne

Doctoral Thesis

Distributionally robust, skeptical inferences in supervised classification using imprecise probabilities

Spécialité : Informatique

Author:
Yonatan Carlos Carranza Alarcón

Supervisor:
Dr. Sébastien Destercke

Jury:

Assoc. Prof.	Willem WAEGEMAN,	Ghent University	Reviewer
Prof.	Ines COUSO,	University of Oviedo,	Reviewer
Prof.	Jesse READ,	École Polytechnique,	Reviewer
Prof.	Thierry DENOEUX,	University of Technology of Compiègne,	Examiner
Assoc. Prof.	Benjamin QUOST,	University of Technology of Compiègne,	Examiner
Prof.	Frederic PICHON,	Artois University,	Examiner
Dr.	Nicolas VERZELEN,	National Institute of Agricultural Research,	Examiner

A thesis submitted in fulfillment of the requirements for the degree of Doctor in the

CID (Connaissances, Incertitudes, Données)
Team Heudiasyc Laboratory

December 08, 2020

"Er zijn mannen die op een dag vechten en goed zijn.
Er zijn anderen die een jaar vechten en beter zijn.
Er zijn mensen die vele jaren vechten, en ze zijn erg goed.
Maar er zijn mensen die hun hele leven vechten: ***dat zijn de essenties.****"*

"Il y a des hommes qui luttent un jour et ils sont bons,
d'autres luttent un an et ils sont meilleurs,
il y a ceux qui luttent pendant de nombreuses années et ils sont très bons,
mais il y a ceux qui luttent toute leur vie et ceux-là sont ***les indispensables.****"*

"There are men that fight one day and are good,
others fight one year and they're better,
and there are those who fight many years and are very good,
but there are the ones who fight their whole lives and those are ***the indispensable one****"*

"Hay hombres que luchan un día y son buenos.
Hay otros que luchan un año y son mejores.
Hay quienes luchan muchos años y son muy buenos.
Pero hay los que luchan toda la vida. Esos son los ***imprescindibles.****"*

Bertolt Brecht

Remerciements

C'est avec une profonde tristesse que je m'apprête à écrire les dernières lignes de ce manuscrit de thèse. Un chapitre de ma vie vient de s'achever et un nouveau commence. Je voudrais donc clore celui-ci en remerciant toutes les personnes qui ont, de près ou de loin, contribué d'une façon ou d'une autre à cet aboutissement.

Sébastien, si vous avez l'occasion de lire ces lignes, vous comprendrez qu'il n'y a pas de mots ou de beaux gestes qui puissent exprimer mon infinie gratitude d'avoir pu bénéficier de votre confiance et de votre inestimable temps. Cela m'a énormément touché et j'espère avoir été à la hauteur. Trois ans se sont déjà écoulés, et je n'oublierai jamais vos premiers conseils. Je me permets d'en citer un :

"Chaque personne a son propre rythme d'apprentissage, et il faut travailler sur des sujets de recherche qui nous plaisent..."

Sans ce conseil, parmi tant d'autres, je n'aurais jamais pu arriver là où je suis aujourd'hui. Vous avez toujours été à l'écoute, très réactif face à mes nombreuses questions tant absurdes que pertinentes, et surtout très soigneux à chaque relecture de mes travaux. Avec votre bonne humeur et votre patience, vous m'avez donné la liberté de poursuivre toutes les directions de recherche que j'envisageais —- aussi délirantes qu'elles puissent être. C'est pour cela et pour de nombreuses autres raisons que je ne cesserai jamais de dire que vous êtes le meilleur encadrant dont un.e doctorant.e puisse rêver.

Je remercie également les membres de mon jury de thèse pour leur lecture attentive et leurs remarques si judicieuses qui aboutiront indéniablement à de nouveaux travaux scientifiques. Ce fut pour moi un grand honneur de vous avoir à ma soutenance, même si pour quelques-uns.e.s, c'était en visioconférence, ainsi que d'avoir été évalué par des chercheurs et chercheuses de grande qualité et réputation.

Je tiens également à remercier mes collègues du laboratoire Heudiasyc, et surtout très chaleureusement Jean-Benoist Leger, Vu-Linh Nguyen, Gabriel Frisch, Soundouss Messoudi, Yves Grandvalet et Mylène Masson avec qui j'ai pu partager des discussions scientifiques si intéressantes et enrichissantes. Je ne voudrais certainement pas oublier tous ceux avec qui j'ai aussi eu d'agréables conversations; Youcef Amarouche, Kandi Mohamed-Ali, Sana Benhamaid, Anthony Welte, Tan-Nhu Nguyen ... ainsi que tous les permanents et les non-permanents que j'ai pu croiser et qui ont contribué à mon épanouissement personnel et professionnel.

Il va de soi que je remercie mes parents et ma famille, car ils sont et seront toujours dans mon coeur — quoi que puisse signifier rationnellement cette phrase qui reste pour le moment incompréhensible de la perspective de mes équations. Je tiens en particulier à remercier ma mère et mon grand frère qui m'ont poussé à faire des études d'ingénieur en informatique qui m'ont ensuite permis de financer mes études de master à l'étranger dans le domaine de mes rêves, les mathématiques

appliquées et, finalement, d'aboutir à une thèse hybride qui m'a beaucoup plue.

L'amitié compte énormément pour moi, notamment car il s'agit d'un sentiment abstrait sans aucune incertitude et comportant donc zéro imprécision. Bien que l'on soit parfois si éloigné, je voudrais vous remercier, mes cher.e.s ami.e.s proches; Jesus Porras, Leslie Guerra, Alfonso Paredes, Alan Angeles, César Melgarejo, Joseph Mena, David Quispe, Aziz Ibazizene, Patricia Fernández-Sánchez, Alexander Contreras... qui ont été présents dans ces moments si difficiles de doutes existentiels et qui m'ont soutenu psychologiquement avec une douce et juste clarté.

La liste de personnes à remercier est si longue qu'il faudrait que je fasse une deuxième thèse pour citer tout le monde. Voici, malgré tout, quelques-uns.e.s dont je me souviens : Sébastien Bernis, Xavier Bry, Benoîte de Saporta, Servajean Maximilien, Louise Baschet ... et si j'ai oublié quelqu'un.e, je tiens à m'en excuser :)....

UNIVERSITY OF TECHNOLOGY OF COMPIÈGNE

Abstract

CID (Connaissances, Incertitudes, Données)
Team Heudiasyc Laboratory

Doctor of Philosophy

Distributionally robust, skeptical inferences in supervised classification using imprecise probabilities

by Yonatan Carlos Carranza Alarcón

Decision makers are often faced with making single hard decisions, without having any knowledge of the amount of uncertainties contained in them, and taking the risk of making damaging, if not dramatic, mistakes. In such situations, where the uncertainty is higher due to imperfect information, it may be useful to provide set-valued but more reliable decisions.

This works thus focuses on making distributionally robust, skeptical inferences (or decisions) in supervised classification problems using imprecise probabilities. By distributionally robust, we mean that we consider a set of possible probability distributions, i.e. imprecise probabilities, and by skeptical we understand that we consider as valid only those inferences that are true for every distribution within this set. Specifically, we focus on extending the Gaussian discriminant analysis and multi-label classification approaches to the imprecise probabilistic setting.

Regarding to Gaussian discriminant analysis, we extend it by proposing a new imprecise classifier, considering the imprecision as part of its basic axioms, based on robust Bayesian analysis and near-ignorance priors. By including an imprecise component in the model, our proposal highlights those hard instances on which the precise model makes mistakes in order to provide cautious decisions in the form of set-valued class, instead.

Regarding to multi-label classification, we first focus on reducing the time complexity of making a cautious decision over its output space of exponential size by providing theoretical justifications and efficient algorithms applied to the Hamming loss. Relaxing the assumption of independence on labels, we obtain partial decisions, i.e. not classifying at all over some labels, which generalize the binary relevance approach by using imprecise marginal distributions. Secondly, we extend the classifier-chains approach by proposing two different strategies to handle imprecise probability estimates, and a new dynamic, context-dependent label ordering which dynamically selects the labels with low uncertainty as the chain moves forwards.

Keyworkds: imprecise probabilities, supervised classification, multi-label classification, multiclass classification, uncertainty

Contents

List of Notations

Symbols	Description
Data related notations	
$\mathscr{X} = \mathbb{R}^p$	Input space of dimension p
$\mathscr{K} = \{m_1, ..., m_K\}$	Output space of size K
$\mathscr{D} \subset \mathscr{X} \times \mathscr{K}$	Training data set
$\mathscr{T} \subset \mathscr{X} \times \mathscr{K}$	Test data set
$\mathbf{x} = (x^1, \ldots, x^p)^\top$	New unlabeled instance to predict
$\hat{y}$	Precise output prediction.
$N := \lvert\mathscr{D}\rvert$	Number of training instances.
Probability and statistics	
X	Multivariate random variable
Y	Discrete random variable
$\mathbb{P}$	Probability distribution
$\mathbb{P}_{Y\lvert X}, \mathbb{P}_{X\lvert Y}$	Conditional probability distributions
$P(A)$	Unconditional probability of event A
$P_{\mathbf{x}}(Y = m_k) := P(Y = m_k \lvert X = \mathbf{x})$	Conditional probability of m_k given $\mathbf{x}$
Imprecise probabilities	
$\mathscr{P}$	Set of probability distributions (Credal set)
$\underline{P}(A)$	Lower probability of event A
$\overline{P}(A)$	Upper probability of event A
$\underline{\mathbb{E}}_{\mathscr{P}}$	Lower expectation operator under $\mathscr{P}$
$\overline{\mathbb{E}}_{\mathscr{P}}$	Upper expectation operator under $\mathscr{P}$
$\hat{\mathbb{Y}}$	Set-valued output predictions.
Decision theory	
$\mathcal{R}(\cdot)$	Theoretical risk
$\mathscr{R}(\cdot)$	Empirical risk
Θ	Parameter space
$\mathscr{F} = \{\varphi := \mathbb{P} \mid \mathbb{P} : \mathscr{X} \times \Theta \to \mathscr{K}\}$	Learning probabilistic models
$\succ$	Partial order operator
$\succ\prec_*$	Incomparable operator
$\sim$	Indifferent operator

Loss functions

$\ell(\cdot,\cdot)$	General loss function
$\ell_{0/1}(y,\hat{y})$	*Zero-one* loss function
ℓ_H	Hamming loss function
ℓ_{F_β}	F-measure loss function

Discriminant Analysis

n_k	Number of observations of label m_k
$(\boldsymbol{x}_{i,k}, y_{i,k})_{i=1}^{n_k} = \left\{(\boldsymbol{x}_{1,k}, y_{1,k}), \ldots, (\boldsymbol{x}_{n_k,k}, y_{n_k,k})\right\}$	Observations of label m_k.
$\overline{\boldsymbol{x}}_k = \frac{1}{n_k}\sum_{i=1}^{n_k} \boldsymbol{x}_{i,k}$	Empirical mean of label m_k
$\widehat{\sigma}^j_{m_k} = \frac{1}{n_k-1}\sum_{i=1}^{n_k}(x^j_{i,k} - \overline{x}^j_k)^2, \forall j \in \{1,\ldots,p\}$	Empirical variance of label m_k
$\widehat{S}_{m_k} = \frac{1}{N-n_k}\sum_{i=1}^{n_k}(\boldsymbol{x}_{i,k} - \overline{\boldsymbol{x}}_k)(\boldsymbol{x}_{i,k} - \overline{\boldsymbol{x}}_k)^T$	Empirical covariance matrice of label m_k
$\widehat{S} = \frac{1}{(N-K)}\sum_{k=1}^{K}\sum_{i=1}^{n_k}(\boldsymbol{x}_{i,k} - \overline{\boldsymbol{x}}_k)(\boldsymbol{x}_{i,k} - \overline{\boldsymbol{x}}_k)^T$	Empirical total covariance matrix
$\widehat{\pi}_y = \left\{ \widehat{\pi}_{y=m_k} \middle\vert \widehat{\pi}_{y=m_k} = \frac{n_k}{N}, \sum_{m_k \in \mathcal{K}} \widehat{\pi}_{y=m_k} = 1 \right\}$	Empirical marginal distribution $\mathbb{P}_Y$.

Multi-label classification

$\mathscr{Y}$	M-dimensional binary or boolean space
$\mathscr{Y}^*$	M-dimensional partial binary space
$\mathscr{X}^*_{j-1}$	Augmented input space
$\boldsymbol{Y}$	Multivariate random binary variable
$\mathscr{I}$	A subset of indices of $[\![m]\!]$
$\boldsymbol{Y}_\mathscr{I}$	The marginals of $\boldsymbol{Y}$ over indices $\mathscr{I}$
$\boldsymbol{y}$	Binary vector of $m \times 1$
$\boldsymbol{y}_\mathscr{I}$	Binary vector over indices $\mathscr{I}$
$\boldsymbol{y}^*$	Partial binary vector
$\mathscr{I}_\mathcal{R}$	Set of indices of relevant labels
$\mathscr{I}_\mathcal{I}$	Set of indices of irrelevant labels
$\mathscr{I}_\mathcal{A}$	Set of indices of abstained labels

Mathematic and other notations

$(\cdot)^T$	Transpose Operator
$\|\cdot\|$	Euclidian norm
$[\![j]\!]$	A set of the first j integers

To my mother

Chapter 1

Introduction

"Begin at the beginning", the King said gravely, "and go on till you come to the end: then stop."

— Lewis Carroll, Alice in Wonderland

Contents

The (imprecise) cautious[1] classification task[2] is a relatively new phenomenon in machine learning, dating back at least to the late nineties, with the referred work of [Zaffalon, 1999]. This approach extends the classical one by allowing us to describe our uncertainty through a set of probability distributions rather than a single one. Besides, it also aims at highlighting those hard cases for which information is insufficient to isolate a single reliable solution (or prediction), proposing then a subset of possible solutions.

Walley (1996) early illustrated an e.g. of cautious inference in a randomized clinical trial, in which a new patient may be classified as life or death or ineligible according to the applied treatment.

Thus, this research work focuses on extending the classical-classification approach to the imprecise probabilistic setting [De Finetti, 1937; Walley, 1991; Troffaes et al., 2014], in order to detect and palliate those unreliable *single* decisions made by it and to propose instead potential set-valued decisions including all those decisions that are not dominated for every probability distribution within the set of probability distributions. We also investigate the robustness of these set-valued

[1] Cautious and imprecise are used throughout this thesis interchangeably.
[2] Throughout this thesis, we will use "supervised learning" term to refer only to classification task.

decisions made by this new extension, or *imprecise supervised-learning approach*, when faced with noisy and missing information (be it on the input or output component), showing that including the imprecision in our model produces a gain in the decision-making process. Such set-valued decisions can be useful in sensitive applications where it can be disastrous to decide wrongly.

In this chapter, we shall introduce and discuss the benefits and drawbacks of the precise and imprecise approaches in general, and then delve into more specific concepts in other chapters[3]. At the end of the chapter, precisely in the two last sections 1.4 and 1.5, I summarize my research work performed during the thesis and provide a self-learning guide allowing readers to focus on topics more related to their interest.

1.1 STATISTICAL LEARNING THEORY

Any learning process is based on knowledge acquisition, be it implicit, explicit, or both. That is how it happens in humans and not too differently in computers, yet certainly with a higher focus on a specific task, in which it learns to generalize repetitive and similar patterns of a well-framed and well-specific experiment, e.g. classification of images.

Laplace (1773) and Gauss (1810) previously work about the uncertainty on the parameter of a model, so its estimate.

In the theory of statistical estimation , this process was deeply studied in the twenties and thirties by [Fisher, 1925; Neyman, 1937; Wald, 1939], introducing a theoretical background that matured in the 50's with concepts of statistical decision theory [Wald, 1950] (inspired by developments in game theory [Neumann et al., 1947]). During the 60-70's V. N. Vapnik and A. Ya. Chervonenkis introduced the *statistical learning theory* (a.k.a. pattern recognition learning theory or inductive learning principle) which can be stated in terms of decision-making, and it was widely disseminated during the nineties and became popular in the machine learning community.

Supervisor is the one that labels or identifies a response $y \in \mathscr{K}$ for every input $x \in \mathscr{X}$, where $\mathscr{X}$ and $\mathscr{K}$ are the input and output spaces.

In formal terms, it aims at estimating or learning, on the basis of empirical evidence (or data), an approximation of the supervisor's response through a function $\varphi \in \mathscr{F}$ (a.k.a. *learning machine*) of a given functional space. To do that, [Vapnik et al., 1974] introduced the risk minimization inductive principle, which aims to measure the discrepancy or loss incurred between the supervisor's response (or true response) and the learning machine's response φ (predicting response) by computing the expected value of a specified loss function $\ell(\cdot,\cdot) : \mathscr{K} \times \mathscr{K} \to \mathbb{R}$ penalising every bad decision, as follows

$$\mathcal{R}(\varphi) = \mathbb{E}_{X\times Y}\left[\ell(Y,\varphi))\right] = \int_{\mathscr{X}\times\mathscr{K}} \ell(y,\varphi(x))d\mathbb{P}(x,y), \tag{1.1}$$

[3] Someone with knowledge in supervised learning approach can directly go to Section 1.3.

where $\mathbb{P}$ is a unknown theoretical probability measure on a measurable space $\mathscr{X} \times \mathscr{K}$ (from which are drawn independently the values of (x, y), under identical distributional conditions, i.e. i.i.d) and $\mathbb{E}$ is the expected value of $\ell(\cdot, \cdot)$ understood as an abstract Lebesgue integral with respect to the measure $\mathbb{P}$.

Wald (1950) named ℓ as rule *to make a decision.*

Therefore the "optimal" (or potential approximation) solution $\hat{\varphi}$ can be obtained by minimizing $\mathcal{R}(\varphi)$ over $\mathscr{F}$

$$\mathcal{R}(\hat{\varphi}) = \inf_{\varphi \in \mathscr{F}} \mathcal{R}(\varphi). \tag{1.2}$$

Under canonical loss functions, such as quadratic and zero-one loss, and assuming that $\ell(\cdot, \cdot)$ is defined instance-wise, we can obtain explicit reductions of Equation (1.2) (c.f. Table 1.1). Besides, if we knew $\mathbb{P}$ we could easily deduce the lowest possible mean squared error and missclassification probability of the desirable risk function. Unfortunately, and for obvious reasons, these last are not implementable since $\mathbb{P}$ is unknown.

The quadratic and zero-one loss function are of classic use in the statistic for its mathematical tractability, introduced by Gauss (1810) and Neyman-Pearson, respectively.

	Regression problem	Classification problem
Risk minimizer	$\int_{\mathscr{X} \times \mathscr{K}} (y - \varphi(x))^2 d\mathbb{P}$	$\int_{\mathscr{X}} \sum_{y \in \mathscr{K}} \mathbb{1}_{(y \neq \varphi(x))} \, d\mathbb{P}$
Explicit solutions	$\hat{\varphi}(x) = \mathbb{E}[Y \mid X = x]$	$\hat{\varphi}(x) = \underset{\varphi(x) \in \mathscr{K}}{\arg\min} \; \mathbb{E}[\mathbb{1}_{(Y \neq \varphi(x))}]$

Table 1.1: Explicit reductions of the risk minimization [Friedman et al., 2001, Eq. 2.13, 2.23].

This means that in most cases, it is impossible to obtain an explicit solution to Equation (1.1), and in practice we use the empirical risk minimization (ERM) principle as follows

$$\mathscr{R}(\varphi) = \frac{1}{N} \sum_{i=1}^{N} \ell(y_i, \varphi(x_i)) \tag{1.3}$$

which has strong theoretical justifications [Vapnik et al., 1982] such as the uniform convergence towards the theoretical risk of the Equation (1.1). Thus, the "optimal" model is the one minimizing the Equation (1.3).

Even if this principle is theoretically seducing to find an "optimal" solution, it has some shortcomings in practice, such as overfitting [Vapnik, 1995], ill-posedness [Hadamard, 1902], sensitivity to data (non-continuity), need for a larger sample size without any guarantee of getting a small risk error, an input dimensionality $\mathscr{X}$ not larger than the empirical sample size (i.e. $N >>> p$), and so on. Furthermore, even if our assumptions about $\mathscr{F}$ and ℓ were right, estimating a *precise* "optimal" solution $\hat{\varphi}$ would be like having a very idealistic thought since the empirical evidence is seldom a

truly, sufficient i.i.d. representative of the population of interest. In which case, it may be better to perform a distributionally robust estimation (or inference), for instance using a neighborhood around $\hat{\varphi}$ [Kuhn et al., 2019; Chen et al., 2018].

In the same vein, using a subjective probability approach [De Finetti, 1937] which describes our ignorance or partial knowledge by means of sets of probabilities, we can mitigate several shortcomings of the ERM principle. For instance, [Mantas et al., 2014] proposes a robust classifier sensitive to noise data (corrupted data).

Since the ERM may lead to poor results when empirical evidence is limited or of poor quality (partial, noisy, ...), this research work focuses on proposing robust cautious classification models on structured outputs (e.g. multi-label and multi-class problems) that are sensitive to data quantity and quality. In what follows, we introduce preliminares about the classical classification based on the ERM principle, and leave the cautious classification for Chapter 2.

1.2 SUPERVISED LEARNING APPROACH

In statistics, the classification task was quietly introduced at the beginning of the 20th century — roughly at the same time as the theory of statistical estimation and hypothesis testing— aiming to minimize the risk of making a wrong decision. Predictive inference[4] in a probabilistic modelling paradigm (PMP)[5] can be divided as a two-step process[6]

1. LEARNING MODEL.- to make assumptions about the parametric form of the unknown probability measure $\mathbb{P}$ and then estimating its parameters on the basis of empirical data (here called training data set), and

"Decision rule" was one of the first names assigned to loss function, by Wald.

2. INFERENCE OR PREDICTION.- to construct an optimal decision rule from the estimated probability measure and a given missclassification cost.

An illustration of the previous setting can be seen at the top of the Figure 1.1. We can also note that in the case of a cautious classification task

[4] Predictive inference is an approach to statistical inference that emphasizes the prediction of future observations (or unobserved observations) of a given population on the basis of past observations of the same population [Salmon, 1957; De Finetti, 1970; Aitchison et al., 1975; Hinkley, 1979; Clarke et al., 2018].

[5] We prefer to calling it PMP at the frequentist or Bayesian statistical decisional paradigm as a way of distinguishing on other non-probabilistic approaches (like score methods).

[6] In a Bayesian predictive inference, we should introduce a intermediary step, i.e. making assumptions on the a priori probability distribution, and then use the posterior distribution estimated instead in the inference step [Berger, 1985, §2.4.4].

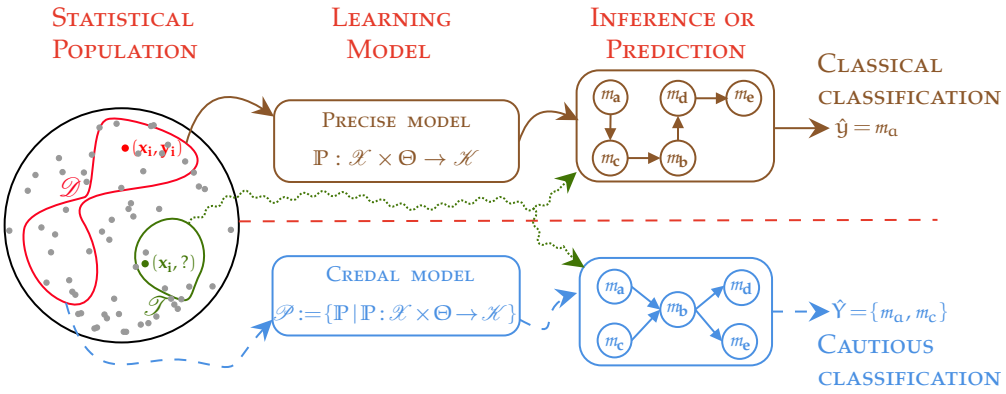


Figure 1.1: Statistical learning in imprecise and precise approach.

(at the bottom) the learning and decision-making process are quite similar, except for the fact that we shall use concepts extending those of the classical setting, i.e., sets of probabilities.

In machine learning (ML), and to the best of our knowledge, this abstract and generic process has been given the name of Decision-Theoretic approach (DTA) by [Lewis, 1995], or more recently with the name of Population Utility (PU) by [Dembczyński et al., 2017], which consist firstly to fit a probabilistic model at training time (learning model step) and then use it in the inference time (inference step). Another paradigm, not too far away from the scheme presented in Figure 1.1, is the *empirical utility maximazation* (EUM) approach [Ye et al., 2012; Dembczyński et al., 2017] that learns a *score function* in the learning step instead and then selects a threshold for the score function to minimize a loss function in the inference step.

Without loss of generality, the practitioners in ML name the scheme of the Figure 1.1 as the *inductive principle* (or ERP), if and only if the metric of evaluation (or loss function) is decomposable on a set of i.i.d. test samples [Joachims, 2005, §2]. This will be the case here, so we will use the ERP principle.

Let $\mathscr{D} = \{(\mathbf{x}_i, y_i) | i = 1, \ldots, N\}$ be a training data set (empirical evidence) issued from $\mathscr{X} \times \mathscr{K}$, such that $\mathbf{x}_i \in \mathscr{X}$ are regressors or features (input space) and $y \in \mathscr{K}$ is the response variable or class (output space). We denote n_k the number of observations that belong to the label[7] m_k, and so $N = \sum_{k=1}^{K} n_k$. Thus, the goal of classical classification is to build a predictive model $\varphi : \mathscr{X} \to \mathscr{K}$ that predicts a label $m_k \in \mathscr{K}$ given a new unlabelled instance $(\mathbf{x}, \cdot) \notin \mathscr{D}$.

By contrast with PMP, but not quite different, the ERM principle –in the classification setting– develops these steps as follow; *learning step* consists in determining the optimal model $\hat{\varphi} \in \mathscr{F}$ which partitions the underlying

[7] Label and class are used throughout this thesis interchangeably when these do not cause ambiguity, e.g. in the case of multi-label problem.

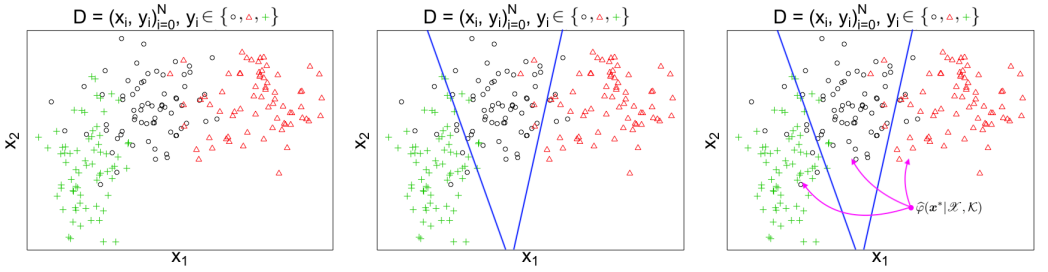
abstract (vector) space, where lives labelled observations $(x_i, y_i)_{i=1}^N$ (i.e. training data) generated from an unknown joint probability distribution $\mathbb{P}_{X,Y}$, into as many disjoint subsets as there are classes (see figure 1.2 for an illustration).

This function space $\mathscr{F}$ can be represented; e.g. by a set of finite set of hyperplanes as it is the case in the model of support vector machine (SVM), or by a family of probability distributions with unknown parameters, for instance

In a PM, we can obtain these estimates e.g. by using the maximum likelihood estimation principle.

$$\mathscr{F} := \{\varphi := \mathbb{P} | \mathbb{P} \sim \mathcal{N}(\mu, \Sigma), (\mu, \Sigma) \in (\mathbb{R}, \mathbb{R}^2)\}. \tag{1.4}$$

In this research work, we focus on this last kind of functional spaces.



(a) Getting Training Data **(b)** Learning model ϕ **(c)** Prediction of unlabeled instance x

Figure 1.2: *Learning model steps.* Figure (a) shows the initial training data, from which are induced the boundaries defining the decision function (b), then used to perform the predictions (c).

Under a set of hyper-parameters, the set of admissible "optimal" models must decide which one best classifies unlabelled instances on a hold-out set of test data

After getting an "optimal" model $\hat{\varphi}$, we must decide what is the label of a new unlabelled instance $(x, \cdot)$. This latter *decision step* can be handled by minimising the risk of getting misclassifications (c.f. Equation (1.1)), and can formally be defined as follows.

Definition 1 (Risk minimizing [Friedman et al., 2001, §2.4]) *Given a general loss function $\ell(\cdot, \cdot)$, the optimal model is defined as the one minimizing the average loss of getting missclassification.*

$$\hat{\varphi} := \operatorname{argmin}_{\varphi(X) \in \mathscr{F}} \mathbb{E}_{X \times Y} [\ell(Y, \varphi(X))] \tag{1.5}$$

If the loss function is defined instance-wise, then, Equation (1.5) can also be expressed as the minimization of conditional expectation [Friedman et al., 2001, eq. 2.21]:

$$\hat{\varphi} := \arg\min_{y \in \mathscr{K}} \mathbb{E}_{Y|X} [\ell(Y, y)] \tag{1.6}$$

Classical accuracy corresponds to a *zero-one* loss function, where all missclassifications are penalised identically, i.e. $\ell_{0/1}(y, \hat{y})$ is equal to 1 if y and $\hat{y}$ are different and 0 otherwise. Therefore, given $\ell_{0/1}$, we can reformulate the risk minimization as the well-known *Bayes classifier*, which would

choose the learning model maximizing the conditional probability (a.k.a. maximum a posterior (MAP) probability, also see the Table 1.1) given a new unlabeled instance $\mathbf{x}$:

$$\hat{\varphi}(\mathbf{x}) = \operatorname{argmax}_{m_k \in \mathscr{K}} P(Y = m_k | X = \mathbf{x}) \tag{1.7}$$

Hence, the main task is to estimate the conditional distribution $\mathbb{P}_{Y|X}$, from which can be obtained the optimal decision.

This last classifier model is called *precise* since it performs pointwise predictions (or precise estimations) in the form of single class labels, even in extreme cases, regardless of the available information we have about an instance. The reliability of this *precise* prediction (or *single* decision) may depend heavily on prior beliefs (e.g. assumptions made by data analysts, such as asymptotically unbiased estimators) and the nature of training data sets (e.g. in small amounts [Kitchin et al., 2015; Dalton et al., 2015] and/or with high degree of *uncertainty*[8]), both will be referred as *imperfect information*[9].

One of our motivations is precisely to investigate means to make robust and cautious predictions. Thus, in the next chapter, we shall present an extension of this decision-making framework resulting sometimes in partial decisions in form of set-valued solutions (see at the bottom of the Figure 1.1) using imprecise probabilities.

1.3 SUBJECTIVE PROBABILITY APPROACH

De Finetti was one of precursors in the subjective probability (SP) approach with his famous work entitled "La prévision: ses lois logiques, ses sources subjectives". In the later years, Walley adopted this approach and extended de Finetti's point of view as follows

There have been other mentions of the subjective probability by [Bertrand, 1889] in pp. 27 and [Borel, 1924] in pp. 332-333.

> *...de Finetti assumes that for each event of interest, there is some betting rate that you regard as fair, in the sense that you are willing to accept either side of a bet on the event at that rate. This fair betting rate is your personal probability for the event. More generally, we take your* **lower probability** *to be the maximum rate at which you are prepared to bet on the event, and your* **upper probability** *to be the minimum rate at which you are prepared to bet against the event. It is not irrational for you to assess an upper probability that is strict1y greater than your lower probability. Indeed, you ought to do so when you have little information on which to base your*

[8] *Uncertainty* can be due to lack of knowledge or to the natural variability in the observed data [Roeser et al., 2012, ch.2] (or a.k.a. *epistemic and aleatoric uncertainty* [Senge et al., 2014]), and it can lead us to biased estimations and high variance models [Braga-Neto et al., 2004].

[9] Imperfect information is here used as a synonym for limited information or/and lack of knowledge or prior beliefs.

> *assessments. In that case we say that your beliefs about the event are indeterminate, and that (for you) the event has imprecise probability.*
> *Statistical reasoning with imprecise Probabilities*, pp. 3

Walley thus coined the term *imprecise probability* (IP) in his published book [Walley, 1991] that was widely disseminated during the nineties and 21st century [Augustin et al., 2014; Troffaes et al., 2014].

1.3.1 *Imprecise probabilities*

Imprecise probability (IP) theory often (and will in our case) consists in representing our uncertainty (or lack of evidence in the empirical data) by a convex set $\mathscr{P}_X$ of probability distributions (i.e. a *credal set* [Levi, 1983]), defined over a space $\mathscr{X}$ rather than by a precise probability measure $\mathbb{P}_X$ [Taylor, 1973]. As they include precise distributions as special cases, such convex sets of distributions provide richer, more expressive models of uncertainty than the latter, and therefore allow us to better describe uncertainty originating from imperfect data. When using them to make decision, they naturally allow one to produce *cautious* set-valued decisions in case of high uncertainty.

Furthermore, whatever the uncertainty model $\mathscr{P}$ chosen, it will always converge to a precise probability estimate $\mathbb{P}$ so long as additional evidence is supplied (e.g. in Figure 2.1, the polytope composed from six extreme points may converge for instance to its centroid). That is why, in some field where it is very difficult to obtain more data (e.g. biology, clinical trials, and so on), one could consider that it should be mandatory to describe our uncertainty by means of a distributionally robust framework such as IP.

Given such a set of distributions $\mathscr{P}_X$ and any measurable event $A \subseteq \mathscr{X}$, we can define the notions of lower and upper probabilities as:

$$\underline{P}_X(A) = \inf_{\mathbb{P} \in \mathscr{P}_X} P(A) \qquad \text{and} \qquad \overline{P}_X(A) = \sup_{\mathbb{P} \in \mathscr{P}_X} P(A) \tag{1.8}$$

where $\underline{P}_X(A) = \overline{P}_X(A)$ only when we have sufficient information about event A. The lower probability is dual to the upper [Augustin et al., 2014], in the sense that $\underline{P}_X(A) = 1 - \overline{P}_X(A^c)$ where A^c is the complement of A. Many authors [Walley, 1991; Zaffalon, 2002] have argued that when information is lacking or imprecise, considering credal sets as our model of information better describes our actual uncertainty.

With such an approach, (1) the estimation of parameters in a parametric or non-parametric approach usually becomes more complicated computationally, since we consider a set of distributions $\mathscr{P}$ instead of a *single* probability distribution $\mathbb{P}$, and (2) the classical decision-making framework presented in the Equation (1.6) needs to be extended to handle sets of distributions $\mathscr{P}$.

For the first issue of model estimation, we propose, in Part I, to rely on previous works providing efficient generalized Bayesian inference (GBI) [Dempster, 1968] methods for exponential families (which include Gaussian distributions), that we will present in Section 3.1.2, and in Part II, to use a well-known classifier that extends the Naive Bayes classifier and that can compute the lower and upper probabilities in polynomial time.

For the latter issue of decision making, we will present and discuss, with some practical examples, several possible extensions in Chapter 2.

1.4 A SELF-LEARNING GUIDE TO THE READER

In order to turn reading into a coherent story, I have decided to organize this dissertation in two connected but distinct topics. To do so, I decided to do a flow diagram, in Figure 1.3, in order for the reader to find his/her way according to his/her preferences.

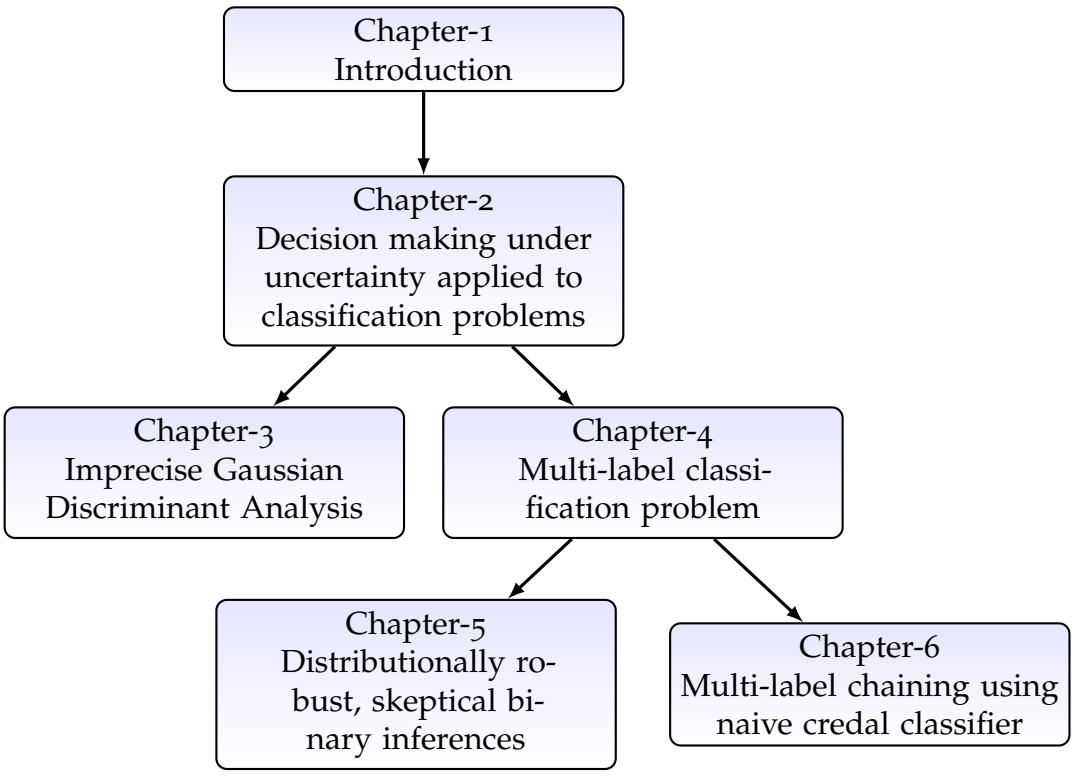


Figure 1.3: Flow diagram: Logical structure of the dissertation.

Firstly, in the Chapter 2, we present the theoretical background about decision making under uncertainty using imprecise probabilities that is necessary to understand the other chapters.

In Part I of this thesis, i.e. the Chapter 3, we present our imprecise Gaussian discriminant classifier using some existing ideas [Zaffalon, 2001; Benavoli et al., 2014].

Part II of this thesis starts by introducing, in the Chapter 4, some background about the multi-label problem in the classical approach. In Chapter 5, we discuss the problem of making cautious inferences in multi-label problems, demonstrating that efficient algorithms can be used for various common settings. Finally, in the Chapter 6, we present some theoretical and experimental results in multi-label chaining using as base classifier the well-known Naive Credal classifier.

1.5 RESEARCH WORKS

The research that led to this dissertation has produced six different papers up to now, of which some of them has been published, or are currently being reviewed, or just submitted, or in preparation. Like in the previous flow diagram 1.3, we present papers according to every topic, in what follows:

1. IMPRECISE CLASSIFICATION USING GAUSSIAN DISTRIBUTIONS

   - Carranza Alarcón et al. (2019). "Imprecise Gaussian Discriminant Classification". In: Proceedings of the Eleventh International Symposium on Imprecise Probabilities: Theories and Applications.
   - Carranza Alarcón et al. (2020). "Imprecise Gaussian Discriminant Classification". *(reviewed)*

   In the context of cautious multi-class classification, our first paper proposes a new imprecise classifier using a set of Gaussian distributions, followed by its extended version in journal format with extended experiments and theoretical results.

2. MULTI-LABEL PROBLEM UNDER IMPRECISE PROBABILITIES

   - Carranza Alarcón et al. (2020). "Distributionally robust, skeptical binary inferences in multi-label problems". *(submitted)*

   In the context of multi-label classification, we firstly propose theoretical procedures: (1) to reduce the complexity time of its inference step when we consider the Hamming loss case, and (2) to generalize the classical binary relevance by using imprecise marginal distributions.

   Second, we propose two different general strategies to adapt the classical multi-label chaining problem to the imprecise probabilistic setting.

   - Carranza Alarcón et al. (2020). "A first glance at multi-label chaining using imprecise probabilities". In: Workshop on Uncertainty in Machine Learning.

This latter has led to an extended paper by adapting the last two strategies to the use of the naive credal classifier and proposing a new dynamic, context-dependent label ordering. As a result of these extensions, we obtained new theoretical contributions presented in Chapter 6.

- Carranza Alarcón et al. (2020). "Multi-label chaining using naive credal classifier". *(in preparation)*

In addition to all this, I also contributed to a conference paper in the context of the label ranking problem, as a side collaboration, that proposes an efficient way to make partial predictions using imprecise probabilities.

- Carranza Alarcón et al. (2020). "Cautious Label-Wise Ranking with Constraint Satisfaction". In: Information Processing and Management of Uncertainty in Knowledge-Based Systems.

Finally, and aiming at a scientific diffusion of my research in France, some of the published conference papers were presented in French national conferences.

- Carranza Alarcón et al. (2018). "Analyse Discriminante Imprécise basée sur l'inférence Bayésienne robuste". In: 27èmes Rencontres francophones sur la Logique Floue et ses Applications.
- Carranza Alarcón et al. (2020). "Apprentissage de rangements prudent avec satisfaction de contraintes". In: 29èmes Rencontres francophones sur la Logique Floue et ses Applications.

Chapter 2

Decision making under uncertainty applied to classification problems

"The intermediate theories do not content themselves with the proper formulation of the statistical theory of estimation, neither do they accept the indispensability of a priori probabilities. For the attainment of any strong conclusion, they are, in my opinion, hopeless trials of eclecticism, intended to avoid particular faults or distasteful points of both alternatives without endeavoring to amalgamate their principles in a superior synthesis."

— Bruno De Finetti, in *Recent suggestions for the reconciliation of theories of probability*

Contents

Often, the decision maker can be faced with unreliable or hard situations where making a *single* decision may lead to damaging, if not dramatic, mistakes. The hardness of such situations can for instance be due to the lack of sufficient evidence or information (i.e. *uncertainty* in data). In such cases of imperfect information, it may be useful to provide set-valued, but more reliable decisions, especially for sensitive applications (e.g. medical diagnosis, control systems, cancer screening, etc.) where we cannot afford to make mistakes.

Hence, in this chapter, we firstly introduce some necessary theoretical background about how to make partial decisions under uncertainty using imprecise probabilities. This latter shall specifically be examined in the

context of the multiclass classification problem. Then, we shall introduce the *imprecise* classification approach, which is mainly an extension of the classical classification using imprecise probabilities and decision making under uncertainty. Finally, we will introduce the well-known imprecise classifier named Naive credal classifier (NCC).

2.1 DECISION MAKING UNDER UNCERTAINTY

When we talk about making a decision under uncertainty, it is always related to the fact that we do not know the state of nature, which is not under the control of the decision maker. This uncertainty may be described by unknown factors, such as missing or noisy information, where we mean by noisy that information has been provided incorrectly without bad faith.

In classical-statistic analysis, the uncertainty is often measured through a probability measure (or distribution), which can in certain cases not be enough to detect the indecision or imprecision. Thus, in this section, we consider that our uncertainty is modelled as a set of probability distributions.

To turn reading into a pleasing experience and not give just abstract theoretical definitions. I decided to illustrate all the different criteria of decision making presented below using the following example.

Example 1 *Let us consider a hard real problem in which a naïve graduating student needs to make a risky decision about his/her future.*

To do so, let us consider a classification problem composed of four labels (a.k.a. actions or decisions available to the decision-maker).

$$\mathcal{K}^* = \{m_a, m_b, m_c, m_d\}, \tag{2.1}$$

which have the following descriptions

m_a : *to understand laws of the universe,*

m_b : *to start a job in computer science,*

m_c : *to co-fund an innovative startup in her/his garage,*

m_d : *to pursue a PhD thesis.*

As our purpose is to illustrate the benefits of making a decision under uncertainty using a set of probability distributions $\mathscr{P}_{Y|x}$ instead of a single probability distribution $\mathbb{P}_{Y|x}$, we then consider the probability estimates described in Table 2.1.

An illustration of the precise probability and credal set estimates is shown in Figure 2.1. On the left side, we can firstly see a probability 3-simplex (or tetrahedron), in which every vertex corner corresponds to a probability distribution $\mathbb{P}_{Y|x}$ such that the probability of a specific event $m_k \in \mathcal{K}^*$ is $P_x(Y = m_k) = 1$.

		$\hat{P}_x(Y=m_a)$	$\hat{P}_x(Y=m_b)$	$\hat{P}_x(Y=m_c)$	$\hat{P}_x(Y=m_d)$
$\hat{\mathbb{P}}_{Y\|x}$		0.225	0.222	0.225	0.328
$\hat{\mathscr{P}}_{Y\|x}$	$\hat{\mathbb{P}}_1$	0.282	0.209	0.194	0.315
	$\hat{\mathbb{P}}_2$	0.208	0.281	0.196	0.315
	$\hat{\mathbb{P}}_3$	0.188	0.229	0.268	0.315
	$\hat{\mathbb{P}}_4$	0.237	0.186	0.262	0.315
	$\hat{\mathbb{P}}_5$	0.186	0.186	0.186	0.442
	$\hat{\mathbb{P}}_6$	0.243	0.243	0.243	0.271

Table 2.1: Conditional probability estimates of credal set and precise distribution.

$$\Delta := \left\{ \mathbb{P}_{Y|x} := (P_x(Y=m_a), \ldots, P_x(Y=m_d)) \in \mathbb{R}^4 \;\middle|\; \sum_{y \in \mathcal{K}^*}^{n} P_x(Y=y) = 1 \right\} \tag{2.2}$$

Inside the tetrahedron, we have our convex credal set (or convex polytope) $\hat{\mathscr{P}}_{Y|x} \subset \Delta$ represented by six extreme points, in which each probability distribution estimation is illustrated on the right side of Figure 2.1. Besides, the precise distribution estimation $\hat{\mathbb{P}}_{Y|x}$ is located in the center of inertia of the polytope.

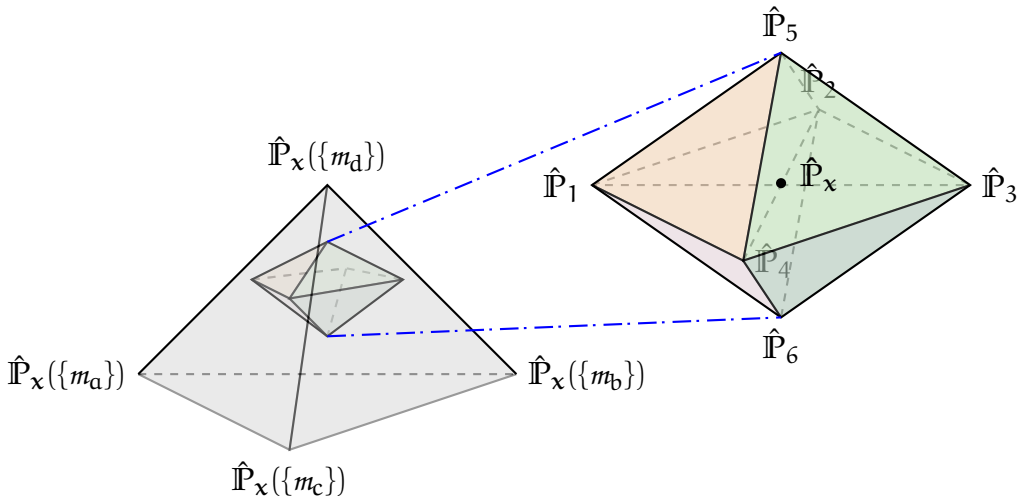


Figure 2.1: Polytope of set of probabilities

Finally, as commonly used in the context of decision making, let us consider a classical cost matrix (c.f. Table 2.2a) obtained from the $\ell_{0/1}$ function and a custom cost matrix (c.f. Table 2.2b) defined according to the risk level of every decision, in other words, the naïve graduating student estimated costs.

In the context of IP, every row is often considered as a gamble function, i.e. $1-\ell_{0/1}$ are betting profits.

In Example 1, we provide two cost matrices (a.k.a loss matrix in classification problems) with the purpose of exemplifying:

1. that the use of an uncertainty model $\mathscr{P}$ does not imply that we always obtain a set-valued predictions, it may depend on the loss function as well as the empirical evidence given to estimate $\hat{\mathscr{P}}$, and

$\ell(\cdot,\cdot)$	m_a	m_b	m_c	m_d
m_a	**0.0**	1.0	1.0	1.0
m_b	1.0	**0.0**	1.0	1.0
m_c	1.0	1.0	**0.0**	1.0
m_d	1.0	1.0	1.0	**0.0**

(a) $\ell_{0/1}$ classic

$\ell(\cdot,\cdot)$	m_a	m_b	m_c	m_d
m_a	**0.00**	0.73	0.56	0.04
m_b	0.55	**0.00**	0.19	0.78
m_c	0.09	0.76	**0.00**	0.18
m_d	0.20	0.88	0.27	**0.00**

(b) ℓ_* contextual

Table 2.2: Loss values incurred

2. situations where one makes partial predictions in the form of set-valued labels in case of hard situations where a single decision is not safe also depends on the loss function and the credal set estimate $\hat{\mathscr{P}}$.

For theoretical developments of next two subsections, we will assume that we know the form of the convex set of distributions $\mathscr{P}_{Y|x}$ and the precise probability distribution $\mathbb{P}_{Y|x}$.

Furthermore, all different calculations given in examples of the next section, i.e. the lower, upper and precise expected values, have been be performed using the *improb-redux* software[1], so it can be reproduced to the reader's liking.

2.1.1 *Decision making under precise probabilities*

The criterion of making-decision introduced in the Definition 1 can also be represented as a strict total order relation[2] $\succ_x$ over $\mathscr{K} \times \mathscr{K}$, meaning that it can be posed as a problem of inferring preferences between labels, as follows:

Definition 2 (Precise ordering [Berger, 1985, pp. 47]) *Given a general loss function $\ell(\cdot,\cdot)$ and a conditional probability distribution $\mathbb{P}_{Y|x}$, m_a is preferred to m_b, denoted by $m_a \succ_x m_b$, if and only if:*

$$\mathbb{E}_{\mathbb{P}_{Y|x}}[\ell(\cdot, m_a)|\mathbf{x}] < \mathbb{E}_{\mathbb{P}_{Y|x}}[\ell(\cdot, m_b)|\mathbf{x}] \tag{2.3}$$

Definition 2 tells us that exchanging m_b for m_a would incur a positive expected loss, due to the fact that expectation loss of m_b is greater than m_a, therefore m_a should be preferred to m_b for a given new unlabelled instance $\mathbf{x}$. In the particular case where we use the loss function $\ell_{0/1}$, it is easy to prove that:

$$m_a \succ_x m_b \iff P_x(Y = m_a) > P_x(Y = m_b) \tag{2.4}$$

[1] Implementation is available in https://github.com/mcmtroffaes/improb-redux
[2] A complete, transitive, and asymmetric relation.

where $P_{\mathbf{x}}(Y = m_a) := P(Y = m_a|X = \mathbf{x})$ is the unknown conditional probability of label m_a given a new unlabeled instance $\mathbf{x}$.

Therefore, given a set of labels $\mathcal{K}$ and a conditional probability estimate $\hat{P}$, we can then establish a strict total order by making pairwise comparisons (see figure 2.2) as follows:

$$m_{i_K} \succ_{\mathbf{x}} m_{i_{K-1}} \succ_{\mathbf{x}} \cdots \succ_{\mathbf{x}} m_{i_1} \iff \hat{P}_{\mathbf{x}}(Y = m_{i_K}) > \cdots > \hat{P}_{\mathbf{x}}(Y = m_{i_1}). \quad (2.5)$$

We can then pick out one of the undominated labels, i.e., one with ***maximal probability***.

In the case where Equation (1.7) returns multiple elements, which is unlikely in practice but not impossible, they can be considered as *indifferent* and chosen randomly without affecting the theoretical performance or risk of the classifier. It should be noted that, whatever the quantity of data used to induce the model or the specific new instance $\mathbf{x}$ we consider (that may come from a poorly populated region), we will always get (up to indifference) a unique undominated label. In contrast, the IP approach where we consider sets $\mathscr{P}_{Y|X}$ of probabilities may, depending on the criterion-of-decision choice, result in strict partial orders having multiple undominated and *incomparable* labels.

Example 2 *Making use of the probability estimates of the conditional distribution $\hat{\mathbb{P}}_{Y|\mathbf{x}}$ given in the Table 2.1, we can compute the strict total order between its labels $\mathcal{K}^*$, which is $m_d \succ_{\mathbf{x}} m_a \sim_{\mathbf{x}} m_c \succ_{\mathbf{x}} m_b$ and can be illustrated graphically in the Figure 2.2*

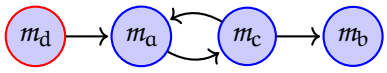


Figure 2.2: Graph of the strict total order on labels $\mathcal{K}^*$

$\{m_d\}$ being the maximal label dominating other ones, it is the predicted one. This means that the most accurate decision, and also smart, is to "to pursue a PhD thesis" (of course in Compiègne).

As shown by Equation (2.5) and Example 2, usual statistics and probability model *uncertainty* with a unique distribution $\mathbb{P}$, canonically and axiomatically ending up in a unique undominated label as a decision. While it is possible to implement decision rules providing set-valued predictions in such settings [Ha, 1997], several authors [Walley, 1991; Dempster, 1968] have argued that a single distribution cannot always faithfully represent lack of information.

So, if we describe our model *uncertainty* by a convex set of distributions $\mathscr{P}$ —which may be a better choice— this implies, at first glance, that we should naively verify Equation (2.3) for every probability distribution $\mathbb{P} \in \mathscr{P}$, which is indeed one of the criteria we shall study next.

2.1.2 *Decision making under imprecise probabilities*

Within IP theories, we can find different methods extending the decision criterion given in Definition 2 (for further details [Walley, 1991; Augustin et al., 2014; Troffaes, 2007, §3.9, §8]). To classify a new instance $\mathbf{x}$, we will introduce five criteria that have strong theoretical justifications and often remain applicable in practice [Zaffalon, 2002; Yang et al., 2017].

Such criteria can be considered more or less conservatives, in the sense of cautiousness, meaning that some will provide more imprecise decisions than others. The most common being used are : (1) Maximality, (2) Interval dominance, (3) E-admissibility, (4) Γ-minimin, (5) Γ-maximin.

Note that there does exist other extensions in the state-of-the-art, e.g. [Destercke, 2010].

Before starting to define such criteria, let us first introduce some definitions. Given a loss function ℓ, we will denote by

$$\overline{\mathbb{E}}_{\mathscr{P}}[\ell(y,\cdot)] := \max_{\mathbb{P}\in\mathscr{P}} \mathbb{E}_{\mathbb{P}}[\ell(y,\cdot)] \quad \text{and} \quad \underline{\mathbb{E}}_{\mathscr{P}}[\ell(y,\cdot)] := \min_{\mathbb{P}\in\mathscr{P}} \mathbb{E}_{\mathbb{P}}[\ell(y,\cdot)] \tag{2.6}$$

the upper and lower expected values of this loss under uncertainty $\mathscr{P}$. They respectively provide an assessment of the worst-case and best-case situations.

We will use two different notations to distinguish the set-valued and single prediction(s), as follows

$$\hat{y} : \text{a single prediction, where } \hat{y} \in \mathscr{K}, \text{ and}$$
$$\hat{\mathbb{Y}} : \text{a set-valued prediction, where } \hat{\mathbb{Y}} \subseteq \mathscr{K}.$$

Besides, we will also use the superscript (the subscript) to denote the type of criterion used (resp. the loss function used), e.g. $\hat{y}^{\Gamma\max}_{\ell_{0/1}}$ is a single prediction using the Γ-minimax criterion and the $\ell_{0/1}$ function.

Moreover, in the case of performing pairwise comparisons over labels that may result in a strict partial order, we will use the following notations

$$m_x \succ_* m_y \quad \text{which denote that } m_x \text{ is preferred to } m_y$$
$$m_x \succ\!\prec_* m_y \quad \text{which denote that } m_x \text{ and } m_y \text{ are } \textit{incomparable}.$$

Note that in this thesis, we do not deal with the *indifference* relation, denoted here as $m_x \sim m_y$, meaning that m_x and m_y are equally desirables.

2.1.2.1 Γ-*minimax and* Γ-*minimin criteria*

We begin by introducing two criteria close of the classical one, since these choose the best- or worst-case solution amongst the existing ones [Berger, 1985, §5].

Definition 3 (Γ-Minimax) Γ-MINIMAX *consists in replacing the expected value of Equation* (1.6) *by its upper expectation*

$$\hat{y}^{\Gamma\max}_{\ell,\mathscr{P}} = \arg\min_{y\in\mathscr{K}} \overline{\mathbb{E}}_{\mathscr{P}}\left[\ell(y,\cdot)\right]. \tag{2.7}$$

It amounts to returning the best worst-case prediction (i.e. a pessimistic attitude), since it consists in minimizing the worst possible expected loss.

Definition 4 (Γ-Minimin) Γ-Minimin, *in contrast, consists in replacing the expected value of Equation (1.6) by its lower expectation*

$$\hat{y}^{\Gamma\min}_{\ell,\mathscr{P}} = \arg\min_{y\in\mathscr{K}} \underline{\mathbb{E}}_{\mathscr{P}}\left[\ell(y,\cdot)\right]. \tag{2.8}$$

It amounts to returning the best best-case prediction (i.e. an optimistic attitude), since it consists in choosing the prediction with the smallest lower expectation.

In the particular case where we consider the loss function $\ell_{0/1}$ and a credal set $\mathscr{P}_{Y|x}$, it is easy to prove that

$$\hat{y}^{\Gamma\max}_{\ell_{0/1},\mathscr{P}_{Y|x}} = \arg\max_{y\in\mathscr{K}} \underline{P}_x(Y=y) \quad \text{and} \quad \hat{y}^{\Gamma\min}_{\ell_{0/1},\mathscr{P}_{Y|x}} = \arg\max_{y\in\mathscr{K}} \overline{P}_x(Y=y). \tag{2.9}$$

In the case of a small output space $\mathscr{K}$, these last optimizations may be easy to solve (especially when the credal set has a finite number of extreme points). Otherwise, they may be hard to solve.

Example 3 *Making use of the credal set $\hat{\mathscr{P}}_{Y|x}$, we can easily obtain the upper and lower expectation with respect to matrix loss values of Table 2.2a*

y	m_a	m_b	m_c	m_d
$\underline{\mathbb{E}}_{\hat{\mathscr{P}}_{Y|x}}\left[\ell_{0/1}(y,\cdot)\right]$	0.718	0.719	0.732	**0.558**
$\overline{\mathbb{E}}_{\hat{\mathscr{P}}_{Y|x}}\left[\ell_{0/1}(y,\cdot)\right]$	0.814	0.814	0.814	**0.729**

applying Equations (2.7) and (2.8), we obtain

$$\hat{y}^{\Gamma\min}_{\ell_{0/1},\hat{\mathscr{P}}_{Y|x}} = m_d \quad \textit{and} \quad \hat{y}^{\Gamma\max}_{\ell_{0/1},\hat{\mathscr{P}}_{Y|x}} = m_d$$

which perfectly matches with the precise prediction.

However, if the matrix loss is the one of Table 2.2b

y	m_a	m_b	m_c	m_d
$\underline{\mathbb{E}}_{\hat{\mathscr{P}}_{Y|x}}\left[\ell_{*}(y,\cdot)\right]$	0.197	0.312	0.234	**0.186**
$\overline{\mathbb{E}}_{\hat{\mathscr{P}}_{Y|x}}\left[\ell_{*}(y,\cdot)\right]$	**0.257**	0.397	0.283	0.263

we obtain a different result in the Γ-minimax criterion

$$\hat{y}^{\Gamma\min}_{\ell_{*},\hat{\mathscr{P}}_{Y|x}} = m_d \quad \textit{and} \quad \hat{y}^{\Gamma\max}_{\ell_{*},\hat{\mathscr{P}}_{Y|x}} = m_a.$$

The last example illustrates perfectly the fact that, regardless of how large is our uncertainty model $\mathscr{P}$, we may obtain an accurate decision that perfectly matches that of its precise counterpart, albeit it does not imply that this one is the ground-truth one.

In contrast to the $\ell_{0/1}$ matrix, if we consider the ℓ_* matrix customized to our preferences, we obtain a best worst-case prediction not far away of the best best-case prediction, since preferences m_a and m_d are strongly related with the science and research.

These criteria are not conservatives (or cautious), since they always generate single predictions, and hence, do not necessarily reflect our lack of information. In other words, they do not represent our indecision [Augustin et al., 2014, p. 193], or imprecision, as they will output a single prediction whatever our uncertainty is. Thus, in what follows, we shall introduce the first of three criteria which produce a set of possible solutions (including incomparable decisions if necessary).

2.1.2.2 *Maximality criterion*

The maximality criterion is the most natural extension of Equation (2.3) as it amounts to comparing decisions pairwisely in a robust way, meaning that every preference of Equation (2.3) holds only if it holds for every model (i.e. every precise distribution in the credal set), as follows.

Definition 5 (Maximality [Walley, 1991, §3.9.5]) *Under maximality criterion m_a is preferred to m_b iff the cost of exchanging m_a with m_b have a positive lower expectation*

$$m_a \succ_\ell^{\mathscr{P}} m_b \iff \underline{\mathbb{E}}_{\mathscr{P}}\left[\ell(\cdot, m_b) - \ell(\cdot, m_a)\right] > 0. \tag{2.10}$$

The prediction is then non-dominated elements of the strict partial order $\succ_\ell^{\mathscr{P}}$

$$\hat{\mathbb{Y}}^{\mathsf{M}}_{\ell,\mathscr{P}} = \left\{ m_a \in \mathscr{K} \,\middle|\, \nexists m_b \in \mathscr{Y} : m_b \succ_\ell^{\mathscr{P}} m_a \right\} \tag{2.11}$$

Since $\succ_\ell^{\mathscr{P}}$ is a strict partial order, $\hat{\mathbb{Y}}^{\mathsf{M}}_{\ell,\mathscr{P}}$ may result in a set of multiple, incomparable elements (i.e. maximal non-dominated elements), in which case the prediction becomes imprecise due to high uncertainty in the model.

Computing $\hat{\mathbb{Y}}^{\mathsf{M}}_{\ell,\mathscr{P}}$ can be a computationally demanding task with at most a quadratic time complexity on output space $\mathscr{O}(|\mathscr{K}|^2)$. So, it may make the inference step critical when considering combinatorial spaces, such as multi-label problems in which getting Equation (5.2) may require at worst to perform $|\mathscr{Y}|(|\mathscr{Y}|-1)/2$ comparisons, where $|\mathscr{Y}| = 2^m$ is the output space with m labels, ending up with a complexity of $\mathscr{O}(2^{2^m})$ that quickly becomes untractable even for small values of m.

Remark 1 *It should be noted that the computational time mentioned previously, namely $\mathscr{O}(|\mathscr{K}|^2)$, may still be reduced using two different strategies:*

1. *removing dominated elements already verified (cf.[Augustin et al., 2014, algo. 16.4]), and*
2. *verifying if maximal elements obtained from the precise probabilistic setting (i.e. Equation (2.5)) are non-dominated (c.f Section 3.6)*

However, in the worst-case scenario, in which all elements are non-dominated, the time complexity remains the same (i.e quadratic time).

Furthermore, if we consider the loss function $\ell_{0/1}$ and credal set $\mathscr{P}_{Y|x}$, Equation (2.10) can be reduced to

$$m_a \succ^{\mathscr{P}_{Y|x}}_{\ell_{0/1}} m_b \iff \inf_{\mathbb{P}_{Y|x} \in \mathscr{P}_{Y|x}} \Big[P_x(Y = m_a) - P_x(Y = m_b) \Big] > 0 \tag{2.12}$$

Equation (2.12) amounts to requiring Equation (2.4) to be true for all possible probability distributions in $\mathscr{P}$.

Example 4 *Applying Equation (2.10) to the credal set $\hat{\mathscr{P}}_{Y|x}$, we can calculate the lower expectation of each pairwise comparison of labels of the output space $\mathscr{K}^*$ (so 12 comparisons) as follows*

y	m_a	m_b	m_c	m_d
$\underline{\mathbb{E}}_{\hat{\mathscr{P}}_{Y|x}}[\ell_{0/1}(y,\cdot) - \ell_{0/1}(m_a,\cdot)]$	·	−0.07	−0.08	−0.26
$\underline{\mathbb{E}}_{\hat{\mathscr{P}}_{Y|x}}[\ell_{0/1}(y,\cdot) - \ell_{0/1}(m_b,\cdot)]$	−0.07	·	−0.08	−0.26
$\underline{\mathbb{E}}_{\hat{\mathscr{P}}_{Y|x}}[\ell_{0/1}(y,\cdot) - \ell_{0/1}(m_c,\cdot)]$	−0.09	−0.09	·	−0.26
$\underline{\mathbb{E}}_{\hat{\mathscr{P}}_{Y|x}}[\ell_{0/1}(y,\cdot) - \ell_{0/1}(m_d,\cdot)]$	**0.03**	**0.03**	**0.03**	·

Following them, we can then produce the partial ordering $\mathscr{B}^{M}_{\ell_{0/1}}$

$$\mathscr{B}^{M}_{\ell_{0/1}} = \left\{ m_d \succ^{\hat{\mathscr{P}}_{Y|x}}_{\ell_{0/1}} m_a, \quad m_d \succ^{\hat{\mathscr{P}}_{Y|x}}_{\ell_{0/1}} m_b, \quad m_d \succ^{\hat{\mathscr{P}}_{Y|x}}_{\ell_{0/1}} m_c \right\} \tag{2.13}$$

where $\hat{Y}^{M}_{\ell_{0/1},\hat{\mathscr{P}}_{Y|x}} = \{m_d\}$ is the predicted set obtained from set $\mathscr{B}^{M}_{\ell_{0/1}}$ of comparisons by the criterion of maximality (Figure 2.3a).

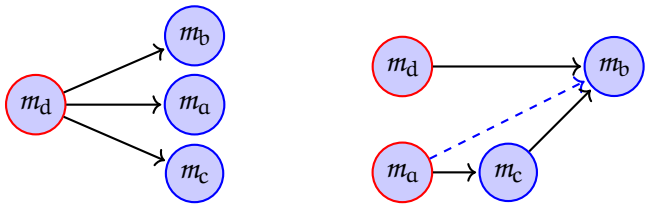
In contrast, if we consider the cost matrix of ℓ_*, it produces the below lower expectations

y	m_a	m_b	m_c	m_d
$\underline{\mathbb{E}}_{\hat{\mathscr{P}}_{Y|x}}[\ell_*(y,\cdot) - \ell_*(m_a,\cdot)]$	·	**0.34**	**0.02**	**−0.02**
$\underline{\mathbb{E}}_{\hat{\mathscr{P}}_{Y|x}}[\ell_*(y,\cdot) - \ell_*(m_b,\cdot)]$	−0.46	·	−0.41	−0.48
$\underline{\mathbb{E}}_{\hat{\mathscr{P}}_{Y|x}}[\ell_*(y,\cdot) - \ell_*(m_c,\cdot)]$	−0.09	**0.32**	·	−0.07
$\underline{\mathbb{E}}_{\hat{\mathscr{P}}_{Y|x}}[\ell_*(y,\cdot) - \ell_*(m_d,\cdot)]$	−0.03	**0.32**	−0.01	·

such that the partial ordering $\mathscr{B}^{M}_{\ell_*}$ (Figure 2.3b) obtained is

$$\mathscr{B}^{\mathrm{M}}_{\ell_*} = \left\{ m_a \succ^{\hat{\mathscr{P}}_{Y|x}}_{\ell_*} m_b,\quad m_a \succ^{\hat{\mathscr{P}}_{Y|x}}_{\ell_*} m_c,\quad m_c \succ^{\hat{\mathscr{P}}_{Y|x}}_{\ell_*} m_b,\quad m_d \succ^{\hat{\mathscr{P}}_{Y|x}}_{\ell_*} m_b,\quad m_d \succ \prec^{\hat{\mathscr{P}}_{Y|x}}_{\ell_*} m_a \right\}, \tag{2.14}$$

where $\hat{\mathbb{Y}}^{\mathrm{M}}_{\ell_*,\hat{\mathscr{P}}_{Y|x}} = \{m_d, m_a\}$ *is the predicted set obtained from set* $\mathscr{B}^{\mathrm{M}}_{\ell_*}$



(a) $\mathscr{B}^{\mathrm{M}}_{\ell_{0/1}}$ **(b)** $\mathscr{B}^{\mathrm{M}}_{\ell_*}$

Figure 2.3: Graphs of partial order of Example 4.

That means, regardless of the $\ell_{0/1}$ *or* ℓ_* *function, the decision* $\{m_a, m_d\}$ *of the subject will always be inclined to pursue a quest of knowledge.*

Note that, in Example 4, specifically in the Figure 2.3b, the comparison $m_a \succ^{\hat{\mathscr{P}}_{Y|x}}_{\ell_*} m_b$ does not need to be verified, since thanks to the property of transitivity of $\succ^{\mathscr{P}}_{\ell}$ it is automatically inferred. Yet, one still has to compute the remaining preferences, i.e $m_a \succ^{\hat{\mathscr{P}}_{Y|x}}_{\ell_*} m_c$ and $m_c \succ^{\hat{\mathscr{P}}_{Y|x}}_{\ell_*} m_b$.

As we will see later, more specifically in Figure 2.5, this criterion is located between the most and the least imprecise one in producing a set of predictions, which are respectively the easiest and hardest to obtain computationally. In other words, maximality is a good compromise, and also the most theoretically justified rule in Walley's framework. That is also one of the reasons why this criterion was chosen to implement our imprecise classifier presented in Chapter 3 and considered as one of the main criteria to optimise the multi-label problem, namely Chapter 5

2.1.2.3 *Interval dominance*

The interval dominance (ID) is a very conservative rule, as one has $\hat{\mathbb{Y}}^{\mathrm{M}}_{\ell,\mathscr{P}} \subseteq \hat{\mathbb{Y}}^{\mathrm{ID}}_{\ell,\mathscr{P}}$, producing the largest set-valued predictions. This is mainly due to the relaxation of Equation (2.11) obtained by using the super-additivity (or super-linearity) property of the lower expectation operator $\underline{\mathbb{E}}_{\mathscr{P}}$ [Walley, 1991, §2.3.3, P3], producing a lower bound of Equation (2.10) as follows

$$\underline{\mathbb{E}}_{\mathscr{P}}\left[\ell(\cdot, m_b) - \ell(\cdot, m_a)\right] > \underline{\mathbb{E}}_{\mathscr{P}}\left[\ell(\cdot, m_b)\right] - \overline{\mathbb{E}}_{\mathscr{P}}\left[\ell(\cdot, m_b)\right]. \tag{2.15}$$

In terms of computations to perform, interval-dominance is more efficient (see Equation 2.18) than maximality, meaning that it can be performed in linear complexity. Formally, interval-dominance criterion can be defined as follows:

Definition 6 (Interval dominace) *Under this criterion, m_a is preferred to m_b, denoted by $\sqsupset_\ell^{\mathscr{P}}$, iff the largest expected loss of m_a is strictly lower than the smallest expected loss of m_b, for all distribution in the credal set $\mathbb{P} \in \mathscr{P}$, as follows*

$$m_a \sqsupset_\ell^{\mathscr{P}} m_b \iff \overline{\mathbb{E}}_{\mathscr{P}}[\ell(m_a, \cdot)] < \underline{\mathbb{E}}_{\mathscr{P}}[\ell(m_b, \cdot)], \tag{2.16}$$

One can see from Equation (2.15) and Equation (2.16) that $m_a \succ m_b$ implies $m_a \sqsupset_\ell^{\mathscr{P}} m_b$, showing that interval dominance could be quite conservative. The prediction is then non-dominated elements of the strict partial order $\sqsupset_\ell^{\mathscr{P}}$

$$\hat{\mathbb{Y}}_{\ell,\mathscr{P}}^{\mathrm{ID}} = \left\{ m_a \in \mathscr{K} \,\middle|\, \nexists m_b \in \mathscr{Y} : m_b \sqsupset_\ell^{\mathscr{P}} m_a \right\}, \tag{2.17}$$

That is, interval dominance retains all these predictions not dominated by the worst-case expected loss situation of another prediction. Besides, Equation (2.16) can be equivalently expressed as

$$\forall m_a \in \hat{\mathbb{Y}}_{\ell,\mathscr{P}}^{\mathrm{ID}} \iff \underline{\mathbb{E}}_{\mathscr{P}}[\ell(m_a, \cdot)] < \min_{m_k \in \mathscr{K}} \overline{\mathbb{E}}_{\mathscr{P}}[\ell(m_k, \cdot)], \tag{2.18}$$

which amounts to linearly compare all labels $\mathscr{O}(2n-1)$ against the Γ-minimax value [Augustin et al., 2014, §16.3.4].

In the same way as other criteria, if we consider the loss function $\ell_{0/1}$ and the credal set $\mathscr{P}_{Y|x}$, Equation (2.16) is equivalent to

$$m_a \sqsupset_{\ell_{0/1}}^{\mathscr{P}_{Y|x}} m_b \iff \underline{P}_x(Y = m_a) > \overline{P}_x(Y = m_b). \tag{2.19}$$

Example 5 *Using the Equation (2.16) and the upper and lower expectations calculated in Example 3, the interval dominance criterion then produces the strict partial order of below (Figure 2.4a)*

$$\mathscr{B}_{\ell_{0/1}}^{\mathrm{ID}} = \left\{ m_a \sqsupset\!\sqsubset_{\ell_{0/1}}^{\mathscr{P}_{Y|x}} m_b, \quad m_a \sqsupset\!\sqsubset_{\ell_{0/1}}^{\mathscr{P}_{Y|x}} m_d, \quad m_b \sqsupset\!\sqsubset_{\ell_{0/1}}^{\mathscr{P}_{Y|x}} m_d, \quad m_d \sqsupset_{\ell_{0/1}}^{\mathscr{P}_{Y|x}} m_c \right\} \tag{2.20}$$

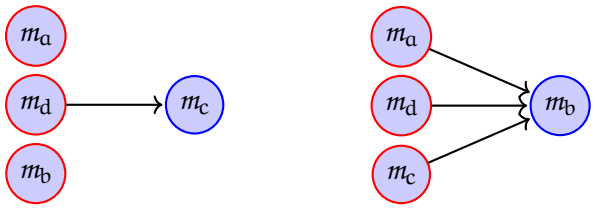
where $\hat{\mathbb{Y}}_{\ell_{0/1},\mathscr{P}_{Y|x}}^{\mathrm{ID}} = \{m_d, m_a, m_b\}$ is the predicted set obtained from set $\mathscr{B}_{\ell_{0/1}}^{\mathrm{ID}}$

In contrast, if we consider the cost matrix of ℓ_*, we obtain partial ordering solution $\mathscr{B}_{\ell_*}^{\mathrm{ID}}$ (Figure 2.4b)

$$\mathscr{B}_{\ell_*}^{\mathrm{ID}} = \left\{ \begin{array}{lll} m_a \sqsupset_{\ell_*}^{\mathscr{P}_{Y|x}} m_b, & m_c \sqsupset_{\ell_*}^{\mathscr{P}_{Y|x}} m_b, & m_d \sqsupset_{\ell_*}^{\mathscr{P}_{Y|x}} m_b, \\ m_a \sqsupset\!\sqsubset_{\ell_*}^{\mathscr{P}_{Y|x}} m_d, & m_a \sqsupset\!\sqsubset_{\ell_*}^{\mathscr{P}_{Y|x}} m_c, & m_d \sqsupset\!\sqsubset_{\ell_*}^{\mathscr{P}_{Y|x}} m_c \end{array} \right\} \tag{2.21}$$

where $\hat{\mathbb{Y}}_{\ell_*,\mathscr{P}_{Y|x}}^{\mathrm{M}} = \{m_d, m_a, m_c\}$ is the predicted set obtained from set $\mathscr{B}_{\ell_*}^{\mathrm{ID}}$

In Example 5, we can first note that the set of predictions is the largest among the previous one, i.e. $\hat{y}_{\ell,\mathscr{P}}^{\Gamma_{\min}}, \hat{y}_{\ell,\mathscr{P}}^{\Gamma_{\max}} \in \hat{\mathbb{Y}}_{\ell,\mathscr{P}}^{\mathrm{M}} \subseteq \hat{\mathbb{Y}}_{\ell,\mathscr{P}}^{\mathrm{ID}}$. Secondly, even



(a) $\mathscr{B}^{\mathrm{ID}}_{\ell_{0/1}}$ **(b)** $\mathscr{B}^{\mathrm{ID}}_{\ell_*}$

Figure 2.4: Graph of partial order $\mathscr{B}$.

if the solutions of $\hat{\mathbb{Y}}^{\mathrm{ID}}_{\ell_{0/1},\hat{\mathscr{P}}_{Y|x}}$ and $\hat{\mathbb{Y}}^{\mathrm{ID}}_{\ell_*,\hat{\mathscr{P}}_{Y|x}}$ are slightly different, because of exchanging m_b to m_c, predictions proposed by the ℓ_* may be considered as more relevant since all of them dominate m_b, whereas m_a and m_b of $\hat{\mathbb{Y}}^{\mathrm{ID}}_{\ell_{0/1},\hat{\mathscr{P}}_{Y|x}}$ are non-dominated maximal elements but incomparable with m_c.

2.1.2.4 *E-admissibility criterion*

In contrast to other criteria, where optimisation problems is always represented by computing a infimum or a supremum value[3], E-admissibility criterion must verify that all probability distributions of the credal set satisfy Equation (1.6).

Definition 7 (E-admissibility [Levi, 1983]) *E-admissibility returns the set of predictions that are optimal for at least one probability within the credal set $\mathscr{P}$. In other words, the E-admissibility rule returns the prediction set*

$$\hat{\mathbb{Y}}^{\mathrm{E}}_{\ell,\mathscr{P}} = \left\{ y \in \mathcal{Y} \;\middle|\; \exists \mathbb{P} \in \mathscr{P} \text{ s.t. } \forall y' \in \mathscr{K}, \mathbb{E}_{\mathbb{P}}\left[\ell(y,\cdot)\right] < \mathbb{E}_{\mathbb{P}}\left[\ell(y',\cdot)\right] \right\}. \tag{2.22}$$

The last equation can be equivalently expressed as

$$\hat{\mathbb{Y}}^{\mathrm{E}}_{\ell,\mathscr{P}} = \bigcup_{\mathbb{P}\in\mathscr{P}} \left\{ \arg\min_{y\in\mathscr{K}} \mathbb{E}_{\mathbb{P}}\left[\ell(y,\cdot)\right] \right\}. \tag{2.23}$$

If we consider the loss function $\ell_{0/1}$ and a credal set $\mathscr{P}_{Y|x}$, this gives:

$$\hat{\mathbb{Y}}^{\mathrm{E}}_{\ell_{0/1},\mathscr{P}_{Y|x}} = \bigcup_{\mathbb{P}_{Y|x}\in\mathscr{P}_{Y|x}} \left\{ \arg\max_{y\in\mathscr{K}} \mathrm{P}_x(Y=y) \right\}. \tag{2.24}$$

E-admissibility is often the hardest to solve amongst the criteria presented in this chapter. Such complexity can be alleviated if we know in

[3] As $\mathscr{P}$ is considered convex and weak*-compact [Walley, 1991, prop. 3.6.1], meaning that is compact in the weak*-topology [Walley, 1991, Appendix D], computing the lower/upper prevision (or expectation) of any loss function is easily performed by using the extreme points of $\mathscr{P}$, namely, $\inf_{\mathscr{P}} \cdot := \inf_{\mathtt{ext}(\mathscr{P})} \cdot$ (in the same way for the supremum).

advance the set of solutions of the maximality criterion, since as we know, that $\hat{\mathbb{Y}}^{E}_{\ell,\mathscr{P}} \subseteq \hat{\mathbb{Y}}^{M}_{\ell,\mathscr{P}}$ (c.f. Figure 2.5), it would be enough to check if any solution of the maximality set is or not E-admissible. In particular, we could simply verify if all solutions of the maximality are E-admissibles by using the extreme points $ext(\mathscr{P})$ of the credal set $\mathscr{P}$. But if one solution does not so, we cannot deduce anything about this one and we should use a linear program [Augustin et al., 2014, §16.3.3] in order to verify if it exists at least one probability distribution $\mathbb{P} \in int(\mathscr{P})$ that does so. Let us see this matter in the next example.

Example 6 *Using Equation (2.24) under the credal set $\hat{\mathscr{P}}_{Y|x}$ and the set of solutions of the maximality criterion $\hat{\mathbb{Y}}^{M}_{\ell_{0/1},\hat{\mathscr{P}}_{Y|x}} = \{m_d\}$. We can easily verify if m_d belongs to E-admissibility criterion if one of probability distributions of Table 2.1 returns m_d. So, E-admissibility criterion produce the following prediction set*

$$\hat{\mathbb{Y}}^{E}_{\ell_{0/1},\hat{\mathscr{P}}_{Y|x}} = \{m_d\}. \tag{2.25}$$

In contrast, and making use of Equation (2.23) , if we consider the cost matrix of ℓ_* and the set of maximality solutions $\hat{\mathbb{Y}}^{M}_{\ell_*,\hat{\mathscr{P}}_{Y|x}} = \{m_a, m_d\}$, we can check if the following expectations obtained from extreme points (or probability distributions) returns $\{m_a, m_d\}$

y	m_a	m_b	m_c	m_d	$\hat{y}$
$\mathbb{E}_{\mathbb{P}_1}[\ell_*(y,\cdot)]$	**0.195**	0.631	0.283	0.209	m_a
$\mathbb{E}_{\mathbb{P}_2}[\ell_*(y,\cdot)]$	**0.235**	0.578	0.255	0.263	m_a
$\mathbb{E}_{\mathbb{P}_3}[\ell_*(y,\cdot)]$	**0.213**	0.618	0.234	0.234	m_a
$\mathbb{E}_{\mathbb{P}_4}[\ell_*(y,\cdot)]$	**0.189**	0.649	0.253	0.202	m_a
$\mathbb{E}_{\mathbb{P}_5}[\ell_*(y,\cdot)]$	0.207	0.666	0.259	**0.186**	m_d
$\mathbb{E}_{\mathbb{P}_6}[\ell_*(y,\cdot)]$	0.210	0.601	0.255	**0.243**	m_a

E-admissibility prediction is then

$$\hat{\mathbb{Y}}^{E}_{\ell_*,\hat{\mathscr{P}}_{Y|x}} = \{m_a, m_d\}. \tag{2.26}$$

In Example 6, we can first note that when the $\ell_{0/1}$ function is considered the optimal decision is still the m_d label, being coherent with the Γ-minmin and Γ-minimax solutions obtained previously. When we consider the ℓ_* cost matrix, the E-admissibility prediction is the set $\{m_a, m_d\}$, meaning that this criterion is more conservative than previous ones, giving us two plausible solutions to decision-makers (i.e. the graduating student).

2.1.2.5 *Summarizing*

In summary, we can see from the previous examples that certain criteria are more or less conservatives than other ones. [Troffaes, 2007] has theoretically proven these implications in a general context. An illustration to

summarize those implications is shown in Figure 2.5, where an implication $A \to B$ means that $A \subseteq B$.

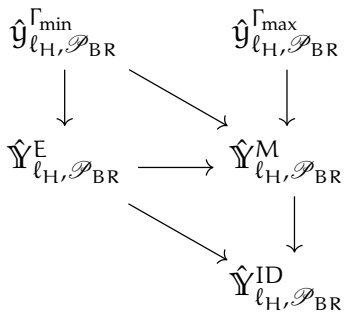


Figure 2.5: Decision relations on all criteria.

Furthermore, it is clear that the more imprecise is our uncertainty model $\mathscr{P}$, the larger will be the set of predictions of **E**-admissibility, **M**aximality and **I**nterval **D**ominance criteria. That means, if $\mathscr{P}$ is vacuous (modelling ignorance), the set of predictions will be all the elements of the output space.

2.2 CLASSIFICATION WITH IMPRECISE PROBABILITIES

Cautious classification does not aim to do "better" than their precise counterparts, nor to implement a rejection option (i.e., not classifying at all) in case of ambiguity [Herbei et al., 2006], but to highlight those hard cases for which information is insufficient to isolate a *single* reliable precise prediction, and to propose a subset of possible predictions. We can find in the literature three "main" ways to build cautious classifier models:

1. using a classical precise classifier but deriving a set-valued predictions from them [Cheng et al., 2012; Mortier et al., 2019] (e.g. partial reject [Ha, 1997], conformal prediction [Shafer et al., 2008]),

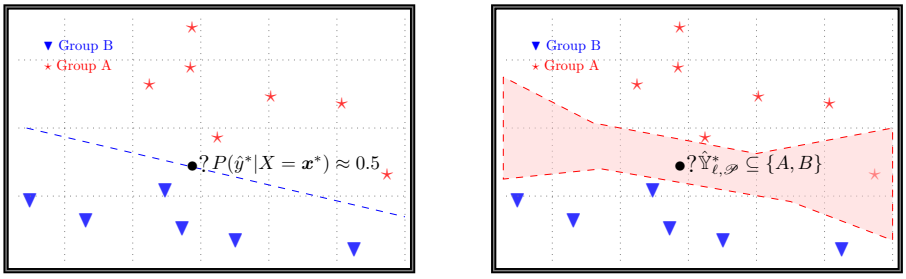
2. making data imperfect (coarse or impartial observations) and then building a corresponding imperfect robust model, and finally

3. learning an imprecise probabilistic classifier from which set-valued predictions follow naturally (using techniques such as robust frequentist methods [Cattaneo, 2007; Cattaneo, 2008] or robust Bayesian inference [Walley, 1991; Walter, 2013; Quaeghebeur et al., 2005]).

In this thesis, we retain the latter, as this one indeed considers imprecision as parts of its basic axioms, in contrast to the other approaches where imprecision is not directly integrated into the learned model.

Depending on the chosen decision criterion, the imprecise classifier may generate a region of imprecision, like in Figure 2.6b. Otherwise, if

it produces a single prediction, like classical classification, the imprecise classifier may generate a decision boundary amongst the labels, like in Figure 2.6a.



(a) Precise decision-making **(b)** Cautious decisions-making

Figure 2.6: *Cautious vs precise decision-making.* Figure (a) shows a precise model, where there are no regions where the model will output set-valued predictions, in contrast with (b) where such a region exists (in red).

In the literature, and to the best of our knowledge, few methods have been developed in the spirit of the third category presented above. The first one of them was the Naive Credal classifier (NCC)[Zaffalon, 1999; Corani et al., 2008b] (for more details in the next section) which is based on the Imprecise Dirichlet Model (IDM) [Walley, 1996] to make statistical inferences from multinomial data and the Naive Bayes assumption.

Corani et al. proposed two new imprecise classifiers [Corani et al., 2008a; Corani et al., 2015] extending the precise version of the Bayesian model averaging (BMA) to the imprecise probabilistic setting, so-called credal model averaging (CMA), by substituting the single prior distribution over all models of BMA by a set of priors.

Another original work, and also in the same spirit as our imprecise classifier, is the one of [Paton et al., 2015] which use the imprecise conjugate prior density [Coolen, 1993; Quaeghebeur et al., 2005] to create an imprecise multinomial logistic model. However, its posterior distribution depends strongly [Paton, 2016, §4.3.2]; (1) on the assumption that it is a unimodal distribution, and (2) on performing an imprecise Monte Carlo simulation, meaning that one needs a Monte Carlo simulation for every distribution in $\mathscr{P}$ (which is already complicated with a single distribution). In addition, the complexity arising from the prior/posterior conjugate depends on the training data size which in practical applications can quickly increase.

Recently [Dendievel et al., 2018] proposed a first instance of an imprecise non-parametric model using the classical kernel density estimation (KDE), albeit this one is not yet adapted for multivariate analysis. Also, [Mauá et al., 2017] proposed to extend the relatively new probabilistic graphical model known as sum-product networks (SPN) to the imprecise

probabilistic setting, namely credal sum-product networks (CSPN), allowing that singleton weights of SPN can vary in some space with constraints (i.e. probability simplexes). Finally, [Basu et al., 2020] has also recently proposed an imprecise binary classifier under sparsity constraints based on a set of likelihood functions.

The remaining other ones were developed borrowing some ideas of the NCC, by extending or adapting them to their own scope, such as those using the IDM. For instance; random forest [Abellán et al., 2017], boosting classifier [Utkin, 2015], credal decision trees [Abellan et al., 2012], credal ensemble of classifier [Corani et al., 2014], tree-augmented naive credal classifier [Zaffalon et al., 2003b], and so on. Unfortunately, there are no review papers summarizing the weaknesses and strengths from all of them, which could be a future contribution.

In what follows, we will focus on describing essential concepts about the naive credal classifier, which is built in the same spirit as ours.

2.3 NAIVE CREDAL CLASSIFIER

The naive bayes classifier (NBC) is based on an assumption of independence, meaning that it assumes the attribute independence given a class, combined with the simple Bayes' theorem [Domingos et al., 1997]. It can formally be written as the marginal probability given the class m_k

$$P(Y = m_k|X = \mathbf{x}) = \frac{P(Y = m_k)\prod_{i=1}^{d} P(X_i = x_i|Y = m_k)}{\sum_{m_l\in\{0,1\}} P(Y = m_l)\prod_{i=1}^{d} P(X_i = x_i|Y = m_l)}. \tag{2.27}$$

Under the $\ell_{0/1}$ loss function [Domingos et al., 1997; Hand et al., 2001], in which the class (i.e. Equation (2.27)) with the highest probability is chosen, NBC achieves a good predictive performance, since it tends to assign an unrealistic high probability to the most probable class which would be the ground truth one (e.g. the class m_a in Equation (2.4)).

Moreover, in terms of the bias-variance tradeoff of the misclassification error [Friedman, 1997], the probability estimate of NBC has high bias but also low variance, which is more significant on larger data sets, and hence with a poorer predictive performance. Yet, with small and medium data sets (or also with noisy information), on which our research focuses, comes with the same problem of high bias and low variance, NBC can be a competitive method under $\ell_{0/1}$ and be unfortunately overperformed by complex classifiers on larger data sets (for further details see [Augustin et al., 2014, §10.2]).

The Naive credal classifier (NCC) [Zaffalon, 2002] is based on the same assumptions as NBC, but instead of only using a single distribution to estimate the probability of Equation (2.27), NCC replaces it with a credal set $\mathscr{P}$, in which case it becomes necessary to choose one of the criteria

presented in Section 2.1 for the inference-step. Therefore, in order to keep a consistency with the scheme presented in Figure 1.1 , we will first detail the inference-step and will then present the imprecise learning model.

NCC has been applied in a vast variety of real applications, such as medicine [Zaffalon et al., 2003a], agriculture [Zaffalon, 2005], geology [Antonucci et al., 2007], or betting for FIFA World cup [Quaeghebeur et al., 2017].

We decided to present two decision-making criteria for this classifier; the interval dominance and the maximality, the latter being the most used in *The Society for Imprecise Probability: Theories and Applications* (SIPTA[4]), under the $\ell_{0/1}$ loss function[5]. Of course, it is straightforward to obtain the optimal solution of Γ-minimax and Γ-minimin criteria from the interval dominance criterion.

2.3.1 *Decision making applied to NCC*

INTERVAL DOMINANCE Under the $\ell_{0/1}$ loss function and a credal set $\mathscr{P}_{Y|x}$, the optimization problem is reduced to computing lower and upper probability bounds $[\underline{P}_x(Y = m_k), \overline{P}_x(Y = m_k)]$ (c.f. Equation (2.19)) over all possible marginals $\mathscr{P}_Y$ and conditional distributions $\mathscr{P}_{X|Y}$. This can be performed by solving the following minimization/maximization problem for Equation (2.27)

$$\underline{P}_x(Y = m_k) = \min_{\mathbb{P}_Y \in \mathscr{P}_Y} \min_{\substack{\mathbb{P}_{X_i|Y} \in \mathscr{P}_{X_i|Y} \\ i \in \{1,\dots,d\}}} \frac{P(Y = m_k) \prod_{i=1}^{d} P(X_i = x_i | Y = m_k)}{\sum_{m_l \in \mathscr{K}} P(Y = m_l) \prod_{i=1}^{d} P(X_i = x_i | Y = m_l)},$$

$$\overline{P}_x(Y = m_k) = \max_{\mathbb{P}_Y \in \mathscr{P}_Y} \max_{\substack{\mathbb{P}_{X_i|Y} \in \mathscr{P}_{X_i|Y} \\ i \in \{1,\dots,d\}}} \frac{P(Y = m_k) \prod_{i=1}^{d} P(X_i = x_i | Y = m_k)}{\sum_{m_l \in \mathscr{K}} P(Y = m_l) \prod_{i=1}^{d} P(X_i = x_i | Y = m_l)},$$

and assuming that the denominator is different from zero, the lower probability can be reduced to

$$\underline{P}_x(Y = m_k) = \min_{\mathbb{P}_Y \in \mathscr{P}_Y} \min_{\substack{\mathbb{P}_{X_i|Y} \in \mathscr{P}_{X_i|Y} \\ i \in \{1,\dots,d\}}} \left(1 + \frac{\sum_{\substack{m_l \in \mathscr{K} \\ m_l \neq m_k}} P(Y = m_l) \prod_{i=1}^{d} P(X_i = x_i | Y = m_l)}{P(Y = m_k) \prod_{i=1}^{d} P(X_i = x_i | Y = m_k)} \right)^{-1} \tag{2.28}$$

[4] Web site: http://www.sipta.org/

[5] Developments presented below are not at all applicable to a generic loss function ℓ.

$$= \min_{\mathbb{P}_Y \in \mathscr{P}_Y} \left(1 + \frac{\sum_{\substack{m_l \in \mathscr{K} \\ m_l \neq m_k}} P(Y = m_l) \prod_{i=1}^{d} \overline{P}(X_i = x_i | Y = m_l)}{P(Y = m_k) \prod_{i=1}^{d} \underline{P}(X_i = x_i | Y = m_k)} \right)^{-1} . \tag{2.29}$$

Note that, before moving to the last equation, we firstly focus on the inner minimization problem, in which minimizing the fraction term $(1 + a/b)^{-1}$ is the same as maximizing a/b where a and b do not share any common term. Additionally, note that we can measure the lower and upper bounds independently of $P(X_i = x_i | Y = m_*)$ because they are defined for different conditioning events, meaning that we can without any problem compute the lower bound of $P(X_i = x_i | Y = m_k)$ without considering quantities $P(X_i = x_i | Y = m_*)$ (c.f. [Zaffalon, 1999, §3.1]). In a similar vein, we can obtain the upper bound

$$\overline{P}_x(Y = m_k) = \max_{\mathbb{P}_Y \in \mathscr{P}_Y} \left(1 + \frac{\sum_{\substack{m_l \in \mathscr{K} \\ m_l \neq m_k}} P(Y = m_l) \prod_{i=1}^{d} \underline{P}(X_i = x_i | Y = m_l)}{P(Y = m_k) \prod_{i=1}^{d} \overline{P}(X_i = x_i | Y = m_k)} \right)^{-1} . \tag{2.30}$$

Now the problem of maximizing/minimizing over all possible marginal distributions $\mathscr{P}_Y$ can be solved in two ways: (1) enumerating the extreme points, if possible, and computing every one of them to obtain the minimum/maximum (i.e. combinatoric problem), or (2) directly resolving an optimization programming problem on constraints obtained in the functional form of the credal set $\mathscr{P}_Y$.

However, in this thesis, when we use this imprecise classifier, for practical purposes and as the number of training data is usually sufficient in our experimental experiences, we shall assume a precise estimation of the marginal distribution[6].

MAXIMALITY In contrast with interval dominance, which is often easier to solve, maximality can be complex. However, thanks to the independence assumption it can simply be solved for the NBC. We first recall here the maximality criterion

$$m_a \succ_\ell^{\mathscr{P}} m_b \iff \inf_{\mathbb{P}_{Y|X} \in \mathscr{P}_{Y|X}} P_x(Y = m_a) - P_x(Y = m_b) > 0$$

applying Equation (2.27) and omitting its denominator, the latter being the same positive constant of normalization of each probability, it can be written

[6] A set of marginal distribution can be strongly recommended if we work with imbalanced datasets, which will not be a focus in this thesis.

$$\inf_{\mathbb{P}_Y \in \mathscr{P}_Y} \inf_{\mathbb{P}_{X|Y} \in \mathscr{P}_{X|Y}} P(X = \mathbf{x}|Y = m_a)P(Y = m_a) - P(X = \mathbf{x}|Y = m_b)P(Y = m_b) > 0,$$

using the same arguments of independence as for the previous criteria, i.e. different conditioning events, we can easily obtain

$$\inf_{\mathbb{P}_Y \in \mathscr{P}_Y} P(Y = m_a) \prod_{i=1}^{d} \underline{P}(X_i = x_i|Y = m_a) - P(Y = m_b) \prod_{i=1}^{d} \overline{P}(X_i = x_i|Y = m_b) > 0,$$

once again this last optimization problem can be solved using the same argument as for the previous criterion.

Now, we have to get the probability bounds using a statistical model, in this case, we shall use, as classically done, the imprecise Dirichlet model.

2.3.2 *Imprecise statistical model applied to NCC*

Imprecise Dirichlet model (IDM) [Walley, 1996] is a statistical model verifying several axiomatic principles which are desirables for (indecision) inference step, such as the *learning from data* and *representation invariance principle* (RIP) .

RIP: "inferences based on observations should not depend on the sample space in which the observations and future events of interest are represented" — Walley (1996).

Firstly, let us start by defining some notations

$$\begin{aligned}
&\mathscr{X} := \mathscr{X}_1 \times \dots \mathscr{X}_d && \text{a finite collection of } d \text{ feature domains,}\\
&\mathbf{x} = (x^1, \dots, x^d) \in \mathscr{X} \\
&x^1 \in \mathscr{X}_1, \dots, x^d \in \mathscr{X}_d && \text{a value from the input space,}\\
&(\mathbf{x}, m_k) \in \mathscr{X} \times \mathscr{K} && \text{an instance } \mathbf{x} \text{ labelled with } m_k \text{ class,}\\
&\mathscr{D} = \{(\mathbf{x}_i, y_i)\}_{i=1}^{N} && \text{a training data set generated i.i.d.}
\end{aligned}$$

IDM is built on a natural Bayesian approach using a conjugate prior distribution. Considering that

$$P(X_i = x_i|Y = m_k) = \theta_{x_i|m_k}, \tag{2.31}$$

$$P(Y = m_k) = \theta_{m_k}, \tag{2.32}$$

so that the joint distribution can be written as follows

$$P(Y = m_k, X = \mathbf{x}) := \theta_{m_k,\mathbf{x}} = \theta_{m_k} \prod_{i=1}^{d} \theta_{x_i|m_k}, \tag{2.33}$$

where $\theta_{m_k,\mathbf{x}}$ is the chance of the multinomial distribution (i.e. the odds that the couple $(m_k, \mathbf{x})$ happens at the same time), such that $\theta_{x_i|m_k}$ denote the chance that $X_i = x_i \in \mathscr{X}_1$ conditional on the class $Y = m_k$, and similarly for θ_{m_k}.

The full likelihood after observing data $\mathbf{n}$ can thus be expressed as follows (as a product of multinomial densities)

$$L(\theta|\mathbf{n}) \propto \prod_{m_k \in \mathcal{K}} \theta_{m_k}^{n(m_k)} \prod_{i=1}^{d} \prod_{x_i \in \mathcal{X}_i} \theta_{x_i|m_k}^{n(x_i|m_k)} \tag{2.34}$$

where $n(\cdot)$ is a count function that counts the number of occurrences of events $x_i|m_k$ and m_k in the observed data set. $n(x_i|m_k)$ is the number of instances in the training set where $X_i = x_i$ and the label value is m_k, such that $\sum_{x_i \in \mathcal{X}_i} n(x_i|m_k) = n(m_k)$, and $n(m_k)$ is the number of instances in the training set where the label value is m_k, such that $\sum_{m_k \in \mathcal{K}} n(m_k) = N$.

A natural conjugate prior distribution, which can be obtained by applying the Proposition 5.4 of [Bernardo et al., 2000], to get a posterior distribution of the same family is a product of Dirichlet prior densities [Zaffalon, 2001]

$$f(\theta|\mathbf{t}, s) \propto \prod_{m_k \in \mathcal{K}} \theta_{m_k}^{st(m_k)-1} \prod_{i=1}^{d} \prod_{x_i \in \mathcal{X}_i} \theta_{x_i|m_k}^{st(x_i|m_k)-1} \tag{2.35}$$

with a real number s and a function $t(\cdot)$ as hyper-parameters, over which we may consider different sets of constraints [Walley, 1996] [Zaffalon, 2001, §2.3]. For practical purpose, we decided to use (a.k.a local IDM [Augustin et al., 2014, §10.4.4])

$$\sum_{m_k \in \mathcal{K}} t(m_k) = 1 \quad \text{and} \quad \sum_{x_i \in \mathcal{X}} t(x_i|m_k) = 1 \tag{2.36}$$

so that $t(m_k) \in [0, 1]$ and $t(x_i|m_k) \in [0, 1]$.

Coupling the likelihood and the prior density, we can easily obtain the posterior distribution which is a product of independent Dirichlet densities[7]

$$\pi(\theta|\mathbf{t}, s, \mathbf{n}) \propto \prod_{m_k \in \mathcal{K}} \theta_{m_k}^{st(m_k)+n(m_k)-1} \prod_{i=1}^{d} \prod_{x_i \in \mathcal{X}_i} \theta_{x_i|m_k}^{st(x_i|m_k)+n(x_i|m_k)-1}. \tag{2.37}$$

Remark 2 *Note that we opted to use an unusual re-parametrization of the Dirichlet distribution, since this one often use the hyper-parameter $\boldsymbol{\alpha}$ with constraints $\sum_i^K = \alpha_i$ such that $\alpha_k = st_k$. Walley conveniently splits it in two parameters with the aim of latter fixing s in the model.*

[7] A product of independent densities may, for instance, be represented by: $\pi := \pi(\beta_1, \ldots, \beta_K) \propto f_1(\beta_1) \ldots f_K(\beta_K)$, and in which the expected value of a single parameter β_k over the full distribution can be reduced to marginal expectation over parameter β_k: $\mathbb{E}_\pi[\beta_k] = \mathbb{E}_{\pi(\beta_k)}[\beta_k]$

So, we can compute the expected value of each probability as follows

$$\mathbb{E}_{\mathbb{P}_\theta}\left[\theta_{x_i|m_k}|\mathbf{t}, s, \mathbf{n}\right] = P(X_i = x_i|Y = m_k) = \frac{n(x_i|m_k) + st(x_i|m_k)}{\sum_{x_i \in \mathscr{X}_i} n(x_i|m_k) + st(x_i|m_k)},$$
$$\mathbb{E}_{\mathbb{P}_\theta}\left[\theta_{m_k}|\mathbf{t}, s, \mathbf{n}\right] = P(Y = m_k) = \frac{n(m_k) + st(m_k)}{\sum_{m_k \in \mathscr{K}} n(m_k) + st(m_k)}$$

resolving the sum of denominators, we have

$$P(X_i = x_i|Y = m_k) = \frac{n(x_i|m_k) + st(x_i|m_k)}{n(m_k) + s}, \tag{2.38}$$
$$P(Y = m_k) = \frac{n(m_k) + st(m_k)}{N + s} \tag{2.39}$$

computing the lower and upper probability bounds when $t(\cdot) \to 0$ for lower one and $t(\cdot) \to 1$ for upper one, we have

$$P(X_i = x_i|Y = m_k) \in \left[\frac{n(x_i|m_k)}{n(m_k) + s}, \frac{n(x_i|m_k) + s}{n(m_k) + s}\right], \tag{2.40}$$
$$P(Y = m_k) \in \left[\frac{n(m_k)}{N + s}, \frac{n(m_k) + s}{N + s}\right] \tag{2.41}$$

Note that the higher s is, the wider the intervals $[\underline{P}(X_i = x_i|Y = m_k), \overline{P}(X_i = x_i|Y = m_k)]$ are (the same for $[\underline{P}_x(Y = m_k), \underline{P}_x(Y = m_k)]$). For $s = 0$, we retrieve the classical NBC with precise predictions, and for high enough values of $s >>> 0$, the NCC model will make vacuous predictions (i.e. abstain for all labels $Y = \mathscr{K}$).

Remark 3 ([Walley, 1996, p. 10]) *Note that the degree of imprecision of the upper and lower posterior probability can be measured by* $\overline{P}(Y = m_k) - \underline{P}(Y = m_k) = \frac{s}{N+s}$*, which does not depend on the event* $Y = m_k$*, but on the imprecision level* s *and on the training data set size.*

Finally, if the input space of training data set is continuous, it shall be discretized in z intervals in order to get values of $n(x_i|m_k)$ and $n(m_k)$. For simplicity, we shall use in this thesis, depending on the case, only two levels of discretization $z = 5$ and $z = 6$ with equal-width intervals. Since our main goal is to compare model behaviours and not to optimize our approach, this seems sufficient.

Remark 4 (Laplace smoothing) *Also called additive smoothing, this correction will be used when we compare the precise case amongst its imprecise version, since using* $s = 0$ *it may be possible to get* $n(x_i|m_k) = n(m_k) = 0$ *because of discretization.*

2.4 CONCLUSION

After having introduced some of the main concepts used in the rest the thesis, we can start presenting our contributions.

Note that preliminaries as well as existing works specific to those contributions have not been discussed here, but will be in each subpart.

Part I

IMPRECISE GAUSSIAN DISCRIMINANT

Gaussian discriminant analysis is a popular classification model, that in the precise case can produce unreliable predictions in case of high uncertainty (e.g., due to scarce or noisy data). We remedy this, by proposing a new Gaussian discriminant analysis based on robust Bayesian analysis and near-ignorance priors. The model delivers cautious predictions, in form of set-valued class, in case of limited or imperfect available information. Our experiments show that including an imprecise component in the Gaussian discriminant analysis produces reasonably cautious predictions, and that set-valued predictions correspond to instances for which the precise model performs poorly.

Chapter 3

Imprecise Gaussian Discriminant Classification

"What we know is not much. What we don't know is enormous."

—Pierre Simon Laplace

Contents

A well-known *precise* generative classifier model used to perform the classification task, that we will consider in this chapter, is the Gaussian discriminant analysis (GDA) [Friedman et al., 2001, §4.3]. Let $\mathscr{X} \times \mathscr{K}$ be the space of observations, with $X \in \mathscr{X} = \mathbb{R}^p$ a random vector and $Y \in \mathscr{K} = \{m_1, ..., m_K\}$ the set of labels. The main goal of GDA is to estimate the theoretical conditional probability distribution (c.p.d) $\mathbb{P}_{Y=m_k|X}$ of the class $Y = m_k$ given an observation x via Bayes' theorem as follows

$$\mathbb{P}_{Y=m_k|X} = \frac{\mathbb{P}_{X|Y=m_k}\mathbb{P}_{Y=m_k}}{\sum_{m_l \in \mathscr{K}} \mathbb{P}_{X|Y=m_l}\mathbb{P}_{Y=m_l}}. \tag{3.1}$$

Thus, quantifying $\mathbb{P}_{Y=m_k|X}$ is equivalent to quantify $\mathbb{P}_{X|Y=m_k}$ and the marginal distribution $\mathbb{P}_Y$.

In *precise* probabilistic approaches, this is typically done by using maximum likelihood estimation (MLE) and by making some parametric assumptions about the probability density $\mathbb{P}_{X|Y=m_k}$ (i.e. Gaussian probability distribution (g.p.d)) in order to find a plausible estimate (see Section 3.1.1). However, such precise estimates usually have trouble differentiating different kinds of uncertainties [Senge et al., 2014], such as uncertainty due to ambiguity (mixed classes in some areas of the input space) and uncertainty due to lack of knowledge or information (limited training data set inducing biases in estimates [Braga-Neto et al., 2004]).

Bayesian methods, in contrast, incorporate some prior beliefs in the form of probability distribution defined on unknown parameters of the model. Such beliefs typically come from expert opinions or persons that are knowledgeable in the context of the problem. However, it is also well-known that the elicitation of prior beliefs can be absent or hard to obtain during the study of a problem, especially when learning classifiers. A classical way out of this problem is to use a so-called non-informative prior, which allow one to obtain a posterior not including any prior knowledge [Dalton et al., 2015].

Yet, the use of such prior is not without problems within the Bayesian theory, as it is often not coherent/proper in the sense of De Finetti, and mainly boils down to use maximum likelihood estimators. Moreover, it may seem strange that an absence of prior should lead to a fully precise, completely informed posterior. Alternatively, it has been argued and shown that using truly vacuous prior information (considering all possible priors, including very extreme ones) while remaining coherent with this information usually lead to vacuous posterior predictions [Walley, 1991; Bernardo et al., 2000, §7.4, §5.6.2] (i.e. our model would not be able to learn from data), so considering that we have no information also seems a poorly sensible approach. Walley has therefore proposed to use a set of non-informative prior distributions, called *near-ignorance priors* [Walley, 1991, §4.6.9], to solve this issue. These near-ignorance priors must respect certain properties [Benavoli et al., 2014, §2] so as not to obtain vacuous predictions, while remaining invariant under a large set of transformations. Hence, one of our motivations in this chapter is to not use a single prior distribution, but a set of prior distributions (or credal set [Levi, 1983]) to reflect our lack of knowledge and obtain cautious predictions.

In Section 3.1, we describe the estimation of the conditional distribution in the case of the precise GDA, using a frequentist inference approach. We then extend this *precise* parametric estimation to *imprecise* estimation in a robust Bayesian inference context, using the IP near-ignorance model proposed by Benavoli *et al.* [Benavoli et al., 2014] to do so and obtaining estimates in the form of a convex set. Coupling this imprecise estimation with the *maximality* criterion, we present our Imprecise GDA (IGDA) model and its different variants in Section 3.2.

In Section 3.3, we perform a set of experiments on different datasets using our imprecise model and compare it to its precise counterparts. We show that the cautious predictions are useful, in the sense that they (1) concern instances for which the precise classifier often makes mistakes, (2) often include the true class within the predicted set on these same instances, and (3) are not overly imprecise.

Furthermore, we briefly discuss (focusing on computational issues) in Section 3.4 the extension of our method to other settings, namely to the case where the class proportions $\mathbb{P}_{Y=m_k}$ are also imprecisely estimated, and where the criterion to minimise is not the raw number of errors (corresponding to a 0/1 loss function) but a generic loss function.

In Section 3.5, we perform supplementary experiments on 4 different synthetic data sets in order to investigate how our imprecise model and its counterparts behave in terms of predictive robustness when: (1) the testing data sets are corrupted by some noise (i.e., do not follow the i.i.d. assumption), and (2) the number of training data considerably decreases.

Finally, we propose in Section 3.6 an "optimal" algorithm for a cautious prediction using the maximality criterion, in the same spirit as the one of Nakharutai et al. (2019). We also provide a comparative benchmark between the naive and "optimal" version in order to empirically to show a computational improvement.

3.1 GAUSSIAN DISCRIMINANT ANALYSIS MODEL

As mentioned above, a classical way to estimate the distribution $\mathbb{P}_{Y|X}$ is by using Bayes' theorem. Sections 3.1 and 3.2 describe its use for the precise and imprecise approaches, respectively.

3.1.1 *Statistical inference with precise probabilities*

Among the many ways to model $\mathbb{P}_{X|Y=m_k}$, this work focus on *parametric* discriminant analysis for which $\mathbb{P}_{X|Y=m_k}$ follows a multivariate Gaussian distribution $\mathcal{N}(\mu_{m_k}, \Sigma_{m_k})$ with unknown mean μ_{m_k} and covariance matrix Σ_{m_k}, i.e.:

$$\mathscr{G}_{m_k} := \mathbb{P}_{X|Y=m_k} \sim \mathcal{N}(\mu_{m_k}, \Sigma_{m_k}) \tag{3.2}$$

whose probability density function is written

$$P(X = x|Y = m_k) = \frac{1}{(2\pi)^{p/2}|\Sigma_{m_k}|^{1/2}} e^{-\frac{1}{2}(x-\mu_{m_k})^T \Sigma_{m_k}^{-1}(x-\mu_{m_k})}. \tag{3.3}$$

The marginal distribution is defined as a multinomial $\pi_y := \mathbb{P}_Y$, where $P(Y = m_k) = \pi_{m_k}$. So, under a 0/1 loss function, the optimal prediction becomes:

$$\hat{\varphi}(x|\theta_{m_k}) := \arg\max_{m_k \in \mathcal{K}} \log \pi_{m_k} - \frac{1}{2}\log|\Sigma_{m_k}| - \frac{1}{2}(x^\top - \mu_{m_k})^\top \Sigma_{m_k}^{-1}(x^\top - \mu_{m_k}) \quad (3.4)$$

where $\Theta = \{\theta_{m_k} | \theta_{m_k} = (\pi_{m_k}, \Sigma_{m_k}, \mu_{m_k}), \forall m_k \in \mathcal{K}\}$ is the parametric space from which comes our estimate. In Table 3.1, we remind different discriminant models corresponding to various constraints imposed to the covariance matrices of the conditional distributions given by Equation (3.2).

Discriminant analysis model	Assumptions ($\forall m_k \in \mathcal{K}$)	Parametric space ($\forall m_k \in \mathcal{K}$)
Parametric Gaussian conditional distribution $\mathbb{P}_{X\|Y=m_k}$		
Linear Discriminant [Friedman et al., 2001, §4.3]	Homoscedasticity: $\Sigma_{m_k} = \Sigma$	$\Theta = \{\theta_{m_k} \| \theta_{m_k} = (\pi_{m_k}, \Sigma, \mu_{m_k})\}$
Quadratic Discriminant [Friedman et al., 2001, §4.3]	Heteroscedasticity: $\Sigma_{m_k} = \Sigma_k$	$\Theta = \{\theta_{m_k} \| \theta_{m_k} = (\pi_{m_k}, \Sigma_k, \mu_{m_k})\}$
Naive Discriminant [Friedman et al., 2001, §6.63]	Feature independence: $\Sigma_{m_k} = \sigma_k^\top \mathbb{I}$	$\Theta = \{\theta_{m_k} \| \theta_{m_k} = (\pi_{m_k}, \sigma_k, \mu_{m_k})\}$
Euclidean Discriminant [Marco et al., 1987]	Unit-variance feature indep.: $\Sigma_{m_k} = \mathbb{I}$	$\Theta = \{\theta_{m_k} \| \theta_{m_k} = (\pi_{m_k}, \mu_{m_k})\}$

Table 3.1: Gaussian discriminant analysis models

In frequentist inference, usual estimation of parameters of Equation (3.4) is obtained by MLE using a subset $\mathcal{D}_{m_k} = \{(x_{i,k}, y_{i,k} = m_k) | i = 1, \ldots, n_k\} \subseteq \mathcal{D}$ of observations of training data. We have $\hat{\pi}_{m_k} = n_k/N$ (frequency of m_k) and $\hat{\mu}_{m_k} = \bar{x}_k$ (sample mean of $\mathcal{D}_{m_k}$). Depending on whether we assume the model to have (1) dependent features, we will have an hetero- or homo-scedastic assumption, with respectively $\hat{\Sigma}_{m_k} = \hat{S}_{m_k}$ (sample covariance matrix of $\mathcal{D}_{m_k}$) or $\hat{\Sigma}_{m_k} = \hat{S}$ (within-class covariance matrix $\mathcal{D}$), or to have (2) independent features, we will have features weighted proportionally to their inverse variance $\hat{\Sigma}_{m_k} = \hat{\sigma}_k^\top \mathbb{I}$ or unweighed with all weights equal to 1, i.e. $\hat{\Sigma}_{m_k} = \mathbb{I}$.

However, those estimates do not account for the quantity of data they are based on, which may be low to start with, and may also vary significantly across classes, especially in case of imbalanced data sets. To solve this issue, we propose in the next section an imprecise discriminant model, based on the use of imprecise probabilities and using results from Benavoli *et al.* [Benavoli et al., 2014].

3.1.2 *Statistical inference with imprecise probabilities*

To estimate $\mathbb{P}_{X|y}$ and $\mathbb{P}_Y$ in the form of convex sets of distributions, we will use robust Bayesian inference under prior near-ignorance models. Before describing our *imprecise* estimation, we make three general assumptions for our *imprecise* Gaussian discriminant model:

1. Normality of conditional probability distribution $\mathbb{P}_{X|Y=m_k} := \mathcal{G}_{m_k}$, as in the classical case.
2. A *precise* estimation of marginal distribution $\mathbb{P}_Y := \hat{\pi}_y$.
3. A *precise* estimation of covariance matrix $\Sigma_k := \hat{\Sigma}_k = \hat{S}_k$ or $\hat{S}$.

In Section 3.4, we will discuss the relaxation of assumption 2, considering a set of distributions $\mathscr{P}_Y$.

3.1.2.1 *Robust Bayesian inference*

The estimation of parameters in Bayesian inference relies mainly on two components; the *likelihood function* and the *prior distribution*, from which posterior inferences can then be made on unknown parameters of the model, in our case θ_{m_k}.

In the particular case of $\mathbb{P}_{X|Y=m_k}$, the *likelihood function* is the product of conditional probabilities $\prod_i^{n_k} P_{x_{i,k}|y_{i,k},\theta_{m_k}}$ and the *prior distribution* $\mathbb{P}_{\theta_{m_k}}$ models our knowledge about $\theta_{m_k} = (\Sigma_{m_k}, \mu_{m_k})$. In this chapter, we focus on estimating imprecise ***mean parameters*** (i.e. $\theta_{m_k} = \mu_{m_k}$), assuming a (precise) estimation of $\hat{\Sigma}_{m_k}$, for reasons of computational complexity that will be discussed in Section 3.7. Thus, the posterior on the mean is such that

$$P(\mu_{m_k} \mid \mathscr{D}_{m_k}) \propto \prod_i^{n_k} P(X = \mathbf{x}_{i,k} \mid \mu_{m_k}, y_{i,k} = m_k) P(\mu_{m_k}). \tag{3.5}$$

To simplify, we will from now on remove the subscript m_k, always bearing in mind that these estimations are related to a group of observations labelled m_k.

3.1.2.2 *Near-ignorance on Gaussian discriminant analysis*

Near-ignorance models allow us to provide an *"objective inference"* approach, representing *ignorance about unknown parameter* and *letting the data speak for themselves*. In their work, Benavoli *et al* in [Benavoli et al., 2014] propose a new near-ignorance model based on a set of distributions $\mathscr{M}$, which aims to reconcile two approaches, namely, re-parametrization invariance and *Walley's* near-ignorance prior. For that, they define four minimal properties, which must be satisfied whenever there is no prior information about the unknown parameter, on the set of distributions $\mathscr{M}$ (more details in [Benavoli et al., 2014, §2]).

(P1) **Prior-invariance,** that states that $\mathscr{M}$ should be invariant under some re-parametrization of the parameter space (translation, scale, permutation, symmetry, etc).

(P2) **Prior-ignorance,** that states that $\mathscr{M}$ should be sufficiently large for reflecting a complete absence of prior information w.r.t. unknown parameters, but no too large to be incompatible with property **(P3)**.

(P3) **Learning from data,** that states that $\mathscr{M}$ should always provide non-vacuous posterior inferences, in other words, it should learn from the observations.

(P4) **Convergence,** that states that the influence of $\mathscr{M}$ on the posterior inference vanishes when increasing number of observations, i.e. $n \to \infty$, requiring consistency with the precise approach at limit.

Benavoli *et al* [Benavoli et al., 2014] provide a set of conjugate priors $\mathscr{M}$ for *regular multivariate exponential families* [Robert, 2005, §3.3.4] ($\mathscr{FE}xp$) that satisfies the last four properties under quite weak assumptions. Borrowing from [Benavoli et al., 2014], we can define this set of prior distribution $\mathscr{M}$ as follows:

Definition 8 (Prior near-ignorance for k-parameter exponential families [Benavoli et al., 2014, §4, eq. 16]) *Let $\mathbb{L}$ be a bounded closed convex subset of $\mathbb{R}^k$ strictly including the origin ([Benavoli et al., 2014, lem. 4.5]).*

$$\mathbb{L} = \left\{ \zeta \in \mathbb{R}^k : \zeta_i \in [-c_i, c_i], c_i > 0, i \in \{1, \dots, k\} \right\} \tag{3.6}$$

Let $W \in \mathscr{W} = \mathbb{R}^k$ be a random variable with probability density function defined, for all $\zeta_i \neq 0$, as:

$$p(w) = \exp(\zeta^T w) \prod_{i=1}^{k} \frac{\zeta_i}{\exp(\zeta_i r_i)} \mathbb{1}_{\mathscr{W}_{r_i}}(w_i) \tag{3.7}$$

with

$$\mathscr{W}_{r_i} = \begin{cases} (-\infty, r_i] & \text{if } \zeta_i > 0 \\ [r_i, \infty) & \text{if } \zeta_i < 0 \end{cases} \tag{3.8}$$

and where $\zeta, r \in \mathbb{R}^k$ are k-real values. Otherwise, for all $\zeta_i = 0$ the density $p(w)$ becomes a multivariate uniform distribution with $\mathscr{W}_{r_i} = [-r_i, r_i]$. Given an $\zeta \in \mathbb{L}$, it can be shown that the following set of prior distributions (c.f. [Benavoli et al., 2014, th. 4.6])

$$\mathscr{M}^w = \left\{ w \in \mathscr{W} \mid p(w) \propto \exp(\zeta^T w), \zeta = [\zeta_1, \dots, \zeta_k]^T \in \mathbb{L} \right\}, \tag{3.9}$$

satisfies (P1)-(P4) properties as well as conjugacy between the likelihood and the set of posterior distributions.

Since our Gaussian probability distribution $\mathbb{P}_{X|y=m_k}$ given by Equation (3.2) belongs to $\mathscr{FE}xp$, we can use for the mean a set $\mathscr{M}^\mu$ of prior distributions satisfying Equation (3.9), in order to get a set of posterior distributions $\mathscr{M}_n^\mu$ having the same functional form ($\mathscr{FE}xp$) [Bernardo et al., 2000, §5.2]:

$$\mathscr{M}_n^\mu = \left\{ (\mu | \bar{x}_n, \zeta) \propto \mathcal{N}\left(\frac{\zeta^T + n\bar{x}_n}{n}, \frac{1}{n}\hat{\Sigma} \right) \middle| \mu \in \mathbb{R}^p, \zeta \in \mathbb{L} \right\} \tag{3.10}$$

where $\bar{x}_n = \frac{1}{n}\sum_{i=1}^{n} x_i$ and $\zeta \in \mathbb{L}$. A sketch of how to get the Equation (3.10) is presented below (further details in [Benavoli et al., 2011, §4.1]).

Proof 1 (Normal prior-near ignorance conjugate) *Let* $\{x_1, \ldots, x_n\} \subseteq \mathscr{X} = \mathbb{R}^p$ *be a sample i.i.d. generated of a random multivariate normal* $\mathcal{N}(\mu, \Sigma)$*, so the likelihood based on a new re-parametrisation* $\eta = \Sigma^{-1}\mu$ *can be written:*

$$L(\mu|x_1, \ldots, x_n, \Sigma) \propto \exp\left\{n\left(\bar{x}_n^T\Sigma^{-1}\mu - \frac{1}{2}\mu^T\Sigma^{-1}\mu\right)\right\} \tag{3.11}$$

$$\propto \exp\left\{n\left(\bar{x}_n^T\eta - \frac{1}{2}\eta^T\Sigma\eta\right)\right\} \tag{3.12}$$

where log-partition function of previous equation is $b(\eta) = \frac{1}{2}\eta^T\Sigma\eta$ *and its derivative* $\nabla_\eta b(\eta) = \Sigma\eta = \mu$*, besides the prior near-ignorance proposed for* Benavoli at al *has the following form:*

$$p(\eta) \propto \exp\left\{\zeta^T\eta\right\}$$

making a transformation of the original parameter space $\eta = \Sigma^{-1}\mu$*, the last equation can be reduced to:*

$$p(\mu) \propto \exp\left\{\zeta^T\Sigma^{-1}\mu\right\}. \tag{3.13}$$

Therefore, we can calculate the posterior distribution combining the Equation (3.13) *and* (3.11)*:*

$$p(\mu|X, \Sigma) \propto \exp\left\{n\left(\bar{x}_n^T\Sigma^{-1}\mu - \frac{1}{2}\mu^T\Sigma^{-1}\mu\right)\right\}\exp\left\{\zeta^T\Sigma^{-1}\mu\right\} \tag{3.14}$$

$$\propto \exp\left\{-\frac{n}{2}\left[\mu^T\Sigma^{-1}\mu - 2\left(\frac{n\Sigma^{-1}\bar{x}_n + \Sigma^{-1}\zeta}{n}\right)^T\mu\right]\right\} \tag{3.15}$$

$$\propto \exp\left\{-\frac{n}{2}\left\|\mu - \frac{n\bar{x}_n + \zeta}{n}\right\|^2_{\Sigma^{-1}}\right\} \tag{3.16}$$

the posterior distribution is thus:

$$\mu|X, \Sigma \sim \mathcal{N}\left(\frac{n\bar{x}_n + \zeta}{n}, \frac{1}{n}\Sigma\right) \tag{3.17}$$

■

We can then estimate the lower and upper values of the unknown μ parameters, giving us for every dimension $i \in \{1, \ldots, p\}$ [Benavoli et al., 2014]:

$$\inf_{\mathscr{M}_n^\mu} \mathbb{E}[\mu_i|\bar{x}_n, \zeta] = \underline{\mathbb{E}}[\mu_i|\bar{x}_n, \zeta] = \frac{-c_i + n\bar{x}_n}{n} \tag{3.18}$$

$$\sup_{\mathscr{M}_n^{\mu}} \mathbb{E}[\mu_i|\overline{x}_n, \zeta] = \overline{\mathbb{E}}[\mu_i|\overline{x}_n, \zeta] = \frac{c_i + n\overline{x}_n}{n} \tag{3.19}$$

As a result, we will obtain for each label m_k a convex space of estimated values for the mean μ_{m_k} which can be represented by the hyper-cube

$$\mathbb{G}_{m_k} = \left\{ \hat{\mu}_{m_k} \in \mathbb{R}^p \,\middle|\, \hat{\mu}_{i,m_k} \in \left[\frac{-c_i + n_k \overline{x}_{i,n_k}}{n_k}, \frac{c_i + n_k \overline{x}_{i,n_k}}{n_k} \right], \forall i \in \{1, ..., p\} \right\}. \tag{3.20}$$

Remark 5 *The convergence property (P4) ensures us that no matter the initial value of our convex space $\mathbb{L}$, when the number of observations tends to infinity, $n \to \infty$, their influence on the posterior inference of $\hat{\mu}$ will disappear, i.e $\mathbb{G}_{m_k} \xrightarrow[n\to\infty]{} \overline{x}_n$, and will become the asymptotic estimator of the precise Gaussian distribution.*

On the basis of the set $\mathbb{G}_{m_k}$ previously calculated, we can simply consider the following set of conditional probability distributions $\mathscr{P}_{X|y=m_k}$ (or set of predictive distributions) for every label m_k on $\mathscr{K}$:

$$\mathscr{P}_{X|y=m_k} = \left\{ \mathbb{P}_{X|Y=m_k} \;\middle|\; \mathbb{P}_{X|Y=m_k} \sim \mathcal{N}(\mu_{m_k}, \hat{\Sigma}_{m_k}), \mu_{m_k} \in \mathbb{G}_{m_k} \right\} \tag{3.21}$$

In what follows, we study how we can incorporate the sets of distributions $\mathscr{P}_{X|Y=m_k}$ in Gaussian discriminant analysis, using maximality (Definition 5) to get our (possibly) *imprecise* classification.

3.2 IMPRECISE CLASSIFICATION WITH $\ell_{0/1}$ LOSS FUNCTION

Let us now discuss the operational aspects (inferences and computations) of our approach to make cautious classification by using sets of conditional distributions given by Equation (3.21) and obtained from a near-ignorance model. Using the maximality criterion, to know whether $m_a \succ_{\ell_{0/1}}^{\mathscr{P}_{Y|x}} m_b$, we need to solve Equation (2.12) by applying Bayes' theorem:

$$\inf_{\mathbb{P}_Y \in \mathscr{P}_Y} \inf_{\substack{\mathbb{P}_{X|m_a} \in \mathscr{P}_{X|m_a} \\ \mathbb{P}_{X|m_b} \in \mathscr{P}_{X|m_b}}} P(X = x|Y = m_a)P(Y = m_a) - P(X = x|Y = m_b)P(Y = m_b) > 0 \tag{3.22}$$

as the marginal $P(X = x)$ can be omitted from the denominator, being the same positive constant of normalisation for each probability.

As conditional distributions sets $\mathscr{P}_{X|Y=m_k}$ are independent of each others, we can rewrite Equation (3.22) as follows (cf. [Zaffalon, 2002, eq. 4.3]):

$$\inf_{\mathbb{P}_Y \in \mathscr{P}_Y} \underline{P}(X = x|Y = m_a)P(Y = m_a) - \overline{P}(X = x|Y = m_b)P(Y = m_b) > 0 \tag{3.23}$$

where $\underline{P}$ ($\overline{P}$) is the infimum (supremum) conditional probability. Also, applying Assumption 2 and the fact that every $\hat{\pi}_y > 0$, solving Equation (3.23) is reduced to finding the two values

$$\underline{P}(X = x|Y = m_a) = \inf_{\mathbb{P}_{X|m_a} \in \mathscr{P}_{X|m_a}} P(X = x|Y = m_a), \tag{3.24}$$

$$\overline{P}(X = x|Y = m_b) = \sup_{\mathbb{P}_{X|m_b} \in \mathscr{P}_{X|m_b}} P(X = x|Y = m_b) \tag{3.25}$$

As $\mathscr{P}_{X|y=m_k}$ is a set of Gaussian distributions, the solutions of Equations (3.24) and (3.25) are respectively obtained for the following values of the means

$$\underline{\mu}_{m_a} = \arg\inf_{\mu_{m_a} \in \mathbb{G}_{m_a}} -\frac{1}{2}(x - \mu_{m_a})^T \hat{\Sigma}_{m_b}^{-1}(x - \mu_{m_a}), \tag{3.26}$$

$$\overline{\mu}_{m_b} = \arg\sup_{\mu_{m_b} \in \mathbb{G}_{m_b}} -\frac{1}{2}(x - \mu_{m_b})^T \hat{\Sigma}_{m_b}^{-1}(x - \mu_{m_b}), \tag{3.27}$$

where $\hat{\Sigma}_{m_b}^{-1}$ is the inverse of the covariance matrix (Assumption 3). Depending on the internal structure of the *precise* covariance matrix $\hat{\Sigma}_k$, solving Equations (3.26) and (3.27) may be more or less computationally challenging. We will consider two main different imprecise discriminant models: (1) with non-diagonal covariance matrix and (2) with diagonal covariance matrix.

3.2.1 *Gaussian discriminant model with dependent features*

Similarly to the distinction made in the precise case, we will consider two different variants of the non-diagonal case.

Case 1 *Imprecise Quadratic discriminant analysis (IQDA): if we suppose that the covariance structures of all groups of observations are different, that is* $\hat{\Sigma}_{m_k} = \hat{S}_{m_k}, \forall m_k \in \mathscr{K}$.

Case 2 *Imprecise linear discriminant analysis (ILDA): if we assume that all groups of observations have the same covariance structure, that is* $\hat{\Sigma}_{m_k} = \hat{S}, \forall m_k \in \mathscr{K}$.

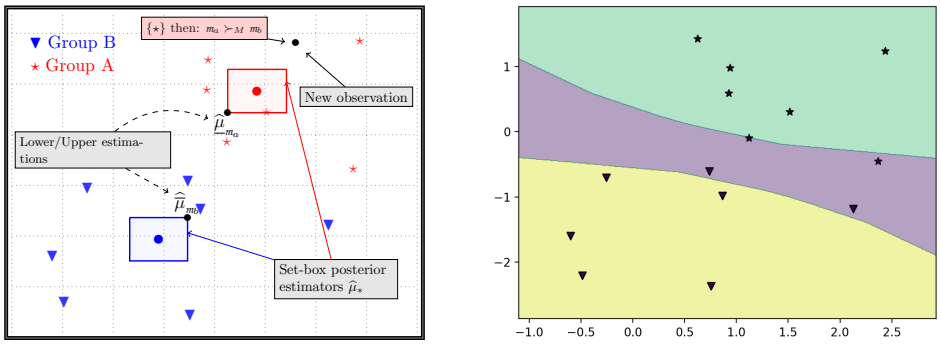
In those cases where the covariance matrix contains collinear columns, $\hat{\Sigma}_{m_k}$ will not be invertible, in which case we use the *singular value decomposition* (SVD) method for computing the *pseudo-inverse* of covariance matrix. Before studying the computational aspects of IQDA and ILDA, i.e. Equations (3.26) and (3.27), we will illustrate the last case (ILDA) in Example 7.

Example 7 *The interest of modelling an imprecise mean is to be able to detect areas where we should be cautious and predict sets of labels rather than a single one. For example, in Figure 3.1, we simulated two groups of observations $x_{m_a *}$ and $x_{m_b *}$ (i.e. binary case), each with two non-correlated regressors and different means:*

$$\begin{pmatrix} x_{m_a 1} \\ x_{m_a 2} \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} 0.25 \\ 0.5 \end{pmatrix}, \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}\right)$$
$$\begin{pmatrix} x_{m_b 1} \\ x_{m_b 2} \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} 0.5 \\ -1.0 \end{pmatrix}, \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}\right)$$
$$\mathbb{L} = \left\{ \zeta \in \mathbb{R}^2 : \zeta_i \in [-c_i, c_i], c_i = 2 \right\}$$

Figure 3.1a illustrates this example and pictures the following things: groups of observations $x_{m_a *}$ and $x_{m_b *}$ with the symbols $\star$ and $\blacktriangledown$, respectively, and the posterior convex estimates $\mathbb{G}$ (solid) of the means after injecting the information contained in the training data.

We also drew the (precise) mean of each group, i.e. μ_{m_a} and μ_{m_b}, as solid points, and a black dot ($\bullet$) representing a new unlabelled instance $\mathbf{x}$ as well as positions of solutions of Equations (3.26) and (3.27). In Figure 3.1, we observe (in purple) an area of uncertainty, where both labels are incomparable, generated by the imprecise mean and the maximality criterion.



(a) Upper/lower estimation of μ_k **(b)** Imprecise decision area

Figure 3.1: *Imprecise boundary area and estimation*. Figure 3.1a shows an example of the imprecise estimation of means μ_*, and Figure 3.1b shows an imprecise decision area of purple colour where the subset $\hat{\mathbb{Y}} = \{m_a, m_b\}$ of labels is the imprecise decision, that is in this region m_a and m_b are incomparable.

Let us now discuss the problem of solving Equations (3.26) and (3.27). Expressing $\mathbb{G}_{m_b}$ as constraints, the solution $\overline{\mu}_{m_b}$ of Equation (3.27) can be written as

$$\begin{aligned}\overline{\mu}_{m_b} = \arg\sup \; & -\frac{1}{2}\hat{\mu}_{m_b}^T \hat{\Sigma}_{m_b}^{-1}\hat{\mu}_{m_b} + q^T\hat{\mu}_{m_b} \\ \text{s.t. } & \frac{-c_j + n_{m_b}\overline{x}_{j,n_{m_b}}}{n_{m_b}} \leqslant \hat{\mu}_{j,m_b} \leqslant \frac{c_j + n_{m_b}\overline{x}_{j,n_{m_b}}}{n_{m_b}}, \\ & q^T = x^T\hat{\Sigma}_{m_b}^{-1}, \quad \forall j \in \{1,...,p\}\end{aligned} \tag{BQP}$$

This optimisation problem is well known as a box-constraint quadratic program (BQP) [De Angelis et al., 1997], as (1) the constraint space $\mathbb{G}_{m_k}$ is a convex space, and (2) $\hat{\Sigma}_{m_k}^{-1}$ is a positive (semi)-definite matrix, pending the fact that the covariance matrix $\hat{\Sigma}_{m_k}$ does not have multicollinearity problems [Johnson, 1970]. Computing an optimal global solution of (BQP) in polynomial time is easy using modern optimisation libraries (e.g. using the CvxOpt python library [Andersen et al., 2018]), as we have to maximize a concave function (or, equivalently, minimise a convex one).

Finding $\underline{\mu}_{m_a}$ in Equation (3.26) is much more difficult, as one seeks to solve the optimization problem

$$\underline{\mu}_{m_a} = \underset{\hat{\mu}_{m_a}\in\mathbb{G}_{m_a}}{\arg\inf} \; -\frac{1}{2}\mu_{m_a}^T \hat{\Sigma}_{m_a}^{-1}\hat{\mu}_{m_a} + q^T\hat{\mu}_{m_a}. \tag{NBQP}$$

That comes down this time to maximizing a convex function over box-constraints ($\mathbb{G}_{m_a}$), which is known to be NP-Hard [Pardalos et al., 1991]. To solve it, we use a brand-and-bound (B&B) algorithm [Burer et al., 2009; Xia et al., 2015], that employs a finite branching based on the first-order Karush-Kuhn-Tucker[1] conditions and polyhedral semidefinite relaxation in each node of the B&B tree (more details in [Burer et al., 2009]).

3.2.2 *Gaussian discriminant model with independence features*

When the number of features becomes high, and the associated optimisation problem quite time consuming to solve, it may be interesting to consider some additional assumptions which will significantly reduce the inference complexity. In what follows, we will assume that features x_i^j are independent conditional on the label m_k. This translates in the fact that covariance matrices become diagonal matrices, i.e. $\Sigma_{m_k} = \boldsymbol{\sigma}_{m_k}^T \mathbb{I}$ with $\boldsymbol{\sigma}_{m_k}^T = (\sigma_{m_k}^1, \ldots, \sigma_{m_k}^p)$ a p-dimensional vector containing the variance of each feature, which can be interpreted as weights of the features. Therefore, we can rewrite Equations (3.26) and (3.27) as follows:

$$\underline{\mu}_{m_a} = \underset{\mu_{m_a}\in\mathbb{G}_{m_a}}{\arg\inf} -\frac{1}{2}\boldsymbol{w}_{m_a}\left\|x - \mu_{m_a}\right\|^2, \tag{3.28}$$

[1] Also known as KKT, which allows to solve problems of optimisation subject to non-linear constraints in the form of inequalities.

$$\overline{\mu}_{m_b} = \arg \sup_{\mu_{m_b} \in \mathbb{G}_{m_b}} -\frac{1}{2} w_{m_b} \|x - \mu_{m_b}\|^2 \tag{3.29}$$

where $w_{m_k} = (w^1_{m_k}, \dots, w^p_{m_k})^T$ such that $w^j_{m_k} = 1/\sigma^j_{m_k}, \forall j \in \{1, \dots, p\}$, in this scenario, we will consider two new models.

Case 3 *Imprecise naive discriminant analysis (INDA): this case is similar to the Naive Bayes classifier, as we simply consider the assumption* $\Sigma_{m_k} = \hat{\sigma}^T_{m_k} \mathbb{I}$ *where* $\hat{\sigma}_{m_k}$ *are the empirical variance estimator obtained from a group of observation belonging to the label* m_k.

Case 4 *Imprecise Euclidian discriminant analysis (IEDA): this is a case more specific than INDA, where we assume that for every* $j \in \{1, \dots, p\}$ *we have* $\hat{\sigma}^j_{m_k} = 1$, *meaning that the measure used to evaluate the probability of a label given a new instance is proportional to the Euclidian distance between the instance and the corresponding mean. The Euclidean classifier is one of the simplest existing classifier, and is the supervised counterpart of the standard k-means method.*

We show below that when the covariance matrix is diagonal, optimisation problems (3.28) and (3.29) become very easy (i.e. linear in p, $\mathcal{O}(p)$) to solve.

Proposition 1 *For two vectors* $x, w \in \mathbb{R}^p$, *and a box-convex space on* $\mathbb{R}^p$:

$$\mathbb{G} = \left\{ \mu \in \mathbb{R}^p \;\middle|\; \mu^j \in [\underline{\mu}^j, \overline{\mu}^j], \forall j \in \{1, \dots, p\} \right\}$$

- *the infimum weighted distance subject to constraints* $\mathbb{G}$ *is:*

$$\inf_{\mu \in \mathbb{G}} -\frac{1}{2} w^T \|x - \mu\|^2 = -\frac{1}{2} \sum_j^p w^j \max_j \{(x^j - \underline{\mu}^j)^2, (x^j - \overline{\mu}^j)^2\} \tag{3.30}$$

- *and the supremum weighted distance subject to same constraints is:*

$$\sup_{\mu \in \mathbb{G}} -\frac{1}{2} w^T \|x - \mu\|^2 = -\frac{1}{2} \sum_j^p w^j \begin{cases} 0 & \text{if } x^j \in [\underline{\mu}^j, \overline{\mu}^j] \\ \min_j \left\{ \begin{matrix} (x^j - \underline{\mu}^j)^2, \\ (x^j - \overline{\mu}^j)^2 \end{matrix} \right\} & \text{otherwise} \end{cases} \tag{3.31}$$

Proof 2 (Proof of Proposition 1) *Since each element of the sum is positive, we can interchange the infimum operator with summation, and calculate the supremum of each component as follows:*

$$\inf_{\mu \in \mathbb{G}} -\frac{1}{2} \sum_j^p w^j (x^j - \mu^j)^2 \quad \Longleftrightarrow \quad -\frac{1}{2} \sum_j^p w^j \sup_{\mu^j \in [\underline{\mu}^j, \overline{\mu}^j]} (x^j - \mu^j)^2 \tag{3.32}$$

where the supremum can be calculated as follows:

$$\sup_{\mu^j \in [\underline{\mu}^j, \overline{\mu}^j]} (x^j - \mu^j)^2 = \max_j \{(x^j - \underline{\mu}^j)^2, (x^j - \overline{\mu}^j)^2\} \tag{3.33}$$

In the second case and for similar reasons, we can also put the supremum operator inside of summation and calculate of infimum value of each component:

$$\sup_{\mu \in G} -\frac{1}{2} \sum_j^p w^j (x^j - \mu^j)^2 \quad \Longleftrightarrow \quad -\frac{1}{2} \sum_j^p w^j \inf_{\mu^j \in [\underline{\mu}^j, \overline{\mu}^j]} (x^j - \mu^j)^2 \tag{3.34}$$

where the infimum of squared subtraction of each element is:

$$\inf_{\mu^j \in [\underline{\mu}^j, \overline{\mu}^j]} (x^j - \mu^j)^2 = \begin{cases} 0 & \text{if } x^j \in [\underline{\mu}^j, \overline{\mu}^j], \\ \min\{(x^j - \underline{\mu}^j)^2, (x^j - \overline{\mu}^j)^2\} & \text{otherwise.} \end{cases} \tag{3.35}$$

■

3.3 EXPERIMENTAL SETTING

Now that we have computational means to learn and infer from our model, we provide experimental results evaluating the performance of our different imprecise Gaussian discriminant models (cf. Section 3.2).

3.3.1 *How can we choose parameter c_i?*

The choice of parameters c_i determines the amount of imprecision in our posterior inference. It should be large enough to guarantee more reliable predictions when missing information, but small enough so as to provide informative predictions when possible. Therefore, in the absence of prior information and for symmetry reasons, we will consider a symmetric box around 0, as follows:

$$\mathbb{L}' = \left\{ \zeta \in \mathbb{R}^k : \zeta_i \in [-c, c], c > 0, i = \{1, \ldots, p\} \right\}. \tag{3.36}$$

In order to fix a value of c, there exist different approaches already mentioned in Section 4.3 of [Benavoli et al., 2014]. One can for example rely on the rate of convergence of the lower and upper posterior expectations [Walley, 1991]:

$$\forall i \quad \left(\overline{E}[\mu_i | \overline{x}_n, \zeta] - \underline{E}[\mu_i | \overline{x}_n, \zeta]\right) = \frac{2c}{n} \xrightarrow[n \to \infty]{} 0 \tag{3.37}$$

meaning that for small values of c, we would reach a faster convergence of Equation (3.37) to a *precise* posterior inference (as precise models). A value of $c \leqslant 0.75$ is recommended [Benavoli et al., 2014, §4.3, §8]. However

since we are in a classification problem, c will be selected through cross-validation. More precisely, we restrict c to the interval $[0.01, 5]$, discretised into $[0.01, 0.02, ..., 5]$, with the optimal value decided by cross validation on the training samples. A typical empirical evolution of the accuracy measures used in the next sections is shown in Figure 3.4 for the four IGDA methods. It clearly shows that performances first increase in average with imprecision, but then degrades as imprecision becomes too large.

3.3.2 *Data sets and experimental setting*

We perform experiments on 12 data sets issued from UCI machine repository [Frank et al., 2010](cf. Table 3.2), following a 10×10-fold cross-validation procedure. We aim to compare the performance of our imprecise Gaussian classifier model approach with the existing precise models (c.f. Table 3.1).

#	name	# instances	# features	# labels
a	iris	150	4	3
b	wine	178	13	3
c	forest	198	27	4
d	seeds	210	7	3
e	glass	214	9	6
f	ecoli	336	7	8
h	dermatology	385	34	6
i	vehicle	846	18	4
j	vowel	990	10	11
k	yeast	1484	8	12
l	wine quality	1599	11	6
n	wall-following	5456	24	4

Table 3.2: Data sets used in the experiments

Owing to small amounts of samples in some groups of observations (belonging to a specific label m_k) of some data sets, the QDA model can suffer from a phenomenon known as ill-posed covariance matrix (i.e. $n_{m_k} < p$), and in such cases even calculating the pseudo-inverse of the estimated covariance matrix $\hat{\Sigma}_{m_m}$ using SVD method cannot solve the problem. This affects the performance of our classifiers that significantly drop (e.g. in Table 3.3, glass and yeast data sets). Therefore, in these specific cases, we used a basic regularized method for the estimated covariance matrix named Regularization QDA (or RQDA)[Friedman, 1989; Friedman et al., 2001]:

$$\Sigma_{m_k}(\alpha) = \alpha\hat{\Sigma}_{m_k} + (1-\alpha)\mathbb{I}, \tag{3.38}$$

where $\hat{\Sigma}_{m_m}$ is the estimated covariance for a group of observations, $\mathbb{I}$ a identity matrix and α the regularization factor.

With respect to ecoli data set, we consider appropriate to take the *imS* and *imL* labels out of the data set, because they only have two instances by label, making it impossible to perform the cross-validation procedure for IQDA and QDA models, as we cannot calculate an empirical covariance matrice with a single instance.

Comparing indeterminate predictions given in the form of a subset $\hat{\mathbb{Y}}$ of plausible labels against just one determinate prediction $\hat{y}$ is a hard problem that mostly depends on the circumstances or the context in which a decision-marker may or may not accept partial predictions (or cautious decision) instead of a unique, risky decision. A good evaluation should reward cautiousness provided by $\hat{\mathbb{Y}}$ when it allows to include the true observed label, but not so much as to systematically privilege imprecision over precision. In other words, we need an evaluation metric that seeks a compromise between cautiousness and informativeness. To do this, we adopt the evaluation metric proposed and theoretically justified in [Zaffalon et al., 2012], called *utility-discounted accuracy*, which makes it possible to reward the imprecision in a more or less strong way. It is written as follows:

$$u(y, \hat{\mathbb{Y}}) = \begin{cases} 0 & \text{if } y \notin \hat{\mathbb{Y}}, \\ \frac{\alpha}{|\hat{\mathbb{Y}}|} - \frac{\alpha-1}{|\hat{\mathbb{Y}}|^2} & \text{otherwise.} \end{cases} \tag{3.39}$$

[Zaffalon et al., 2012] shows that a value $\alpha = 1$ amounts to not reward cautiousness and to confuse it with randomness, while $\alpha \to \infty$ does not penalize non-informativeness, as the vacuous prediction (i.e. $\hat{\mathbb{Y}} = \mathscr{K}$) would always get a full, guaranteed reward. We will use the usual values u_{65} with $\alpha = 1.6$ and u_{80} with $\alpha = 2.2$ (as in [Yang et al., 2017]). To have an intuition about these measures, let us simply recall that the u_{65} (u_{80}) measure rewards a binary correct prediction with 0.65 (0.80), while a purely random, non-cautious guesser picking one of the two possible labels would reward it with 0.50. It therefore gives a "reward" of 0.15 (0.30) for rightful cautiousness.

3.3.3 *Experimental results*

The average results obtained according to u_{65} and u_{80} utilities, and the average execution time to predict the label of a new unlabeled instance are shown in Table 3.3. It should be noted that, to compute the predictor $\hat{\mathbb{Y}}$, we used an algorithm detailed in Section 3.6 and inspired by the work of Nakahrutai *et al.*[Nakharutai et al., 2019].

It should be noted that while allowing for imprecision gives more flexibility in terms of prediction than standard, precise methods, using u_{65} and u_{80} may either penalize or reward such flexibility in the final accuracy. Indeed, if the imprecision is added to an instance for which the precise

	NDA	INDA		Avg. Time
#	acc.	u_{80}	u_{65}	
a	95.07 ± 0.44	$\mathbf{95.87 \pm 3.58}$	95.77 ± 3.71	1.25×10^{-3}
b	$\mathbf{97.70 \pm 0.58}$	96.42 ± 22.99	96.34 ± 5.55	1.79×10^{-3}
c	95.26 ± 0.33	$\mathbf{95.42 \pm 3.78}$	95.42 ± 3.78	1.42×10^{-3}
d	90.38 ± 0.19	$\mathbf{90.95 \pm 3.25}$	90.57 ± 3.66	1.30×10^{-3}
e	43.92 ± 1.36	49.17 ± 13.83	$\mathbf{52.30 \pm 13.89}$	2.03×10^{-3}
f	$\mathbf{82.39 \pm 1.22}$	61.91 ± 22.96	61.87 ± 23.03	1.65×10^{-3}
h	85.52 ± 0.98	$\mathbf{92.70 \pm 3.37}$	92.38 ± 3.43	1.76×10^{-3}
i	45.63 ± 0.89	$\mathbf{46.50 \pm 9.71}$	42.68 ± 10.78	0.89×10^{-3}
j	67.26 ± 0.39	$\mathbf{73.59 \pm 3.67}$	68.57 ± 3.61	1.45×10^{-3}
k	43.36 ± 0.51	$\mathbf{48.57 \pm 5.67}$	48.49 ± 5.74	1.11×10^{-3}
l	54.83 ± 0.34	$\mathbf{62.10 \pm 4.44}$	58.52 ± 2.92	0.90×10^{-3}
n	52.55 ± 0.12	$\mathbf{52.74 \pm 1.35}$	52.64 ± 1.34	0.61×10^{-3}
avg.	71.16 ± 0.61	72.16 ± 8.21	71.30 ± 6.79	1.35×10^{-3}

(a) NDA versus INDA

	EDA	IEDA		Avg. Time
#	acc.	u_{80}	u_{65}	
a	91.60 ± 0.61	$\mathbf{95.20 \pm 5.56}$	93.40 ± 6.37	0.26×10^{-3}
b	46.65 ± 0.85	$\mathbf{61.48 \pm 6.11}$	51.45 ± 5.91	0.37×10^{-3}
c	81.09 ± 0.39	$\mathbf{83.40 \pm 5.15}$	80.27 ± 5.82	0.71×10^{-3}
d	90.38 ± 0.36	$\mathbf{89.90 \pm 4.19}$	89.48 ± 4.03	0.29×10^{-3}
e	46.26 ± 1.68	$\mathbf{55.91 \pm 5.91}$	47.31 ± 6.68	0.64×10^{-3}
f	42.59 ± 0.04	$\mathbf{43.11 \pm 11.44}$	41.93 ± 13.43	0.76×10^{-3}
h	51.22 ± 0.92	$\mathbf{55.08 \pm 9.56}$	52.99 ± 10.53	1.11×10^{-3}
i	28.03 ± 0.19	$\mathbf{46.89 \pm 2.10}$	36.93 ± 2.35	0.53×10^{-3}
j	58.08 ± 0.90	$\mathbf{64.65 \pm 5.50}$	60.73 ± 6.79	1.04×10^{-3}
k	31.27 ± 0.13	$\mathbf{31.30 \pm 2.40}$	31.28 ± 2.38	0.95×10^{-3}
l	19.72 ± 0.19	$\mathbf{23.76 \pm 1.84}$	22.06 ± 3.56	0.61×10^{-3}
n	57.90 ± 0.11	$\mathbf{58.65 \pm 1.36}$	58.26 ± 1.34	0.55×10^{-3}
avg.	53.73 ± 0.53	59.11 ± 5.09	55.51 ± 5.77	0.65×10^{-3}

(b) EDA versus IEDA

Table 3.3: Average utility-discounted accuracies (%) and time to predict in seconds.

model was right, the imprecise model will be penalized, as the reward will go from 1 (for the precise model) to a lower value given by Equation (3.39). On the contrary, if the imprecise prediction adds the true label to a wrong precise prediction, the score of this prediction will go from zero (for the precise model) to a positive value. Hence a higher accuracy for u_{65} and u_{80} means that, on average, the additional imprecision (1) concerns instances for which the precise method was wrong and (2) allows to include the true class within the prediction. Of course, since $u_{80} > u_{65}$, it will in the majority of times achieve a higher accuracy, albeit not always (e.g., data set glass for the INDA model, line e of Table 3.3a).

Given this, we can see that including some cautiousness can increase our accuracies on most data sets, by picking the right values of c. This increase is sometimes noticeable, for example in the vehicle (i), wine-quality (l), wall-following (o) and vowel data sets (j). All of this, keeping a time

#	LDA	ILDA		Avg. Time
	acc.	u_{80}	u_{65}	
a	97.96 ± 0.05	$\mathbf{98.38 \pm 0.21}$	97.16 ± 0.27	0.56
b	98.85 ± 0.36	$\mathbf{98.99 \pm 1.17}$	98.95 ± 1.26	1.49
c	94.61 ± 0.60	$\mathbf{94.56 \pm 1.08}$	94.05 ± 1.02	12.14
d	96.35 ± 0.25	$\mathbf{96.59 \pm 0.23}$	96.51 ± 0.23	1.50
e	62.15 ± 0.76	$\mathbf{66.78 \pm 0.73}$	58.87 ± 0.77	17.59
f	87.14 ± 0.37	$\mathbf{89.74 \pm 1.43}$	88.23 ± 1.42	12.40
h	96.58 ± 0.35	$\mathbf{97.06 \pm 0.62}$	96.94 ± 0.61	19.24
i	77.96 ± 0.48	$\mathbf{81.98 \pm 0.91}$	79.59 ± 0.82	3.10
j	60.10 ± 0.68	$\mathbf{67.45 \pm 0.48}$	62.41 ± 0.40	4.95
k	58.92 ± 0.17	$\mathbf{61.50 \pm 3.09}$	59.20 ± 3.37	10.81
l	59.25 ± 0.27	$\mathbf{65.83 \pm 0.26}$	60.31 ± 0.63	34.85
n	67.96 ± 0.07	$\mathbf{71.34 \pm 0.23}$	66.65 ± 0.19	10.77
avg.	79.82 ± 0.37	82.52 ± 0.87	79.91 ± 0.92	10.78

(a) LDA versus ILDA

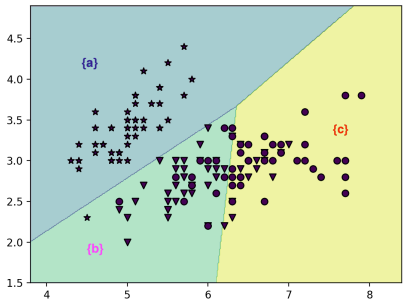
#	QDA	RQDA	IQDA		Avg. Time
	acc.	acc.	u_{80}	u_{65}	
a	97.29 ± 0.44	96.66 ± 4.47	$\mathbf{98.08 \pm 0.41}$	97.13 ± 0.42	0.71
b	99.03 ± 0.45	98.89 ± 2.22	$\mathbf{99.39 \pm 0.14}$	99.09 ± 0.13	2.94
c	89.43 ± 1.34	$\mathbf{97.47 \pm 3.37}$	91.77 ± 1.38	88.90 ± 1.32	6.54
d	94.64 ± 0.47	94.29 ± 2.86	$\mathbf{95.20 \pm 0.26}$	94.72 ± 0.24	1.52
e	7.15 ± 2.39	51.40 ± 9.79	$\mathbf{64.38 \pm 1.36}$	58.36 ± 1.30	14.74
f	46.19 ± 2.97	$\mathbf{88.25 \pm 5.97}$	87.57 ± 3.99	87.12 ± 4.54	5.01
h	82.47 ± 0.42	$\mathbf{96.92 \pm 0.88}$	84.24 ± 0.87	84.05 ± 0.88	26.27
i	85.07 ± 0.86	85.11 ± 2.63	$\mathbf{87.96 \pm 0.34}$	86.13 ± 0.27	3.17
j	87.83 ± 0.49	87.07 ± 3.49	$\mathbf{89.96 \pm 0.67}$	88.40 ± 0.70	3.60
k	13.18 ± 2.37	$\mathbf{56.27 \pm 2.29}$	49.28 ± 5.02	48.34 ± 4.90	10.85
l	55.62 ± 0.47	55.79 ± 5.35	$\mathbf{65.85 \pm 0.55}$	60.36 ± 0.62	28.05
n	65.87 ± 0.17	70.56 ± 2.63	$\mathbf{71.79 \pm 0.12}$	69.75 ± 0.12	9.32
avg.	68.65 ± 1.07	81.56 ± 3.83	82.12 ± 1.26	80.20 ± 1.29	9.39

(b) QDA versus IQDA

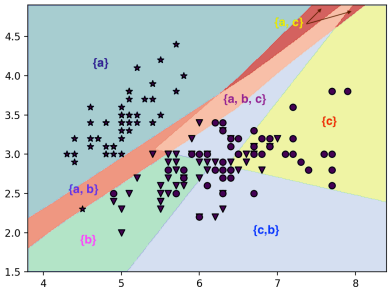
Table 3.4: Average utility-discounted accuracies (%) and time to predict in seconds.

execution reasonable in view of the problems to be solved (e.g. a non-convex, NP-hard problem), and without an optimized implementation. As expected, assuming independence between the features (i.e., diagonal covariance matrices) significantly reduces the computational time, making it negligible, but overall reduces performances, as the assumptions are often violated in a stronger way.

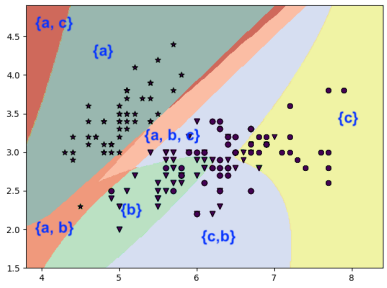
In order to highlight the major role of cautiousness of an imprecise classifier model, we show in Figure 3.2c and 3.2b how, in the IRIS data set, our IQDA and ILDA models create different areas of decision boundaries (not to be confused with rejection area), where each area has a different combination of subset of labels $\hat{\mathbb{Y}} \subseteq \mathscr{K}$, in contrast to precise classifier model (LDA), in Figure 3.2a, where it creates one area per label. We can clearly see that the two classifiers behave quite differently. In particular,



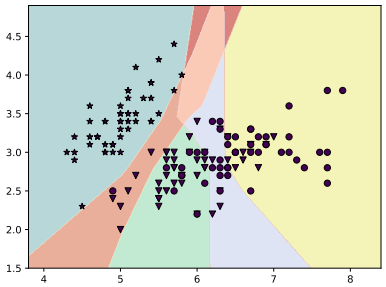
(a) Precise boundary decision



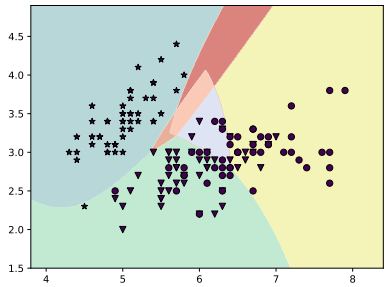
(b) ILDA zone decision



(c) IQDA zone decision



(d) IEDA zone decision



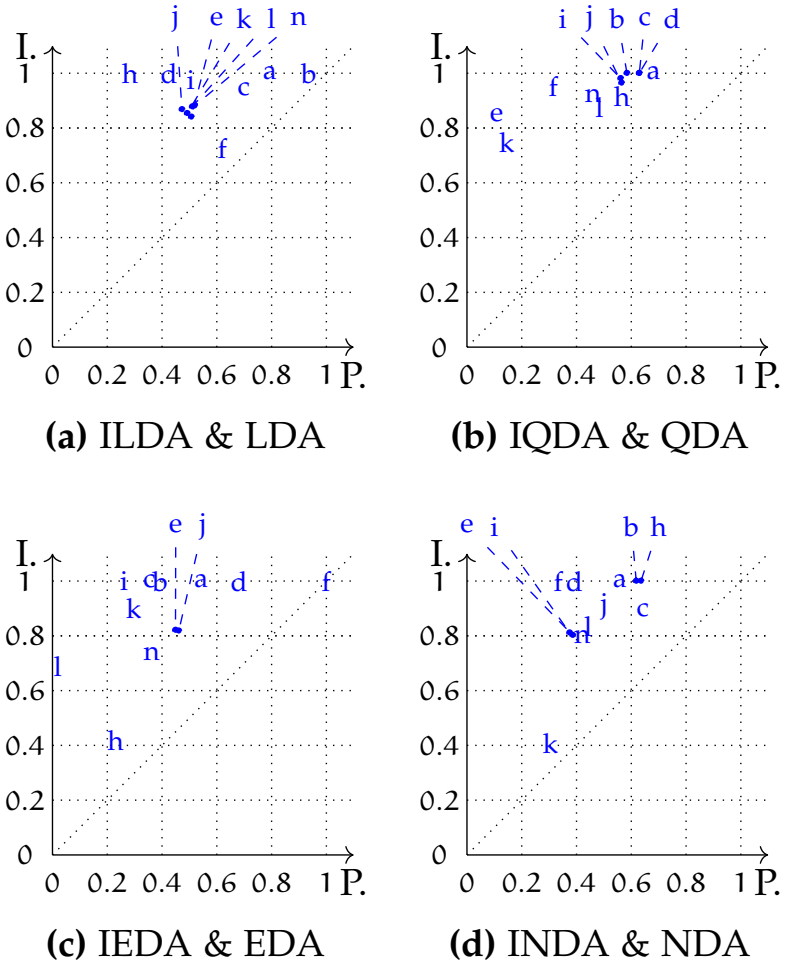
(e) INDA zone decision

Figure 3.2: Figure 3.2a shows how a precise model divides the instance space in three single different zones by label (i.e $\{a\},\{b\},\{c\}$), the Figure 3.2b shows how an ILDA model divides the instance space in different zones as much as different combinations of a subset of labels (i.e $\{a\},\{b\},\{c\},\{a,b\},\{b,c\}$, and so on), Figure 3.2c shows how IQDA model can also divide in different zones with smooth curves instead, Figure 3.2d shows IEDA model, and finally, Figure 3.2e shows INDA model.

ILDA (and IEDA) will induce regions delimited by piece-wise linear functions, while IQDA (and INDA) will induces regions delimited by piece-wise quadratic functions.

Also, in Figure 3.4, we show the evolution of utility-discounted accuracy (i.e. u_{65} and u_{80} of vowel dataset), with a standard deviation calculated by a 10-fold cross-validation on the training dataset, according to the imprecision of estimators μ. As expected we notice that when c reaches a too high value, the overall model performances decrease, as it becomes

too imprecise with respect to our attitude towards cautiousness (modelled through utility (3.39)). The rest of experiments are in Appendix A.1.



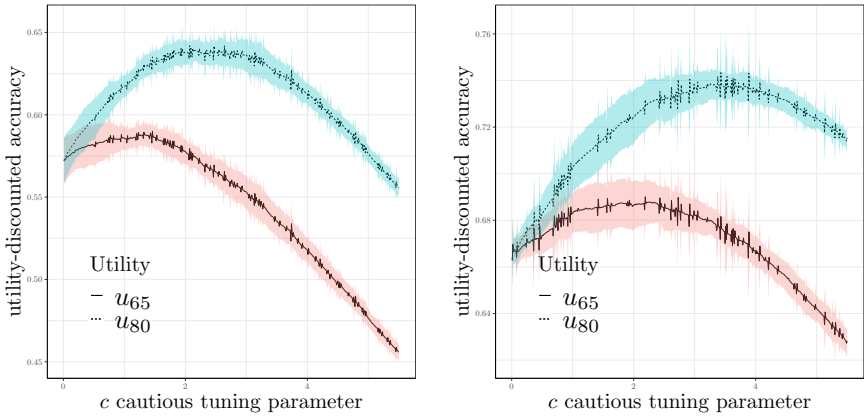
(a) ILDA & LDA **(b)** IQDA & QDA

(c) IEDA & EDA **(d)** INDA & NDA

Figure 3.3: Correctness of the different methods in the case of abstention versus accuracy of their precise counterparts, only on those instances for which an indeterminate prediction was given. Graphs are given for the u_{80} accuracies.

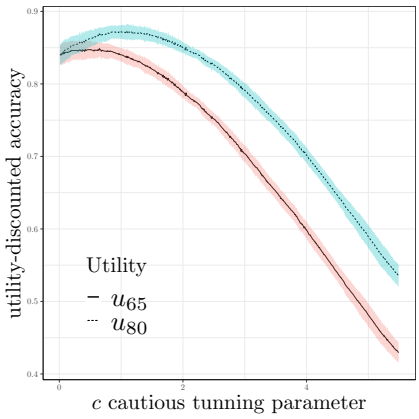
Another desirable feature of an imprecise classifier is that it should abstain (i.e. by providing a set of plausible choices) on hard instances, that is the instances where the *precise* classifier makes an unusual high amount of mistakes. In Figure 3.3, we verify that our imprecise classifiers follow this desirable behaviour on most data sets, for the u_{80} measures (conclusions for the u_{65} are similar, but not displayed to gain some space). Figure 3.3a displays the percentage of time the true label is in the prediction of ILDA, given that the prediction was imprecise, versus the accuracy of LDA on those same instances. The same graphs for the QLDA, IEDA and INDA methods are given by Figure 3.3b, Figure 3.3c and Figure 3.3d, respectively. We notice that on those hard instances where precise classifiers are

wrong, our imprecise classifiers successfully overcome them, getting the ground-truth value into partial predictions (most often $> 80\%$). A typical and quite remarkable example of this is the dermatology data set (h) for the linear case, where the accuracy on the imprecisely classified instances drop to 30% for the precise classifier (to be compared to an average of 96% on all instances), while the imprecise classifier always includes the true class. Moreover, the fact that u_{80} is higher indicates that the overall amount of imprecision remains acceptable. Our approach therefore seems to be able to well robustify the very simple, linear decision frontiers of the ILDA models.

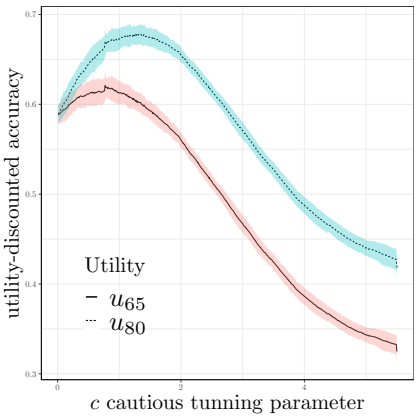


(a) IEDA MODEL

(b) INDA MODEL



(c) IQDA MODEL



(d) ILDA MODEL

Figure 3.4: Figures shows performance evolution with a standard deviation region of three principales methods, (1) Figure 3.4d for ILDA model, (2) Figure 3.4c for IQDA model, (3) Figure 3.4b for INDA model, and (4) Figure 3.4a for IEDA model, w.r.t. utility-discount accuracy u_{65}, u_{80} and c tuning parameter on vowel dataset

Before considering some generalisation of the presented methods, we would also like to mention that the imprecise probabilistic approach will in

general induces decision frontiers that are different from classical rejection rule. Figure 3.1b illustrates this well: rejection regions in a binary setting are most often equivalent to require to predict $\{a, b\}$ whenever $\hat{P}(\{a\}|x) \in [0.5 - \epsilon, 0.5 + \epsilon]$ for some ϵ. This means that in the case of LDA, the rejection regions will be delimited by two parallel lines, corresponding to the iso-density points $\mathbf{x}$ for which $\hat{P}(\{a\}|\mathbf{x}) = 0.5 - \epsilon$ and $\hat{P}(\{a\}|\mathbf{x}) = 0.5 + \epsilon$. In contrast, we can clearly see in Figure 3.1b that the boundaries are not linear, but piece-wise linear.

3.4 IMPRECISE PRIOR MARGINAL AND GENERIC LOSS FUNCTIONS

In this section, we will discuss two new variants of IGDA model: (1) relaxing Assumption 2, i.e. $\mathbb{P}_Y := \hat{\pi}$, with the purpose of putting a set of probability distributions $\mathscr{P}_Y$ instead, and (2) dealing with generic loss functions instead of the classical $\ell_{0/1}$ loss function. We will evaluate the impact of this two new variants in our IGDA model in terms of added computational complexity.

3.4.1 *Imprecise prior marginal*

The first extension we will consider is to make imprecise the marginal distribution, considering a set $\mathscr{P}_Y$ rather than a precise distribution, in the same vein as we have made the conditional distribution $\mathscr{P}_{X|Y}$ imprecise. For the time being, we will still work with the $\ell_{0/1}$ loss function. Since the conditionals are still independent of each other, solving the maximality criterion amounts to solve Equation (3.23), that we recall here

$$\inf_{\mathbb{P}_Y \in \mathscr{P}_Y} \underline{P}(X = \mathbf{x}|Y = m_a)P(Y = m_a) - \overline{P}(X = \mathbf{x}|Y = m_b)P(Y = m_b) \tag{3.40}$$

with $m_a \succ^{\mathscr{P}_{Y|\mathbf{x}}}_{\ell_{0/1}} m_b$ if this is positive. This equation can be solved easily, as it is a linear form in $P(Y = m_a), P(Y = m_b)$, meaning that we can either use linear programming over the constraints induced by $\mathscr{P}_Y$, or find the extreme point (e.g., by enumeration) of $\mathscr{P}_Y$ for which the solution is obtained.

The problem then amounts to estimate $\mathscr{P}_Y$. A quite popular choice to do so is to use an Imprecise Dirichlet Model (IDM) [Bernard, 2005]. However, as Benavoli *et al.* have already mentioned in [Benavoli et al., 2014, §4.2], the set of prior distributions of IDM does not correctly satisfy (P1) Prior-invariance property and permutation invariance of near-ignorance model. So, to remain consistent with our previous estimates, we retain a solution proposed by Benavoli *et al.*.

Let Y be a discrete random variable on a finite space of labels $\mathscr{K}$ with probability distribution $\mathbb{P}_Y$ and let the parameters $\pi_{m_k}, \forall m_k \in \mathcal{K}$ be the

unknown non-negative chances, i.e. $P(Y = m_k)$. The Corollary 4.10 in [Benavoli et al., 2014] proposes adding some constraints in the space $\mathbb{L}$ in order not to favour some chances π_{m_k} over others. They then consider the following set of prior distributions:

$$\mathscr{P}_\pi = \left\{ \pi_{m_1}^{\zeta_1 - 1} \pi_{m_2}^{\zeta_2 - 1} \dots \pi_{m_K}^{-\sum_{i=1}^{K-1} \zeta_i - 1}, \|\zeta\|_1 \leqslant 2c, \sum_{i=1}^{K-1} \zeta_i \in [-c, c], K = |\mathscr{K}| \right\}. \tag{3.41}$$

It is also shown [Benavoli et al., 2014, Eq. 24] that, after combining this set with the likelihood, the lower and upper expectations of the chances of observing a given subset A of labels result in

$$\underline{\mathbb{E}} \left[\sum_{m_k \in A} \pi_{m_k} \middle| n, \hat{\mathbf{y}}_\mathbf{n} \right] = \min \left(1, \frac{1}{n} \left[\sum_{m_k \in A} n_k + c \right] \right) := \underline{P}_Y(A), \tag{3.42}$$

$$\overline{\mathbb{E}} \left[\sum_{m_k \in A} \pi_{m_k} \middle| n, \hat{\mathbf{y}}_\mathbf{n} \right] = \max \left(0, \frac{1}{n} \left[\sum_{m_k \in A} n_k - c \right] \right) := \overline{P}_Y(A), \tag{3.43}$$

where n is the total number of observations in the data set, i.e $n = |\mathscr{D}| = N$. We will then consider the probability set

$$\mathscr{P}_Y = \left\{ P \middle| \underline{P}_Y(A) \leqslant P(A) \leqslant \overline{P}_Y(A), \forall A \subseteq \mathscr{K} \right\}. \tag{3.44}$$

Such a model, which corresponds to take a neighbourhood around the empirical distribution using the total variation distance (i.e., L_∞ norm) has been recently investigated by Miranda *et al.* [Miranda et al., 2019], showing for instance that it induced a 2-monotone lower probability, but was not a specific case of probability intervals, in contrast with the IDM model. Using this fact, we know that the result of Equation (3.40) will be obtained by the Choquet integral, which results in this particular case in

$$\begin{aligned}
&\inf_{\mathbb{P}_Y \in \mathscr{P}_Y} \underline{P}(X = \mathbf{x}|Y = m_a) P(Y = m_a) - \overline{P}(X = \mathbf{x}|Y = m_b) P(Y = m_b) \\
&\iff \underline{P}(X = \mathbf{x}|Y = m_a) \underline{P}_Y(\{m_a\}) - \overline{P}(X = \mathbf{x}|Y = m_b) \overline{P}_Y(\{m_b\}) \\
&\iff \underline{P}(X = \mathbf{x}|Y = m_a) \max \left(0, \frac{n_a - c}{n} \right) - \overline{P}(X = \mathbf{x}|Y = m_b) \min \left(1, \frac{n_b + c}{n} \right)
\end{aligned}$$

In particular, this shows that there would be no differences if we considered only the projections of $\mathscr{P}_Y$ over its singletons, which amounts to consider the bigger set

$$\mathscr{P}'_Y = \left\{ P(Y = m_k) = \pi_{m_k} \middle| \pi_{m_k} \in \left[\max \left(0, \frac{n_k - c}{n} \right), \min \left(1, \frac{n_k + c}{n} \right) \right], \forall m_k \in \mathscr{K} \right\}. \tag{3.45}$$

3.4.2 *Generic loss function*

Function $\ell_{0/1}$ is the default choice in classification problems, considering that every mistake should be penalized in the same way. However, in many practical problems different errors will have different impacts, and this is especially true for sensitive applications in which imprecise probabilistic approaches could be useful. This is why studying generic loss function is useful. With such functions, Equation (2.10) can be written, $m_a \succ_\ell^{\mathscr{P}_{Y|x}} m_b$

$$\iff \inf_{\substack{\mathbb{P}_{X|m_*} \in \mathscr{P}_{X|m_*} \\ \mathbb{P}_Y \in \mathscr{P}_Y}} \sum_{m_k \in \mathcal{K}} (\ell(m_k, m_b) - \ell(m_k, m_a))\, P(Y = m_k | X = \mathbf{x}) > 0,$$

which, if we use the notation $c_{m_k}^{b-a} := \ell(m_k, m_b) - \ell(m_k, m_a)$, gives

$$\iff \inf_{\substack{\mathbb{P}_{X|m_*} \in \mathscr{P}_{X|m_*} \\ \mathbb{P}_Y \in \mathscr{P}_Y}} \sum_{m_k \in \mathcal{K}} c_{m_k}^{b-a} P(X = \mathbf{x}|Y = m_k) P(Y = m_k) > 0 \tag{3.46}$$

$$\iff \inf_{\mathbb{P}_Y \in \mathscr{P}_Y} \sum_{\{k | c_{m_k}^{b-a} > 0\}} c_{m_k}^{b-a} \underline{P}(X = \mathbf{x}|Y = m_k) P(Y = m_k) + \sum_{\{k | c_{m_k}^{b-a} \leqslant 0\}} c_{m_k}^{b-a} \overline{P}(X = \mathbf{x}|Y = m_k) P(Y = m_k) > 0 \tag{3.47}$$

that uses the fact that the conditional probabilities $P(X = \mathbf{x}|Y = m_k)$ are all independent. As Equation (3.47) remains a linear form of the probabilities $P(Y = m_k)$, it can be solved as previously, i.e., through the use of linear programming or the identification of the extreme point for which the bound is reached. If we now consider the specific credal set given by Equation (3.44) and induced by the constraints (3.42)-(3.43), we still have that this induces a 2-monotone lower probability, meaning that we can estimate (3.47) by using the Choquet integral.

All these remarks show that making the marginal probabilities imprecise or considering generic loss functions preserves the model tractability, as the computational complexity is not increased by much, especially when $\mathscr{P}_Y$ has mathematical properties making computations easier (which is luckily the case for most IP models over multinomial distributions).

3.5 SYNTHETIC DATA EXPLORING NON I.I.D. CASE

In this section, we perform some additional empirical experiments on synthetic data sets, with the goal of exploring the capabilities of our approach when training and test data are non identically distributed. Indeed, while the imprecise probabilistic approaches presented in this chapter are not especially aimed to solve such an issue, they may be interesting to do so, as

they are akin to distributionally robust approaches that consider imprecise neighbourhoods around the empirical distribution on training data [Kuhn et al., 2019].

Sections 3.5.1 and 3.5.2 respectively describe the procedures used to generate the training and test sets of the experiments[2], and Section 3.5.3 discusses the obtained results.

3.5.1 *Synthetic datasets generation*

We artificially generate 4 synthetic data sets, each one composed with different number of features and labels (cf. Table 3.5), with the aim of exploring certain special aspects related to the robustness and cautiousness of prediction performed by (im)precise classifier models. These aspects will be explained in detail in the following two subsections, but we first explain how synthetic data sets $\mathscr{D}_l, \forall\, l \in \{1, \ldots, 4\}$ are generated.

Simply put, for each data sets $\mathscr{D}_l$, we generate one sub-population $\mathscr{D}_l^{m_k}$ per label m_k, for which attributes follows a Gaussian distribution, that is

$$x_i | y_i = m_k \sim \mathcal{N}(\mu_{m_k}, \Sigma_{m_k}).$$

We will consider different settings, going from the easiest classification problems where all populations are spherical, homoscedastic, well separated and where the number of features and classes are small, to the most difficult ones where populations are mixed, heteroscedastic and with a high number of features and classes.

To generate the distributions, we will use a "root" label m_r centered at the origin, i.e., $\mu_{m_r} = (0, \ldots, 0)$, and will generate the other populations by setting them in different quadrants. More precisely, we will have

$$\begin{aligned} x_i | y_i = m_r &\sim \mathcal{N}\left(\mu_{m_r} = \mathbf{0}, \Sigma_{m_r}\right) \qquad \text{(root sub-population)}, \\ \forall m_r \neq m_k,\ x_i | y_i = m_k &\sim \mathcal{N}\left(\mu_{m_r} + \boldsymbol{\omega}_k^T * \rho^q_{m_r, m_k}, \Sigma_{m_k}\right),\ \boldsymbol{\omega}_k \in \{-1, 1\}^p. \end{aligned}$$

with all covariance matrix being equal in the homoscedastic case. We require every $\boldsymbol{\omega}_i \neq \boldsymbol{\omega}_j$, so that populations do not overlap, and we define the distance factors $\rho^q_{m_r, m_k}$ as

$$\begin{aligned} \rho^q_{m_r, m_k} &= 2 * \sqrt{\chi^2_{p,q} * \boldsymbol{\lambda}_{m_r}} \qquad , \qquad \text{(Homoscedastic)} \\ \rho^q_{m_r, m_k} &= 0.8 * \sqrt{\boldsymbol{\lambda}_{m_r} * \chi^2_{p,q}} + 0.8 * \sqrt{\boldsymbol{\lambda}_{m_k} * \chi^2_{p,q}} \qquad \text{(Heteroscedastic)} \end{aligned}$$

where $\boldsymbol{\lambda}_{m_r}$ contains the eigenvalue of covariance matrix Σ_{m_r} and $\chi^2_{p,q}$ is q-quantile of a Chi-square distribution with p-degree of freedom. Roughly

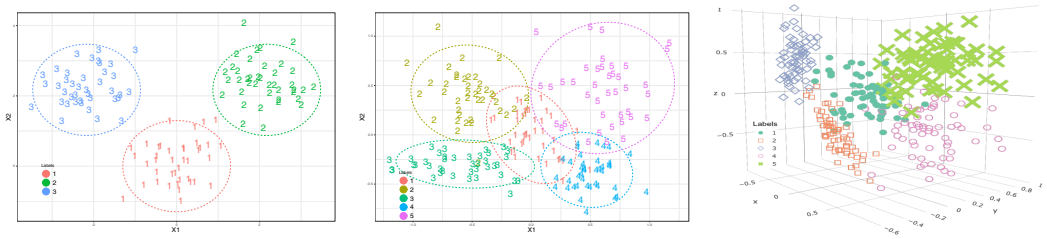
[2] It is also available in R code on https://github.com/salmuz/synthetic_noise_igda

speaking, $2 * \sqrt{\chi^2_{p,q} * \lambda_{m_r}}$ can be thought as the length of the ellipsoid covering q% of the population m_r. This means that the higher is q, the more separated are the classes. q can therefore be seen as a good proxy to measure how difficult is a generated classification problem. In Table 3.5, we summarize 4 different synthetic data sets which will be used in experiments where we will corrupt test instances by some noise.

Dataset	#features	#labels	Σ_*	Variability	q-CV
$\mathscr{D}_1$	2	3	sphere	Homoscedastic	0.80
$\mathscr{D}_2$	2	5	ellipse	Heteroscedastic	0.60
$\mathscr{D}_3$	3	5	ellipse	Heteroscedastic	0.35
$\mathscr{D}_4$	6	8	ellipse	Heteroscedastic	0.10

Table 3.5: Synthetic datasets used in the experiments

A user-friendly visualisation of three first synthetic datasets of Table 3.5 with confidence region at 90% is plotted in Figure 3.5.



(a) $\mathscr{D}_1$ **(b)** $\mathscr{D}_2$ **(c)** $\mathscr{D}_3$

Figure 3.5: Synthetic datasets of three first dataset of Table 3.5.

3.5.2 *Experiments setting with corrupted test data*

The reliability of a classifier is directly related to the unknown underlying distribution of the test data set (a.k.a. out-of-sample test data), which is usually assumed to be the same one as for the training data set. However, this assumption is unlikely to hold in many applications where models may be applied to an evolving population, or to a population different from the one it was initially trained for. As imprecise probabilistic approaches consider sets of distributions, we thought it interesting to study their behaviour when the test data distribution is perturbed or modified.

We performed such a disturbance in two ways: (1) we move instances of the test data set away from its center of inertia (i.e. μ_{m_k} mean) using some noise parameter, and (2) we randomly corrupt the initial inertia of instances of the test data set with a random dispersion matrix generated from a Wishart distribution.

We will denote by $\mathbb{T}_l^\epsilon$ and $\mathbb{T}_l^\psi$ the collections of perturbed test data of data set $\mathscr{D}_l$, respectively for the first and second disturbance. ϵ and ψ are parameters within $[0, 1]$, such that 0 corresponds to no disturbance and 1 to a maximal disturbance. To ensure that our results are close to asymptotic behaviours of classifiers, each test data set is composed of 10^4 instances. We now define the disturbance procedures:

Shifting mean in this case, each test data set $\mathscr{T}(\epsilon)$ of the finite collection $\mathbb{T}_l^\epsilon$ is corrupted with a ϵ noise parameter which basically works as a force of attraction towards the center of gravity of the whole observed population. So, the strategy here is to move away test instances from its ground-truth sub-population towards the center of gravity, corrupting their ground-truth gravitational center μ_{m_k} as follows:

$$\mathbb{T}_l^\epsilon = \left\{ \mathscr{T}_l(\epsilon) \;\middle|\; \epsilon = \{0.00, 0.02, \ldots, 0.98, 1.00\} \right\},$$

$$\mathscr{T}_l(\epsilon) = \left\{ \bigcup_{m_k \in \mathscr{K}} \mathscr{T}^{m_k} \;\middle|\; \begin{array}{l} \forall m_k \in \mathscr{K}, \mathscr{T}^{m_k} \sim \mathcal{N}(\tilde{\mu}_{m_k}, \Sigma_{m_k}), \\ \tilde{\mu}_{m_k} = (1-\epsilon)\mu_{m_k} + \epsilon\mu_G, \quad \mu_G = \dfrac{1}{K}\displaystyle\sum_{k=1}^{K} \mu_{m_k} \end{array} \right\}$$

where $\mathscr{T}^{m_k}$ denotes the corrupted test data of label m_k with $\tilde{\mu}_{m_k}$.

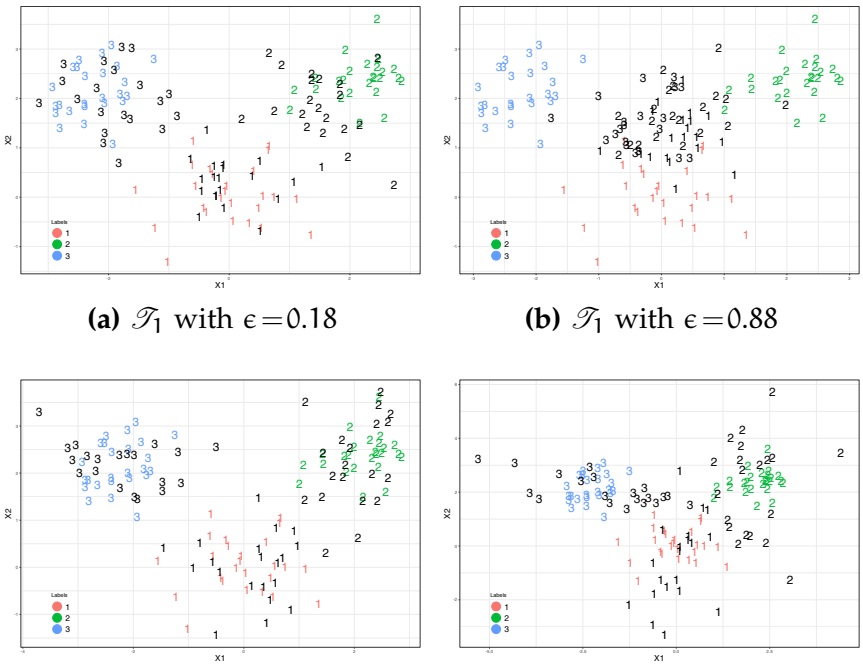
Noise dispersion in contrast to previous case, each test data set $\mathscr{T}_l(\psi)$ of the finite collection $\mathbb{T}_l^\psi$ is corrupted with a ψ noise parameter which basically disturbs the initial inertia of each sub-population with some proportion ψ of the Γ random covariance matrix generated from a Wishart distribution, as follows:

$$\mathbb{T}_l^\psi = \left\{ \mathscr{T}_l(\psi) \;\middle|\; \psi = \{0.00, 0.02, \ldots, 0.98, 1.00\} \right\},$$

$$\mathscr{T}_l(\psi) = \left\{ \bigcup_{m_k \in \mathscr{K}} \mathscr{T}^{m_k} \;\middle|\; \begin{array}{l} \forall m_k \in \mathscr{K}, \mathscr{T}^{m_k} \sim \mathcal{N}(\mu_{m_k}, \tilde{\Sigma}_{m_k}), \\ \tilde{\Sigma}_{m_k} = (1-\psi)\Sigma_{m_k} + \psi\Gamma, \quad \Gamma \sim \mathcal{W}(\mathbb{I}, p) \end{array} \right\}$$

where $\mathscr{T}^{m_k}$ denotes the corrupted test data of label m_k with $\tilde{\Sigma}_{m_k}$.

Illustrations providing some intuition about those settings can be seen in Figures 3.6a and 3.6b for the **Shifting means** and in Figures 3.6c and 3.6d for the **Noise dispersion**. In the first case, we can observe a strong concentration of test instances around the gravity center of the whole population at higher values of ϵ, whereas in the second one, a steep dispersion in test instances of each sub-population can be observed at higher values of ψ.



(a) $\mathscr{T}_1$ with $\epsilon = 0.18$ **(b)** $\mathscr{T}_1$ with $\epsilon = 0.88$

(c) $\mathscr{T}_1$ with $\psi = 0.08$ **(d)** $\mathscr{T}_1$ with $\psi = 0.98$

Figure 3.6: Noise-corrupted test instances of synthetic data sets (in black: corrupted instances) Table 3.5.

While we can set the number of test instances high enough to get asymptotic results, the number of training data should be kept reasonable, otherwise we would just obtain perfect and precise estimations. Results should also be averaged over many training instances, so as to display average trends. For this purpose, we randomly generate 50 different training data sets $\mathbb{D}_l^N = \{\mathscr{D}_l^1, \ldots, \mathscr{D}_l^{50}\}$ with a fixed number of instances $N = \{10, 25, 50\}$, for every synthetic data set l described in Table 3.5.

Metrics used for evaluating the performance of classifiers are: u_{65} and u_{80} utility-discounted accuracies (already been mentioned in 3.3.2), the classical classification accuracy (acc) which determine the percentage of correctly classified instances, and the set-accuracy (set) which is equal to 1 if ground-truth label y is in the cautious prediction $y \in \mathbb{Y}$, and 0 otherwise.

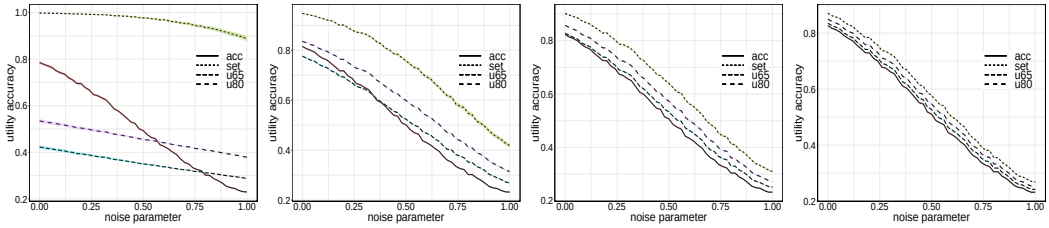
In what follows, we will show the performance of imprecise and precise classifiers in function of each noise parameter and different synthetic datasets (cf. Figures 3.8 and 3.10).

3.5.3 *Experimental results on synthetic data sets*

While we could have optimised hyper-parameter $\hat{c}$ per training data set, in order to not have an "unfair" advantage compared to the precise approach, we decided to fix the value of hyper-parameter $\hat{c} = 0.75$ as Benavoli et al propose in [Benavoli et al., 2014, §8] ($\hat{c} \leqslant 0.75$).

3.5.3.1 *Results on shifting mean*

We first begin by providing in Figure 3.7 some insight on how corrupt test instances behave when the number of instances of training data set increases. For this purpose, we start fitting the (I)QDA model on each random synthetic training data set of the collection of sets $\{\mathbb{D}_2^{10}, \mathbb{D}_2^{25}, \mathbb{D}_2^{50}, \mathbb{D}_2^{100}\}$ (with different number of instances each set), and then evaluate the performance on $\mathbb{T}_2^{\epsilon}$.



(a) $\mathbb{D}_2^{10}$: 10 instances **(b)** $\mathbb{D}_2^{25}$: 25 instances **(c)** $\mathbb{D}_2^{50}$: 50 instances **(d)** $\mathbb{D}_2^{100}$: 100 instances

Figure 3.7: Utility accuracies (%) with confidence intervals of the (I)QDA model on corrupt test data sets $\mathbb{T}_2^{\epsilon}$ using 50 training data sets with different number of instances, i.e. $\{\mathbb{D}_2^{10}, \mathbb{D}_2^{25}, \mathbb{D}_2^{50}, \mathbb{D}_2^{100}\}$.

The results show three important things:

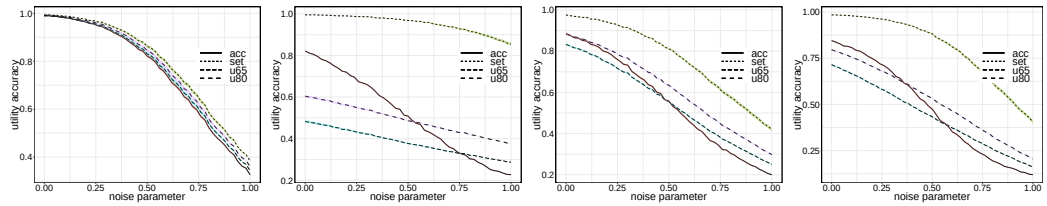
(1) as the number of instances increase the performance of the precise and imprecise classifiers converge to similar values (even if the imprecise case achieves just slightly better results). This result also confirms what was mentioned in Remark 5, and is also partly due to c remaining constant;

(2) the confidence intervals, given by coloured regions around the curves, are considerably tight even when the evaluation of performance has only been repeated 50 times for each set $\mathbb{D}_2^*$, and finally

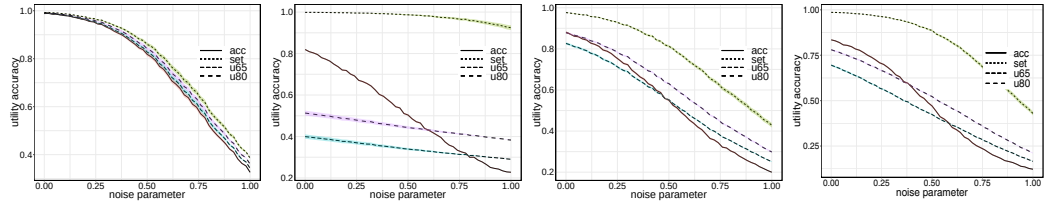
(3) we can see, in particular for a small number of training data ($\mathbb{D}_2^{10}$ or $\mathbb{D}_2^{25}$), that the imprecise approaches are quite robust to change in the distributions. This is shown by the fact that the decrease of performance when ϵ increases is much slower in the imprecise case than in the precise one. This is mostly noticeable in $\mathbb{D}_2^{25}$. In $\mathbb{D}_2^{10}$, the model is quite imprecise, hence very stable, and in $\mathbb{D}_2^{50}$, $\mathbb{D}_2^{100}$, the low value of c makes the precise and imprecise models almost identical, at least in terms of trends when the disturbance increases.

As the behaviour of the imprecise approach appears particularly interesting for small training data sets (which is precisely the situations for which imprecise methods are built for), in Figure 3.8 we show the results for 10 instances on all data sets, each column corresponding to one data

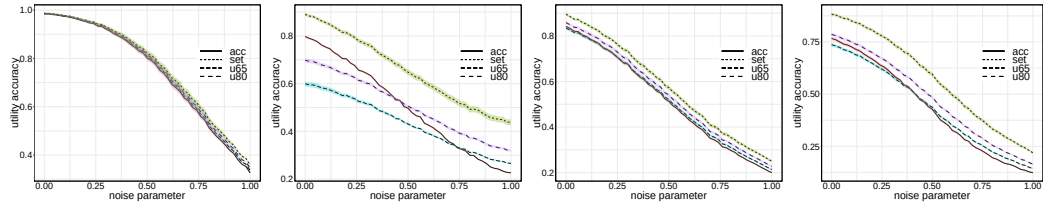
set, and each line to a particular model. The rest of experiments are in Appendix A.2.



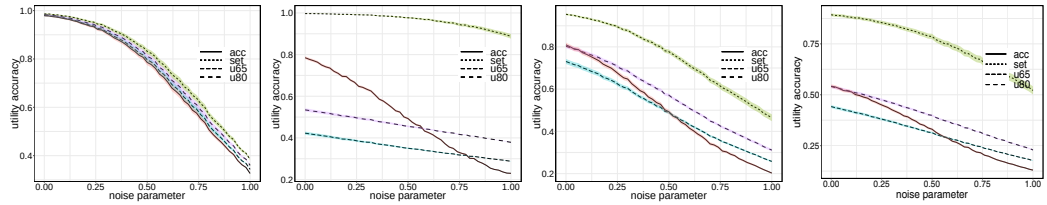
(a) (Imprecise) Euclidian discriminant analysis



(b) (Imprecise) Linear discriminant analysis



(c) (Imprecise) Naive discriminant analysis



(d) (Imprecise) Quadratic discriminant analysis

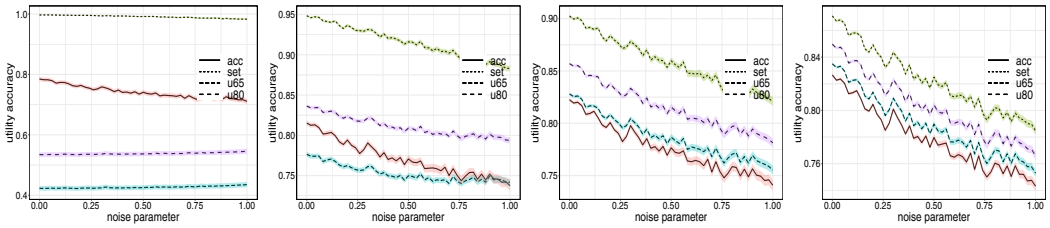
Figure 3.8: Utility accuracies (%) with confidence intervals on corrupt test data sets $\mathbb{T}_\iota^\epsilon$. The first column ($\mathbb{D}_1^{10}$), the second column ($\mathbb{D}_2^{10}$), and so on. In each row a different Gaussian classifier model is fitted.

Overall, the results obtained in the Figure 3.8 with respect to u_{65} and u_{80} remains coherent with our previous findings, except for $\mathbb{D}_1^{10}$ where the easiness of the data sets makes all methods alike.

3.5.3.2 *Results on noise dispersion*

In the same way as in the previous subsection, but with a different set of corrupt test data set $\mathbb{T}_2^\psi$, we provide in Figure 3.9 some insight about the performance of (I)QDA classifier fitted to each data set of the collection $\{\mathbb{D}_2^{10}, \mathbb{D}_2^{25}, \mathbb{D}_2^{50}, \mathbb{D}_2^{100}\}$.

Roughly speaking, we can see the same trends as in the case of the shifted mean, except that now the most noticeable case is the one where 25 instances are used to train the model. Indeed, with 10 instances the imprecise remains more robust than the precise one, as this latter does decrease as noise increases, but the imprecise approach is too imprecise, as we can guess from the big gap between the u_{80} and the set accuracy (which almost always one) curves. The fact that the curves are noisier than in the previous case can be easily explained by the fact that the disturbance is here random, while the previous one was deterministic for each data set.



(a) $\mathbb{D}_2^{10}$: 10 instances **(b)** $\mathbb{D}_2^{25}$: 25 instances **(c)** $\mathbb{D}_2^{50}$: 50 instances **(d)** $\mathbb{D}_2^{100}$: 100 instances

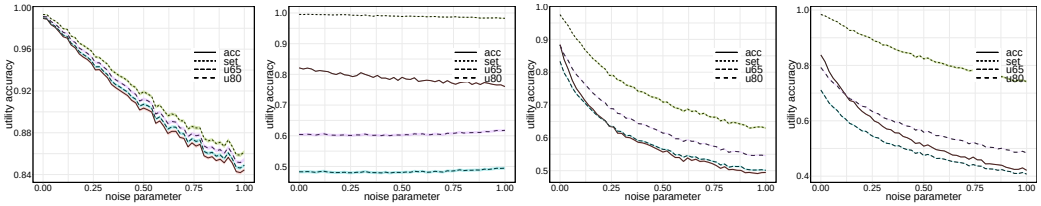
Figure 3.9: Utility accuracies (%) with confidence intervals of the (I)QDA model on corrupt test data sets $\mathbb{T}_2^{\psi}$ using 50 training data sets with different number of instances, i.e. $\{\mathbb{D}_2^{10}, \mathbb{D}_2^{25}, \mathbb{D}_2^{50}, \mathbb{D}_2^{100}\}$.

Similarly to the case of the shifting mean, we provide in Figure 3.10 the evolution of the curves for the different data sets, in the case of 10 instances. The results are similar to the previous case, with a precise model that degrades more quickly than the imprecise approach as the level of noise increases. This is especially clear for the third and fourth data sets, for which the precise model is initially better than the imprecise ones, but then become worse. The rest of experiments are in Appendix A.2

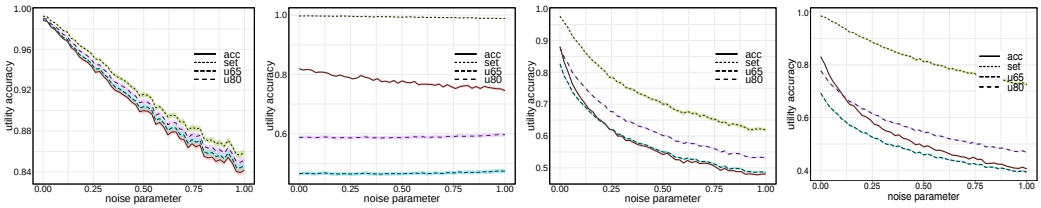
Overall, we can conclude that, although the imprecise method we have presented is not specifically designed to deal with the problem of non identically distributed data, it does provide some protection against it. Fully studying such an aspect is out of the scope of the present chapter, but these experiments are certainly encouraging enough to pursue in this direction.

3.6 OPTIMAL ALGORITHM FOR A CAUTIOUS PREDICTION USING THE MAXIMALITY CRITERION

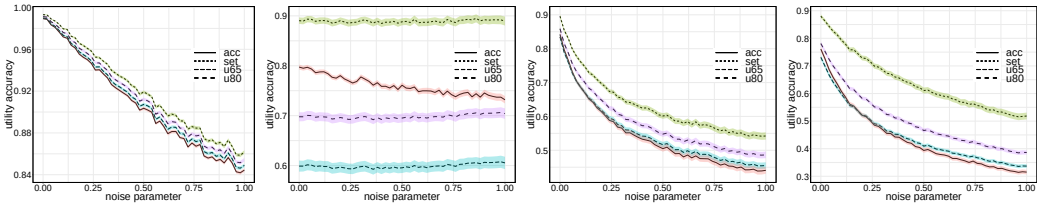
In order to decrease the average number of computer operations performed for getting the set-valued prediction $\hat{\mathbb{Y}}_M$, we propose a specific algorithm very close in spirit to the one of Nakharutai *et al.* [Nakharutai et al., 2019]. Indeed, rather than making the comparisons required by the maximality criterion in any order, we focus on those classes that we know to have high probability for the precise model. In particular, the first of



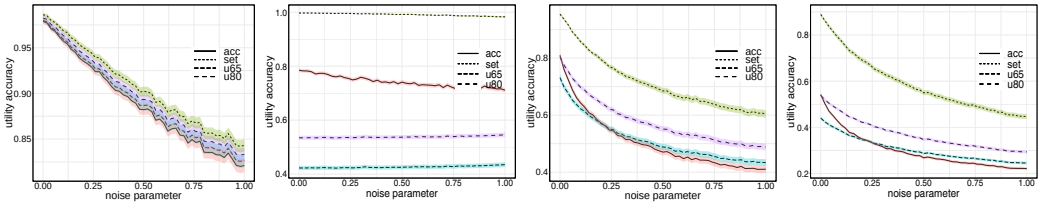
(a) (Imprecise) Euclidian discriminant analysis



(b) (Imprecise) Linear discriminant analysis



(c) (Imprecise) Naive discriminant analysis



(d) (Imprecise) Quadratic discriminant analysis

Figure 3.10: Utility accuracies (%) with confidence intervals on corrupt test data sets $\mathbb{T}_l^{\psi}$. The first column ($\mathbb{D}_1^{10}$), the second column ($\mathbb{D}_2^{10}$), and so on. In each row a different Gaussian classifier model is fitted.

those class will be a maximal one, as it corresponds to a Bayes optimal model for one of the probabilities included in our imprecise model.

Algorithm 1 details this idea, resulting in the set of maximal elements. It should be noted that in the worst case, the complexity remains quadratic, as all pairwise comparison will have to be made. Nevertheless, if $\hat{\mathbb{Y}}_M$ is small enough, we can expect a significant gain in performances (in particular, if $|\hat{\mathbb{Y}}_M| = 1$, the method becomes linear and has to perform $|\mathscr{K}|$ tests).

To know if such an algorithm is really interesting in our specific case, we perform a comparison of the worst case ($\mathcal{V}_1$) that compute all lower and

Algorithm 1: Maximal elements from maximality criterion

Require: $\mathscr{K} = \{m_1, \ldots, m_K\}$ ▷ Set labels
Require: $\boldsymbol{p}^{\boldsymbol{x}} = \{p^{\boldsymbol{x}}_{m_1}, \ldots, p^{\boldsymbol{x}}_{m_K}\}$ ▷ Precise probabilities of GDA
Require: $\boldsymbol{\pi} = \{\pi_{m_1}, \ldots, \pi_{m_K}\}$ ▷ Precise marginal probability $\mathbb{P}_Y$
$\{\overline{p}^{\boldsymbol{x}}_{m_1}, \ldots, \overline{p}^{\boldsymbol{x}}_{m_k}\} :=$ compute_upper_probabilities($\mathscr{K}$)
$\mathscr{C} = \mathscr{K}$, $\mathscr{Z} = \emptyset$, $\mathscr{N}_{opt} = \emptyset$
while $|\mathscr{C} \setminus \mathscr{Z}| > 0$ **do** ▷ Subset of labels not yet compared
  $m_z = \arg\max_{m_i \in \mathscr{C} \setminus \mathscr{Z}} p_{m_i}$ ▷ Pick out the maximal element among a sub-set
  $\underline{p}^{\boldsymbol{x}}_{m_z} :=$ compute_lower_probabilities(m_z)
  $\mathscr{M}_{opt} = \emptyset$
  for $m_k \in \mathscr{C} \setminus m_z$ **do**
    if $\pi_{m_z} \underline{p}^{\boldsymbol{x}}_{m_z} - \pi_{m_k} \overline{p}^{\boldsymbol{x}}_{m_k} > 0$ **then** ▷ $m_z \succ_M m_k$
      $\mathscr{N}_{opt} = \mathscr{N}_{nopt} \cup m_k$ ▷ Subset of not-optimal labels
    else
      $\mathscr{M}_{opt} = \mathscr{M}_{opt} \cup m_k$ ▷ Subset of non-comparable labels for m_z
    end if
  end for
  $\mathscr{Z} = \mathscr{Z} \cup m_z$ ▷ Subset of label already compared with others
  $\mathscr{C} = (\mathscr{M}_{opt} \cup m_z) \setminus \mathscr{N}_{opt}$ ▷ Partition of possible optimal labels in this step
end while
return $\mathscr{C}$

upper probabilities necessary to perform the $K(K-1)$ comparisons for getting the set-valued $\hat{\mathbb{Y}}_M$ and then perform those comparisons, against the proposed Algorithm 1 ($\mathcal{V}_2$), for every imprecise model seen in the Section 3.2 and for every data set of Table 3.2. In both cases, we use the optimal imprecise model (i.e. $\hat{c}$) w.r.t. discount-utility measure u_{80} obtained in Section 3.3. So we calculate the average empirical time complexity on the set of prediction times obtained from applying both cases on 20% of the data set, then we randomly repeat this experiment 10 times and we show the overall mean and the percentage of increase (blue up-arrow) or decrease (red down-arrow) in the Table 3.6.

Overall, the results obtained on the ILDA and IQDA models are quite satisfying, as computational time decrease by $\geqslant\sim 45\%$ and $\geqslant\sim 30\%$, respectively. In contrast, the overall empirical time complexity of INDA and IEDA models have increased by $\leqslant\sim 7\%$ and $\leqslant\sim 30\%$, respectively. This is due to the step of picking out the maximal element and the additional loop added (*while*) in order to evaluate the maximality criterion. However, this increased time does not significantly affect the inference time since it is in milliseconds, whereas the reduction time obtained in (IQ)ILDA models is clearly significant, as the involved NP-hard optimisation problems run in seconds.

#	ILDA		Inc.(%)	IQDA		Inc.(%)
a	0.72 ± 0.06	0.40 ± 0.05	44.44↑	0.73 ± 0.06	0.45 ± 0.06	38.36↑
b	3.72 ± 0.46	1.63 ± 0.29	56.18↑	3.71 ± 0.49	2.15 ± 0.38	42.05↑
c	8.78 ± 2.08	3.11 ± 2.17	64.58↑	8.10 ± 0.87	4.01 ± 0.67	50.49↑
d	0.94 ± 0.08	0.38 ± 0.06	59.57↑	1.92 ± 0.18	1.37 ± 0.17	28.65↑
e	21.56 ± 0.81	5.57 ± 0.46	74.17↑	13.55 ± 2.50	7.96 ± 2.47	41.25↑
f	6.04 ± 1.65	3.00 ± 1.54	50.33↑	6.04 ± 0.76	3.28 ± 0.62	45.70↑
h	11.31 ± 1.25	2.41 ± 0.31	78.69↑	28.07 ± 1.80	16.31 ± 1.10	41.90↑
i	3.36 ± 0.32	1.09 ± 0.30	67.56↑	3.36 ± 0.20	1.23 ± 0.10	63.39↑
j	4.35 ± 0.14	0.71 ± 0.03	83.68↑	3.92 ± 0.24	0.57 ± 0.04	85.46↑
k	13.91 ± 2.94	2.66 ± 0.78	80.88↑	10.74 ± 0.76	5.11 ± 0.64	52.42↑
l	38.60 ± 2.99	9.73 ± 0.68	74.79↑	26.88 ± 2.30	8.18 ± 0.88	69.56↑
n	8.23 ± 0.39	3.04 ± 0.20	63.06↑	9.83 ± 0.50	3.49 ± 0.27	64.49↑

(a) ILDA and IQDA models

#	INDA$\times 10^{-3}$		Inc.(%)	IEDA$\times 10^{-3}$		Inc.(%)
	$\mathcal{V}_1$	$\mathcal{V}_2$		$\mathcal{V}_1$	$\mathcal{V}_2$	
a	1.21 ± 0.02	1.27 ± 0.02	4.96↓	1.19 ± 0.02	1.27 ± 0.05	6.72↓
b	1.44 ± 0.04	1.50 ± 0.12	4.17↓	1.38 ± 0.02	1.78 ± 0.04	28.99↓
c	2.24 ± 0.01	2.24 ± 0.03	0.00	2.25 ± 0.04	2.41 ± 0.08	7.11↓
d	1.23 ± 0.03	1.31 ± 0.01	6.50↓	1.35 ± 0.20	1.34 ± 0.04	0.74↑
e	2.62 ± 0.07	2.58 ± 0.04	1.53↑	2.58 ± 0.02	2.71 ± 0.05	5.03↑
f	2.42 ± 0.13	2.46 ± 0.03	1.65↓	2.37 ± 0.03	2.67 ± 0.07	12.65↓
h	4.36 ± 0.15	4.16 ± 0.05	4.59↑	4.09 ± 0.06	4.07 ± 0.06	0.49↑
i	2.02 ± 0.07	2.03 ± 0.07	0.50↓	2.07 ± 0.07	2.66 ± 0.03	28.50↓
j	4.73 ± 0.10	4.40 ± 0.07	6.98↑	5.02 ± 0.08	4.50 ± 0.06	10.36↑
k	4.17 ± 0.04	4.05 ± 0.06	2.88↑	3.97 ± 0.14	4.14 ± 0.10	4.28↓
l	2.79 ± 0.09	2.66 ± 0.10	4.66↑	2.61 ± 0.05	2.62 ± 0.05	0.38↓
n	2.03 ± 0.07	2.06 ± 0.05	1.48↓	2.15 ± 0.05	2.06 ± 0.06	4.18↑

(b) INDA and IEDA models

Table 3.6:A benchmark of average empirical time complexity in seconds of two approach:($\mathcal{V}_1$)worst case and ($\mathcal{V}_2$)Algorithm 1.

3.7 CONCLUSION

In this chapter, we have generalized classical Gaussian discriminant models to the imprecise setting, mainly by allowing the estimated means of the conditional Gaussian distributions to become imprecise. This was achieved by a robust Bayesian procedure using sets of prior satisfying near-ignorance properties.

We have explored the computational issues associated to the predictions of such models, essentially showing that considering general covariance matrices ended up in practically manageable yet computationally difficult to solve problems, while considering diagonal covariance matrices essentially made the problem much easier to solve.

Experiments on various data sets show that the method is providing quite satisfactory results, in the sense that the induced imprecision in the

predictions is reasonable and mostly concerns instances that were wrongly classified by the precise methods. We have also discussed some possible extensions of our approaches to the case of imprecise priors and generic loss functions, showing that such extensions would not add a prohibitive computational cost.

Finally, some first experiments concerning the case of non-identically distributed data suggest that investigating the potential of our model, and of imprecise probabilistic models in general to solve this issue could be an interesting topic. Indeed, imprecise models appear more robust to such changes in the data, even when they do not seek to address this specific issue.

A natural next step is to also make the covariance matrix estimate imprecise, possibly leaving the mean estimate precise in a first step. Computationally, this would be attractive, as the objective functions could be made linear by fixing the mean and using an eigenvalue decomposition of the covariance matrix [Bensmail et al., 1996]. The main problem would then be to derive a principled approach (i.e., using near-ignorance prior) that would deliver an easy-to-deal convex set of inverse covariance matrices.

Part II

MULTI-LABEL CLASSIFICATION

Part II of the manuscript is focused on the multi-label classification problem, starting in Chapter 4 with a brief reminder focusing on cautious predictions.

In Chapter 5, we consider the problem of making distributionally robust, skeptical binary inferences for the multi-label problem, or more generally for Boolean vectors. We study in particular the Hamming loss case, a common loss function in multi-label problems, showing how skeptical inferences can be made in this setting. We also provide a generalization of Binary relevance model by using imprecise marginal distributions, and experimental results.

In Chapter 6, we present two different ways to extend the classical multi-label chaining approach and a new dynamic, context-dependent label ordering by using imprecise probability estimates. The main reason one could have for using such estimates are (1) to make cautious predictions when a high uncertainty is detected in the chaining and (2) to make better precise predictions by avoiding biases caused in early decisions in the chaining. Our experimental results are encouraging when the minimax approach adopts our new dynamic label ordering that selects the labels with low uncertainty.

Chapter 4

Multi-label classification

"One should never try to prove anything that is not almost obvious."

—Alexander Grothendieck

Contents

In contrast to multi-class problems where each instance is associated to one label, multi-label classification (MLC) consists in associating an instance to a subset of relevant labels from a set of possible labels. That is why MLC can be considered a generalization of traditional multi-class problems, as well as a special case of the multi-task learning.

Such problems can arise in different research fields, such as the classification of proteins in bioinformatics [Tsoumakas et al., 2007], text classification in information retrieval [Fürnkranz et al., 2008], object recognition in computer vision [Boutell et al., 2004], and so on.

In this chapter, we introduce a light but necessary background knowledge about multi-label classification problems in the precise probabilistic setting. Such backgrounds are necessary, but not essential, to understand the next two chapters[1].

In Section 4.1, we introduce the problem setting of multi-label problem using precise probabilities as well as a brief description of problems which we will tackle in the next chapters.

In Section 4.2, we briefly introduce the different kinds of losses existing in the multi-label setting, and finally in Section 4.3, we present the few works we know of dealing with cautious multi-label classification,

[1] Someone with knowledge in multi-label problems can directly go to Chapter 5 or 6

whereby cautious we mean that we may abstain from predicting one label of the set of relevant labels by providing a set-valued predictions.

4.1 MULTI-LABEL PROBLEM SETTING

In multi-label problem, an instance $\mathbf{x}$ of an input space $\mathscr{X} = \mathbb{R}^p$ is no longer associated with a single label m_k of an output space $\mathcal{K} = \{m_1, \ldots, m_m\}$, but with a subset of labels $\Lambda_x \subseteq \mathcal{K}$ often called the set of relevant labels while its complement $\mathcal{K} \backslash \Lambda_x$ is considered as irrelevant for $\mathbf{x}$. Let $\mathscr{Y} = \{0, 1\}^m$ be a m-dimensional binary space and $\mathbf{y} = (y_1, \ldots, y_m) \in \mathscr{Y}$ be any element of $\mathscr{Y}$ such that $y_i = 1$ if and only if $m_i \in \Lambda_x$. (Example 4.1).

X_1	X_2	X_3	X_4	y_1	y_2	y_3
107.1	25	Blue	60	1	0	0
-50	10	Red	40	1	0	1
200.6	30	Blue	58	1	1	0
107.1	5	Green	33	0	1	0
...	...	...	...	...	...	...

Table 4.1: An example of a multi-label data set

From a decision theoretical approach (c.f. Figure 1.1), the goal of the multi-label problem is the same as the usual classification problem (c.f. Definition 1). This goal is to learn a classifier $\mathbf{h} : \mathscr{X} \to \mathscr{Y}$ that generalizes the behaviour of the empirical evidence $\mathscr{D}$ in the sense of minimizing the risk of getting missclassification with respect to a specified loss function $\ell(\cdot, \cdot)$

$$\mathscr{R}_\ell(Y, \mathbf{h}(X)) = \arg\min_{\mathbf{h}} \mathbb{E}_{\mathbf{X} \times Y} \left[\ell(Y, \mathbf{h}(X)) \right]. \tag{4.1}$$

where the classifier $\mathbf{h}$ outputs a m-dimensional vector as predictive output for a given new instance $\mathbf{x}$

$$\hat{\mathbf{y}} = \mathbf{h}(\mathbf{x}) = (h_1(\mathbf{x}), \ldots, h_m(\mathbf{x})). \tag{4.2}$$

Under similar conditions presented in Definition 1, this minimization can also be expressed as the minimization of conditional expected risk of a given unlabeled instance $\mathbf{x}$ (cf. [Dembczyński et al., 2012, eq. 3])

$$\mathbf{h}(\mathbf{x}) = \arg\min_{\mathbf{y} \in \mathscr{Y}} \mathbb{E}_{\hat{\mathbb{P}}_{Y|\mathbf{x}}} \left[\ell(\mathbf{y}, \cdot) \right] = \arg\min_{\mathbf{y} \in \mathscr{Y}} \sum_{\mathbf{y}' \in \mathscr{Y}} \hat{P}(Y = \mathbf{y}' | X = \mathbf{x}) \ell(\mathbf{y}', \mathbf{y}). \tag{4.3}$$

Moreover, in an equivalent way to what was presented in Definition 2, the maximum element of Equation (4.3) can be obtained picking it from the strict total order relation $\succ$ over $\mathscr{Y} \times \mathscr{Y}$, where $\mathbf{y}^1 \succ \mathbf{y}^2$ ($\mathbf{y}^1$ is preferred to $\mathbf{y}^2$) if

$$\mathbb{E}_{\hat{\mathbb{P}}_{Y|\mathbf{x}}}\left(\ell(\mathbf{y}^2,\cdot)-\ell(\mathbf{y}^1,\cdot)\right)=\mathbb{E}_{\hat{\mathbb{P}}_{Y|\mathbf{x}}}\left(\ell(\mathbf{y}^2,\cdot)\right)-\mathbb{E}_{\mathbb{P}}\left(\ell(\mathbf{y}^1,\cdot)\right)\geqslant 0. \tag{4.4}$$

In the state-of-the-art [Zhang et al., 2013; Mena et al., 2016; Moyano et al., 2018; Read et al., 2019], we can find a variety of approaches which can be divided in; (1) methods focusing on the reduction of the time complexity of Equation (4.3) with respect to different types of loss functions and assuming that the probability distribution $\hat{\mathbb{P}}$ is known, or (2) methods focusing on the optimisation of Equation (4.1) in order to get an estimated probability distribution that generalizes well beyond the observations of a training data set over a specified loss function.

Concerning to the second approach, obviously, the minimisation of Equation (4.1) will be untreatable if we approach it as a classical classification issue, since the output space $\mathscr{Y}$ grows up exponentially over the number of classes in $\mathscr{K}$, i.e. $2^{|\mathscr{K}|}$. Among other things, the probability of each distinct output would be too small and hard to estimate (as some outputs may even be never observed). Such an approach is known as *label powerset (LP)* method [Boutell et al., 2004] and is tailored to minimize the subset zero-one loss $\ell_{0/1}$. One way to circumvent this issue is to adopt decomposition techniques [Tsoumakas et al., 2007; Menon et al., 2019], in which the initial difficult problem is split into a set of simpler problems [Zhang et al., 2018] (e.g. Binary relevance), or the classifier-chains approach which computes the full joint probability estimate of a greedy way. The first approach is briefly discuss in Section 4.2.

In this part of the manuscript, we will contribute in each of such approaches, but on an imprecise probability setting. In Chapter 5, we focus on reducing the time complexity of criterial presented in Chapter 2 on different settings, and in Chapter 6, we focus on learning an imprecise classifier-chains method based on different strategies and applying to the NCC as base classifier.

Concerning the *precise* learning approaches used and extended in this manuscript, namely classifier-chains [Read et al., 2011] and probabilistic classifier-chains [Cheng et al., 2010] approaches, we prefer to recall and detail of each one of them in its respective chapter (Chapter 6 and 5, respectively) to make the reading easier.

4.2 LOSS FUNCTIONS

In contrast to the multi-class (or binary) classification problem, where the usual metrics measure the loss incurred from a given inferred univariate response Y, the multi-label classification problem is confronted not to measure only a univariate response but a set-valued responses $\mathbf{Y}=\{Y_1,\ldots,Y_m\}$ (or multivariate responses), so that these usual metrics are not directly tailored at all (except for the zero/one loss, $\ell_{0/1}$).

In the state-of-the-art, a variety of loss functions [Gibaja et al., 2014; Díez et al., 2015] has been proposed to satisfy different real-application problems modelled as a multi-label classification. Depending on the nature and structure of such loss functions (e.g. linear as Hamming or non-linear as F-measure over labels), Equation (4.3) can be simplified according to whether or not its optimisation requires the knowledge of the full joint probability distribution.

Such loss functions can be classified in two different kinds, so-called label-wise: (1) decomposable and (2) non-decomposable [Dembczyński et al., 2012].

1. DECOMPOSABLE METRICS roughly speaking, a decomposable loss is the one which can evaluate the incurred loss of each label Y_i *"independently"* from all other ones, so that it can be expressed in the form

$$\ell(\mathbf{y}, \mathbf{h}(X)) = \sum_{i=1}^{m} \ell_i(y_i, h_i(X)), \tag{4.5}$$

   where $\ell_i : \{0,1\}^2 \to \mathbb{R}$ is a binary loss function over i-th label. A particular case is the Hamming loss (c.f. Equation (5.3)) with an image set $\{0,1\}$ instead of $\mathbb{R}$ (i.e. zero-one loss function $\ell_{0/1}$), but also precision@k, DCG@k, squared-error label-wise, as well as others [Dembczyński et al., 2012; Jasinska Kalina, 2018; Vu-Linh Nguyen, 2019].

   It is well known that the optimization of Equation (4.3) on these types of losses generate an "optimal" (prediction) decision requiring only the knowledge of the conditional marginal (single-label) distribution. In other words, the optimal binary vector prediction $\hat{\mathbf{y}}$ on these losses is the one for which $\hat{y}_j = 1$ if $P_{\mathbf{x}}(Y_j = 1) \geqslant 0.5$ and $\hat{y}_j = 0$ otherwise.

   However, as will be seen later on in Chapter 5, in an imprecise probabilistic setting it is not enough to know the conditional marginal distribution, but the full joint probability distribution to get an "optimal" set of predictions (i.e. a set of binary vectors), even in the case of decomposable losses.

2. NON-DECOMPOSABLE METRICS in contrast to previous metrics, non-decomposable loss does not allow to evaluate the incurred loss independently. Roughly speaking, these losses can account for label-dependencies [Dembczyński et al., 2012], which implies that it is necessary to know in some case the partial or full conditional probability distribution $\hat{\mathbb{P}}_{\mathbf{Y}|\mathbf{x}}$. Evidently, it computationally quickly becomes untractable, due to the fact that the output predictive space increases exponentially with m (e.g. $|\mathscr{Y}| = 32768$ for $m = 15$)

Among such losses, we have the well-known F-measure and Jaccard losses, given in Equation 4.6 and 4.7, respectively.

$$\ell_{F_\beta}(\mathbf{y}, \hat{\mathbf{y}}) = 1 - \frac{(1+\beta^2)\sum_{i=1}^{k} y_i \hat{y}_i}{\beta^2 \sum_{i=1}^{k} y_i + \sum_{i=1}^{k} \hat{y}_i} \tag{4.6}$$

$$\ell_J(\mathbf{y}, \hat{\mathbf{y}}) = 1 - \frac{\sum_{i=1}^{k} y_i \hat{y}_i}{\sum_{i=1}^{k} y_i + \sum_{i=1}^{k} \hat{y}_i - \sum_{i=1}^{k} y_i \hat{y}_i} \tag{4.7}$$

Despite the inherent complexity of these sophisticated losses, [Dembczynski et al., 2011] proposed for the F-measure loss an exact algorithm of complexity $\mathcal{O}(m^3)$ assuming that the probability distribution $\hat{\mathbb{P}}_{Y|X}$ is known. [Ramón Quevedo et al., 2012], on the other hand, proposed for the Jaccard loss an algorithm of complexity $\mathcal{O}(m^2)$ under the assumption of label independence.

Owing precisely to the non-decomposition of these losses, proposing procedures in an imprecise probabilistic setting becomes very hard and challenging. Indeed, we have to consider the fact that with general sets and in the worst of cases, Equation (4.8)

$$\mathbf{y}^1 \succ_\ell^{\mathscr{P}} \mathbf{y}^2 \iff \mathbb{E}_{\mathscr{P}}[\ell_{F_\beta}(\mathbf{y}^2, \mathbf{y}) - \ell_{F_\beta}(\mathbf{y}^1, \mathbf{y})] \tag{4.8}$$

remains with a time complexity $\mathcal{O}(2^{2m})$. Solving such challenges, even approximatively, is one of our future works, for instance by using a truncated formulation of Harmony mean (F-measure) with Bernstein functions [Qi et al., 2017].

4.3 CAUTIOUS MODELS IN MULTI-LABEL PROBLEMS

In the literature, there are only a few works on multi-label classification producing cautious predictions, they can be classified as

1. PARTIAL REJECTION RULES. an interesting work is the one of Pillai et al. [Pillai et al., 2013], in which the decision of abstaining is based on threshold parameters adapted on the F-measure as a performance metric.

2. PARTIAL ABSTENTION. a new approach of abstaining has recently been proposed in [Vu-Linh Nguyen, 2019] based on a generalization of loss function, adding it a new term in order to penalize the abstention, as follows:

$$\mathcal{L}(\mathbf{y}, \hat{\mathbf{y}}) = \ell(\mathbf{y}, \hat{\mathbf{y}}) + f(|A(\hat{\mathbf{y}})|) \tag{4.9}$$

where $\ell(\cdot,\cdot)$ is the usual loss function and $f(\cdot)$ is the penalty for abstaining on $A(\hat{\mathbf{y}})$ the set of indices i for which $y_i = *$. However, this

penalty function should always be chosen according to the context problem, making it strongly dependent on the data set and imposing a restriction even if [Vu-Linh Nguyen, 2019] recommended two types of functions in experiments section. In contrast, our approach presented in Chapter 5 is based on axiomatic bases and theoretical justifications [De Finetti, 1937; Walley, 1991] to model uncertainties, and does not require an adaptation of a loss function (it does, however, require to settle the amount of imprecision).

3. INDETERMINATE CLASSIFICATION. [Destercke, 2014] proposed two outer approximations for the Hamming and Ranking loss functions, respectively, which have a polynomial time complexity and are therefore efficient in the inference step. Albeit these approximations are efficient, we have been able to prove that in the Hamming case, its approximation is based on the assumption of label independence, and that in the general case, it can degrade or become a gross approximation if the number of labels is huge, $m > 14$.

   Finally, Antonucci and Corani [Antonucci et al., 2017]'s work is closer to our framework since it also uses credal sets to quantify the uncertainty, although it focuses on a specific model and another loss function (the zero/one loss). This work might be compared to our approach presented in Chapter 6, since both are focused in the zero/one loss, in one of our future works.

4.4 SUMMARY

This chapter has provided a brief reminder of the multi-label setting and its whereabouts, with a light focus on cautious predictions in this setting. Next chapters will detail our contributions.

Chapter 5

Distributionally robust, skeptical binary inferences in multi-label problems

"Der Mensch kann wohl tun was er will, aber er kann nicht wollen was er will."

—Arthur Schopenhauer

Contents

Considering all possible subsets of labels as possible predictions make the estimation and decision steps of a learning problem significantly more difficult: partial observations are more likely to occur, especially when the number of labels increases, and the output space over which the probability needs to be estimated grows exponentially with the number of labels.

This means that in some applications where guaranteeing the robustness and reliability of predictions is of particular importance, one may consider being cautious about such predictions, by predicting a set of possible answers rather than a single one when uncertainties are too high.

In this chapter, we consider the problem of making such set-valued predictions by performing skeptic inferences when our uncertainty is described by a set of probabilities (the more uncertainties, the bigger the set). By skeptic inference, we understand the logical procedure that consists, in the presence of multiple models, to accept only those inferences that are true for every possible model. Such approaches are different from thresholding approaches [Vu-Linh Nguyen, 2019; Pillai et al., 2013], and

are closer in spirit to distributionally robust approaches, even if these later typically consider precise, minimax inferences, that are cautious yet not skeptic [Hu et al., 2018; Chen et al., 2018]. We also make no assumptions about the considered set of probabilities, thus departing from usual distributionally robust approaches, that typically consider precise predictions, or from existing works dealing with sets of probabilities and multi-label problems [Antonucci et al., 2017], that considered specific probability sets and zero/one loss function (seldom used in multi-label problems).

Section 5.1 introduces some basic notations that we will use more specifically in this chapter, and gives the necessary reminders about skeptic inferences made with sets of probabilities defined on (binary) tree structures.

In Section 5.2, we provide novel theoretical results, coupling the hamming loss and the maximality decision criterion, showing that our exact procedure for making skeptical inferences has an almost linear time complexity with respect to the size of the output space of the multi-label classification. Furthermore, we show that under some specific independence conditions, the set-valued predictions induced by the E-admissible decision criterion match perfectly with the one of the maximality. Finally, we provide additional results on remaining criterial introduced in Section 2.1.2.

Finally, in Section 5.3, we perform a set of experiments on simulated and real data sets. The first set aims to study the exactness of the existing outer-approximation [Destercke, 2014] against our exact optimal procedure. The second aims to make a first study of the behaviour of the skeptical inference against its precise counterpart; (1) under an assumption of independence as specified in Section 5.2.2 and (2) on missing and noisy labels.

5.1 PROBLEM SETTING

We denote by $\mathbf{Y} = (Y_1, \ldots, Y_m)$ the random binary vector over $\mathscr{Y}$. Given a subset $\mathscr{I} \subseteq \{1, \ldots, m\}$ of indices, we denote by $\mathscr{Y}_\mathscr{I}$ the space of binary vectors over those indices, and by $Y_\mathscr{I}$ and $Y_{-\mathscr{I}}$ the marginals of $\mathbf{Y}$ over these indices $\mathscr{I}$ and over the complementary indices $\{1, \ldots, m\} \setminus \mathscr{I}$, respectively. In particular, $Y_{\{i\}}$ will denote the marginal random variable over the ith label. Similarly, we will denote by $\mathbf{y}_\mathscr{I}$ the values of a vector restricted to elements indexed in $\mathscr{I}$, and by $\mathbf{b}_\mathscr{I}$ a particular assignment over these elements. The associated marginal probability will be

$$P_x(\mathbf{b}_\mathscr{I}) = \sum_{\mathbf{y} \in \mathscr{Y}, \mathbf{y}_\mathscr{I} = \mathbf{b}_\mathscr{I}} P_x(\mathbf{Y} = \mathbf{y}).$$

We will also consider the complement of a given vector or assignment over a subset of indices. These will be denoted by $\overline{\mathbf{b}}_{\mathscr{I}}$ and $\overline{\mathbf{b}}_{\mathscr{I}}$, respectively.

Given two vectors $\mathbf{y}^1$ and $\mathbf{y}^2$, we will denote by $\mathscr{I}_{\mathbf{y}^1 \neq \mathbf{y}^2} := \{i \in \{1, \ldots, m\} : y_i^1 \neq y_i^2\}$ the set of indices over which two vectors are different, and similarly by $\mathscr{I}_{\mathbf{y}^1 = \mathbf{y}^2} := \{i \in \{1, \ldots, m\} : y_i^1 = y_i^2\}$ the sets of indices for which they will be equal.

Example 8 *Consider the probabilistic tree developed in Figure 5.1 defined over $\mathscr{Y} = \{0,1\}^2$ describing a full joint distribution over two labels. In such trees, the probability of any vector $\mathbf{y}$ is simply the product of the probabilities along its path. We also have that the partial vector $(\cdot, 1)$ has probability*

$$P((\cdot,1)) = P((0,1)) + P((1,1)) = 0.5 \cdot 0.2 + 0.5 \cdot 0.7 = 0.45.$$

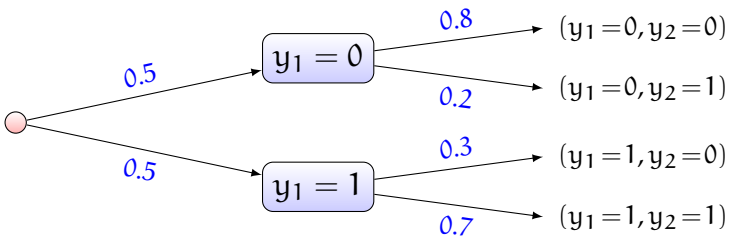


Figure 5.1: Probabilistic binary tree of two labels

In the sequel of this chapter, we will use such trees to illustrate our results, replacing the precise probabilities on the branches by intervals. An example will be provided later. The resulting set of probabilities over $\mathscr{Y}$ will then simply be the set of all joint probabilities obtained by taking precise values within those intervals.

As in this chapter we are interested in making set-valued predictions for the multi-label problems, we will use the notation $\mathbb{Y} \subseteq \mathscr{Y}$ for generic subsets of $\mathscr{Y}$. We will use the notation $\mathscr{Y}^* = \{0, 1, *\}^m$ for the specific subsets induced by partially specified binary vectors $\mathbf{y}^* \in \mathscr{Y}^*$, where the symbol $*$ stands for a label on which we abstain. Denoting by $\mathscr{I}^*$ the indices of such labels, we will also use $\mathbf{y}^*$ and $\mathscr{Y}^*$ for the corresponding family of subsets over $\mathscr{Y}$, i.e.,

$$\mathbf{y}^* := \{\mathbf{y} \in \mathscr{Y} : \forall i \notin \mathscr{I}^*, y_i = y_i^*\}.$$

Such subsets are indeed often used to make partial multi-label predictions, and we will refer to them on multiple occasions, calling them partial vectors. However, using only subsets within $\mathscr{Y}^*$ may be insufficient if one wants to express complex partial predictions. For instance, in the case where $m = 2$, the partial prediction $\mathbb{Y} = \{(0,1), (1,0)\}$ cannot be expressed as an element of $\mathscr{Y}^*$, as approximating $\mathbb{Y}$ with an element of $\mathscr{Y}^*$ would be empty.

5.1.1 *Skeptic inferences with distribution sets*

In this chapter, we assume that our uncertainty is described by a convex set of probabilities $\mathscr{P}$, defined over $\mathscr{Y}$ (for more details we refer to Section 1.3).

SKEPTIC INFERENCE AND DECISION Once our uncertainty is described by a credal set $\mathscr{P}$, the next step according to the scheme presented in Figure 1.1 is to deliver an optimal prediction (or set-valued ones), depending on how large is the credal set $\mathscr{P}$.

When our credal set is reduced to a precise estimate $\hat{\mathbb{P}}$, it is based on the decision theoretic approach already defined in Section 4.1 (c.f. Definition 1). Otherwise, we will consider in this chapter two main decision rules among those introduced in Section 2.1.2, that may return more than one decision in case of insufficient information: E-admissibility and maximality. We quickly recall these rules and for further details, we refer to Definitions 7 and 5, respectively.

$$\hat{\mathbb{Y}}^{E}_{\ell,\mathscr{P}} = \left\{ \mathbf{y} \in \mathscr{Y} \,\middle|\, \exists \mathbb{P} \in \mathscr{P} \text{ s.t. } \forall \mathbf{y}' \in \mathscr{Y}, \mathbb{E}_{\mathbb{P}}\left[\ell(\mathbf{y},\cdot)\right] < \mathbb{E}_{\mathbb{P}}\left[\ell(\mathbf{y}',\cdot)\right] \right\} \tag{5.1}$$

$$\hat{\mathbb{Y}}^{M}_{\ell,\mathscr{P}} = \left\{ \mathbf{y} \in \mathscr{Y} \,\middle|\, \forall \mathbb{P} \in \mathscr{P}, \nexists \mathbf{y}' \in \mathscr{Y} : \mathbf{y}' \succ^{\mathscr{P}}_{\ell} \mathbf{y}, \text{s.t. } \mathbb{E}_{\mathbb{P}}\left[\ell(\mathbf{y},\cdot) - \ell(\mathbf{y}',\cdot)\right] > 0 \right\} \tag{5.2}$$

Also, we recall that computing $\hat{\mathbb{Y}}^{E}_{\ell,\mathscr{P}}$ and $\hat{\mathbb{Y}}^{M}_{\ell,\mathscr{P}}$ can be computationally significantly harder than the precise case, regardless of the functional ℓ chosen. For instance, obtaining $\hat{\mathbb{Y}}^{M}_{\ell,\mathscr{P}}$ may require at worst to perform $\mathcal{O}(2^{2m})$ comparisons (c.f. Section 2.1.2.2), which can be intractable for extreme multi-label problems, with $|\mathscr{Y}| > 10^3$.

Example 9 *Figure 5.2 illustrates the computation of an expected loss in the case of a probabilistic tree and the 0/1 loss function ($\ell(\mathbf{y}',\mathbf{y}) = 1$ if $\mathbf{y} \neq \mathbf{y}'$, 0 else), when comparing the two items $\mathbf{y}' = (0,1)$ and $\mathbf{y}'' = (1,0)$. The global expected value is then retrieved by computing local expectations recursively at each node, starting from the leaves of the tree to get at the root. In this case, we have that $(1,0) \succ^{\mathbb{P}}_{\ell_{0/1}} (0,1)$, since the expectation of the difference $\ell_{0/1}((0,1),\cdot) - \ell_{0/1}((1,0),\cdot)$ is positive.*

Figure 5.3 pictures an imprecise probabilistic tree for the same situation, where the probabilities in each branch are replaced by intervals (that in the binary case are sufficient to represent any convex set). The computation of the corresponding lower expectation is done in the same way as in the precise case, starting from the leaves and picking the right interval bounds to obtain lower values of the local expectations. In the example, we still have that $(1,0) \succ^{\mathscr{P}}_{\ell_{0/1}} (0,1)$, as the final lower expectation is positive.

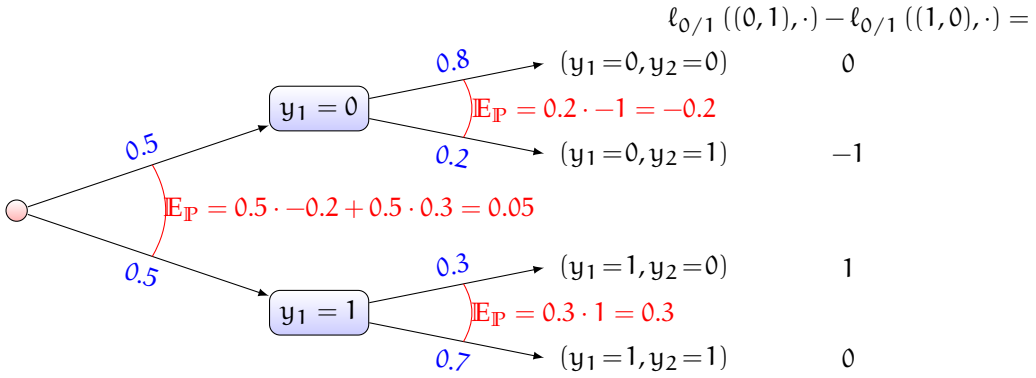


Figure 5.2: Probabilistic tree and expected loss

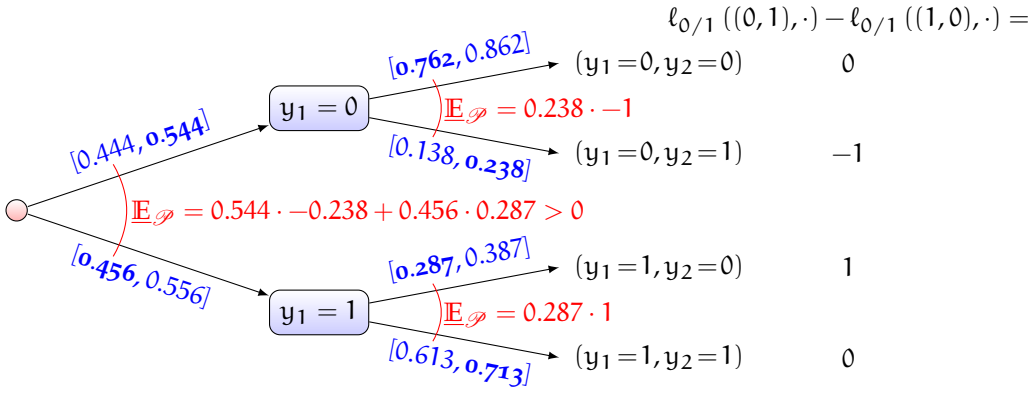


Figure 5.3: Imprecise probabilistic tree and lower expected loss

All of this means that simply enumerating elements of $\mathscr{Y}$ is not practically possible, and other strategies need to be adopted. In the next section, we show that in the case of Hamming loss, we can use efficient algorithmic procedures to perform skeptic inferences, both for general sets $\mathscr{P}$ and for specific probability sets induced from binary relevance models. We also show that some previous results giving rough outer-approximations of skeptic inferences in the general case turn out to be exact for such binary relevance models.

5.2 SKEPTIC INFERENCE FOR THE HAMMING LOSS

The hamming loss, that we will denote ℓ_H, is a commonly used loss in multi-label problems. It simply amounts to compute the Hamming distance between the ground truth $\mathbf{y}$ and a prediction $\hat{\mathbf{y}}$, that is

$$\ell_H(\hat{\mathbf{y}}, \mathbf{y}) = \sum_{i=1}^{m} \mathbb{1}_{(\hat{y}_i \neq y_i)} = |\mathscr{I}_{\hat{\mathbf{y}} \neq \mathbf{y}}| \tag{5.3}$$

where $\mathbb{1}_{(A)}$ denotes the indicator function of the event A. Note that in contrast with the subset loss $\ell_{0/1}(\hat{\mathbf{y}}, \mathbf{y}) = \mathbb{1}_{(\hat{\mathbf{y}} \neq \mathbf{y})}$, the Hamming loss differentiates the situations where only some mistakes are made from the ones where a lot of them are made (being maximum when $\hat{\mathbf{y}}$ is the complement of $\mathbf{y}$).

In the case of precise probabilities, it is also useful to recall that the optimal prediction for the Hamming loss [Dembczyński et al., 2012], i.e. the vector $\hat{\mathbf{y}}_{\ell_H, \mathbb{P}}$ satisfying Equation (1.6) is

$$\hat{y}_{i, \ell_H, \mathbb{P}} = \begin{cases} 1 & \text{if } P(Y_{\{i\}} = 1) \geqslant \frac{1}{2}, \\ 0 & \text{else.} \end{cases} \tag{5.4}$$

When considering a set $\mathscr{P}$ of distribution, one is immediately tempted to adopt the following partial vector as a solution:

$$\hat{y}^*_{i, \ell_H, \mathscr{P}} = \begin{cases} 1 & \text{if } \underline{P}(Y_{\{i\}} = 1) > \frac{1}{2}, \\ 0 & \text{if } \underline{P}(Y_{\{i\}} = 0) > \frac{1}{2}, \\ * & \text{if } \frac{1}{2} \in [\underline{P}(Y_{\{i\}} = 1), \overline{P}(Y_{\{i\}} = 1)]. \end{cases} \tag{5.5}$$

It has however been proven that $\hat{\mathbf{y}}^*_{\ell_H, \mathscr{P}}$ is in general an outer-approximation of $\hat{\mathbb{Y}}^E_{\ell, \mathscr{P}}$ and $\hat{\mathbb{Y}}^M_{\ell, \mathscr{P}}$, thus only providing a quick heuristic to get an approximate answer [Destercke, 2014].

The next sections study the problem of providing exact skeptic inferences, first for any possible probability set $\mathscr{P}$, then for the specific case where $\mathscr{P}$ is built from marginal models on each label, that corresponds to binary relevance approaches.

5.2.1 *General case*

In this section, we demonstrate that for the Hamming loss, we can use inference procedures that are much more efficient than an exhaustive, naive enumeration. Let us first simplify the expression of the expected value.

Lemma 1 *In the case of Hamming loss and given $\mathbf{y}^1, \mathbf{y}^2$, we have*

$$\mathbb{E}\left[\ell_H(\mathbf{y}^2, \cdot) - \ell_H(\mathbf{y}^1, \cdot)\right] = \sum_{i=1}^{m} P(Y_i = y_i^1) - P(Y_i = y_i^2) \tag{5.6}$$

Proof 3 (Proof of Lemma 1)
Let us first develop $\mathbb{E}\left[\ell_H(\mathbf{y}^2, \cdot) - \ell_H(\mathbf{y}^1, \cdot) | X = x\right]$:

$$\sum_{\mathbf{y} \in \mathscr{Y}} \left(\sum_{i=1}^{m} \mathbb{1}_{y_i \neq y_i^2} - \sum_{i=1}^{m} \mathbb{1}_{y_i \neq y_i^1} \right) P_x(Y = \mathbf{y})$$

$$\sum_{y_1\in\{0,1\}}\sum_{y_2\in\{0,1\}}\cdots\sum_{y_m\in\{0,1\}}\left(\sum_{i=1}^{m}\mathbb{1}_{y_i\neq y_i^2}-\sum_{i=1}^{m}\mathbb{1}_{y_i\neq y_i^1}\right)P_x(Y=\mathbf{y})$$

For a given $k\in\{1,\dots,m\}$*, let us consider the rewriting*

$$\overbrace{\sum_{y_1\in\{0,1\}}\sum_{y_2\in\{0,1\}}\cdots\sum_{y_m\in\{0,1\}}}^{m-1}\left[\sum_{y_k\in\{0,1\}}\left(\sum_{i=1}^{m}\mathbb{1}_{y_i\neq y_i^2}-\sum_{i=1}^{m}\mathbb{1}_{y_i\neq y_i^1}\right)\right]P_x(Y=\mathbf{y}). \tag{5.7}$$

Developing the sum between brackets, we get

$$\sum_{y_k\in\{0,1\}}\sum_{i=1}^{m}\mathbb{1}_{y_i\neq y_i^2}P_x(Y=\mathbf{y})-\sum_{y_k\in\{0,1\}}\sum_{i=1}^{m}\mathbb{1}_{y_i\neq y_i^1}P_x(Y=\mathbf{y})\quad\textit{(by linearity)} \tag{5.8}$$

Developing again the left term, we obtain

$$\begin{aligned}\sum_{y_k\in\{0,1\}}\sum_{i=1}^{m}\mathbb{1}_{y_i\neq y_i^2}P_x(Y=\mathbf{y}) &= \sum_{y_k\in\{0,1\}}\left(\mathbb{1}_{y_1\neq y_1^2}+\mathbb{1}_{y_2\neq y_2^2}+\cdots+\mathbb{1}_{y_m\neq y_m^2}\right)P_x(Y=\mathbf{y})\\ &= \mathbb{1}_{y_1\neq y_1^2}\sum_{y_k\in\{0,1\}}P_x(Y^k=y_k)+\cdots+\sum_{y_k\in\{0,1\}}\mathbb{1}_{y_k\neq y_k^2}P_x(Y^k=y_k)+\\ &\quad\cdots+\mathbb{1}_{y_m\neq y_m^2}\sum_{y_k\in\{0,1\}}P_x(Y^k=y_k)\\ &= \mathbb{1}_{y_1\neq y_1^2}P_x(Y_{\{-k\}})+\cdots+\sum_{y_k\in\{0,1\}}\mathbb{1}_{y_k\neq y_k^2}P_x(Y^k=y_k)+\\ &\quad\cdots+\mathbb{1}_{y_m\neq y_m^2}P_x(Y_{\{-k\}})\\ &= \sum_{y_k\in\{0,1\}}\mathbb{1}_{y_k\neq y_k^2}P_x(Y^k=y_k)+\sum_{i=1,i\neq k}^{m}\mathbb{1}_{y_i\neq y_i^2}P_x(Y_{\{-k\}}),\end{aligned}$$

where

$$P_x(Y^k=y_k):=P_x(Y_1,\dots,Y_k=y_k,\dots,Y_m) \tag{5.9}$$

and

$$P_x(Y_{\{-k\}}):=P_x(Y_1,\dots,Y_{k-1},Y_{k+1},\dots,Y_m) \tag{5.10}$$

Similarly, we get for the right term

$$\sum_{y_k\in\{0,1\}}\sum_{i=1}^{m}\mathbb{1}_{y_i\neq y_i^1}P_x(Y=\mathbf{y})=\sum_{y_k\in\{0,1\}}\mathbb{1}_{y_k\neq y_k^1}P_x(Y^k=y_k)+\sum_{i=1,i\neq k}^{m}\mathbb{1}_{y_i\neq y_i^1}P_x(Y_{\{-k\}})$$

We put back these rewritten sums in Equation (5.7)

$$\overbrace{\sum_{y_1\in\{0,1\}}\sum_{y_2\in\{0,1\}}\cdots\sum_{y_m\in\{0,1\}}}^{m-1}\left[\sum_{y_k\in\{0,1\}}\mathbb{1}_{y_k\neq y_k^2}P_x(Y^k=y_k)-\sum_{y_k\in\{0,1\}}\mathbb{1}_{y_k\neq y_k^1}P_x(Y^k=y_k)\right.$$
$$\left.+\sum_{i=1,i\neq k}^{m}\mathbb{1}_{y_i\neq y_i^2}P_x(Y_{\{-k\}})-\sum_{i=1,i\neq k}^{m}\mathbb{1}_{y_i\neq y_i^1}P_x(Y_{\{-k\}})\right]=$$
$$\sum_{y\in\mathcal{Y}}(\mathbb{1}_{y_k\neq y_k^2}-\mathbb{1}_{y_k\neq y_k^1})P_x(Y=y)+\overbrace{\sum_{y_1\in\{0,1\}}\cdots\sum_{y_m\in\{0,1\}}}^{m-1}\left[\sum_{i=1,i\neq k}^{m}\mathbb{1}_{y_i\neq y_i^2}-\mathbb{1}_{y_i\neq y_i^1}\right]P_x(Y_{\{-k\}}) \tag{5.11}$$

The left term can be reduced in the following way:

$$\begin{aligned}\sum_{y\in\mathcal{Y}}(\mathbb{1}_{y_k\neq y_k^2}-\mathbb{1}_{y_k\neq y_k^1})P_x(Y=y)&=\sum_{y\in\mathcal{Y}}\mathbb{1}_{y_k\neq y_k^2}P_x(Y=y)-\sum_{y\in\mathcal{Y}}\mathbb{1}_{y_k\neq y_k^1}P_x(Y=y)\\&=P_x(Y_k\neq y_k^2)-P_x(Y_k\neq y_k^1)\\&=P_x(Y_k=y_k^1)-P_x(Y_k=y_k^2)\end{aligned}$$

since we have $P_x(Y_k\neq y_k)=1-P_x(Y_k=y_k)$. *We can apply the same operations we just did on the right term of Equation* (5.11), *and do so recursively, to finally obtain*

$$\sum_{i=1}^{m}P(Y_i=y_i^1)-P(Y_i=y_i^2) \tag{5.12}$$

■

If we consider a set of indices $\mathcal{I}_{y^1=y^2}$ on which the Equation (5.6) is cancelled, it can be rewritten

$$\sum_{i\in\mathcal{I}_{y^1\neq y^2}}P(Y_i=y_i^1)-P(Y_i=y_i^2). \tag{5.13}$$

The next proposition shows that this expression can be leveraged to perform the maximality check of Equation (5.2) on a limited number of vectors.

Proposition 2 *For a given set* $\mathcal{I}$ *of indices, let us consider an assignment* $a_{\mathcal{I}}$ *and its complement* $\overline{a}_{\mathcal{I}}$. *Then, for any two vectors* y^1, y^2 *such that* $y^1_{\mathcal{I}}=a_{\mathcal{I}}$, $y^2_{\mathcal{I}}=\overline{a}_{\mathcal{I}}$ *and* $y^1_{-\mathcal{I}}=y^2_{-\mathcal{I}}$, *we have*

$$y^1\succ_M y^2\iff\inf_{P\in\mathcal{P}}\sum_{i\in\mathcal{I}}P(Y_i=a_i)>\frac{|\mathcal{I}|}{2} \tag{5.14}$$

Proof 4 (Proof of Proposition 2) *Using Equation* (5.13), *one can readily see that*

$$\mathbf{y}^1 \succ_M \mathbf{y}^2 \iff \inf_{P\in\mathscr{P}} \sum_{i\in\mathscr{I}_{\mathbf{y}^1\neq\mathbf{y}^2}} P(Y_i = y_i^1) - P(Y_i = y_i^2) > 0 \tag{5.15}$$

$$\iff \inf_{P\in\mathscr{P}} \sum_{i\in\mathscr{I}} P(Y_i = a_i) - P(Y_i = \overline{a}_i) > 0 \tag{5.16}$$

Accounting for the fact that $P(Y_i = a_i) + P(Y_i = \overline{a}_i) = 1$, *we get*

$$\iff \inf_{P\in\mathscr{P}} \sum_{i\in\mathscr{I}} 2P(Y_i = a_i) - 1 > 0 \tag{5.17}$$

$$\iff \inf_{P\in\mathscr{P}} \sum_{i\in\mathscr{I}} P(Y_i = a_i) > \frac{|\mathscr{I}|}{2} \tag{5.18}$$

■

In the remaining of the chapter, given a partial assignment $\mathbf{b}_{\mathscr{I}}$ over a subset of indices $\mathscr{I}$, we will define the partial Hamming loss between $\mathbf{b}_{\mathscr{I}}$ and an observation $\mathbf{y}$ as

$$\ell_H^*(\mathbf{b}_{\mathscr{I}}, \mathbf{y}) = \sum_{i\in\mathscr{I}} \mathbb{1}_{(b_i\neq y_i)} \,. \tag{5.19}$$

It is clear that when $\mathscr{I} = \{1, \ldots, m\}$, we simply retrieve the usual Hamming loss. The next proposition shows that the condition of Proposition 2 actually comes down to minimize the expected partial Hamming loss.

Proposition 3 *For a given set* $\mathscr{I}$ *of indices, let us consider an assignment* $\mathbf{a}_{\mathscr{I}}$ *and its complement* $\overline{\mathbf{a}}_{\mathscr{I}}$. *We have*

$$\inf_{P\in\mathscr{P}} \sum_{i\in\mathscr{I}} P(Y_i = a_i) = \underline{\mathbb{E}}[\ell_H^*(\overline{\mathbf{a}}_{\mathscr{I}}, \cdot)] \tag{5.20}$$

Proof 5 (Proof of Proposition 3) *First, let us simply notice that* $P(Y_i = a_i) = \sum_{y\in\mathscr{Y}} \mathbb{1}_{y_i=a_i} P(Y = \mathbf{y})$ *and* $\mathbb{1}_{y_i=a_i} = \mathbb{1}_{y_i\neq\overline{a}_i}$. *Putting these together, we get*

$$\begin{aligned}
\sum_{i\in\mathscr{I}} P(Y_i = a_i) &= \sum_{i\in\mathscr{I}} \sum_{\mathbf{y}\in\mathscr{Y}} \mathbb{1}_{y_i\neq\overline{a}_i} P(Y = \mathbf{y}) \\
&= \sum_{\mathbf{y}\in\mathscr{Y}} \sum_{i\in\mathscr{I}} \mathbb{1}_{y_i\neq\overline{a}_i} P(Y = \mathbf{y}) \qquad \text{(by linearity)} \\
&= \mathbb{E}[\ell_H^*(\overline{\mathbf{a}}_{\mathscr{I}}, \cdot)]
\end{aligned}$$

where $\ell_H^*(\overline{\mathbf{a}}_{\mathscr{I}}, \cdot)$ *is the hamming loss calculated in the set of indices* $\mathscr{I} = \{i_1, \ldots, i_q\}$ *of vector* $\overline{\mathbf{a}}_{\mathscr{I}}$, *which is created in the line 5 of the Algorithm 2. Thus, we apply infimum,* $\inf_{\mathbb{P}\in\mathscr{P}}$, *to each side of the last equation and get what we sought.* ■

This allows us to use Algorithm 2 to find $\hat{\mathbb{Y}}^M_{\ell_H,\mathscr{P}}$. The following result provides the time complexity of the algorithm.

Algorithm 2: Maximal solutions under Hamming loss and general set

Data: $\mathscr{P}$ (convex set of distributions)
Result: $\hat{\mathbb{Y}}^{M}_{\ell_H,\mathscr{P}}$ (set of undominated solutions)

```
1  S = 𝒴;
2  for i in 1:m do
3  |  𝒵_i = {ℐ : ℐ ⊆ {1,...,m}, |ℐ| = i} ;          // Index sets of size i
4  |  forall z ∈ 𝒵_i do
5  |  |  forall a_z ∈ 𝒴_z ;                        // Binary vectors over indices in z
6  |  |  do
7  |  |  |  if inf_{P∈𝒫} Σ_{j∈z} P(Y_j = a_j) > i/2 then
   |  |  |  |  S = S \ {y ∈ 𝒴 : y_z = ā_z};
8  |  |  end
9  |  end
10 end
```

Proposition 4 *Algorithm 2 has to perform $3^m - 1$ computations, and its complexity is in $\mathcal{O}(3^m)$*

Proof 6 (Proof of Proposition 4) *Let us simply analyze the number of computations needed. We will need to perform m times the loop of Line 2. For a given i, we have that $\mathcal{Z}_i = \binom{m}{i}$, meaning that this is the number of elements to check in the loop starting Line 4. Finally, there 2^i elements to check in the loop starting Line 5. The table below summarise the different steps.*

Index Line 2	$i=1$	$i=2$	$\dots$	$i=m-2$	$i=m-1$	$i=m$
$\lvert\mathcal{Z}_i\rvert$	$\frac{m!}{1!(m-1)!}$	$\frac{m!}{2!(m-2)!}$	$\cdots$	$\frac{m!}{(m-2)!2!}$	$\frac{m!}{(m-1)!1!}$	$\frac{m!}{m!0!}$
$\lvert\mathscr{Y}_z\rvert$	$\{0,1\}^1$	$\{0,1\}^2$	$\dots,$	$\{0,1\}^{m-2}$	$\{0,1\}^{m-1}$	$\{0,1\}^m$

Overall, the number of checks to perform amounts to

$$\sum_{k=1}^{m} 2^{m-k} \frac{m!}{k!(m-k)!} = 3^m - 1 \tag{5.21}$$

■

Proposition 4 tells us that, in the case of Hamming loss, finding $\hat{\mathbb{Y}}^{M}_{\ell,\mathscr{P}}$ can be done almost linearly with respect to the size of $\mathscr{Y}$. This is to be compared to a naive enumeration, that requires $(2^m)(2^m-1)$ computations. Figure 5.4 plots the two curves as a function of the number m of labels, demonstrating that our result allows a significant gain in computations.

In the experiments of Section 5.3, we shall study the differences between $\hat{\mathbb{Y}}^{\mathrm{M}}_{\ell,\mathscr{P}}$ and the crude approximation of Equation (5.5).

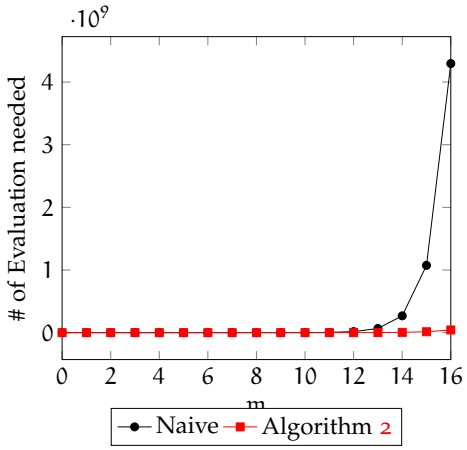


Figure 5.4: Comparison of Algorithm 2 with naive enumeration.

It should be noted that the time complexity obtained in Proposition 2 can still be reduced using two different strategies: (1) removing dominated elements already verified (cf. [Augustin et al., 2014, algo. 16.4]), and (2) using the precise prediction as a solution of the set of solutions E-admissible (c.f [Carranza Alarcón et al., 2020e, appx. A]). However, in the worst-case scenario, in which all elements are non-dominated, the time complexity remains the same.

As said before, the set $\hat{\mathbb{Y}}^{\mathrm{M}}_{\ell_{\mathrm{H}},\mathscr{P}}$ will in general not be exactly described by a partial vector within $\mathscr{Y}^*$, as shows the next example.

Example 10 *Consider again the tree provided in Figure 5.3. The result of applying Algorithm 2 provides the following results:*

$$
\begin{aligned}
\underline{\mathbb{E}}\left[\ell_{\mathrm{H}}((1,*),\cdot)\right] = 0.444 > 0.5 &\implies (0,*) \not\succ^{\mathscr{P}}_{\ell_{\mathrm{H}}} (1,*),\\
\underline{\mathbb{E}}\left[\ell_{\mathrm{H}}((0,*),\cdot)\right] = 0.456 > 0.5 &\implies (1,*) \not\succ^{\mathscr{P}}_{\ell_{\mathrm{H}}} (0,*),\\
\underline{\mathbb{E}}\left[\ell_{\mathrm{H}}((*,1),\cdot)\right] = 0.498 > 0.5 &\implies (*,0) \not\succ^{\mathscr{P}}_{\ell_{\mathrm{H}}} (*,1),\\
\underline{\mathbb{E}}\left[\ell_{\mathrm{H}}((*,0),\cdot)\right] = 0.354 > 0.5 &\implies (*,1) \not\succ^{\mathscr{P}}_{\ell_{\mathrm{H}}} (*,0),\\
\underline{\mathbb{E}}\left[\ell_{\mathrm{H}}((1,1),\cdot)\right] = 0.942 > 1.0 &\implies (0,0) \not\succ^{\mathscr{P}}_{\ell_{\mathrm{H}}} (1,1),\\
\underline{\mathbb{E}}\left[\ell_{\mathrm{H}}((1,0),\cdot)\right] = 0.846 > 1.0 &\implies (0,1) \not\succ^{\mathscr{P}}_{\ell_{\mathrm{H}}} (1,0),\\
\underline{\mathbb{E}}\left[\ell_{\mathrm{H}}((0,1),\cdot)\right] = 1.001 > 1.0 &\implies (\mathbf{1},\mathbf{0}) \succ^{\mathscr{P}}_{\ell_{\mathbf{H}}} (\mathbf{0},\mathbf{1}),\\
\underline{\mathbb{E}}\left[\ell_{\mathrm{H}}((0,0),\cdot)\right] = 0.810 > 1.0 &\implies (1,1) \not\succ^{\mathscr{P}}_{\ell_{\mathrm{H}}} (0,0),
\end{aligned}
$$

where for two partial vectors $\mathbf{y}^1, \mathbf{y}^2$ such that $\mathscr{I}^*_{\mathbf{y}^1} = \mathscr{I}^*_{\mathbf{y}^2}$, we use the short-hand notation $\mathbf{y}^1 \succ^{\mathscr{P}}_{\ell_{\mathrm{H}}} \mathbf{y}^2$ to say that the dominance relation given by Definition 5 holds for any fixed replacement of the abstained labels.

About this example, we can first note that only $3^2 - 1 = 8$ *comparisons are performed (in accord with Proposition 4). Secondly, also note that the final solution which is the set*

$$\hat{\mathbb{Y}}^{M}_{\ell_H,\mathscr{P}} = \{(1,0),(0,0),(1,1)\}$$

does not belong to $\mathscr{Y}^*$.

Remark 6 *Note that if for some partial vectors* $\mathbf{y}$ *Proposition 2 holds, the preference also holds for any completion of such a vector. More precisely, if we denote by* $\overline{\mathscr{I}}^*$ *the indices of non-abstained labels, and* $\mathbf{a}_{\mathscr{J}}$ *an assignment over indices* $\mathscr{J} \subseteq \mathscr{I}^*$ *(where* $\mathscr{I}^* = \{1,\ldots,m\} \setminus \overline{\mathscr{I}}^*$*), one can deduce from* $\mathbf{y} \succ^{\mathscr{P}}_{\ell_H} \overline{\mathbf{y}}$ *that* $(\mathbf{y}_{\overline{\mathscr{I}}^*}, \mathbf{a}_{\mathscr{J}}) \succ^{\mathscr{P}}_{\ell_H} (\overline{\mathbf{y}}_{\overline{\mathscr{I}}^*}, \mathbf{a}_{\mathscr{J}})$*. For instance if* $(0,*,*) \succ^{\mathscr{P}}_{\ell_H} (1,*,*)$*, then we can additionally deduce* $(0,*,0) \succ^{\mathscr{P}}_{\ell_H} (1,*,0)$.

Remark 7 *A key finding of the results of this section, illustrated by Example 10, is that when considering sets of distributions and skeptic inferences, it is not sufficient to consider marginal probabilities in order to get optimal, exact predictions. This contrasts heavily with the case of precise distributions, in which having only the marginal information allows to get optimal predictions for a number of loss functions, including the Hamming loss, but also precision@k, micro- and macro-F measure, as well as others [Kotlowski et al., 2016; Koyejo et al., 2015].*

5.2.2 *Binary relevance and partial vectors*

The previous section looked at the very general case where the set $\mathscr{P}$ is completely arbitrary and proposed some efficient inference methods for this case. In this section, we are interested in conditions imposed upon $\mathscr{P}$ that guarantee the sets $\hat{\mathbb{Y}}^{M}_{\ell_H,\mathscr{P}}$ and $\hat{\mathbb{Y}}^{E}_{\ell_H,\mathscr{P}}$ to be partial vectors, that is to belong to $\mathscr{Y}^*$. In particular, we show that this is the case when considering models that generalize binary relevance notions by using imprecise marginals with an assumption of independence. The interest of studying such models is that they constitute the basic models when it comes to multi-label problems.

Before proceeding to the generalization of binary relevance models producing partial binary vectors, we first prove an intermediate useful result characterising partial vectors in terms of the vector set they represent. More precisely, we first express a condition for a subset $\mathbb{Y}$ of $\mathscr{Y}$ to be a partial vector, in terms of its elements.

Lemma 2 *A subset* $\mathbb{Y}$ *belongs to the space* $\mathscr{Y}^*$ *if and only if*

$$\forall \mathbf{y},\mathbf{y}' \in \mathbb{Y}, \text{ we have that all } \mathbf{y}'' \in \mathscr{Y} \text{ s.t. } y''_i = y'_i \quad \forall i \in \mathscr{I}_{\mathbf{y}=\mathbf{y}'} \text{ are also in } \mathbb{Y}$$

Proof 7 (Proof of Lemma 2) ***Only if****: Immediate, since by assumption* $\mathscr{I}_{\mathbf{y}\neq\mathbf{y}'} \subseteq \mathscr{I}^*$*, the set of label indices on which we abstain.*

***If**: Consider the set $D_{\mathbb{Y}} = \{j | \exists \mathbf{y}, \mathbf{y}' \in \mathbb{Y}, y_j \neq y'_j\}$ of indices for which at least two elements of $\mathbb{Y}$ disagree. What we have to show is that under the condition of Lemma 2, any completion of $D_{\mathbb{Y}}$ is within $\mathbb{Y}$.*

Without loss of generality, as we can always permute the indices, let us consider that $D_{\mathbb{Y}}$ are the $|D_{\mathbb{Y}}|$ first indices. We can then find a couple $\mathbf{y}, \mathbf{y}' \in \mathbb{Y}$ such that the k first elements are distinct, that is $\mathscr{I}_{\mathbf{y} \neq \mathbf{y}'} = \{1, \ldots, k\}$. It follows that the subset of vectors

$$(\underbrace{*, \ldots, *}_{k \text{ times}}, y_{k+1}, \ldots, y_{|D_{\mathbb{Y}}|}, y_{|D_{\mathbb{Y}}|+1}, \ldots, y_m) \tag{5.22}$$

is within $\mathbb{Y}$, by assumption. If $k < |D_{\mathbb{Y}}|$, we can find a vector $\mathbf{y}''$ such that its k' next elements (after the kth first) are different from $\mathbf{y}$, i.e., $y_j \neq y''_j$ for $j = k+1, \ldots, k+k'$ with $k + k' \leqslant |D_{\mathbb{Y}}|$. Note that $k' \geqslant 1$ by assumption. Since the vector (5.22) is in $\mathbb{Y}$, we can always consider the vector y such that its k first elements are different from those of $\mathbf{y}''$, that is in $\mathbb{Y}$. Since $\mathscr{I}_{\mathbf{y} \neq \mathbf{y}'} = \{1, \ldots, k+k'\}$, the subset of vectors

$$(\underbrace{*, \ldots, *}_{k+k' \text{ times}}, y_{k+k'+1}, \ldots, y_{|D_{\mathbb{Y}}|}, y_{|D_{\mathbb{Y}}|+1}, \ldots, y_m)$$

is also in $\mathbb{Y}$. Since we can repeat this construction until having two vectors with the $|D_{\mathbb{Y}}|$ first labels different, this finishes the proof. ■

We now consider that the joint probability p over $\mathscr{Y}$ and its imprecise extension are built in the following way: we have some information on the marginal probability $p_i \in [0, 1]$ of y_i being positive, and define the probability of a vector $\mathbf{y}$ as

$$p(\mathbf{y}) = \prod_{\{i | y_i = 1\}} p_i \prod_{\{i | y_i = 0\}} (1 - p_i). \tag{5.23}$$

Without loss of generality, the imprecise version then amounts to consider that the information we have is an interval $[\underline{p}_i, \overline{p}_i]$, as every convex set of probabilities on a binary space (here, $\{0, 1\}$) is an interval. We then consider that a probability set $\mathscr{P}_{BR}$ over $\mathscr{Y}$ amounts to consider the robust version of Equation (5.23), that is

$$p(\mathbf{y}) \in \left\{ \prod_{\{i | y_i = 1\}} p_i \prod_{\{i | y_i = 0\}} (1 - p_i) \middle| p_i \in [\underline{p}_i, \overline{p}_i] \right\}. \tag{5.24}$$

In this specific case, we can show that $\hat{\mathbb{Y}}^E_{\ell_H, \mathscr{P}}$ can be exactly described by a partial vector.

Proposition 5 *Given a probability set $\mathscr{P}_{BR}$ and the Hamming loss, the set $\hat{\mathbb{Y}}^E_{\ell_H, \mathscr{P}_{BR}} \in \mathscr{Y}^*$*

Proof 8 (Proof of Proposition 5) *Let us first notice that, after Equation (5.4), we have that*

$$\mathbf{y} \in \hat{\mathbb{Y}}^{E}_{\ell_H,\mathscr{P}} \iff \begin{cases} \underline{p}_i \leqslant 0.5 & \text{for } i \in \mathscr{I}_{\mathbf{y}=0} \\ \overline{p}_i \geqslant 0.5 & \text{for } i \in \mathscr{I}_{\mathbf{y}=1} \end{cases} \tag{5.25}$$

where $\mathscr{I}_{\mathbf{y}=0}$, $\mathscr{I}_{\mathbf{y}=1}$ are the indices of labels for which $y_i = 0$ and $y_i = 1$. Indeed, since here we start from the marginals, $\mathbf{y}$ is optimal according to Hamming loss and a distribution in $\mathscr{P}_{BR}$ iff we can fix p_i to be lower than 0.5 if $y_i = 0$, and higher else.

Now, let us consider two vectors $\mathbf{y}^1, \mathbf{y}^2$ and the indices $\mathscr{I}_{\mathbf{y}^1 \neq \mathbf{y}^2}$. Given the first part of this proof, if $\mathbf{y}^1, \mathbf{y}^2 \in \hat{\mathbb{Y}}^{E}_{\ell_H,\mathscr{P}}$, this means that $0.5 \in [\underline{p}_i, \overline{p}_i]$ for any $i \in \mathscr{I}_{\mathbf{y}^1 \neq \mathbf{y}^2}$. Therefore, given any vector $\mathbf{y}''$ such that $y''_i = y^1_i$ for $i \in \mathscr{I}_{\mathbf{y}^1 = \mathbf{y}^2}$, for the other indices $i \in \mathscr{I}_{\mathbf{y}^1 \neq \mathbf{y}^2}$, we can always fix a precise value $p_i \in [\underline{p}_i, \overline{p}_i]$ such that $\mathbf{y}''$ is also optimal w.r.t. $\mathbf{p}$. More precisely, assume the assignments p^1_i and p^2_i result in $\mathbf{y}^1, \mathbf{y}^2$ being optimal predictions for the Hamming loss, respectively. Then $\mathbf{y}''$ is optimal for the assignment

$$p''_i = \begin{cases} p^1_i & \text{if } y''_i = y^1_i \\ p^2_i & \text{if } y''_i = y^2_i \end{cases}$$

that is by definition within $\mathscr{P}_{BR}$. ■

Proposition 5 shows that in the specific yet important case of binary relevance models, $\hat{\mathbb{Y}}^{E}_{\ell_H,\mathscr{P}}$ can be computed efficiently and easily presented to users. In particular, Equation (5.5) is in this case exact, and can be used to compute $\hat{\mathbb{Y}}^{E}_{\ell_H,\mathscr{P}_{BR}}$. We will denote by $\hat{\mathbf{y}}^{*}_{\ell_H,\mathscr{P}_{BR}}$ the partial vector corresponding to $\hat{\mathbb{Y}}^{E}_{\ell_H,\mathscr{P}_{BR}}$, as the next proposition shows that it is also an exact estimation of $\hat{\mathbb{Y}}^{M}_{\ell_H,\mathscr{P}_{BR}}$.

Proposition 6 *Given a probability set $\mathscr{P}_{BR}$ and the Hamming loss, we have*

$$\hat{\mathbb{Y}}^{E}_{\ell_H,\mathscr{P}_{BR}} = \hat{\mathbb{Y}}^{M}_{\ell_H,\mathscr{P}_{BR}}.$$

Proof 9 (Proof of Proposition 6) *As Proposition 5 shows, the E-admissible set is given by the partial vector $\hat{\mathbf{y}}^{*}_{\ell_H,\mathscr{P}_{BR}}$. To show that it also coincides with $\hat{\mathbb{Y}}^{M}_{\ell_H,\mathscr{P}_{BR}}$, we will consider the fact that $\hat{\mathbf{y}}^{*}_{\ell_H,\mathscr{P}_{BR}} \subseteq \hat{\mathbb{Y}}^{M}_{\ell_H,\mathscr{P}_{BR}}$, and will demonstrate that any vector outside $\hat{\mathbf{y}}^{*}_{\ell_H,\mathscr{P}_{BR}}$ is dominated (in the sense of Equation (5.2)) by a vector within $\hat{\mathbf{y}}^{*}_{\ell_H,\mathscr{P}_{BR}}$.*

Let us consider a vector $\mathbf{y}' \notin \hat{\mathbf{y}}^{*}_{\ell_H,\mathscr{P}_{BR}}$, and the indices

$$\mathscr{I}_{\mathbf{y}' \neq \hat{\mathbf{y}}^*} = \{i : \hat{y}^{*}_{i,\ell_H,\mathscr{P}_{BR}} \neq *, \hat{y}^{*}_{i,\ell_H,\mathscr{P}_{BR}} \neq y'_i\}$$

on which they necessarily differ (as we can always set the labels for which $\hat{y}^{*}_{i,\ell_H,\mathscr{P}_{BR}} = *$ to be equal to y'_i). By Proposition 1, we have that

$$\hat{\mathbf{y}}^*_{\ell_H,\mathscr{P}_{BR}} \succ_M \mathbf{y}' \iff \inf_{P\in\mathscr{P}} \sum_{i\in\mathscr{I}_{\mathbf{y}'\neq\hat{\mathbf{y}}^*}} P(Y_i = \hat{y}^*_i) > \frac{|\mathscr{I}|}{2}$$

and since we have that $\hat{y}^*_i = 1 \Rightarrow \underline{P}(Y_i = 1) > 0.5$ *and* $\hat{y}^*_i = 0 \Rightarrow \underline{P}(Y_i = 0) > 0.5$*, the right hand side inequality is satisfied. Hence, we can show that any vector outside* $\hat{\mathbf{y}}^*_{\ell_H,\mathscr{P}_{BR}}$ *is maximally dominated by another vector in* $\hat{\mathbf{y}}^*_{\ell_H,\mathscr{P}_{BR}}$*, meaning that* $\hat{\mathbf{y}}^*_{\ell_H,\mathscr{P}_{BR}} \supseteq \hat{\mathbb{Y}}^M_{\ell_H,\mathscr{P}_{BR}}$*. Combined with the fact that* $\hat{\mathbf{y}}^*_{\ell_H,\mathscr{P}_{BR}} \subseteq \hat{\mathbb{Y}}^M_{\ell_H,\mathscr{P}_{BR}}$*, this finishes the proof.* ■

Remark 8 *As the optimal prediction for the o/1 or subset loss* $\ell_{0/1}$ *in the precise case is the same as Equation (5.4), Proposition 5 is also true for this loss, as well as Proposition 6.*

In Section 5.2.3, we provide a couple of complementary results with respect to other decision criteria, which are either very conservative (i.e., interval dominance) or not skeptic (i.e., minimax and minimin), in the sense that their inferences are always precisely valued, not matter how big $\mathscr{P}$ is.

5.2.3 *On Binary relevance and other decision criterions*

So far, we considered only the most common skeptic decision criteria (maximality and E-admissibility) in order to get a set of predictions (either partial or not). However, there are other decision criteria using probability sets and extending the classical expected loss criterion. The most common being (1) Interval dominance, (2) Γ-minimin, (3) Γ-maximin (at work in distributionally robust approaches).

For a complete reminder about definitions of these last three criteria, we refer to Section 2.1.2. We simply remind that in general one has that $\hat{\mathbf{y}}^{\Gamma_{min}}_{\ell,\mathscr{P}} \neq \hat{\mathbf{y}}^{\Gamma_{max}}_{\ell,\mathscr{P}}$, yet when considering probability sets satisfying the hypothesis of Section 5.2.2, we have the following result regarding Γ-minimin and Γ-maximin criteria:

Proposition 7 *Given a probability set* $\mathscr{P}_{BR}$ *and the Hamming loss* ℓ_H*, we have:*

$$\hat{\mathbf{y}}^{\Gamma_{max}}_{\ell_H,\mathscr{P}_{BR}} = \hat{\mathbf{y}}^{\Gamma_{min}}_{\ell_H,\mathscr{P}_{BR}} \tag{5.26}$$

Proof 10 (proof of Proposition 7) *Let us prove first how get the prediction for each decision criteria, by harnessing the facts that the Hamming loss is decomposable and that* $\mathbb{E}\left[\ell_H(\mathbf{y},\cdot)\right] = m - \sum_i P(Y_i = y_i)$*.*

1. Γ-MINIMAX.— *as the Hamming loss is decomposable we can easily reduce Equation (2.7) as follows:*

$$\arg\min_{\mathbf{y}\in\mathscr{Y}} \overline{\mathbb{E}}_{\mathscr{P}}\left[\ell_H(\mathbf{y},\cdot)\right] \iff \arg\max_{\mathbf{y}\in\mathscr{Y}} \inf_{P\in\mathscr{P}} \sum_{i=1}^{m} P(Y_i = \hat{y}_i) \tag{5.27}$$

by using a probability set $\mathscr{P}_{\mathrm{BR}}$, we have

$$\hat{\mathbf{y}}^{\Gamma_{\max}} = \arg\max_{\mathbf{y}\in\mathscr{Y}} \sum_{i=1}^{m} \underline{P}(Y_i = y_i) \iff \hat{\mathbf{y}}^{\Gamma_{\max}} = \begin{cases} 1 & \text{if } \underline{P}(Y_i = 1) > \underline{P}(Y_i = 0) \\ 0 & \text{if } \underline{P}(Y_i = 1) < \underline{P}(Y_i = 0) \\ * & \text{otherwise} \end{cases} \tag{5.28}$$

where $*$ can here be replaced by 0 or 1, as all those predictions are deemed indifferent by the Minimax *principle.*

2. Γ-Minimin.— *in the same way as previously, we easily have :*

$$\hat{\mathbf{y}}^{\Gamma_{\min}} = \arg\max_{\mathbf{y}\in\mathscr{Y}} \sum_{i=1}^{m} \overline{P}(Y_i = y_i) \iff \hat{\mathbf{y}}^{\Gamma_{\min}} = \begin{cases} 1 & \text{if } \overline{P}(Y_i = 1) > \overline{P}(Y_i = 0) \\ 0 & \text{if } \overline{P}(Y_i = 1) < \overline{P}(Y_i = 0) \\ * & \text{otherwise} \end{cases} \tag{5.29}$$

where $*$ is to be understood as in the previous case.

By using the fact the the lower and upper probabilities are dual, $\overline{P}(Y_i = 1) = 1 - \underline{P}(Y_i = 0)$, it is easy to see that the criteria to choose the $\hat{\mathbf{y}}^{\Gamma_{\min}}$ is equal to $\hat{\mathbf{y}}^{\Gamma_{\max}}$, therefore, $\hat{\mathbf{y}}^{\Gamma_{\max}}_{\ell_H,\mathscr{P}_{\mathrm{BR}}} = \hat{\mathbf{y}}^{\Gamma_{\min}}_{\ell_H,\mathscr{P}_{\mathrm{BR}}}$. ■

It is well known that the set $\hat{Y}^{\mathrm{ID}}_{\ell,\mathscr{P}}$ is a superset of $\hat{\mathbb{Y}}^{\mathrm{M}}_{\ell_H,\mathscr{P}}$, due to its conservative nature. The next simple example shows that even in the case of binary relevance models, this inclusion can be strict.

Example 11 *Consider the simple case where we have two labels with the following bounds: $P(Y_1 = 1) \in [0.6, 1]$ and $P(Y_2 = 1) \in [0, 1]$. We then have the following expectation bounds for the various predictions and a Hamming loss*

$\mathbf{y}$	$(1,1)$	$(1,0)$	$(0,1)$	$(0,0)$
$\underline{\mathbb{E}}[\ell_H(\mathbf{y},\cdot)]$	0	0	0.6	0.6
$\overline{\mathbb{E}}[\ell_H(\mathbf{y},\cdot)]$	1.4	1.4	2	2

from which we deduce that $\hat{Y}^{\mathrm{ID}}_{\ell_H,\mathscr{P}} = (*,*)$, while $\hat{\mathbb{Y}}^{\mathrm{M}}_{\ell_H,\mathscr{P}} = (1,*)$.

Corollary 1 *Given a probability set $\mathscr{P}_{\mathrm{BR}}$ and the Hamming loss ℓ_H, in the Figure 5.5, we can show the following implications for the different decision criteria that are* **M***aximality,* **E***-admissibility, Γ-minimax, Γ-minimin, and* **I***nterval* **D***ominance. As usual with sets, an implication $A \to B$ means that $A \subseteq B$.*

5.3 EXPERIMENTS

In this section, we perform some empirical experiments[1] showing the interest of using skeptical inferences rather than precisely-valued inferences

[1] Implemented in Python, see https://github.com/sdestercke/classifip.

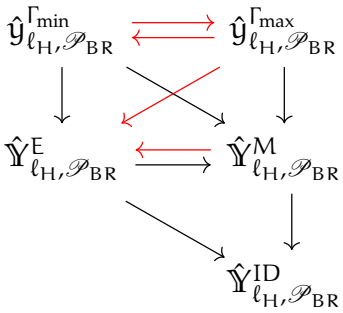


Figure 5.5: Decision relation under a $\mathscr{P}_{BR}$ and a ℓ_H. In red arrow, the new implications.

when uncertainties are too high. First, in Section 5.3.1, we formalize the procedure we used in Example 9 to compute the lower expectation in binary tree structures used for instance to verify Equation of Proposition 2. Section 5.3.2 and 5.3.3 respectively describe experimental results on simulated and real-word data sets. The first experiment aims to evaluate the quality of our proposal against the outer-approximation described in [Destercke, 2014]. The second aims, under the assumption of independence detailed in Section 5.2.2 on labels, to verify how skeptical and precise inferences cope with the following two different settings: missing and noisy labels.

5.3.1 *Inference in imprecise binary trees*

As we saw in Proposition 2 and Algorithm 2, estimating $\hat{Y}^{M}_{\ell_H,\mathscr{P}}$, given an observed instance $\mathbf{x}$, implies the calculation of the infimum expectation $\underline{\mathbb{E}}_{\mathbf{Y}|\mathbf{X}=\mathbf{x}}[\ell_H(\cdot,\overline{a}_{\mathscr{I}})]$ given an assignment $a_{\mathscr{I}}$. One possibility to compute it is to write it as an iterated conditional expectation over the chain of labels, i.e.,

$$\underline{\mathbb{E}}_{\mathbf{Y}|\mathbf{X}=\mathbf{x}}[\ell_H(\cdot,\overline{\mathbf{a}}_{\mathscr{I}})] = \inf_{P\in\mathscr{P}} \mathbb{E}_{Y_1}\left[\mathbb{E}_{Y_2}\left[\dots\mathbb{E}_{Y_m}\left[\ell_H(\cdot,\overline{\mathbf{a}}_{\mathscr{I}})\middle|\mathbf{X}=\mathbf{x},Y_{\mathscr{I}_{[\![m-1]\!]}}=\mathbf{y}_{\mathscr{I}_{[\![m-1]\!]}}\right]\dots\right]\middle|\mathbf{X}=\mathbf{x}\right], \tag{5.30}$$

where $[\![j]\!]=\{1,2,\dots,j-1,j\}$ is a set of previous indices and $Y_{\mathscr{I}_{[\![m-1]\!]}}=\{Y_1,\dots,Y_{m-1}\}$ is a random binary vector. While such an expectation has to be computed globally, it has been shown by Hermans and De

Cooman [Hermans et al., 2009] that in the specific case of tree structures, it can be computed recursively [2] using the law of iterated lower expectations[3]

$$\underline{\mathbb{E}}_{\mathbf{Y}|\mathbf{X}=\mathbf{x}}\left[\ell_H(\cdot,\overline{\mathbf{a}}_{\mathscr{I}})\right] = \underline{\mathbb{E}}_{Y_1}\left[\underline{\mathbb{E}}_{Y_2}\left[\ldots\underline{\mathbb{E}}_{Y_m}\left[\ell_H(\cdot,\overline{\mathbf{a}}_{\mathscr{I}})\middle|\mathbf{X}=\mathbf{x},Y_{\mathscr{I}_{[\![m-1]\!]}}=\mathbf{y}_{\mathscr{I}_{[\![m-1]\!]}}\right]\ldots\right]\middle|\mathbf{X}=\mathbf{x}\right]. \tag{5.31}$$

Equation (5.31) computes the global infimum expectation by backward recursion, i.e., we first compute the local lower expectation starting from the leaves of the tree and proceed iteratively (for further details see [Yang et al., 2014]). Example 12 provides us an illustration of this backward recursion.

Example 12 *Let us consider a multi-label problem with two labels* $\{Y_1, Y_2\}$ *with the credal set* $\mathscr{P}$ *over* $\mathscr{Y}$ *defined by the tree pictured in Figure 5.6. Consider* $\mathbf{y}^1 = (\cdot, 1)$ *and* $\mathbf{y}^2 = (\cdot, 0)$ *two binary vectors which have the same value of the label* Y_1*. According to Proposition 2, the assignment of these vectors is* $\mathbf{a}_{\{2\}} = (1)$ *(and its complement* $\overline{\mathbf{a}}_{\{2\}} = (0)$*). In order to verify whether* $\mathbf{y}^1$ *dominates* $\mathbf{y}^2$ *(in the sense of the maximality criterion), we have to check whether*

$$\underline{\mathbb{E}}_{\mathbf{Y}|\mathbf{X}=\mathbf{x}}\left[\ell^*_H(\overline{\mathbf{a}}_{\mathscr{I}},\cdot)\right] > 0.5, \tag{5.32}$$

where the cost vector of the partial Hamming loss is $= (0,1,1,0)$ *as can be verified in Figure 5.6.*

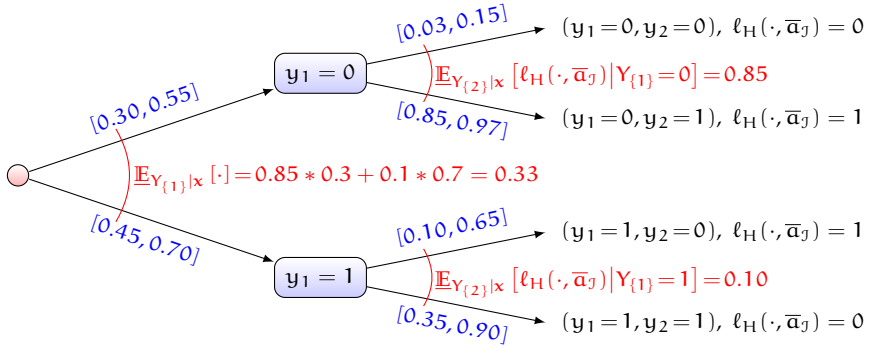


Figure 5.6: Example of computing the infimum expectation.

Thus, applying recursively Equation (5.31), we obtain an infimum expectation $\underline{\mathbb{E}}_{\mathbf{Y}|\mathbf{X}=\mathbf{x}}\left[\ell_H(\cdot,\overline{\mathbf{a}}_{\mathscr{I}})\right] = 0.33$*. As it is lower than* 0.5*, we cannot conclude that* $\mathbf{y}^1 \succ_M \mathbf{y}^2$.

[2] A backward recursive efficient algorithm was implemented by Gen et al in [Yang et al., 2014].

[3] In general, there is only an inequality between Equations (5.30) and (5.31)

Finally, let us note that computing marginals $\underline{P}(Y_{\{i\}} = 0)$ and $\underline{P}(Y_{\{i\}} = 1)$ used in Equation (5.5) is equally easy, as it amounts to compute the lower expectation of the indicator functions $\mathbb{1}_{(y_i=0)}$ and $\mathbb{1}_{(y_i=1)}$, respectively.

5.3.2 *Exact vs approximate skeptic inference*

In this section, we want to assess how good is the outer-approximation proposed in [Destercke, 2014] and given by Equation (5.5), by comparing it to an exact estimation of the set $\hat{\mathbb{Y}}^{M}_{\ell_{H,\mathscr{P}}}$. Such an estimate is essential to know in which situation Equation (5.5) is likely to give a too conservative outer-approximation, and in which cases it can safely be used.

To perform this study, we simulate credal sets $\mathscr{P}$ over $\mathscr{Y}$ by generating binary trees in the following way: we choose an $\epsilon \in [0, 0.5]$, and for a label Y_i and a path $y_1, \ldots, y_{i-1}$, we generate a random $\theta \sim \mathcal{U}([0,1])$ to obtain the interval

$$\begin{aligned}
\underline{P}_x(Y_{\{i\}} = 1 | y_1, \ldots, y_{i-1}) &= \max(0, \theta - \epsilon) = 1 - \overline{P}_x(Y_{\{i\}} = 0 | y_1, \ldots, y_{i-1}),\\
\overline{P}_x(Y_{\{i\}} = 1 | y_1, \ldots, y_{i-1}) &= \min(\theta + \epsilon, 1) = 1 - \underline{P}_x(Y_{\{i\}} = 0 | y_1, \ldots, y_{i-1}),
\end{aligned}$$

where $\mathcal{U}([0,1])$ is a uniform distribution and ϵ is a parameter representing the imprecision level of our interval. The value of parameter ϵ impacts directly on the width of the interval and therefore on the precision of the obtained prediction. The tree in Figure 5.6 is of this kind.

In order to ensure the truthfulness and completeness of the comparison of two skeptic inferences, we evaluate them on five different samples of 2000 binary trees, each sample having a fixed ϵ (i.e. 10^3 instances). For each instance, we evaluate the quality of the outer-approximation by computing the number of added elements in the corresponding set of binary vectors, i.e.

$$d^{\epsilon}_{(\hat{\mathbf{y}},\hat{\mathbb{Y}})} = |\hat{\mathbf{y}}^{*}_{\ell_{H,\mathscr{P}}}| - |\hat{\mathbb{Y}}^{M}_{\ell_{H,\mathscr{P}}}|. \tag{5.33}$$

As we have that $\hat{\mathbf{y}}_{\ell_{H,\mathscr{P}}} \supseteq \hat{\mathbb{Y}}^{M}_{\ell_{H,\mathscr{P}}}$, Equation 5.33 will never be negative. Also, since different number of labels will induce different upper bounds for Equation (5.33), we uniformize the results across different numbers by partitioning the results in four bins:

$$\begin{aligned}
q_0 &= \#\left\{(\hat{\mathbf{y}}, \hat{\mathbb{Y}})_i^{(2000)} \;\middle|\; d^{\epsilon}_{(\hat{\mathbf{y}},\hat{\mathbb{Y}})_i} = 0\right\},\\
q_{\leqslant 0.25} &= \#\left\{(\hat{\mathbf{y}}, \hat{\mathbb{Y}})_i^{(2000)} \;\middle|\; 0 < d^{\epsilon}_{(\hat{\mathbf{y}},\hat{\mathbb{Y}})_i} \leqslant 2^{|\Omega|}/4\right\},\\
q_{\leqslant 0.5} &= \#\left\{(\hat{\mathbf{y}}, \hat{\mathbb{Y}})_i^{(2000)} \;\middle|\; 2^{|\Omega|}/4 < d^{\epsilon}_{(\hat{\mathbf{y}},\hat{\mathbb{Y}})_i} \leqslant 2^{|\Omega|}/2\right\},\\
q_{\leqslant 1} &= \#\left\{(\hat{\mathbf{y}}, \hat{\mathbb{Y}})_i^{(2000)} \;\middle|\; 2^{|\Omega|}/2 < d^{\epsilon}_{(\hat{\mathbf{y}},\hat{\mathbb{Y}})_i} \leqslant 2^{\Omega}\right\}.
\end{aligned}$$

Finally, we perform the computer simulations on a discretization of the parameter $\epsilon \in \{0.05, 0.15, \ldots, 0.45\}$. Thus, the results obtained, in percentage and with confidence interval (of the five repetitions), for each ϵ value and partitions q_* are shown in the Table 5.1. We omitted the results of $\epsilon = 0.45$ since it always yields $q_0 = 100\%$ for all labels.

We can summarise the main findings of those simulations as follows:

- globally, $\hat{\mathfrak{y}}^*_{\ell_H,\mathscr{P}}$ provides a quite accurate approximation of the true set, as it is exact (i.e., in q_0) most of the time;
- the quality of $\hat{\mathfrak{y}}^*_{\ell_H,\mathscr{P}}$ decreases as the number of labels increases, making it unfit for applications involving a very high number of labels such as extreme multi-label [Jain et al., 2016];
- the quality of $\hat{\mathfrak{y}}^*_{\ell_H,\mathscr{P}}$ seems to be the worst for moderate imprecision, probably because a high imprecision will tend to provide more vacuous (i.e., empty vectors) predictions;
- there are a few cases where $\hat{\mathfrak{y}}^*_{\ell_H,\mathscr{P}}$ provides bad (i.e., are in $q_{\leqslant 0.5}$) to really bad approximation (i.e., are in $q_{\leqslant 1}$). This indicates that having exact inference methods may be helpful to identify those cases.

All these last findings are confirmed by Figures 5.7 that display the evolutions of the partitions q_0 (left) and $q_{\leqslant 0.25}$(right).

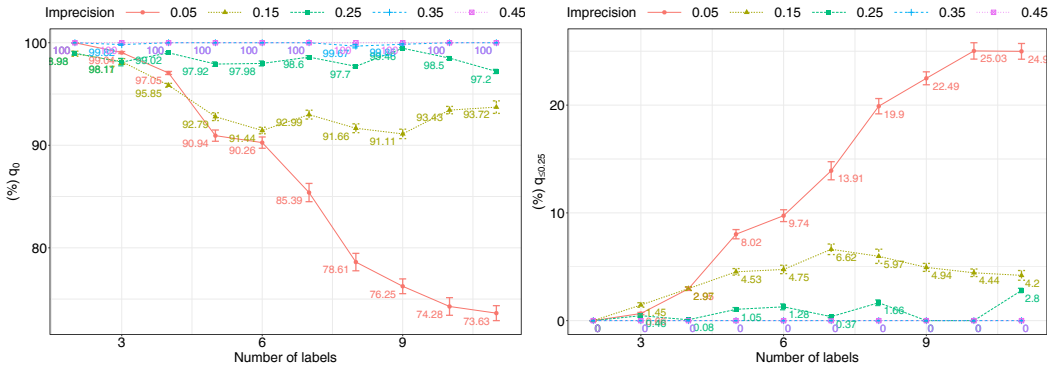


Figure 5.7: Evolution of average partitions amounts q_* (%) (with confidence interval) of the partition q_0 (left) and $q_{\leqslant 0.25}$(right)

In what follows, we perform other experimental studies on real data sets in order to check how skeptic inferences for multi-label problems behave in presence of noisy or missing labels.

5.3.3 *Skeptic inference with Binary relevance*

In this subsection, we perform a set of experiments to investigate the usefulness of using skeptic inferences in multi-label problems. In particular,

#label	ϵ	$d^{\epsilon}_{\hat{y},\hat{Y}}$			
		q_0	$q_{\leqslant 0.25}$	$q_{\leqslant 0.5}$	$q_{\leqslant 1}$
2	0.05	**100.0 ± 0.00%**	0.00 ± 0.00%	0.00 ± 0.00%	0.00 ± 0.00%
	0.15	**98.93 ± 0.11%**	0.00 ± 0.00%	**1.07 ± 0.11%**	0.00 ± 0.00%
	0.25	**98.98 ± 0.18%**	0.00 ± 0.00%	**1.02 ± 0.18%**	0.00 ± 0.00%
	0.35	**100.0 ± 0.00%**	0.00 ± 0.00%	0.00 ± 0.00%	0.00 ± 0.00%
3	0.05	**99.04 ± 0.06%**	**0.66 ± 0.07%**	**0.30 ± 0.09%**	0.00 ± 0.00%
	0.15	**98.17 ± 0.27%**	**1.45 ± 0.29%**	**0.38 ± 0.10%**	0.00 ± 0.00%
	0.25	**98.11 ± 0.17%**	**0.46 ± 0.08%**	**1.43 ± 0.17%**	0.00 ± 0.00%
	0.35	**99.82 ± 0.04%**	0.00 ± 0.00%	**0.18 ± 0.04%**	0.00 ± 0.00%
4	0.05	**97.05 ± 0.25%**	**2.95 ± 0.25%**	0.00 ± 0.00%	0.00 ± 0.00%
	0.15	**95.85 ± 0.38%**	**2.97 ± 0.24%**	**1.17 ± 0.17%**	**0.01 ± 0.02%**
	0.25	**99.02 ± 0.17%**	**0.08 ± 0.05%**	**0.90 ± 0.18%**	0.00 ± 0.00%
	0.35	**100.0 ± 0.00%**	0.00 ± 0.00%	0.00 ± 0.00%	0.00 ± 0.00%
5	0.05	**90.94 ± 0.65%**	**8.02 ± 0.51%**	**1.04 ± 0.23%**	0.00 ± 0.00%
	0.15	**92.79 ± 0.18%**	**4.53 ± 0.42%**	**2.01 ± 0.37%**	**0.67 ± 0.21%**
	0.25	**97.92 ± 0.05%**	**1.05 ± 0.20%**	**0.73 ± 0.15%**	**0.30 ± 0.09%**
	0.35	**100.0 ± 0.00%**	0.00 ± 0.00%	0.00 ± 0.00%	0.00 ± 0.00%
6	0.05	**90.26 ± 0.44%**	**9.74 ± 0.44%**	0.00 ± 0.00%	0.00 ± 0.00%
	0.15	**91.44 ± 0.63%**	**4.75 ± 0.35%**	**2.79 ± 0.19%**	**1.02 ± 0.23%**
	0.25	**97.98 ± 0.18%**	**1.28 ± 0.06%**	**0.71 ± 0.12%**	**0.03 ± 0.02%**
	0.35	**100.0 ± 0.00%**	0.00 ± 0.00%	0.00 ± 0.00%	0.00 ± 0.00%
7	0.05	**85.39 ± 0.53%**	**13.91 ± 0.47%**	**0.70 ± 0.08%**	0.00 ± 0.00%
	0.15	**92.99 ± 0.61%**	**6.62 ± 0.58%**	**0.36 ± 0.08%**	**0.03 ± 0.02%**
	0.25	**98.60 ± 0.15%**	**0.37 ± 0.07%**	**1.03 ± 0.13%**	0.00 ± 0.00%
	0.35	**100.0 ± 0.00%**	0.00 ± 0.00%	0.00 ± 0.00%	0.00 ± 0.00%
8	0.05	**78.61 ± 1.35%**	**19.9 ± 1.31%**	**1.49 ± 0.25%**	0.00 ± 0.00%
	0.15	**91.66 ± 0.33%**	**5.97 ± 0.31%**	**1.78 ± 0.15%**	**0.59 ± 0.14%**
	0.25	**97.70 ± 0.21%**	**1.66 ± 0.21%**	**0.64 ± 0.20%**	0.00 ± 0.00%
	0.35	**99.67 ± 0.04%**	0.00 ± 0.00%	0.33 ± 0.04%	0.00 ± 0.00%
9	0.05	**76.25 ± 0.60%**	**22.49 ± 0.57%**	**1.26 ± 0.16%**	0.00 ± 0.00%
	0.15	**91.11 ± 0.76%**	**4.94 ± 0.58%**	**3.30 ± 0.33%**	**0.65 ± 0.19%**
	0.25	**99.46 ± 0.08%**	0.00 ± 0.00%	**0.54 ± 0.08%**	0.00 ± 0.00%
	0.35	**99.85 ± 0.09%**	0.00 ± 0.00%	**0.15 ± 0.09%**	0.00 ± 0.00%
10	0.05	**74.28 ± 0.92%**	**25.03 ± 0.96%**	**0.69 ± 0.07%**	0.00 ± 0.00%
	0.15	**93.43 ± 0.32%**	**4.44 ± 0.34%**	**1.38 ± 0.33%**	**0.75 ± 0.25%**
	0.25	**98.50 ± 0.15%**	0.00 ± 0.00%	**1.50 ± 0.15%**	0.00 ± 0.00%
	0.35	**100.0 ± 0.00%**	0.00 ± 0.00%	0.00 ± 0.00%	0.00 ± 0.00%
11	0.05	**73.63 ± 0.60%**	**24.99 ± 0.66%**	**1.38 ± 0.13%**	0.00 ± 0.00%
	0.15	**93.72 ± 0.64%**	**4.20 ± 0.55%**	**2.08 ± 0.56%**	0.00 ± 0.00%
	0.25	**97.20 ± 0.20%**	**2.80 ± 0.20%**	0.00 ± 0.00%	0.00 ± 0.00%
	0.35	**100.0 ± 0.00%**	0.00 ± 0.00%	0.00 ± 0.00%	0.00 ± 0.00%

Table 5.1: Average partitions amounts q_* (%) with confidence interval.

we investigate what happens when some labels are noisy or missing. To that end, we use a set of standard real-word data sets from the MULAN

repository[4] (c.f. Table 5.2), following a 10×10 cross-validation procedure to fit the model.

Data set	#Features	#Labels	#Instances	#Cardinality	#Density
emotions	72	6	593	1.90	0.31
scene	294	6	2407	1.07	0.18
yeast	103	14	2417	4.23	0.30

Table 5.2: Multi-label data sets summary

EVALUATION As we perform set-valued predictions, usual measures used in multi-label problems cannot be adopted here. We thus consider it appropriate to use an incorrectness measure (IC), coupled with a completeness (CP) measure [Destercke, 2014, §4.1], defined as follows

$$\mathrm{IC}(\hat{\mathbb{Y}}, \mathbf{y}) = \frac{1}{|Q|} \sum_{\hat{y}_i \in Q} \mathbb{1}_{(\hat{y}_i \neq y_i)}, \tag{5.34}$$

$$\mathrm{CP}(\hat{\mathbb{Y}}, \mathbf{y}) = \frac{|Q|}{m}, \tag{5.35}$$

where Q denotes the set of predicted labels such that $\hat{y}_i = 1$ or $\hat{y}_i = 0$ (in other words any abstained predicted label $\hat{y}_i = *$ is not in Q). When predicting complete vectors, then $\mathrm{CP} = 1$ and IC equals the Hamming loss (5.3), and when predicting the empty vector, i.e. all labels equals to $\hat{y}_i = *$, then $\mathrm{CP} = 0$ and by convention $\mathrm{IC} = 0$. Since those measures are adapted to partial vectors, we will use a simple binary relevance strategy in the experiments.

NAIVE CREDAL CLASSIFIER To obtain intervals over each label, we use the naïve credal classifier presented in Section 2.3. In our case, we will consider each label as a simple binary classification problem, and will estimate the marginal probability using the model

$$P(Y_{\{j\}} = y_j | X = \mathbf{x}) = \frac{P(Y_{\{j\}} = y_j) \prod_{i=1}^{d} P(X_i = x_i | Y_{\{j\}} = y_j)}{\sum_{y_l \in \{0,1\}} P(Y_{\{j\}} = y_l) \prod_{i=1}^{d} P(X_i = x_i | Y_{\{j\}} = y_l)}. \tag{5.36}$$

Computing lower and upper probability bounds $[\underline{P}_{Y_{\{j\}}|X}, \overline{P}_{Y_{\{j\}}|X}]$ over all conditional distributions $\mathscr{P}_{X|Y_{\{j\}}}$ (since we assume a precise estimation of the marginal distribution) can be performed using Equations (2.28) and (2.30), as follows

$$\underline{P}(Y_{\{j\}} = y_j | X = \mathbf{x}) = \left(1 + \frac{P(Y_{\{j\}} = \overline{y}) \prod_{i=1}^{d} \overline{P}(X_i = x_i | Y_{\{j\}} = \overline{y}_j)}{P(Y_{\{j\}} = y_j) \prod_{i=1}^{d} \underline{P}(X_i = x_i | Y_{\{j\}} = y_j)}\right)^{-1}, \tag{5.37}$$

[4] http://mulan.sourceforge.net/datasets.html

$$\overline{P}(Y_{\{j\}} = y_j | X = \mathbf{x}) = \left(1 + \frac{P(Y_{\{j\}} = \overline{y}) \prod_{i=1}^{d} \underline{P}(X_i = x_i | Y_{\{j\}} = \overline{y})}{P(Y_{\{j\}} = y_j) \prod_{i=1}^{d} \overline{P}(X_i = x_i | Y_{\{j\}} = y_j)}\right)^{-1}, \quad (5.38)$$

where $\overline{y}_j$ is the complement to y_j. To obtain the other bounds of those equations, we use Equation (2.40) obtained by Imprecise Dirichlet model (IDM) described in Section 2.3.2.

In this chapter, we restrict the values of the hyper-parameter of the imprecision of IDM to $s \in \{0, 0.5, 1.5, 2.5, 3.5, 4.5\}$. Our purpose here is not to find the "optimal" value of s, but to show the effectiveness of injecting imprecision (i.e. to provide robust and skeptical inferences).

MISSING LABELS In this setting, we proceed to choose uniformly at random a percentage of missing labels, with five different levels of missingness: $\{0, 20, 40, 60, 80\}$. Missing values are removed from the training data. Table 5.3 illustrates a data set with missing values (or partially labelled instances).

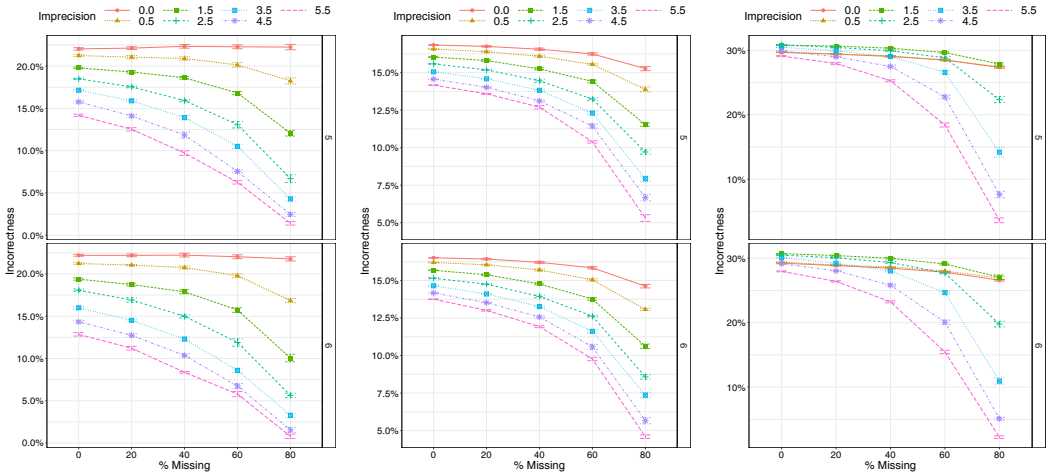
Features					**Missing**			**Noise-Reversing**			**Noise-Flipping**		
X_1	X_2	X_3	X_4	X_5	Y_1	Y_2	Y_3	Y_1	Y_2	Y_3	Y_1	Y_2	Y_3
107.1	25	Blue	60	1	1	0	*	1	0→1	0	1	$1 \wedge_\beta 0$	0
-50	10	Red	40	0	1	*	1	1	0	1→0	1	0	$1 \wedge_\beta 0$
200.6	30	Blue	58	1	*	1	0	0→1	0	0	0	0	0
107.1	5	Green	33	0	*	1	0	1	1→0	0	$1 \wedge_\beta 0$	$1 \wedge_\beta 0$	0
...	...	...	...	...	...	...	...	...	...	...	...	...	...

Table 5.3: Missing and Noise representation of labels

In the Figures 5.8 and 5.9, we provide the results of the incorrectness and completeness measures, respectively, obtained by fitting the NCC model on different percentages of missing labels and data sets of the Table 5.2.

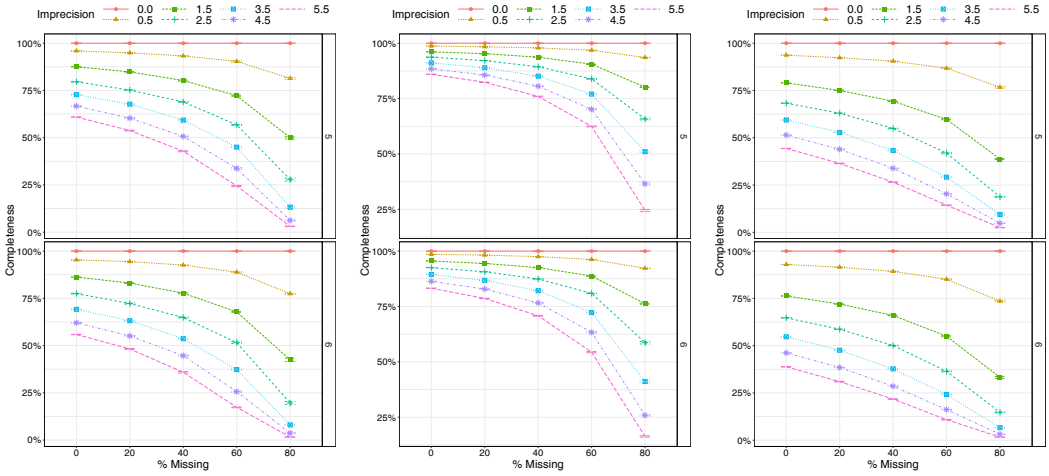
The results show that as the percentage of missing labels increases the incorrectness and the completeness both decrease, especially on Emotions and Scene data sets. This means that the more imprecise we get, the more accurate are those predictions we retain. The effect is less significant on the yeast data set, where one needs a high amount of imprecision to witness a gain in correctness. One quite noticeable result is that for the Emotions data set, even with 80% of missing label, a slight imprecision ($s = 0.5$) allows us to reach a reasonable completeness of about 80% with a gain of 5% in terms of correct predictions. Also, as the confidence intervals displayed in Figure 5.8 are very small, and remain so in the other settings where labels are non-missing but noisy.

Results obtained are sufficient to show that skeptic inferences with probability sets may provide additional benefits when dealing with missing labels. Those results could, of course, be improved by picking other classifiers, such as the NCC2 [Corani et al., 2008b], an extension of the



(a) EMOTIONS **(b)** SCENE **(c)** YEAST

Figure 5.8: **Missing labels**. Evolution of the average incorrectness (%) for each level of imprecision (a curve for each one) and discretization $z = 5$ (top) and $z=6$ (down), with respect the percentage of missing labels.



(a) EMOTIONS **(b)** SCENE **(c)** YEAST

Figure 5.9: **Missing labels**. Evolution of the average completeness (%) for each level of imprecision (a curve for each one) and discretization $z = 5$ (top) and $z=6$ (down), with respect the percentage of missing labels..

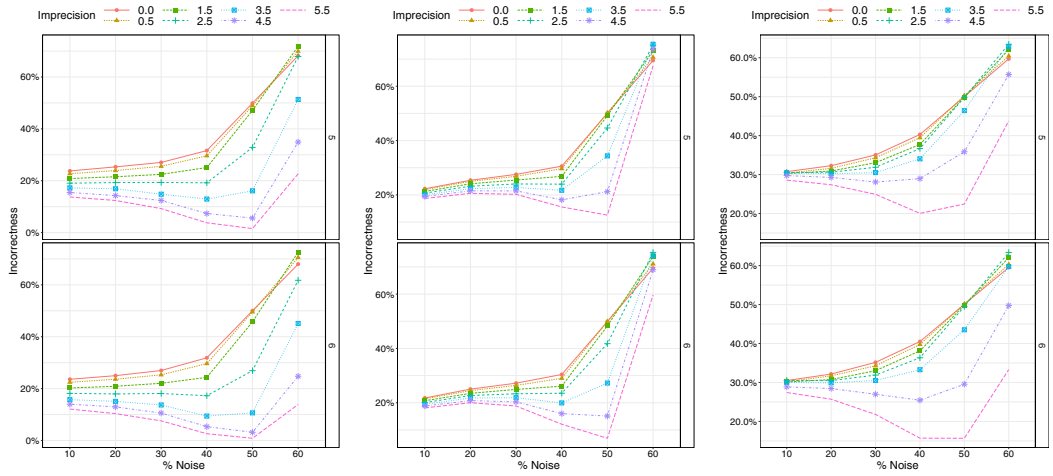
NCC tailored for missing values, or imprecise classifiers able to cope with continuous attributes (c.f. Chapter 3).

NOISY LABELS In this setting, we proceed in the same way as with the missing setting, except that the value of selected labels are not assigned to $*$, but are modified according to some noise scheme. We consider two different ways to modify the assignments:

1. **Reversing:** in this case, we reverse the current value of the selected label. In other words, if $Y_{j,i} = 1$, the label j of the instance i becomes $Y_{j,i} = 0$ (similar for the case of $Y_{j,i} = 0 \rightarrow Y_{j,i} = 1$), with six different levels of noisy $\{10, 20, 30, 40, 50, 60\}$,

2. **Flipping:** in contrast to the previous case, for each selected label $Y_{j,i}$, we replace it by the result of a Bernoulli trial with probability $\beta := P(Y_{j,i} = 1)$, i.e. $Y_{j,i} \sim \mathcal{B}er(\beta)$, with $\beta \in \{0.2, 0.5, 0.8\}$, with three levels of noisy $\{40, 60, 80\}$

Table 5.3 provides an illustration of these two noise settings, in the columns "Noise-Reversing" and "Noise-Flipping", respectively.

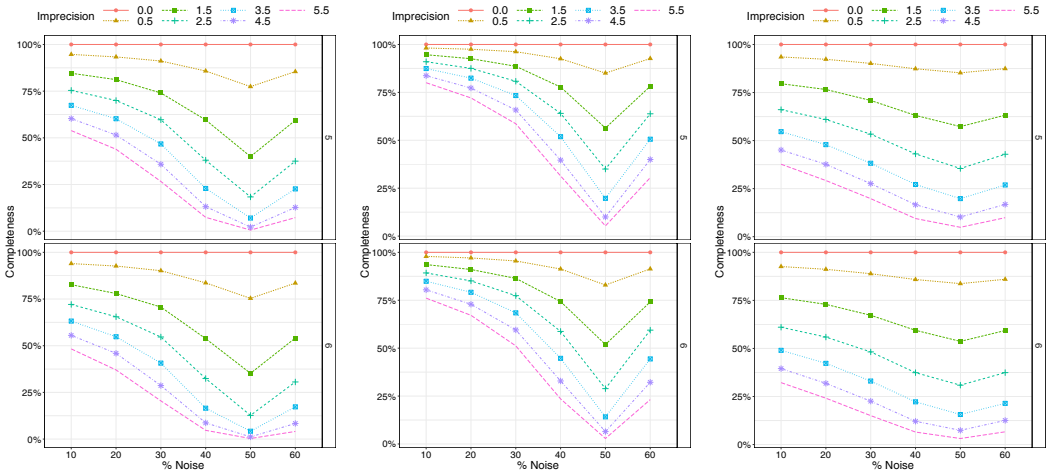


(a) EMOTIONS **(b)** SCENE **(c)** YEAST

Figure 5.10: **Noise-Reversing**. Evolution of the average incorrectness (%) for each level of imprecision (a curve for each one) and two levels of discretization $z = 5$ (top) and $z = 6$ (down), with respect to the percentage of noise.

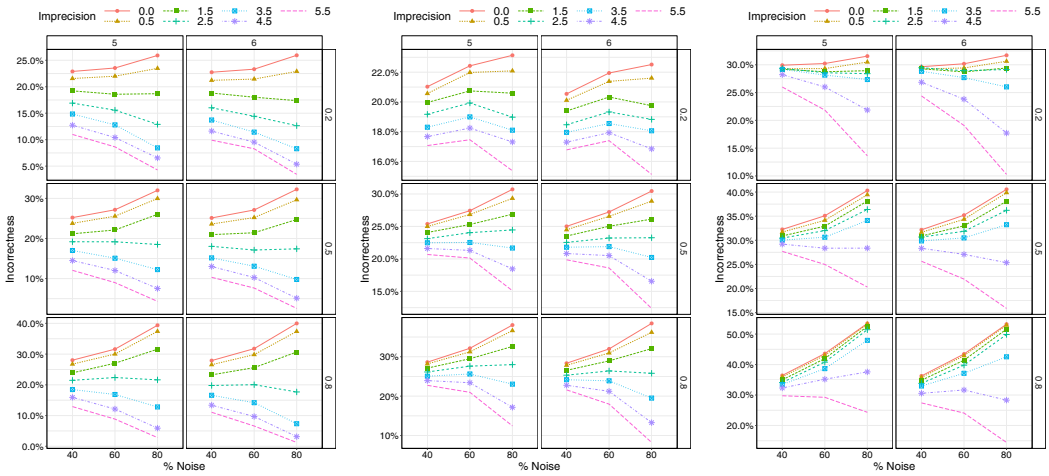
In Figures 5.10 and 5.12, we provide the results of the incorrectness measure obtained by fitting the NCC model on different percentages of **Reversing** and **Flipping** settings applied to the data sets of the Table 5.2. Results about completeness are given in Figures 5.11 and 5.13

Concerning the **Reversing**, or adversarial setting, it is clear from the graphs that allowing for imprecision and skeptical inferences provides some level of protection, which can be witnessed by the fact that at a given level of noise, including some imprecision limits the increase in incorrectness, and sometimes even improves the quality of the made predictions by abstaining on those instances where adversarial noise was introduced. Of course, this goes hand-in-hand with a corresponding decrease of completeness, but this seems a fair price to pay to protect against adversarial noise.



(a) EMOTIONS **(b)** SCENE **(c)** YEAST

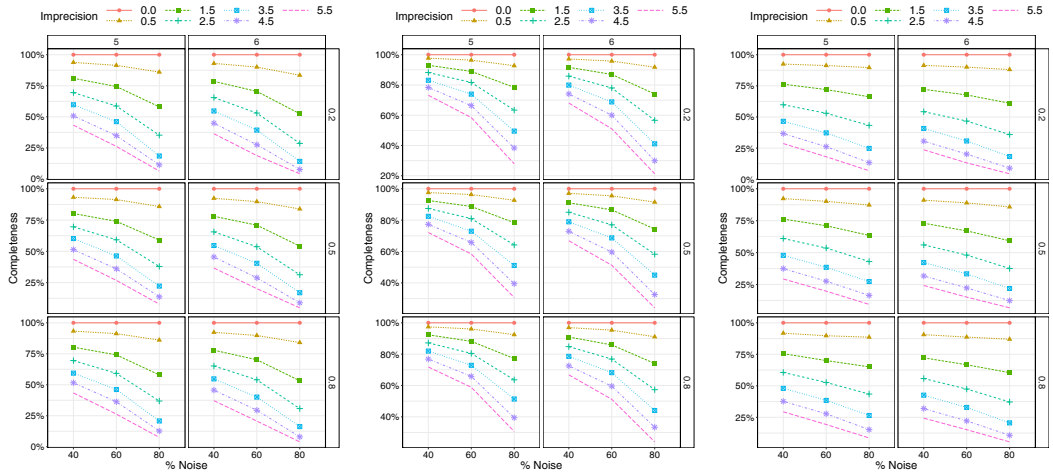
Figure 5.11: **Noise-Reversing**. Evolution of the average completeness (%) for each level of imprecision (a curve for each one) and two levels of discretization $z = 5$ (top) and $z = 6$ (down), with respect to the noise percentage.



(a) EMOTIONS **(b)** SCENE **(c)** YEAST

Figure 5.12: **Noise-Flipping**. Evolution of the average incorrectness (%) for each level of imprecision (a curve for each one), two levels of discretization $z = 5$ (left) and $z = 6$ (right), and three different probabilities $\beta = 0.2$ (top), $\beta = 0.5$ (middle) and $\beta = 0.8$ (down) of replacing the selected label with a 1.

Results obtained on the **Flipping** setting are overall similar to those found in the missing label and the **Reversing** label. Notable small differences are that (1) skeptical inferences are uniformly more robust (provide more accurate predictions) than their precise counter-part, whatever the



(a) EMOTIONS **(b)** SCENE **(c)** YEAST

Figure 5.13: **Noise-Flipping**. Evolution of the average completeness (%) for each level of imprecision (a curve for each one), two levels of discretization $z = 5$ (left) and $z = 6$ (right), and three different probabilities $\beta = 0.2$ (top), $\beta = 0.5$(middle) and $\beta = 0.8$(down), with respect to the percentage of noise.

level and nature of noise, and (2) the evenness of the noise (β value) obviously has an impact on performances, but has little impact on the overall trends.

Similarly to the case of the missing label, it would be interesting to experiment with other imprecise classifier, as well as with other different noise settings (e.g. using other probability distributions as $\beta \sim \mathcal{U}([0,1])$ then $Y_{i,j} = 1$ if $\beta > \tau$ otherwise $Y_{i,j} = 0$, where τ is a threshold parameter).

5.4 CONCLUSION AND DISCUSSION

Describing our uncertainty by a set of probabilities over combinatorial domains such as binary vectors usually leads to difficult optimisation problems at the decision step. In this chapter, we investigated those problems when considering the well-known Hamming loss, providing efficient inference methods and, when considering the binary relevance scheme, connecting it to the zero/one loss function.

In essence, we significantly reduced the complexity of computing exact skeptic, cautious predictions for general probability sets, and showed that in the Binary relevance scheme, those same predictions were reduced to partial vectors computable from marginal probability bounds over the labels.

Experiments on the simulated data sets show that this last solution, when used as an outer-approximation in the general case, degrades in quality as the number of labels increase and the level of imprecision is

mild. On the other hand, experiments on various real data sets show that making skeptical inferences generally provide quite satisfactory results on different scenarios, involving missing or noisy labels.

Our experiments clearly demonstrate a potential interest of the use of skeptic inferences for multi-label problems. In future works, it would be interesting to compare our skeptical inference approach against those rejection and abstaining existing approaches (to the best of our knowledge, there are few works). As for instance those cautious approaches that were mentioned in Section 4.3. Such comparisons would nevertheless require a deep analysis of the models, decision rules as well as instances on which each approach abstains, but due to lack of time, an in-depth analysis could not be carried out.

We also left open a number of theoretical and experimental issues, such as finding out new heuristic approaches to reduce the current complexity time $\mathcal{O}(3^m - 1)$. Another first natural next step will be to solve the maximality criterion using the ranking loss, since it has been proved in [Dembczyński et al., 2012, th.1] that is only necessary to know the marginal distribution to minimize the expected loss of Equation (4.3). Then, we will focus on more sophisticated loss as Jaccard loss, F-measure, and so on. As noticed in Remark 7, such problems are likely to be much more intricate when considering sets of probabilities.

Finally, let us notice that while this chapter focused on the issue of multi-label learning problems, our results readily apply to any Boolean vectors of m items. As Boolean vectors and structures as well as probability bounds naturally appear in a number of other applications, including occupancy grids [Mouhagir et al., 2017] or data bases [Gatterbauer et al., 2014], a future work would be to investigate how our present findings can help in such problems.

Chapter 6

Multi-label chaining using naive credal classifier

"El modo de dar una vez en el clavo, es dar cien veces en la herradura."

—Miguel de Unamuno

Contents

A classical issue in multi-label learning techniques is how to integrate the possible dependencies between labels while keeping the inference task tractable. Indeed, while decomposition techniques [Tsoumakas et al., 2007; Fürnkranz et al., 2008] such as Binary relevance or Calibrated ranking allow to speed up both the learning and inference tasks, they roughly ignore the label dependencies, while using a fully specified model such as probabilistic chains require, at worst, to scan all possible predictions (that grow exponentially in the number of labels). A popular technique to solve this issue, at least for the inference task, is to use a chain model [Read et al., 2011]: this consists in using, incrementally, the predictions made on previous labels to help better predict the relevance of a current label.

To the best of our knowledge, there are few works of multi-label classification producing cautious predictions (c.f. Section 4.3), but none of these have studied this issues in the chain model (or classifier-chains approach).

In this chapter, we consider the problem of extending such an approach to the imprecise probabilistic case, and propose two different ways to extend it, based on the fact that some labels are too uncertain to be used in

the chaining. The first treats the uncertain labels in a robust way, exploring all possible paths in order not to propagate early uncertain decisions, whereas the latter marginalizes the probabilistic model over the uncertain labels, in other words, the uncertain labels are not considered to infer the current label. In addition to last two strategies, we propose a dynamic, context-dependent label ordering, which selects dynamically and in priority those labels for which the decision is the least uncertain. The main goal of such an ordering is to limit the final imprecision and bias.

Section 6.1 introduces the notations which we will use more specifically in this chapter. In Section 6.2, we remind the classical classifier-chains approach and then we present our extended approaches based on imprecise probabilities.

By means of the use of the naive credal classifier [Zaffalon, 1999], in Section 6.3, we propose efficient procedures to solve the strategies presented in Section 6.2, having a time complexity almost polynomial on the number of labels.

Finally, in Section 6.4, we perform a set of experiments on real data sets, which are perturbed with missing and noisy labels, in order to investigate how accurate (when we exchange abstained labels for precise ones) and how cautious (when we abstain on labels difficult to predict) is our approach.

6.1 PROBLEM SETTING

As in the previous chapter, we are also interested in making set-valued predictions when uncertainty is too high (e.g. due to insufficient evidence to include or discard a relevant label, see Example 13). The set-valued prediction will here be described as a partial binary vector $\mathbf{y}^* \in \mathscr{Y}^*$ where $\mathscr{Y}^* = \{0, 1, *\}^m$ is the new output space with a new element $*$ representing the abstention. For instance, a partial prediction $\mathbf{y}^* = (*, 1, 0)$ correspond to two plausible binary vector solutions $\{(0, 1, 0), (1, 1, 0)\} \subseteq \mathscr{Y}$, where $\mathscr{Y}$ is the m-dimensional binary space (for further details see Section 4.1).

In the sequel, we will denote by $\mathscr{I}$ subsets of label indices (and by $[\![j]\!] = \{1, \ldots, j\}$ set of the first j integers). Given a prediction made in the j first labels, we will denote by

1. (relevant labels) $\mathscr{I}^j_{\mathcal{R}} \subseteq [\![j]\!]$ the indices of the labels predicted as relevant among the j first, i.e. $\forall i \in \mathscr{I}^j_{\mathcal{R}},\ y_i = 1$,

2. (irrelevant labels) $\mathscr{I}^j_{\mathcal{I}} \subseteq [\![j]\!], \mathscr{I}^j_{\mathcal{I}} \cap \mathscr{I}^j_{\mathcal{R}} = \emptyset$ the indices of the labels predicted as irrelevant among the j first, i.e. $\forall i \in \mathscr{I}^j_{\mathcal{I}},\ y_i = 0$, and

3. (abstained labels) $\mathscr{I}^j_{\mathcal{A}} = [\![j]\!] \backslash (\mathscr{I}^j_{\mathcal{R}} \cup \mathscr{I}^j_{\mathcal{I}})$ the indices of the labels on which we abstained among the j first, i.e. $\forall i \in \mathscr{I}^j_{\mathcal{A}},\ y_i = \{0, 1\} := *$.

Besides, for the sake of simplicity and when it is not ambiguous, we will henceforth denote probabilities conditioned on previous labels by

$$P^j_x(Y_j = 1) := P_x(Y_j = 1 | Y_{\mathscr{J}^{j-1}} = \hat{y}_{\mathscr{J}^{j-1}}), \tag{6.1}$$

where $\hat{y}_{\mathscr{J}^{j-1}}$ is a $(j-1)$-dimensional vector that contains the previously inferred precise and/or abstained values of labels having indices $\mathscr{J}^{j-1}$.

Example 13 *We consider an output space of two labels* $\mathscr{K} = \{m_1, m_2\}$*, a single binary feature* x_1 *and Table 6.1 with imprecise estimations of the joint distribution* $\mathbf{P}(X_1, Y_1, Y_2)$.

x_1	y_1	y_2	$\hat{\mathbf{P}}$	x_1	y_1	y_2	$\hat{\mathbf{P}}$
0	0	0	[0.4,0.7]	1	0	0	0.00
0	0	1	[0.3,0.6]	1	0	1	0.00
0	1	0	0.00	1	1	0	[0.6,0.8]
0	1	1	0.00	1	1	1	[0.2,0.4]

Table 6.1: Estimated joint probability distribution

Based on the probabilities of Table 6.1, we have that $\hat{P}_0(y_1 = 0) = \hat{P}(y_1 = 0 | x_1 = 0) = 1$ *and* $\hat{P}_0(y_2 = 0) \in [0.4, 0.7]$*, therefore not knowing whether* $\hat{P}_0(y_2 = 0) > 0.5$*. This leads to propose as a prediction* $\hat{\mathbf{y}}^* = (0, *)$*. On the contrary, the imprecision on the right hand-side is such that* $\hat{P}_1(y_2 = 0) \in [0.6, 0.8]$*, leading to the precise prediction* $\hat{\mathbf{y}}^* = (1, 0)$.

Handling partial binary predictions requires a well-founded strategy to do so, as well as efficient procedures to deal with the increased complexity of the prediction space $|\mathscr{Y}^*| = 3^m$. An efficient way to perform it, and already reviewed in Chapter 5, is by making use of the assumption of independence on labels, but unfortunately it does not integrate the dependence amongst them. Moreover, if we approach it using the maximality or interval dominance criterion without this assumption of independence (clearly not at work in chaining approaches), it will not yield partial binary vector either, but set-valued predictions (c.f. Example 10).

So, in the same vein as Chapter 5, we will also describe our uncertainty here by means of a set of probability distributions $\mathscr{P}$ instead of a single probability distribution $\mathbb{P}$, as usually done (for more details about imprecise probabilities, we refer to Section 1.3), jointly with the chaining approach in order to produce partial binary predictions.

As mentioned in Section 1.3, IP comes up with certain additional difficulties in the learning and inference step. Thus, in this chapter, for the former issue, we will use the NCC classifier (see Section 2.3) in order to boost the inference step by proposing efficient procedures. For the decision step, we will use a classical-imprecise-binary inference approach; in which if we consider $\mathscr{K} = \{0, 1\}$ as the output space and Y as a univariate

random variable on $\mathscr{K}$, it is easy to prove that if we use the maximality or interval dominance criterion over a zero/one loss as inference procedures (c.f. Equation (2.12), also described in [Destercke, 2014, Prop. 1]), it is reduced to infer the univariate binary output Y on a credal set $\mathscr{P}$ by

$$\hat{y} = \begin{cases} 1 & \text{if } \underline{P}_x(Y=1) > 0.5, \\ 0 & \text{if } \overline{P}_x(Y=1) < 0.5, \\ * & \text{if } 0.5 \in \left[\underline{P}_x(Y=1), \overline{P}_x(Y=1)\right] \end{cases} . \tag{6.2}$$

This inference procedure will be adapted to the case of mutli-label chaining, by adding subscript j to the current label to infer Y_j and augmenting the input space of regressors with previous inferred labels $\{\hat{Y}_1, \ldots, \hat{Y}_{j-1}\}$.

6.2 MULTILABEL CHAINING WITH IMPRECISE PROBABILITIES

We first recall the classical precise chaining and then propose two different strategies to extend chaining to the imprecise probabilistic case, and a new procedure to dynamically select the order of labels in the chain.

6.2.1 *Precise probabilistic chaining*

Classifier chains is a well-known approach exploiting dependencies among labels by fitting at each step of the chain (see Figure 6.1) a new classifier model $h_j : \mathscr{X} \times \{0,1\}^{j-1} \to \{0,1\}$ predicting the relevance of the jth label. This classifier combines the original input space attribute and all previous predictions in the chain in order to create a new input space $\mathscr{X}^*_{j-1} = \mathscr{X} \times \{0,1\}^{j-1}, j \in \mathbb{N}^{>0}$. In brief, we consider a chain $\mathbf{h} = (h_1, \ldots, h_m)$ of binary classifiers resulting in the full prediction $\hat{\mathbf{y}}$ obtained by solving each single classifier as follows

$$\hat{y}_j := h_j(x) = \arg\max_{y \in \{0,1\}} P^j_x(Y_j = y). \tag{6.3}$$

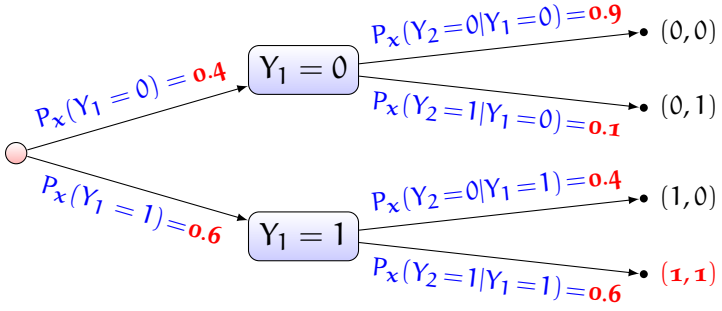
The classical multi-label chaining then works as follows:

1. RANDOM ORDER OF LABELS.-. We randomly pick an order between labels $\mathscr{I}^*$ (possibly different from the original indices $\mathscr{I} = [\![m]\!]$) and assume that the indices are relabelled in an increasing order.

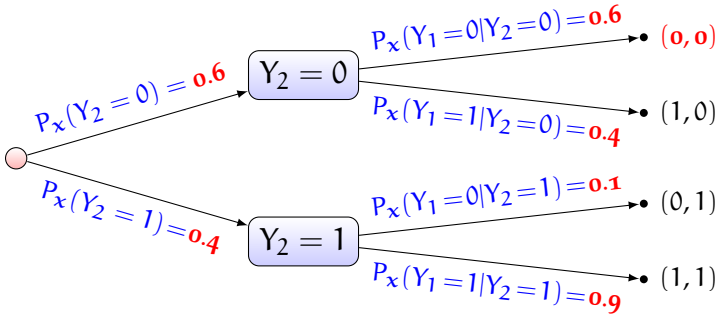
2. PREDICTION j^{th} LABEL.- For a given label y_j, let us assume that we have previously predicted labels of lower index $y_1, \ldots, y_{j-1}$ and let $\mathscr{I}^{j-1}_{\mathcal{R}}, \mathscr{I}^{j-1}_{\mathcal{I}} \subseteq [\![j-1]\!]$ be set of indices of relevant and irrelevant labels, such that $\mathscr{I}^{j-1}_{\mathcal{R}} \cap \mathscr{I}^{j-1}_{\mathcal{I}} = \emptyset$. Then, the prediction of $\hat{y}_j$ (or $h_j(\mathbf{x})$) for a new instance $\mathbf{x}$ is

$$\hat{y}_j = \begin{cases} 1 & \text{if } P_x(Y_j = 1 | Y_{\mathscr{J}_R^{j-1}} = 1, Y_{\mathscr{J}_J^{j-1}} = 0) \geqslant 0.5 \\ 0 & \text{if } P_x(Y_j = 0 | Y_{\mathscr{J}_R^{j-1}} = 1, Y_{\mathscr{J}_J^{j-1}} = 0) < 0.5 \end{cases} \tag{6.4}$$

Figure 6.1 summarizes the procedure presented above, as well as the obtained predictions for a specific case (in bold red predicted labels and probabilities).



(a) Chaining with $\{Y_1, Y_2\}$



(b) Chaining with $\{Y_2, Y_1\}$

Figure 6.1: Precise chaining

From Figure 6.1, it is clear that the ordering and the fact of choosing a single branch at each step can have a significant impact on the final predictions, as in our example it shifts from one prediction to its opposite. Intuitively, adding some robustness and cautiousness in the process could help to avoid unwarranted biases.

In what follows, we propose two different extensions of precise chaining based on imprecise probability estimates. By this, we mean that it is based on binary cautious classifiers, which consider a new output space $\mathscr{Y} = \{0, 1, *\}^m$ from which to pick the prediction.

6.2.2 *Imprecise probabilistic chaining*

When considering imprecise probabilities, the estimates $P_x^j(Y_j = 1)$ become imprecise, that is, we now have $[\hat{P}_x^j](Y_j = y_j) := [\underline{P}_x^j(Y_j = y_j), \overline{P}_x^j(Y_j = y_j)]$.

The basic idea of using such estimates is that in the chaining, we should be cautious when the classifier is unsure about which is the most probable prediction. In this section, we describe two different strategies (or extensions) in a general way, and we will efficiently adapt them by applying the naive credal classifier (an extension of the Naive Bayes classifier) in the next section.

Let us first formulate the generic procedure to calculate the probability bound of j^{th} label,

1. RANDOM ORDER OF LABELS.- As before (in precise version), randomly pick an order between labels, assuming again that indices are relabelled in increasing order.

2. PREDICTION j^{th} LABEL.- For a given label y_j, let us assume we have made possibly imprecise predictions for $y_1, \ldots, y_{j-1}$ such that $\mathscr{I}_{\mathcal{A}}^{j-1}$ contains the set of indices of labels on which we abstained $\{*\}$, and hence, $\mathscr{I}_{\mathcal{R}}^{j-1}$ and $\mathscr{I}_{\mathcal{J}}^{j-1}$ are the set of indices of relevant and irrelevant labels, such that $\mathscr{I}_{\mathcal{A}}^{j-1} \cup \mathscr{I}_{\mathcal{R}}^{j-1} \cup \mathscr{I}_{\mathcal{J}}^{j-1} = \mathscr{I}^{j-1}$. Then, we calculate $[P_x^j](Y_j = 1)$ (we will show after the possible ways to obtain this interval) in order to predict the label $\hat{y}_j$ as

$$\hat{y}_j = \begin{cases} 1 & \text{if } \underline{P}_x^j(Y_j = 1) > 0.5, \\ 0 & \text{if } \overline{P}_x^j(Y_j = 1) < 0.5, \\ * & \text{if } 0.5 \in \left[\underline{P}_x^j(Y_j = 1), \overline{P}_x^j(Y_j = 1)\right], \end{cases} \tag{6.5}$$

where this last equation is a slight variation of Equation (6.2) by using the new input space $\mathscr{X}_{j-1}^*$.

We then propose two different extensions of how to calculate $[P_x^j](Y_j = 1)$ at each inference step of the imprecise chaining.

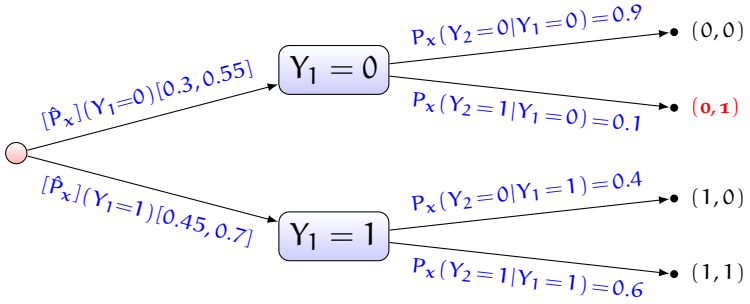
1. IMPRECISE BRANCHING The first strategy treats unsure predictions in a robust way, considering all possible branching in the chaining as soon as there is an abstained label. Thus, the estimation of $[\underline{P}_x^j(Y_j = 1), \overline{P}_x^j(Y_j = 1)]$ (for $Y_j = 0$, it directly obtains as $\underline{P}_x^j(Y_j = 1) = 1 - \overline{P}_x^j(Y_j = 0)$, and similarly for the upper bound) comes down to compute

$$\begin{aligned} \underline{P}_x^j(Y_j = 1) &= \min_{y \in \{0,1\}^{|\mathscr{I}_{\mathcal{A}}|}} P_x(Y_j = 1 | Y_{\mathscr{I}_{\mathcal{R}}^{j-1}} = 1, Y_{\mathscr{I}_{\mathcal{J}}^{j-1}} = 0, Y_{\mathscr{I}_{\mathcal{A}}^{j-1}} = y), \\ \overline{P}_x^j(Y_j = 1) &= \max_{y \in \{0,1\}^{|\mathscr{I}_{\mathcal{A}}|}} \overline{P}_x(Y_j = 1 | Y_{\mathscr{I}_{\mathcal{R}}^{j-1}} = 1, Y_{\mathscr{I}_{\mathcal{J}}^{j-1}} = 0, Y_{\mathscr{I}_{\mathcal{A}}^{j-1}} = y). \end{aligned} \tag{IB}$$

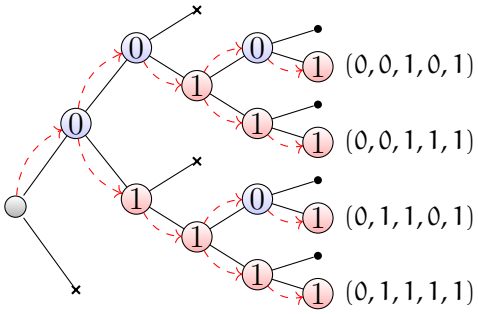
That is to consider every possible replacements of variables for which we have abstained so far. This corresponds to a very robust version

of the chaining, where every possible path is explored. It will therefore propagate imprecision along the tree, and may produce quite imprecise evaluations, especially if we abstain on the first labels.

Illustrations providing some intuition about this strategy can be seen in Figure 6.2b where we have abstained on labels (Y_2, Y_4) and we want to compute lower and upper probability bounds of the label $Y_5 = 1$.



(a) Evaluatin $Y_2 = 1$ labels $\{Y_1, Y_2\}$



(b) Evaluating $Y_5 = 1$ label with $\{0, *, 1, *, ?\}$

Figure 6.2: Imprecise branching strategy

In Figure 6.2a, we will consider the previous example (see Figure 6.1) in order to study in details how we should calculate probability bounds $[\underline{P}^j_x(Y_j = 1), \overline{P}^j_x(Y_j = 1)]$. For the sake of simplicity, we assume that probabilities about Y_2 are precise and that probability bounds of $Y_1 = 1$ is $\hat{P}^j_x(Y_1 = 1) \in [0.45, 0.70]$. This last result would correspond to the following tree where we would consider the first two branches as possibles paths hence

$$\underline{P}^j_x(Y_2 = 1) = \min_{y_1 \in \{0,1\}} P_x(Y_2 = 1 | Y_1 = y_1) = \min(0.1, 0.6) = 0.1, \quad (6.6)$$

$$\overline{P}^j_x(Y_2 = 1) = \max_{y_1 \in \{0,1\}} P_x(Y_2 = 1 | Y_1 = y_1) = \max(0.1, 0.6) = 0.6, \quad (6.7)$$

which means that in this case we would abstain on both labels, i.e. $(\hat{Y}_1, \hat{Y}_2) = (*, *)$.

2. MARGINALIZATION The second strategy simply ignores unsure predictions in the chaining. Its interest is that it will not propagate imprecision in the tree. Thus, we begin by presenting the general formulation (which will after lead to the formulation without unsureness) which takes into account unsure predicted labels conditionally, so the estimation of probability bounds $[\underline{P}^j_x(Y_j = 1), \overline{P}^j_x(Y_j = 1)]$ comes down to compute

$$\begin{aligned}\underline{P}^j_x(Y_j=1)&=\underline{P}_x(Y_j=1|Y_{\mathscr{I}^{j-1}_{\mathcal{R}}}=1, Y_{\mathscr{I}^{j-1}_{j}}=0, Y_{\mathscr{I}^{j-1}_{\mathcal{A}}}=\{0,1\}^{|\mathscr{I}^{j-1}_{\mathcal{A}}|}),\\ \overline{P}^j_x(Y_j=1)&=\overline{P}_x(Y_j=1|Y_{\mathscr{I}^{j-1}_{\mathcal{R}}}=1, Y_{\mathscr{I}^{j-1}_{j}}=0, Y_{\mathscr{I}^{j-1}_{\mathcal{A}}}=\{0,1\}^{|\mathscr{I}^{j-1}_{\mathcal{A}}|}),\end{aligned} \tag{MAR}$$

where $\mathscr{I}^{j-1}_{\mathcal{A}} = \{i_1, \ldots, i_k\}$ denotes the set of indices of abstained labels and the last conditional term of probability bounds can be defined as

$$\left(Y_{\mathscr{I}^{j-1}_{\mathcal{A}}} = \{0,1\}^{|\mathscr{I}^{j-1}_{\mathcal{A}}|}\right) := (Y_{i_1} = 0 \cup Y_{i_1} = 1) \cap \cdots \cap (Y_{i_k} = 0 \cup Y_{i_k} = 1). \tag{6.8}$$

The MAR formulation can be reduced by using Bayes's theorem in conjunction with the law of total probability. That is, for instance, given abstained labels $(Y_1 = *, Y_3 = *)$ and the precise prediction $(Y_2 = 1)$, inferring $Y_4 = 1$ comes down to compute $P_x(Y_4 = 1|(Y_1 = 0 \cup Y_1 = 1), Y_2 = 1, (Y_3 = 0 \cup Y_3 = 1))$ as follows

$$\begin{aligned}\frac{\sum_{y_3,y_1\in\{0,1\}^2} P_x(Y_4=1, Y_1=y_1, Y_2=y_2, Y_3=1)}{\sum_{y_3,y_1\in\{0,1\}^2} P_x(Y_1=y_1, Y_2=y_2, Y_3=1)} &= \frac{P_x(Y_4=1, Y_2=1)}{P_x(Y_2=1)}\\ &= P_x(Y_4=1|Y_2=1),\end{aligned}$$

An illustration providing some intuition about this last example can be seen in Figure 6.3, in which we draw the possible path to infer the label Y_4 (considering a third branch in the chain to represent abstained labels).

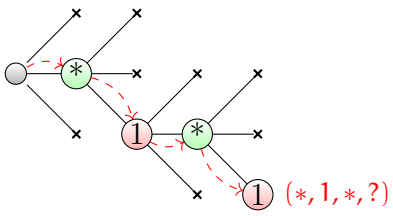


Figure 6.3: Marginalization strategy for four labels $\{Y_1, Y_2, Y_3, Y_4\}$

The results of the last example can easily be generalized, and hence, MAR comes down to calculate the new formulation called (MAR*)

$$\underline{P}^j_x(Y_j = 1) = \min_{P \in \mathscr{P}^*} P_x(Y_j = 1 | Y_{\mathscr{I}_R^{j-1}} = 1, Y_{\mathscr{I}_j^{j-1}} = 0), \tag{6.9}$$

$$\overline{P}^j_x(Y_j = 1) = \max_{P \in \mathscr{P}^*} P_x(Y_j = 1 | Y_{\mathscr{I}_R^{j-1}} = 1, Y_{\mathscr{I}_j^{j-1}} = 0). \tag{6.10}$$

where $\mathscr{P}^*$ is simply the set of joint probability distributions described by the imprecise probabilistic tree (we refer to de Cooman and Herman [De Cooman et al., 2008] for a detailed analysis of those). In general, such an optimisation can be computationally quite intensive, but remains easy in the case of the Naive credal classifier, thanks to its independence assumption (see Section 6.3).

Note that, once any of the two strategies has been applied, we can either keep the prediction as it is, producing an incomplete vector where labels having indices $\mathscr{I}_{\mathcal{A}}$ become imprecise, or we can consider precise estimations of labels $j \in \mathscr{I}_{\mathcal{A}}$ by considering a minimax robust strategy, i.e., picking $\hat{y}_j = \arg\max_{y \in \{0,1\}} \underline{P}^j_x(Y_j = y)$ to replace the label Y_j by the corresponding prediction.

6.2.3 *Safety imprecise chaining*

In contrast to what we presented in the two previous subsections, where the order of labels is obtained randomly, here we propose a new way to order them by dynamically selecting a label as the chain moves forwards.

Our idea aims basically to choose a dynamic, context-dependent label ordering. By context-dependent, firstly, we mean that we consider a metric that measures the level of uncertainty associated to the credal set of the next plausible label to infer. Secondly, and jointly with the latter metric, we borrow, and adapt to our context, the first heuristic proposed in [Sucar et al., 2014]. This heuristic chooses at each step of the chaining the next plausible label with the higher predictive probability (or the one with the best accuracy), in order to minimize the propagation of error in the final joint probability estimate of the chain.

So, concerning the first metric, the label chosen is the one with a lower level of uncertainty[1], and which will be expressed in terms of how small its credal set is (i.e. the length of a one-dimensional interval). In what follows, we will formalize it.

Let us define the function $\phi : [\![m]\!] \to \mathbb{R}_{\geqslant 0}$ which measures the level of uncertainty of the label $Y_j = 1$ (in the same way for $Y_j = 0$), as follows

$$\phi(j) = \overline{P}^j_x(Y_j = 1) - \underline{P}^j_x(Y_j = 1), \tag{6.11}$$

[1] It may be considered as a new perspective where the decision of chosen a label is expressed in terms of gambles, for further details [De Finetti, 2017; Walley, 1991; Shafer et al., 2019]).

where the calculation of probability bounds can be performed using any strategy presented in previous subsections, i.e. IB or MAR strategy.

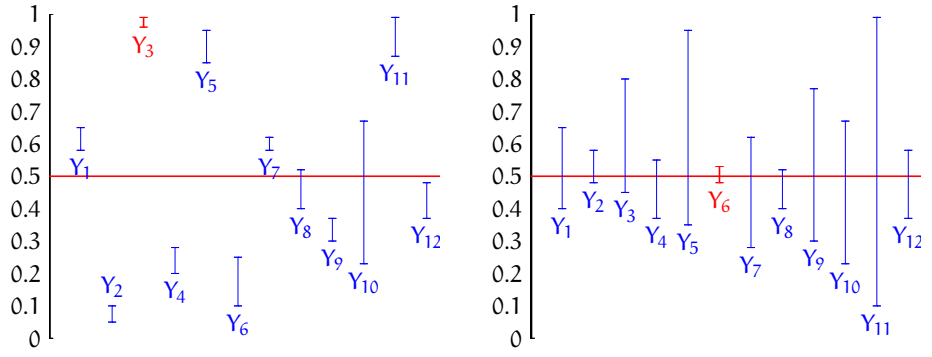
In each inference step (or node of the chain), we perform a slight optimization problem of time complexity $\mathcal{O}(|\mathscr{I}_\mathcal{U}|)$, where $\mathscr{I}_\mathcal{U} = [\![m]\!] \setminus (\mathscr{I}_\mathcal{R} \cup \mathscr{I}_\mathcal{J} \cup \mathscr{I}_\mathcal{A})$ is the set of indices of labels not yet predicted. This optimization chooses the *optimal* index $\hat{j}$ of those labels that has a low uncertainty (i.e. with a $\phi(\cdot)$ minimal) and a high predictive probability, as follows

$$\hat{j} = \arg\min_{j} \left[\phi\left(\arg\max_{\substack{j \in \mathscr{I}_\mathcal{U} \\ 0.5 \notin [\underline{P}^j_x, \overline{P}^j_x]}} \underline{P}^j_x(Y_j = 1) \right), \phi\left(\arg\min_{\substack{j \in \mathscr{I}_\mathcal{U} \\ 0.5 \notin [\underline{P}^j_x, \overline{P}^j_x]}} \overline{P}^j_x(Y_j = 1) \right) \right]. \tag{6.12}$$

This last optimization may in some circumstances give back: (1) two different solutions as long as the level uncertainty $\phi(\cdot)$ of both is equal, even though it is unlikely in practice but not impossible, we select the first element sent by the minimization function of the programming language used, and (2) an empty solution (i.e. no solution at all), it happens when every interval of probability bounds of the set of indices $\mathscr{I}_\mathcal{U}$ contains 0.5, i.e. $0.5 \in [\underline{P}, \overline{P}]$ (see Figure 6.4b). In the latter case, we propose to choose the interval with lower uncertainty, as follows

$$\hat{j} = \arg\min_{j \in \mathscr{I}_\mathcal{U}} \phi(j). \tag{6.13}$$

Once the *optimal* $\hat{j}$ index is selected, be it using Equation (6.12) or (6.13), we can proceed by applying Equation (6.5) for the former equation, in order to obtain the predicted value of $Y_{\hat{j}}$, and for the latter, we can directly assign the abstained value $\{*\}$ since $0.5 \in [\underline{P}, \overline{P}]$.



(a) Optimal w.r.t Equation (6.12) **(b)** Optimal w.r.t. Equation (6.13)

Figure 6.4: Example illustration of optimization problems of the SAFETY IMPRECISE CHAINING.

Illustrations providing some insights on the above-described procedure can be found in Figure 6.4. We here study two different examples;

1. in the Figure 6.4a, we first compute inside optimization problems in order to get the most likely labels (i.e. with a higher probability); Y_3 with the highest lower probability bound and Y_2 with the lowest upper probability bound. Then, we measure the level of uncertainty of each one, using the function ϕ, so the *optimal* label Y_3 as the one with a low uncertainty and a high probability. Finally, applying Equation (6.5) we can obtain its predictive value $\hat{Y}_3 = 1$[2].

2. For the latter in Figure 6.4b, we optimize Equation (6.13) which selects the lowest uncertainty label $\hat{Y}_6 = *$.

Finally, it should easily be noted that although the overall complexity of this procedure can add an increment of $\mathcal{O}(m^2)$ with respect to a random or static ordering, it remains competitive as it selects the less uncertain and the most probable label at each step of the chain in a greedy way.

6.3 NAIVE CREDAL BAYES APPLIED IMPRECISE CHAINING

When we describe our uncertainty by means of a set of distributions $\mathscr{P}^j_{Y_k|\mathbf{x}}$ in lieu of a single distribution $\mathbb{P}^j_{Y_k|\mathbf{x}}$, we can not directly use any existing classical classifier model since they are not tailored at all to use $\mathscr{P}^j_{Y_k|\mathbf{x}}$. However, in the state-of-the-art, we can find a whole variety of them, extending a precise classical one to the imprecise probabilistic setting (c.f. Section 2.2).

Naive credal classifier (NCC)[3][Zaffalon, 2002] is one of these, and the one which we adopt in this chapter for solving optimization problems of strategies presented in the previous section as well as for our experiments (see Section 6.4). NCC extends the classical naive Bayes classifier (NBC) (for further details, we refer to Section 2.3)

As the purpose of the imprecise chaining is to compute binary conditional dependence models, we need only get conditional probability bounds of the probability $P(Y_j = y_j|\mathbf{X} = \mathbf{x}, Y_{\mathscr{I}_{[\![j-1]\!]}} = \hat{y}_{\mathscr{I}_{[\![j-1]\!]}})$, so by using Bayes' theorem and naive Bayes' attribute independence assumption, it can be written as follows

$$\frac{P(Y_j = y_j)\prod_{i=1}^{d} P(X_i = x_i|Y_j = y_j)\prod_{k=1}^{j-1} P(Y_k = \hat{y}_k|Y_j = y_j)}{\sum_{y_l\in\{0,1\}} P(Y_j = y_l)\prod_{i=1}^{d} P(X_i = x_i|Y_j = y_l)\prod_{k=1}^{j-1} P(Y_k = \hat{y}_k|Y_j = y_l)}. \quad (6.14)$$

[2] If the label Y_2 would have had the low uncertainty, then its predictive value would have been $\hat{Y}_2 = 0$.

[3] Bearing in mind that it can be replaced by any other (credal) imprecise classifiers, see [Augustin et al., 2014, §10] or our imprecise classifier presented in Chapter 3.

Computing lower and upper probability bounds $[\underline{P}, \overline{P}]$ over all possible marginals $\mathscr{P}_{Y_j}$ and conditional distributions $\mathscr{P}_{X_i|Y_j}, \mathscr{P}_{Y_k|Y_j}$ can be performed by solving the following minimization/maximization problem of Equation (6.14) as follows

$$\underline{P}(Y_j = y_j | X{=}\mathbf{x}, Y_{\mathscr{I}_{[\![j-1]\!]}}{=}\hat{\mathbf{y}}_{\mathscr{I}_{[\![j-1]\!]}}) = \min_{P \in \mathscr{P}_{Y_j}} \min_{P \in \left\{ \substack{\mathscr{P}_{X_i|Y_j}, \mathscr{P}_{Y_k|Y_j} \\ i=1,\ldots,d \\ k=1,\ldots,j-1} \right\}} P(Y_j = y_j | X{=}\mathbf{x}, Y_{\mathscr{I}_{[\![j-1]\!]}}{=}\hat{\mathbf{y}}_{\mathscr{I}_{[\![j-1]\!]}}), \quad (6.15)$$

$$\overline{P}(Y_j = y_j | X{=}\mathbf{x}, Y_{\mathscr{I}_{[\![j-1]\!]}}{=}\hat{\mathbf{y}}_{\mathscr{I}_{[\![j-1]\!]}}) = \max_{P \in \mathscr{P}_{Y_j}} \max_{P \in \left\{ \substack{\mathscr{P}_{X_i|Y_j}, \mathscr{P}_{Y_k|Y_j} \\ i=1,\ldots,d \\ k=1,\ldots,j-1} \right\}} P(Y_j = y_j | X{=}\mathbf{x}, Y_{\mathscr{I}_{[\![j-1]\!]}}{=}\hat{\mathbf{y}}_{\mathscr{I}_{[\![j-1]\!]}}). \quad (6.16)$$

In practice, we assume a precise estimation of the marginal distribution $\mathbb{P}_{Y_j}$ in lieu of a credal set $\mathscr{P}_{Y_j}$, so optimization problems over the credal set of marginal distributions $\mathscr{P}_{Y_j}$ can be ignored. Therefore, one can easily show that last equations evaluated to $Y_j = 1$ ($Y_j = 0$ can be directly calculated using duality) are equivalent to (cf. Equations (2.30) and (2.28))

$$\underline{P}(Y_j{=}1 | \mathbf{X}{=}\mathbf{x}, Y_{\mathscr{I}_{j-1}} = \hat{\mathbf{y}}_{\mathscr{I}_{j-1}}) = \left(1 + \frac{P(Y_j{=}0)\overline{P}_0(\mathbf{X}{=}\mathbf{x})\overline{P}_0(Y_{\mathscr{I}_{j-1}}{=}\hat{\mathbf{y}}_{\mathscr{I}_{j-1}})}{P(Y_j{=}1)\underline{P}_1(\mathbf{X}{=}\mathbf{x})\underline{P}_1(Y_{\mathscr{I}_{j-1}}{=}\hat{\mathbf{y}}_{\mathscr{I}_{j-1}})}\right)^{-1} \quad (6.17)$$

$$\overline{P}(Y_j{=}1 | \mathbf{X}{=}\mathbf{x}, Y_{\mathscr{I}_{j-1}} = \hat{\mathbf{y}}_{\mathscr{I}_{j-1}}) = \left(1 + \frac{P(Y_j = 0)\underline{P}_0(\mathbf{X}{=}\mathbf{x})\underline{P}_0(Y_{\mathscr{I}_{j-1}}{=}\hat{\mathbf{y}}_{\mathscr{I}_{j-1}})}{P(Y_j{=}1)\overline{P}_1(\mathbf{X} = \mathbf{x})\overline{P}_1(Y_{\mathscr{I}_{j-1}}{=}\hat{\mathbf{y}}_{\mathscr{I}_{j-1}})}\right)^{-1} \quad (6.18)$$

where probability bounds $[\underline{P}_1, \overline{P}_1]$ and $[\underline{P}_0, \overline{P}_0]$ of each different conditional event are defined as follows

$$\overline{P}_0(\mathbf{X} = \mathbf{x}) := \prod_{i=1}^{d} \overline{P}(X_i = x_i | Y_j{=}0) \text{ and } \overline{P}_0(\mathbf{Y}_{\mathscr{I}_{j-1}} = \mathbf{y}_{\mathscr{I}_{j-1}}) := \prod_{k=1}^{j-1} \overline{P}(Y_k = \hat{y}_k | Y_j{=}0), \quad (6.19)$$

$$\underline{P}_1(\mathbf{X} = \mathbf{x}) := \prod_{i=1}^{d} \underline{P}(X_i = x_i | Y_j{=}1) \text{ and } \underline{P}_1(\mathbf{Y}_{\mathscr{I}_{j-1}} = \mathbf{y}_{\mathscr{I}_{j-1}}) := \prod_{k=1}^{j-1} \underline{P}(Y_k = \hat{y}_k | Y_j{=}1). \quad (6.20)$$

The last conditional probability bounds are derived using the Imprecise Dirichlet model (IDM) [Walley, 1996], more precisely using Equation (2.40)

$$\underline{P}(X_i{=}x_i | Y_j{=}y_j) = \frac{n(x_i|y_j)}{n(y_j) + s} \text{ and } \overline{P}(X_i{=}x_i | Y_j{=}y_j) = \frac{n(x_i|y_j) + s}{n(y_j) + s} \quad (6.21)$$

and in the same way, we obtain probability bounds $[\underline{P}(Y_k = \hat{y}_k|Y_j = y_j), \overline{P}(Y_k = \hat{y}_k|Y_j = y_j)]$. For more details about the count function $n(\cdot)$ and the hyper-parameter s, we refer to Section 2.3.2.

In what follows, we propose efficient procedures to solve the strategies presented in the previos section using the properties of the imprecise classifier described above.

6.3.1 *Imprecise branching*

In the specific case where we use the naive credal classifier, we can efficiently reduce optimization problems of the imprecise branching strategy, namely Equations (IB), as expressed in the proposition below.

Proposition 8 *Optimisation problems of the imprecise branching* (IB) *can be reduced by using probability bounds obtained from the naive credal classifier, namely Equations* (6.17) *and* (6.18)*, as follows*

$$\min_{y\in\{0,1\}^{|\mathscr{I}_{\mathcal{A}}|}} \underline{P}_x(Y_j=1|Y_{\mathscr{I}_{\mathcal{R}}^{j-1}}=1, Y_{\mathscr{I}_j^{j-1}}=0, Y_{\mathscr{I}_{\mathcal{A}}^{j-1}}=y) \propto \max_{y\in\{0,1\}^{|\mathscr{I}_{\mathcal{A}}|}} \frac{\overline{P}_0(Y_{\mathscr{I}_{\mathcal{A}}^{j-1}}=y)}{\underline{P}_1(Y_{\mathscr{I}_{\mathcal{A}}^{j-1}}=y)}, \tag{6.22}$$

$$\max_{y\in\{0,1\}^{|\mathscr{I}_{\mathcal{A}}|}} \overline{P}_x(Y_j=1|Y_{\mathscr{I}_{\mathcal{R}}^{j-1}}=1, Y_{\mathscr{I}_j^{j-1}}=0, Y_{\mathscr{I}_{\mathcal{A}}^{j-1}}=y) \propto \min_{y\in\{0,1\}^{|\mathscr{I}_{\mathcal{A}}|}} \frac{\underline{P}_0(Y_{\mathscr{I}_{\mathcal{A}}^{j-1}}=y)}{\overline{P}_1(Y_{\mathscr{I}_{\mathcal{A}}^{j-1}}=y)}. \tag{6.23}$$

Besides, applying Equation (6.21) *derived from the imprecise Dirichlet model, we have that the values of abstained labels for which the previous optimisation problems are solved are, respectively*

$$\underline{\hat{y}}_{\mathscr{I}_{\mathcal{A}}^{j-1}} := \arg\max_{y\in\{0,1\}^{|\mathscr{I}_{\mathcal{A}}|}} \prod_{y_i\in y} \frac{n(y_i|y_j=0)+s}{n(y_i|y_j=1)} \tag{6.24}$$

$$\hat{\overline{y}}_{\mathscr{I}_{\mathcal{A}}^{j-1}} := \arg\min_{y\in\{0,1\}^{|\mathscr{I}_{\mathcal{A}}|}} \prod_{y_i\in y} \frac{n(y_i|y_j=0)}{n(y_i|y_j=1)+s} \tag{6.25}$$

where $\mathscr{I}_{\mathcal{A}}^{j-1}$ *is the set of indices of* $(j-1)$th *first predicted abstained labels.*

Proof 11 (Proof of Proposition 8) *Let us begin to prove the optimization problem of the lower probability of* (IB) *evaluated to* $Y_j = 1$

$$\underline{P}_x(Y_j=1) = \min_{y\in\{0,1\}^{|\mathscr{I}_{\mathcal{A}}|}} \underline{P}_x(Y_j=1|Y_{\mathscr{I}_{\mathcal{R}}^{j-1}}=1, Y_{\mathscr{I}_j^{j-1}}=0, Y_{\mathscr{I}_{\mathcal{A}}^{j-1}}=y). \tag{6.26}$$

Let us to define $\mathscr{I}_*^{j-1} = \mathscr{I}_{\mathcal{R}}^{j-1} \cup \mathscr{I}_j^{j-1}$ *as the set of indices of relevant and irrelevant predicted labels down to* $(j-1)$th *index. By applying Equation* (6.17) *to the right side of last equation, we get*

$$\min_{y\in\{0,1\}^{|\mathscr{I}_{\mathcal{A}}|}}\left(1+\frac{P(Y_j=0)\overline{P}_0(\mathbf{X}=x)\overline{P}_0(Y_{\mathscr{I}_*^{j-1}}=\hat{y}_{\mathscr{I}_*^{j-1}},Y_{\mathscr{I}_{\mathcal{A}}^{j-1}}=y)}{P(Y_j=1)\underline{P}_1(\mathbf{X}=x)\underline{P}_1(Y_{\mathscr{I}_*^{j-1}}=\hat{y}_{\mathscr{I}_*^{j-1}},Y_{\mathscr{I}_{\mathcal{A}}^{j-1}}=y)}\right)^{-1}, \tag{6.27}$$

where $\hat{y}_{\mathscr{I}_*^{j-1}}$ *is the binary vector with predicted relevant and irrelevant values. So, using the fact that minimizing* $\frac{1}{1+x}$ *is equal to maximize* x*, we therefore get*

$$\max_{y\in\{0,1\}^{|\mathscr{I}_{\mathcal{A}}|}}\frac{P(Y_j=0)\overline{P}_0(\mathbf{X}=x)\overline{P}_0(Y_{\mathscr{I}_*^{j-1}}=\hat{y}_{\mathscr{I}_*^{j-1}})\overline{P}_0(Y_{\mathscr{I}_{\mathcal{A}}^{j-1}}=y)}{P(Y_j=1)\underline{P}_1(\mathbf{X}=x)\underline{P}_1(Y_{\mathscr{I}_*^{j-1}}=\hat{y}_{\mathscr{I}_*^{j-1}})\underline{P}_1(Y_{\mathscr{I}_{\mathcal{A}}^{j-1}}=y)}. \tag{6.28}$$

The first three terms of the numerator (and of the denominator) can be considered as constants (and omitted) and by applying Equation (6.21) to the last term, we get what we sought

$$\begin{aligned}&\max_{y\in\{0,1\}^{|\mathscr{I}_{\mathcal{A}}|}}\frac{\overline{P}_0(Y_{\mathscr{I}_{\mathcal{A}}^{j-1}}=y)}{\underline{P}_1(Y_{\mathscr{I}_{\mathcal{A}}^{j-1}}=y)}\\ \iff&\max_{y\in\{0,1\}^{|\mathscr{I}_{\mathcal{A}}|}}\left[\frac{n(y_j=1)+s}{n(y_j=0)+s}\right]^{|\mathscr{I}_{\mathcal{A}}|}\prod_{y_i\in y}\frac{n(y_i|y_j=0)+s}{n(y_i|y_j=1)},\end{aligned}$$

in which it is easy see that: (1) the term $[\cdots]^{|\mathscr{I}_{\mathcal{A}}|}$ *can be omitted, and hence, we can get* $\underline{\hat{y}}_{\mathscr{I}_{\mathcal{A}}^{j-1}}$*, and (2) using the similar arguments above we can easily get* $\overline{\hat{y}}_{\mathscr{I}_{\mathcal{A}}^{j-1}}:=\overline{P}_x(Y_j=1)$*.* ■

Proposition 8 amounts to saying that is not necessary to know the original input features $\mathscr{X}$ and neither the $(j-1)$th first precise predicted labels, in order to get the lower and upper probability bound of Equations (IB). However, on the other hand, it is necessary to know the number of bits $n(\cdot)$ interchanged on all possibles different paths of abstained labels, which remains consistent with the fact that we want to capture the optimal lower and upper bounds of the conditional probability (i.e. the "optimal" dependence interaction between labels) over all possible paths on which we have abstained.

Proposition 8 allows us to propose an algorithm that can calculate Equations (6.24) and (6.25) linearly in the number of abstained labels. It is shown in the following proposition.

Proposition 9 *The bounds* $\underline{\hat{y}}_{\mathscr{I}_{\mathcal{A}}^{j-1}}$ *and* $\overline{\hat{y}}_{\mathscr{I}_{\mathcal{A}}^{j-1}}$ *can be obtained in a time complexity of* $\mathcal{O}(|\mathscr{I}_{\mathcal{A}}^{j-1}|)$ *by a dichotomic search.*

Proof 12 (proof of Proposition 9) *This proof can be performed using a dichotomy algorithm (equivalent to a binary search tree), starting with* y_k *last abstained label (i.e.* $k=|\mathscr{I}_{\mathcal{A}}|-1$*) and calculating the values* $\frac{n(y_k=1|\cdot)+s}{n(y_k=1|\cdot)}$ *and*

$\frac{n(y_k=0|\cdot)+s}{n(y_k=0|\cdot)}$, *then we retain the maximal value of these last two terms (or the minimal value, whichever applies) and we go forward with second-to-last label* y_{k-1}, *but this time multiplied by the last term retained, and so on. After having obtained the lower binary path* $\underline{\hat{\mathbf{y}}}_{\mathscr{I}_{\mathcal{A}}^{j-1}}$ *(or the upper binary path* $\overline{\hat{\mathbf{y}}}_{\mathscr{I}_{\mathcal{A}}^{j-1}}$*), we can directly calculate the values* $\underline{P}_x^j(Y_j = 1)$ *and* $\overline{P}_x^j(Y_j = 1)$ *(and for duality of the lower and upper probability bounds* $[\underline{P}_x^j(Y_j = 0), \overline{P}_x^j(Y_j = 0)]$*).* ■

The following proposition provides the time complexity of the inference step of the imprecise branching strategy, jointly with the naive credal classifier and previous results.

Proposition 10 *The global time complexity of the* IMPRECISE BRANCHING *strategy in the worst-case is* $\mathcal{O}(m^2)$ *and in the best-case is* $\mathcal{O}(m)$.

Proof 13 (Proof of Proposition 10) *The proof for the best-case is straighforward, because if there is not any abstained labels, the time complexity is the same than precise chaining* $\mathcal{O}(m)$. *The worst-case complexity, in which all inferred labels are abstained, is also easy to calculate since; the first label performs a single operation, i.e.* $\mathcal{O}(1)$, *then second label is also inferred in a single operation due to the number of previous abstained labels is equal to* 1 *(c.f. Proposition 9), then the third label takes in account two previous abstained labels and performs two operations (c.f. Proposition 9), and the fourth label performs three operations, and so on. We therefore obtain* $\mathcal{O}(m(m-1)/2+1)$ *operations which is equal* $\mathcal{O}(m^2)$ *asymptotically.* ■

6.3.2 *Marginalization*

When the naive credal classifier is considered, nothing needs to be optimized in the marginalization strategy, thanks to assumption of independence applied to each binary conditional model of the chain.

We recall that the marginalization strategy needs to compute the conditional models described in Equations (MAR). These latter can be solved using Equations (6.17) and (6.18) of the NCC. We thus focus on Equation (6.18) (Equation (6.17) can be treated in a similar way), in order to show that the abstained labels can be removed of the conditioning and to get the expression presented in Equation (6.10).

Based on Equation (6.18), we can only focus on the conditional probability on labels, namely Equation (6.19), and rewrite it as follows:

$$\overline{P}_0(\mathbf{Y}_{\mathscr{I}_{j-1}} = \mathbf{y}_{\mathscr{I}_{j-1}}) := \overline{P}_0(\mathbf{Y}_{\mathscr{I}_*^{j-1}} = \hat{\mathbf{y}}_{\mathscr{I}_*^{j-1}}, \mathbf{Y}_{\mathscr{I}_{\mathcal{A}}^{j-1}} = \{0,1\}^{|\mathscr{I}_{\mathcal{A}}^{j-1}|}) \tag{6.29}$$

where $\mathscr{I}_*^{j-1} = \mathscr{I}_{\mathcal{R}}^{j-1} \cup \mathscr{I}_{\mathcal{J}}^{j-1}$ is the set of indices of relevant and irrelevant inferred labels, and the right side of last equation can be stated as

$$\max_{\substack{P\in\left\{\mathscr{P}_{Y_k|Y_j},\mathscr{P}_{Y_a|Y_j}\right\}\\ k\in\mathscr{I}_*^{j-1},a\in\mathscr{I}_{\mathcal{A}}^{j-1}}}\prod_{k\in\mathscr{I}_*^{j-1}}P(Y_k=\hat{y}_k|Y_j=0)\prod_{a\in\mathscr{I}_{\mathcal{A}}^{j-1}}P(Y_a=0\cup Y_a=1|Y_j=0).$$

Thanks to the assumption of independence, each credal set is also independent of the others, making possible to decouple the multiplication in two parts, as follows

$$\max_{\substack{P\in\mathscr{P}_{Y_k|Y_j}\\ k\in\mathscr{I}_*^{j-1}}}\prod_{k\in\mathscr{I}_*^{j-1}}P(Y_k=\hat{y}_k|Y_j=0)\times\max_{\substack{P\in\mathscr{P}_{Y_a|Y_j}\\ a\in\mathscr{I}_{\mathcal{A}}^{j-1}}}\prod_{a\in\mathscr{I}_{\mathcal{A}}^{j-1}}P(Y_a=0\cup Y_a=1|Y_j=0),$$

where $P(Y_a=0\cup Y_a=1|Y_j=0)=1$, and therefore, we finally get

$$\overline{P}_0(\mathbf{Y}_{\mathscr{I}_{j-1}}=\mathbf{y}_{\mathscr{I}_{j-1}}):=\max_{\substack{P\in\mathscr{P}_{Y_k|Y_j}\\ k\in\mathscr{I}_*^{j-1}}}\prod_{k\in\mathscr{I}_*^{j-1}}P(Y_k=\hat{y}_k|Y_j=0). \tag{6.30}$$

By replacing the last expression to Equation (6.18), we get what we sought

$$\overline{P}_x^j(Y_j=1)=\max_{P\in\mathscr{P}_{Y_j|Y_{\mathscr{R}_j^{j-1}},Y_{\mathscr{I}_j^{j-1}}}}P_x(Y_j=1|Y_{\mathscr{R}_j^{j-1}}=1,Y_{\mathscr{I}_j^{j-1}}=0). \tag{6.31}$$

Therefore, at each inference step, we can directly apply Equations (6.17) and (6.18) on the reduced new formulation of the marginalization strategy (MAR*).

An illustration providing some intuition about this reduction performed by applying the NCC, and followed by what was presented in Figure 6.3, can be seen in Figure 6.5.

Figure 6.5: Marginalization strategy applied to NCC for four labels $\{Y_1, Y_2, Y_3, Y_4\}$

Furthermore, one can be tempted to perform an approximation of the general formulation of the marginalization strategy by using the same arguments (i.e. the law of total probability and Bayes' theorem), as follows.

Using similar arguments[4] as in [Augustin et al., 2014, §9.2.2], the marignalized credal set $\mathscr{P}^*$ can be obtained by considering extreme points of the global credal set as follows

[4] Note that it can fail on certain conditions, further details [Augustin et al., 2014, §2.3.4].

$$\mathscr{P}_x(Y_j|\hat{y}_{\mathscr{I}_*^{j-1}}, \hat{y}^{\{0,1\}}_{\mathscr{I}_A^{j-1}}) := \\ \mathrm{CH}\left\{ P_x(Y_j|\hat{y}_{\mathscr{I}_*^{j-1}}) \middle| \begin{array}{l} P_x(Y_j = y_j|\hat{y}_{\mathscr{I}_*^{j-1}}, \hat{y}^{\{0,1\}}_{\mathscr{I}_A^{j-1}}) := P_x(Y_j = y_j|\hat{y}_{\mathscr{I}_*^{j-1}}), \\ \forall y_j \in \{0,1\}, \forall P_x(Y_j|Y_{\mathscr{I}_*^{j-1}}) \in \mathrm{ext}\left[\mathscr{P}_x(Y_j|Y_{\mathscr{I}_*^{j-1}})\right] \end{array} \right\}$$

where $\mathscr{I}_*^{j-1} = \mathscr{I}_{\mathcal{R}}^{j-1} \cup \mathscr{I}_{\mathcal{J}}^{j-1}$ set of indices of relevant/irrelevant inferred labels, where $\hat{y}_{\mathscr{I}_*^{j-1}}$ and $\hat{y}^{\{0,1\}}_{\mathscr{I}_A^{j-1}} = \{0,1\}^{|\mathscr{I}_A^{j-1}|}$ are the previous predicted precise and abstained values of labels, $\mathrm{ext}[\mathscr{P}]$ is the set of extreme points of the credal set, and $\mathrm{CH}\{\cdot\}$ is the convex hull. However, the number of such extreme points grow exponentially with the size of the tree, and providing efficient algorithms to work with those will be the matter of future works.

6.4 EXPERIMENTS

In this section, we perform experiments[5] on 3 data sets issued from the MULAN repository[6] (c.f. Table 5.2), following a 10×10 cross-validation procedure (at every jth-fold, we proceed randomly to shuffle the set of labels).

Table 6.2: Multi-label data sets summary

Data set	#Features	#Labels	#Instances	#Cardinality	#Density
emotions	72	6	593	1.90	0.31
scene	294	6	2407	1.07	0.18
yeast	103	14	2417	4.23	0.30

6.4.1 *Evaluation and setting*

The usual metrics used in multi-label problems are not adapted at all when we infer set-valued predictions. Thus, we consider appropriate to use the set-accuracy (SA) and completeness (CP) [Destercke, 2014, §4.1], as follows

$$\mathrm{SA}(\hat{y}, y) = \mathbb{1}_{(y \in \hat{y})} \quad \text{and} \quad \mathrm{CP}(\hat{y}, y) = \frac{|Q|}{m},$$

where $\hat{y}$ is the partial binary prediction (i.e. the set of all possibles binary vectors) and Q denote the set of non-abstained labels. When predicting complete vectors, then $\mathrm{CP} = 1$ and SA equals the $0/1$ loss function and when predicting the empty vector, i.e. all labels $\hat{y}_i = *$, then $\mathrm{CP} = 0$ and by convention $\mathrm{SA} = 1$. The reason for using SA is that chaining is used as an approximation of the optimal prediction for a $0/1$ loss function.

[5] Implemented in Python, see https://github.com/sdestercke/classip
[6] http://mulan.sourceforge.net/datasets.html

6.4.2 *Imprecise classifier*

As was mentioned earlier, in Section 6.3, we chose to use the so-called *naive credal classifier* (NCC)[Zaffalon, 2002] in order to compute the class-conditional probability bounds. Note that NCC is not adapted at all to work with a continuous input space, so we discretize data sets to $z = 5$ and $z = 6$ intervals. Besides, we restrict the values of the hyper-parameter of the imprecision to $s \in \{0.0, 0.5, 1.5, \ldots, 4.5, 5.5\}$ (when $s = 0.0$, NCC becomes the precise classifier NBC). At higher values of s $>>>>$ 0, the NCC model will make mostly vacuous predictions (i.e. abstain in all labels $\forall i, Y_i = *$) for the data sets we consider here.

6.4.3 *Missing and Noise labels*

In this chapter, we consider the same settings as in the previous chapter (c.f. Section 5.3.3) for missing and noisy labels. We quickly recall the different levels of missingness and noisiness:

1. **Missing** percentage of missing labels $\{0, 20, 40, 60, 80\}$
2. **Noise**
   (a) **Reversing** percentage of noisy labels $\{10, 20, 30, 40, 50, 60\}$
   (b) **Flipping** percentage of noisy labels $\{20, 40, 60, 80\}$ and $\beta \in \{0.2, 0.5, 0.8\}$.

6.4.4 *Experimental results*

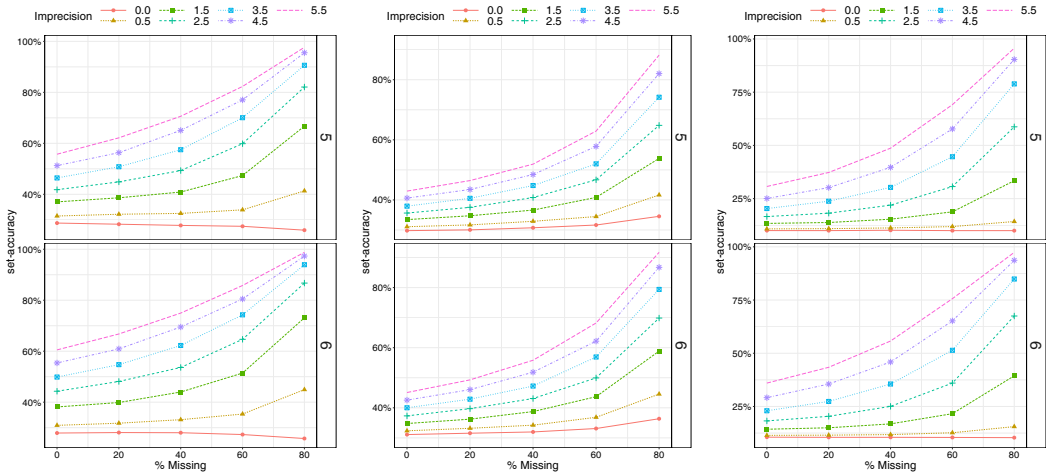
We present the results of this section separated into two parts: (1) to evaluate if we obtain more accurate and precise predictions by injecting the imprecision, and (2) as the minimax approach applied to our imprecise model is compared to its precise counterpart.

6.4.4.1 *Set-accuracy versus Completeness*

The confidence intervals obtained on the results presented in this section are very small so that we prefer not to display in the figures in order not to overcharge them.

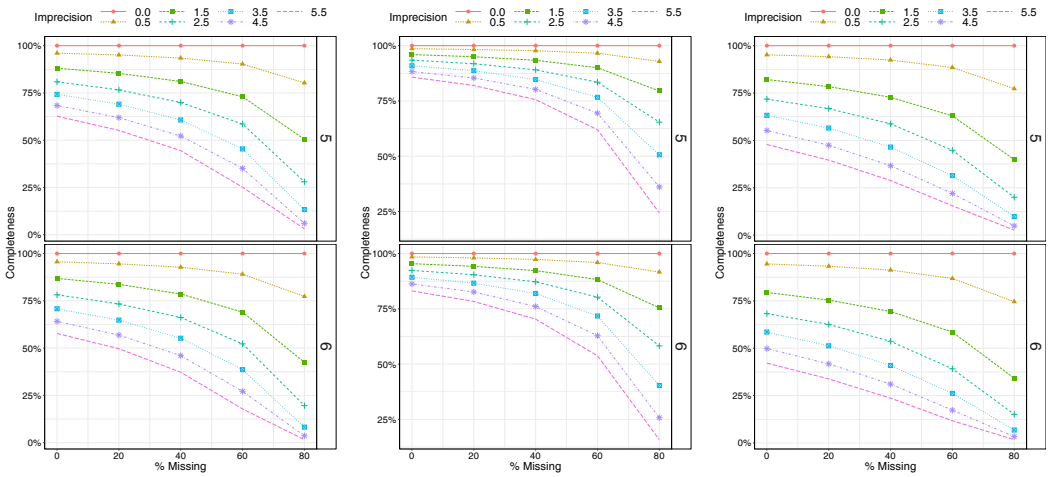
In Figures 6.6 and 6.7, we provide the results of the set-accuracy and completeness measures in average (%), respectively, obtained by fitting the NCC model on different percentage of missing labels applied to the data sets of Table 6.2 and using the imprecise branching strategy[7].

[7] The results of the marginalization strategy have been placed in Appendix B.1 in order to simplify the narrative.



(a) EMOTIONS **(b)** SCENE **(c)** YEAST

Figure 6.6: **Missing labels - Imprecise Branching** Evolution of the average set-accuracy (%) for each level of imprecision (a curve for each one) and discretization $z = 5$ (top) and $z = 6$ (down), with respect the the percentage of missing labels.



(a) EMOTIONS **(b)** SCENE **(c)** YEAST

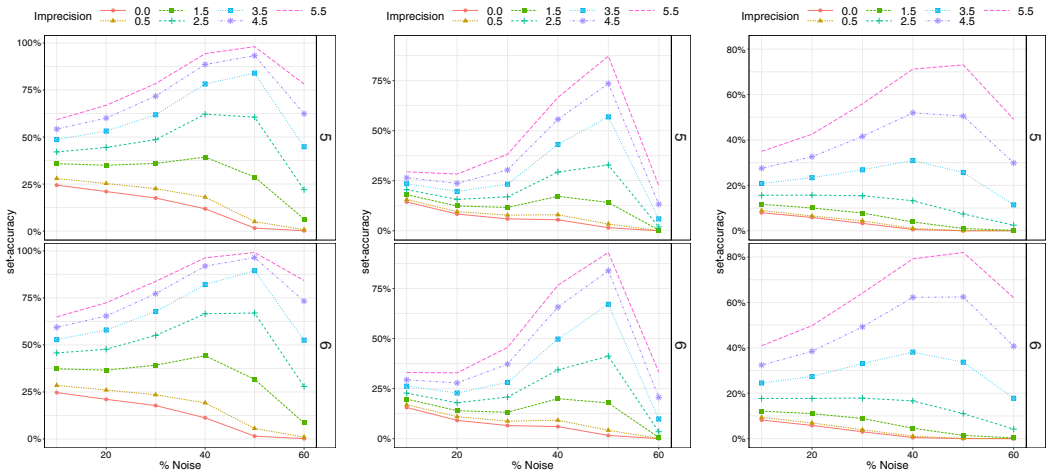
Figure 6.7: **Missing labels - Imprecise Branching** Evolution of the average completeness (%) for each level of imprecision (a curve for each one) and discretization $z = 5$ (top) and $z = 6$ (down), with respect the the percentage of missing labels.

The results displayed are those that we expect, and are at the same time roughly similar to those presented in the previous chapter. Indeed, when s increases, the set-accuracy (SA) increases as we forget more and more (as completeness or CP decreases), meaning that the more imprecision we get, the more accurate are those predictions we retain.

A significative difference, in contrast to the results of missingness on Yeast dataset of the previous chapter, is that this latter does not need anymore a high amount of imprecision to witness a little gain in set-accuracy. This is certainly due to dependency information which the chain provides to the future inferred labels.

One noticeable result shows that, in contrast to those results presented in Chapter 5, a high amount of imprecision is required so that the ground-truth solution to be within the set-valued of predictions (it is certainly due to 0/1 loss metric). For instance, with $s = 5.5$, $z = 4$ and 40% of missingness, we get a $> 65\%$ of set-accuracy versus a $< 50\%$ of completeness, in Emotions data set.

In Figures 6.8 and 6.10, we provide the results of the set-accuracy measure in average (%) obtained by fitting the NCC model on different percentage of **Reversing** and **Flipping** settings applied to the data sets of Table 6.2 and using the imprecise branching strategy.[8]. Results about its completeness are given in Figures 6.9 and 6.11.

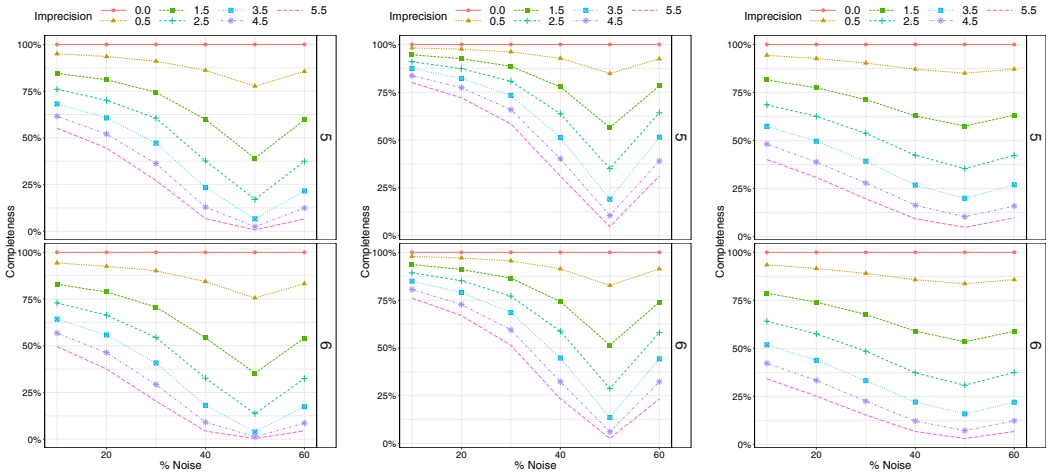


(a) EMOTIONS **(b)** SCENE **(c)** YEAST

Figure 6.8: **Reversing - Imprecise Branching** Evolution of the average set-accuracy (%) for each level of imprecision (a curve for each one) and two levels of discretization $z = 5$ (top) and $z = 6$ (down), with respect to the percentage of noise.

Concerning the **Reversing** and the **Flipping**, we encounter roughly the same findings as in the experiments of Chapter 5, with the only exception that getting the ground-truth solution amongst the set-valued predictions is also the hardest than getting a subset (or partial) solution of the ground-truth one. We can see for instance that the set-accuracy of imprecise setting ($s > 0$) produces a poor performance compared with those of the previous

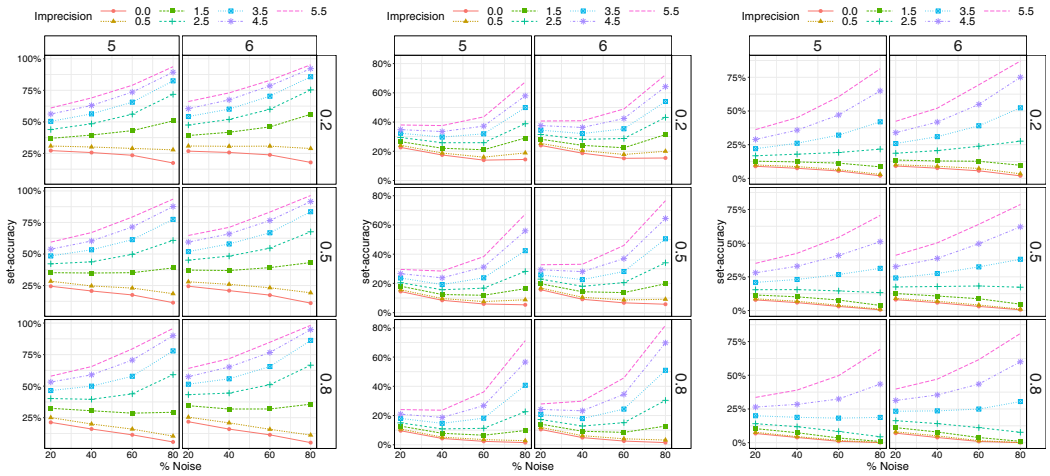
[8] The results of the marginalization strategy had been placed in Appendix B.2 and B.3 in order to simplify the narrative.



(a) EMOTIONS **(b)** SCENE **(c)** YEAST

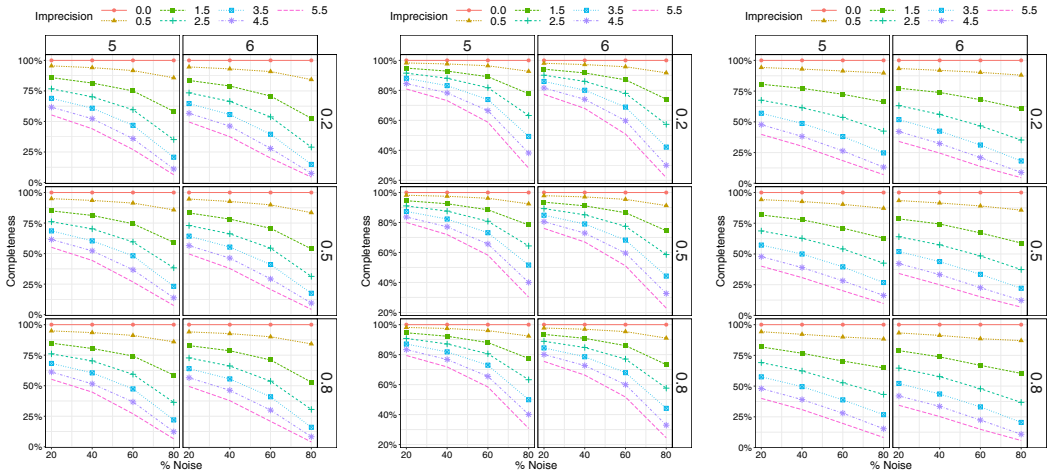
Figure 6.9: **Reversing - Imprecise Branching** Evolution of the average completeness (%) for each level of imprecision (a curve for each one) and two levels of discretization $z = 5$ (top) and $z = 6$ (down), with respect to the percentage of noise.

chapter, but that is quite normal because the incorrectness (IC) penalizes the individual errors produced by the partially inferred vector (as the Hamming loss), whereas the set-accuracy (SA) does not so, this latter does not reward the set-valued predictions if the ground-truth is not within.



(a) EMOTIONS **(b)** SCENE **(c)** YEAST

Figure 6.10: **Flipping - Imprecise Branching** Evolution of the average set-accuracy (%) for each level of imprecision (a curve for each one) and two levels of discretization $z = 5$ (top) and $z = 6$ (down), with respect to the percentage of noise.



(a) EMOTIONS **(b)** SCENE **(c)** YEAST

Figure 6.11: **Flipping - Imprecise Branching** Evolution of the average completeness (%) for each level of imprecision (a curve for each one) and two levels of discretization $z = 5$ (top) and $z = 6$ (down), with respect to the percentage of noise.

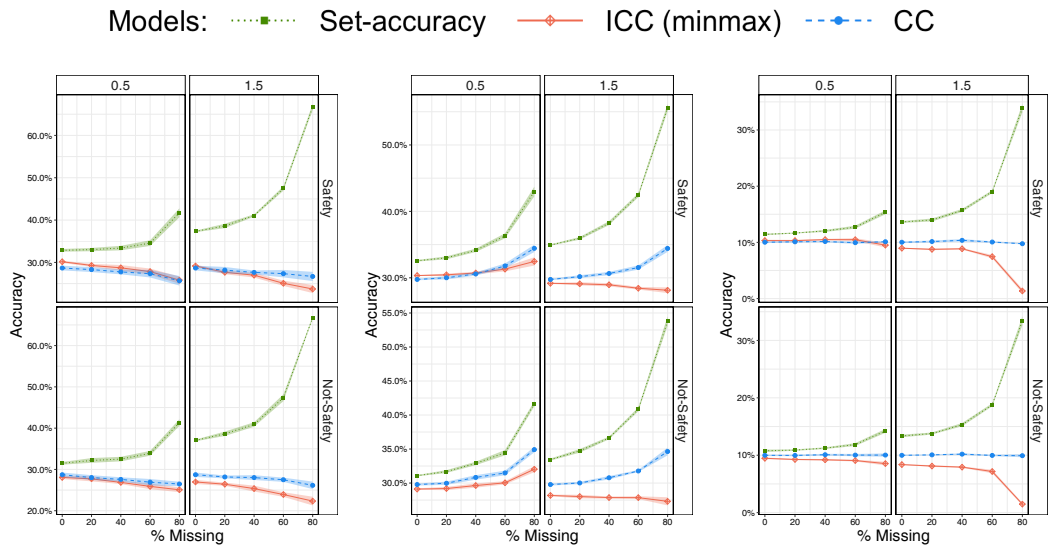
Finally, as regards the comparison of performances on the safety and not-safety (presented above) approach when they make cautious inferences, we can note a slight difference among them in terms of the set-accuracy metric (and, consequently, the completeness as well). The performance of the safety approach is slightly better (or worse) as the amount of imprecision increases, depending on the data set, than the not-safety one. This difference can also be found in all different settings, either with missing or noisy labels, or with imprecise-branching or maginalization strategy. Hence, in order not to overload this section with more illustrations, and besides the interest of using the safety approach is to infer more precise-valued inferences with low uncertainty, we thus prefer to put a single set-up, namely marginalization strategy with missing labels, in Figures B.3 and B.4 (the set-accuracy and completeness, respectively) of Appendix B.1.

6.4.4.2 *CC versus ICC using minimax strategy*

The average performance of the minimax approach for the imprecise branching strategy (IB) obtained in terms of the SE measure and using the **safety imprecise chaining**[9] or not are shown in Figure 6.12, 6.13 and 6.14 for the **missing, reversing and flipping** settings respectively, with two

[9] As results obtained with the maginalization strategy are roughly similar, we preferred to put them in Appendix B.1, B.2 and B.3, with all different imprecise levels, in order to simplify the interpretation.

imprecise levels[10] $s \in \{0.5, 1.5\}$, applied to our imprecise approach (ICC) (resp. precise approach (CC)).



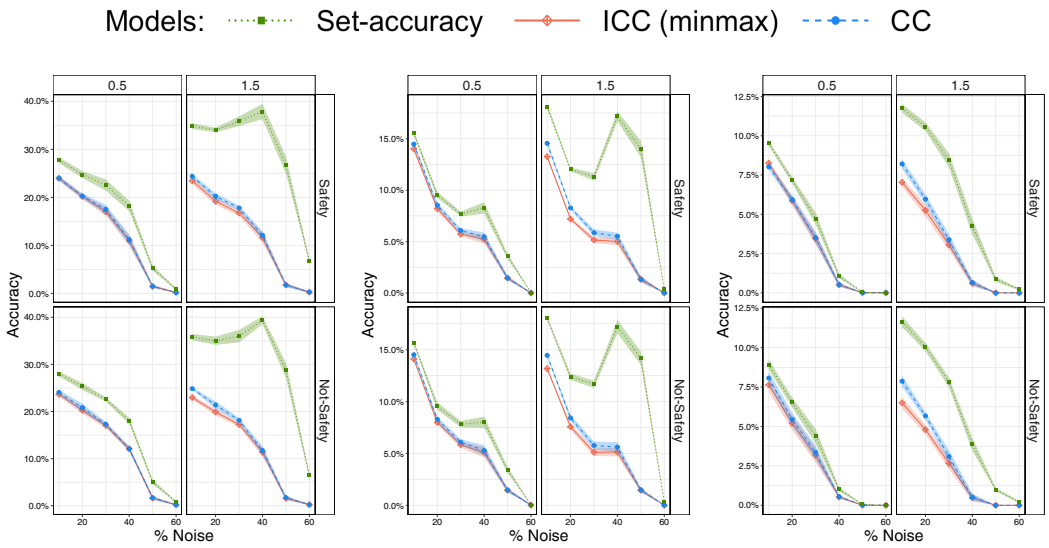
(b) EMOTIONS **(c)** SCENE **(d)** YEAST

Figure 6.12: **Missing - ICC versus CC - Imprecise Branching.** Figures show performance evolution (%) of the imprecise and precise classifier-chains approaches for 0.5 (left) and 1.5 (right) levels of imprecisions and using the **safety imprecise chaining** (top) and not (down), with respect to the percentage of missing (x-axis).

The overall results showed on all different settings give a clear evidence that including a light level of imprecision in our imprecise approach using the **safety imprecise chaining** gives overall comparable results when considering precise, minimax predictions, while significantly increasing accuracy when considering partial predictions. In contrast, when we do not use the **safety imprecise chaining**, the precise approach overcomes ours, more specifically in the case of the missing labels, it seems that the minimax strategy and the addition of imprecision actually degrade the results, which is surprising and worthy of further investigations (e.g. using other strategies as maximin or more complex ones with theoretical justifications). In addition, interestingly though, our strategy seems to be more robust to the presence of high noise in the data, as we systematically outperform the precise chaining when the labels are affected by $\leqslant 60\%$ (depending on the level of imprecision and strategy used (MAR) or (IB)). Future works will aim at achieving a deeper investigation of this result, but we can already see that chosing the right ordering may be even more important in an imprecise setting.

One quite noticeable result that we can also note with respect the minimax approach is that the more the set-accuracy increases, the more the

[10] We could have optimised on s, but it seemed unfair compared to the precise approach that does not benefit from this hyper-parameter.



(b) EMOTIONS **(c)** SCENE **(d)** YEAST

Figure 6.13: **Reversing - ICC versus CC - Imprecise Branching** Figures show performance evolution (%) of the imprecise and precise classifier-chains approaches for 0.5 (left) and 1.5 (right) levels of imprecisions, and using the **safety imprecise chaining** (top) and not (down), with respect to the percentage of noisy (x-axis).

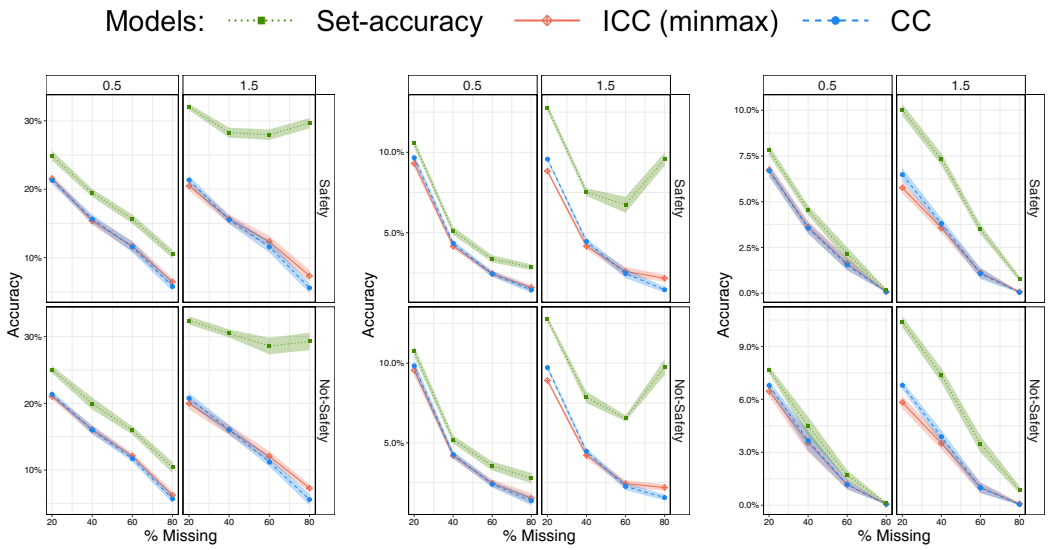
minimax accuracy worsen in the missing setting, and an inverse result is found in the noise setting. Finally, another interesting result about the noisy setting is that when the labels are more noisy with irrelevant values $Y_{i,j} = 0$, i.e. $\beta = 0.2$ a low probability to be relevant $Y_{i,j} = 1$, our proposal ICC is roughly comparable to CC with a low amount of imprecision and it begins to degrade in performance as the imprecision increases (for more details, see Figure B.13 of Appendix B.3).

All those results, however, only provide a proof of concept for our methodology, and are also obtained with a classifier which, through its independence assumption, makes imprecise chaining computationally efficient but limits the benefits of using a chaining approach.

6.5 CONCLUSIONS

In this chapter, we have introduced initial ideas to adapt the classical chaining algorithms of multi-label problems to the case of imprecise or set-valued probabilities. Such an idea is indeed promising to tamper the usual biases of picking a particular branch in the chain.

However, much remains to be done, as how come up with a decision criterion, as the minimax approach, with theoretical results that guarantee a better precise prediction. Indeed, while the Naive credal classifier makes them easy to solve thanks to its assumptions, the same assumptions may be the reason for our mitigated results. It seems therefore essential, in



(b) EMOTIONS **(c)** SCENE **(d)** YEAST

Figure 6.14: **Flipping - ICC versus CC - Imprecise Branching** Figures show performance evolution (%) of the imprecise and precise classifier-chains approaches for 0.5 (left) and 1.5 (right) levels of imprecisions, and $\beta = 0.8$, and using the **safety imprecise chaining** (top) and not (down), with respect to the percentage of noisy (x-axis).

future works, to investigate other classifiers as well as to solve optimisation issues.

Note that for practical purposes, we consider a precise marginal distribution $\mathbb{P}_Y$ in the set of our experiments, but it is well known that multi-label data sets come with a higher class-imbalance among labels, e.g with 7% relevant and 93% irrelevant labels. But due to lack of time, we could not relax this constraint and carry out a in-depth analysis.

An open issue of particular interest is about whether the binary decision of Equation (6.2) is not too penalizing when the lower or upper probabilities are too close of 0.5. For instances, for two labels Y_i and Y_j with intervals of probabilities $[0.51, 0.59]$ and $[0.49, 0.70]$, respectively, Y_i will be inferred as relevant label whereas Y_j will be inferred as abstained label, yet Y_j may also be considered as relevant since the odds is higher. So, a future work might just be focused in as handling this matter, by using other decision criteria or a hyper-parameter on the model.

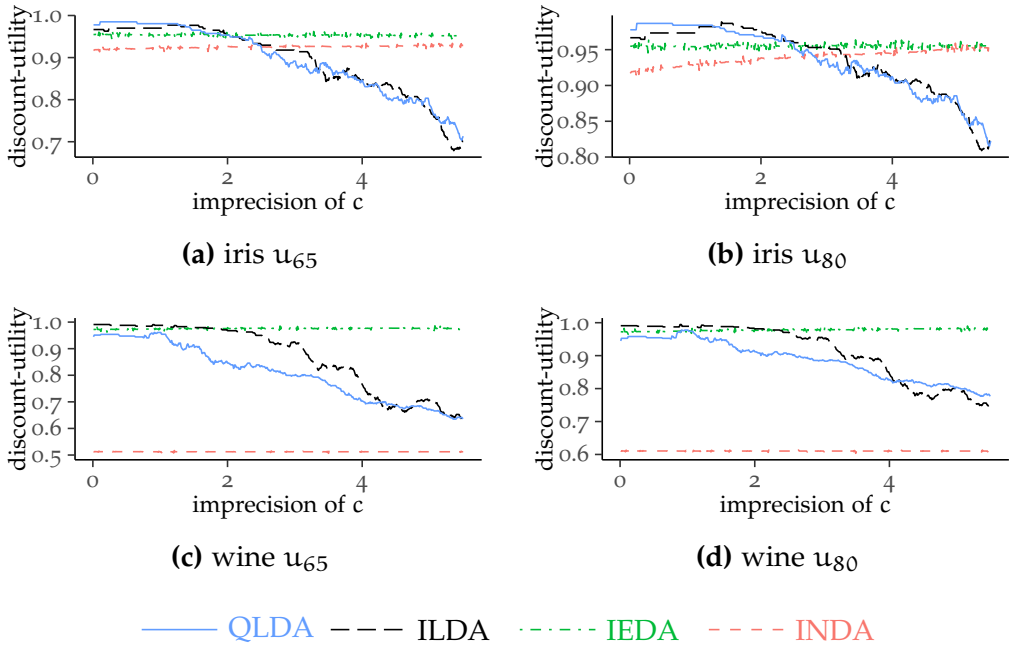
Finally, a last open issue is how we can use or extend the existing heuristics of probabilistic classifier approaches on our proposal strategies, such as epsilon-approximate inference, A^* and beam search methods [Mena et al., 2016; Kumar et al., 2013]. These heuristics explore multiple path in the tree, so it might be interesting to extend to an imprecise probabilistic setting.

Appendix A

Complementary experimental results of IGDA model

A.1 Performance evolution w.r.t. utility-discount and c parameter

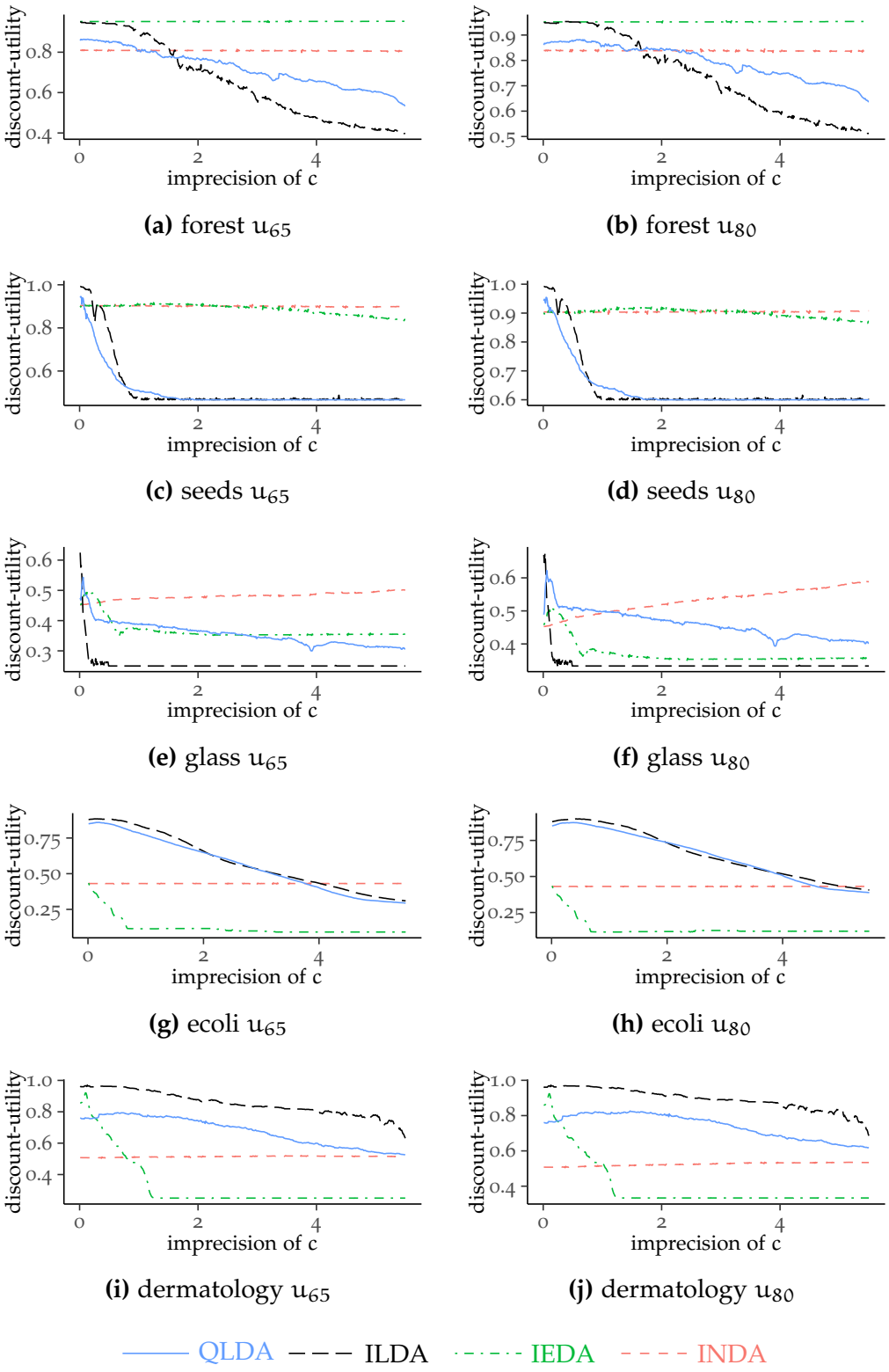
Complementary experimental results are shown in the Figure A.1, A.2 and A.3



(a) iris u_{65} **(b)** iris u_{80}

(c) wine u_{65} **(d)** wine u_{80}

Figure A.1: Experiments for IGDA model (left:utility-discount u_{65}, right:utility-discount u_{80})



(a) forest u_{65} **(b)** forest u_{80}

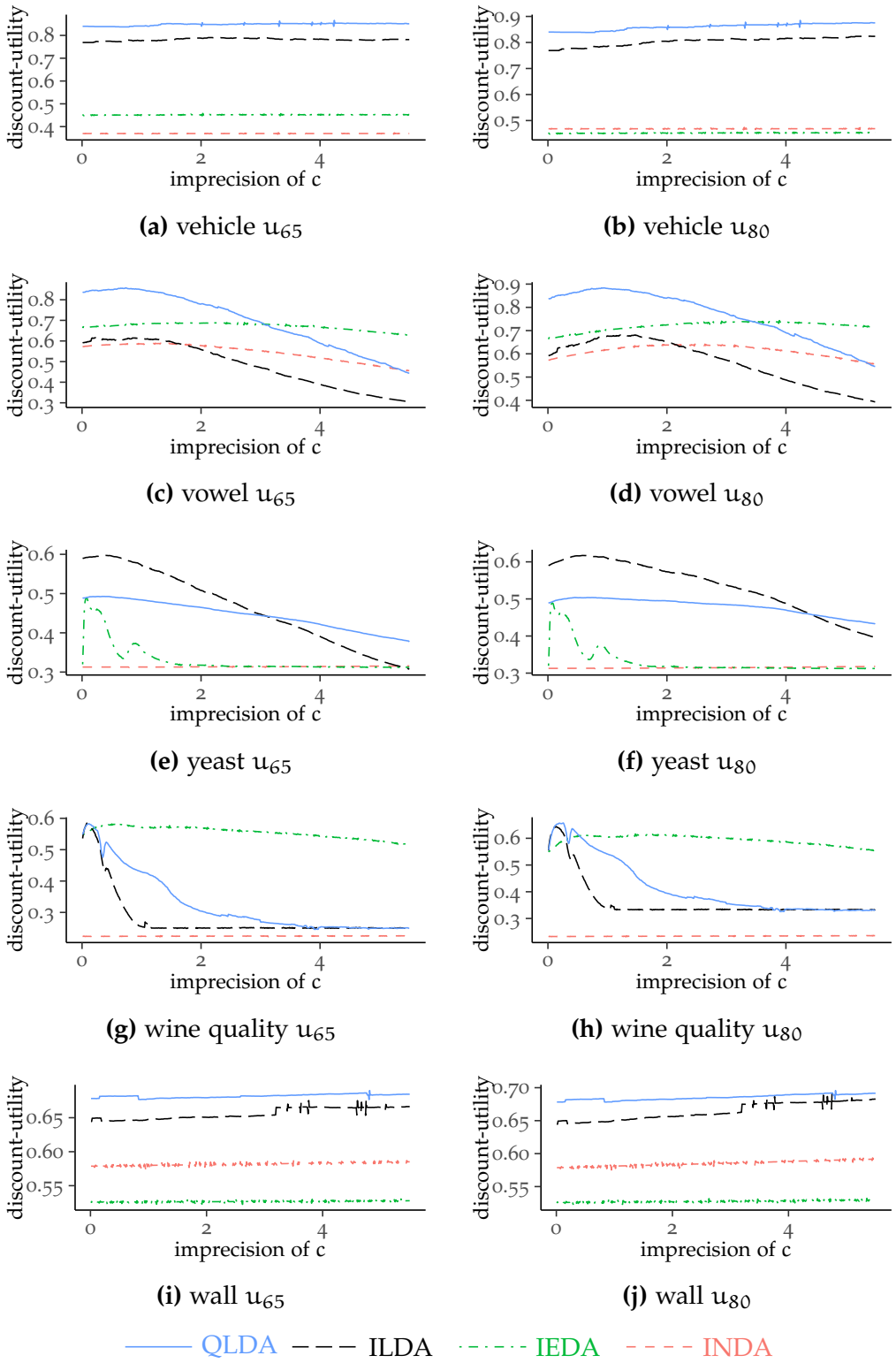
(c) seeds u_{65} **(d)** seeds u_{80}

(e) glass u_{65} **(f)** glass u_{80}

(g) ecoli u_{65} **(h)** ecoli u_{80}

(i) dermatology u_{65} **(j)** dermatology u_{80}

Figure A.2: Experiments for IGDA model (left:utility-discount u_{65}, right:utility-discount u_{80})



(a) vehicle u_{65} **(b)** vehicle u_{80}

(c) vowel u_{65} **(d)** vowel u_{80}

(e) yeast u_{65} **(f)** yeast u_{80}

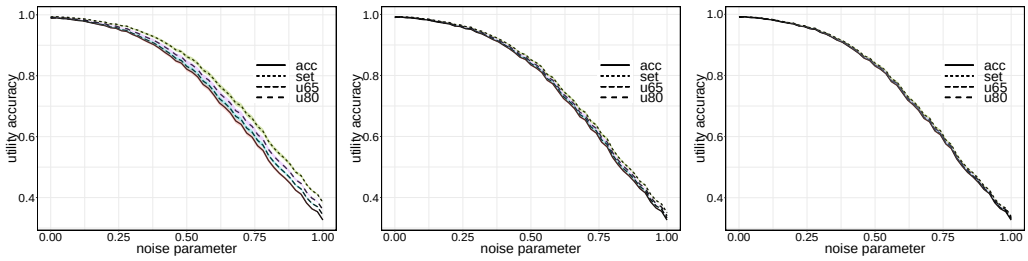
(g) wine quality u_{65} **(h)** wine quality u_{80}

(i) wall u_{65} **(j)** wall u_{80}

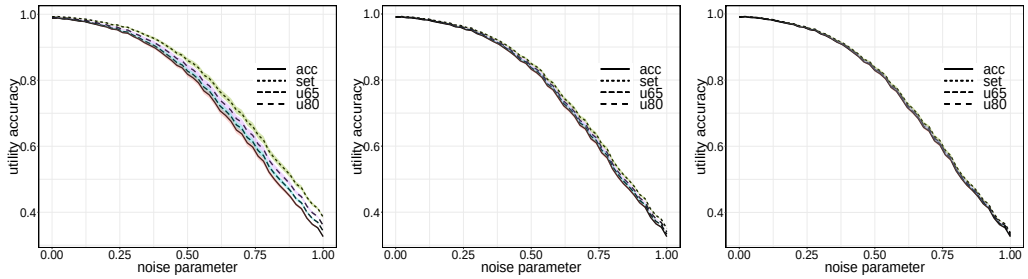
Figure A.3: Experiments for IGDA model (left:utility-discount u_{65}, right:utility-discount u_{80})

A.2 COMPLEMENTARY EXPERIMENTS RESULTS ON DISTURBED SYNTHETIC TEST DATA

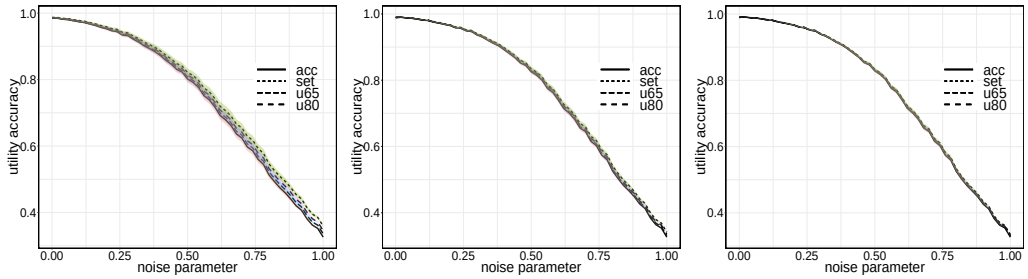
Complementary experimental results using the ϵ noise parameter to corrupt test instances are shown in the Figure A.4, A.5, A.6 and A.7.



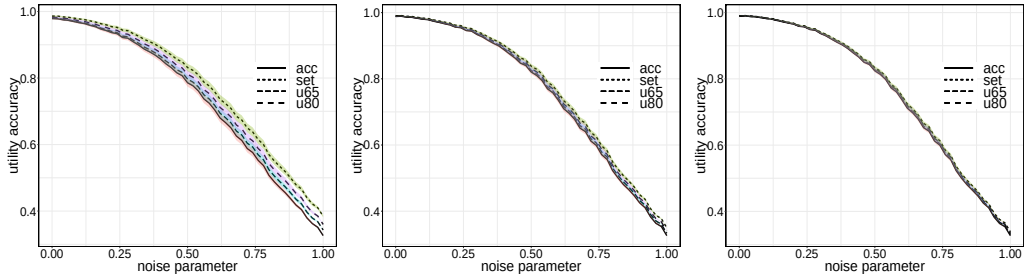
(a) (Imprecise) Euclidian discriminant analysis



(b) (Imprecise) Linear discriminant analysis

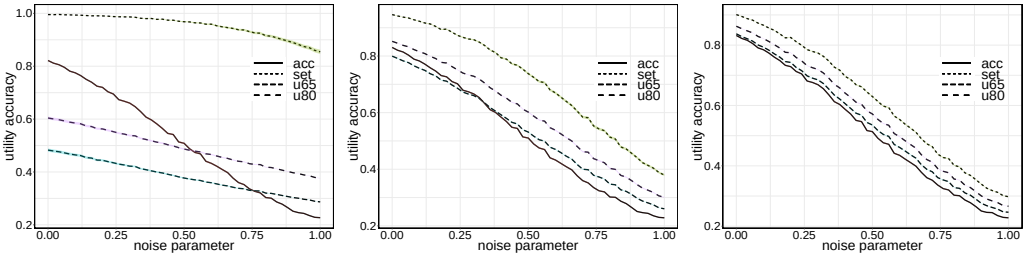


(c) (Imprecise) Naive discriminant analysis

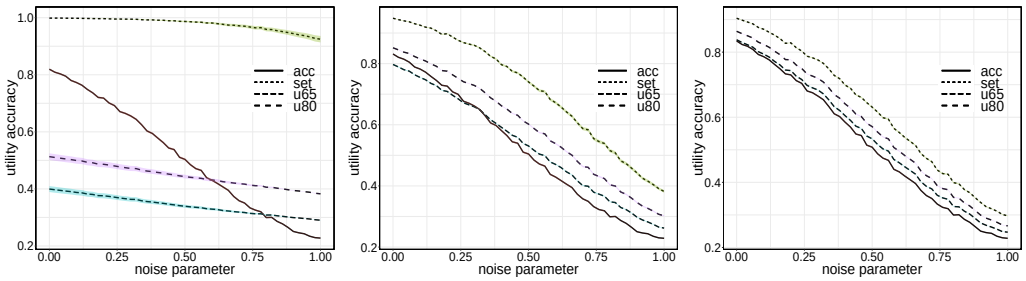


(d) (Imprecise) Quadratic discriminant analysis

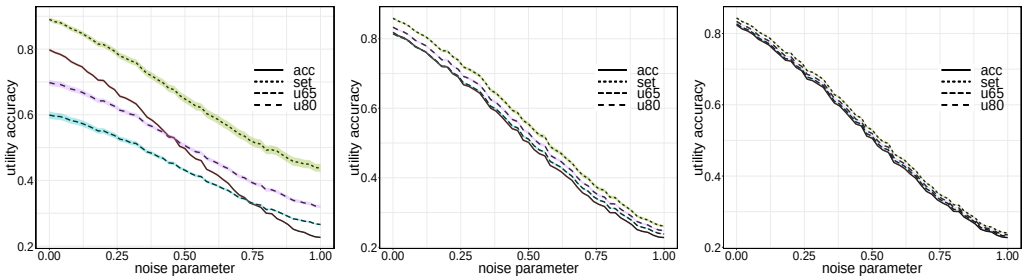
Figure A.4: Utility accuracies (%) with confidence intervals on corrupt test data sets $\mathbb{T}_1^{\epsilon}$. The first column ($\mathbb{D}_1^{10}$), the second column ($\mathbb{D}_1^{25}$) and the third column ($\mathbb{D}_1^{50}$). In each row a different Gaussian classifier model is fitted.



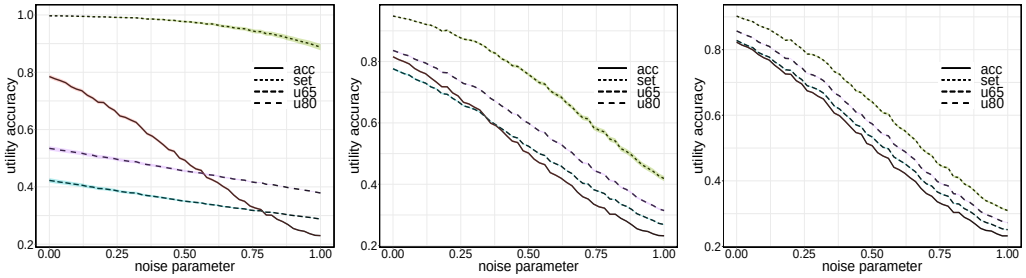
(a) (Imprecise) Euclidian discriminant analysis



(b) (Imprecise) Linear discriminant analysis

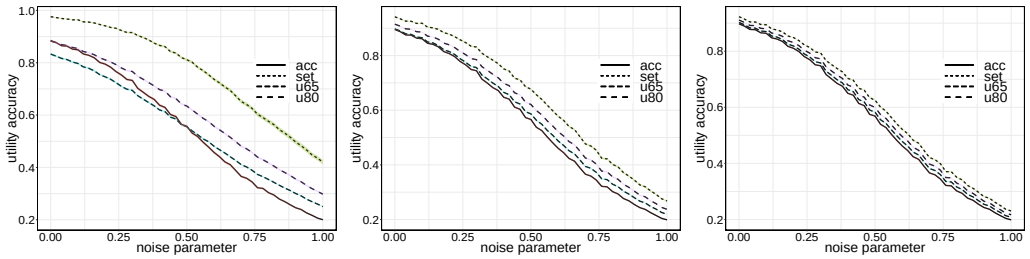


(c) (Imprecise) Naive discriminant analysis

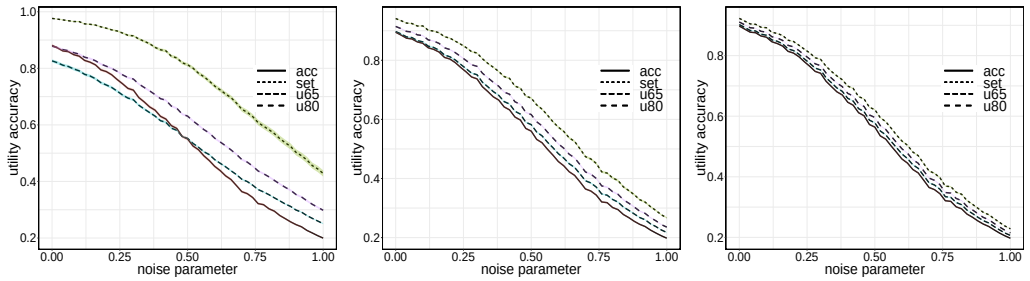


(d) (Imprecise) Quadratic discriminant analysis

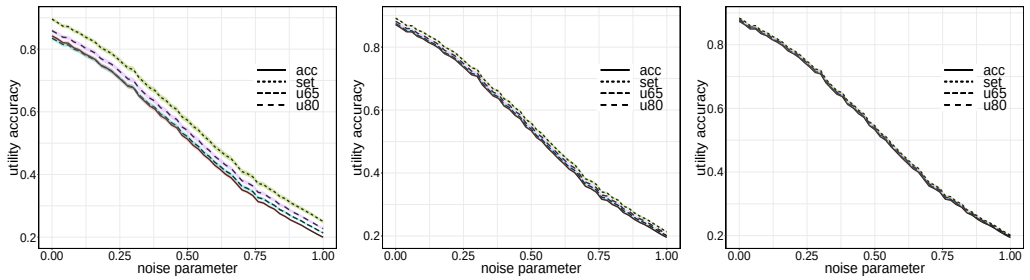
Figure A.5: Utility accuracies (%) with confidence intervals on corrupt test data sets $\mathbb{T}_2^\epsilon$. The first column ($\mathbb{D}_2^{10}$), the second column ($\mathbb{D}_2^{25}$) and the third column ($\mathbb{D}_2^{50}$). In each row a different Gaussian classifier model is fitted.



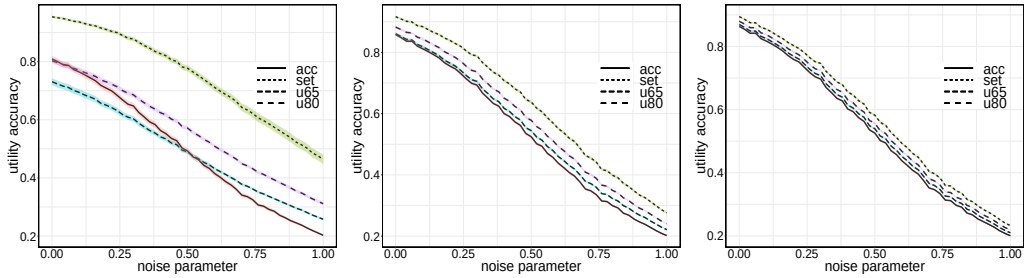
(a) (Imprecise) Euclidian discriminant analysis



(b) (Imprecise) Linear discriminant analysis

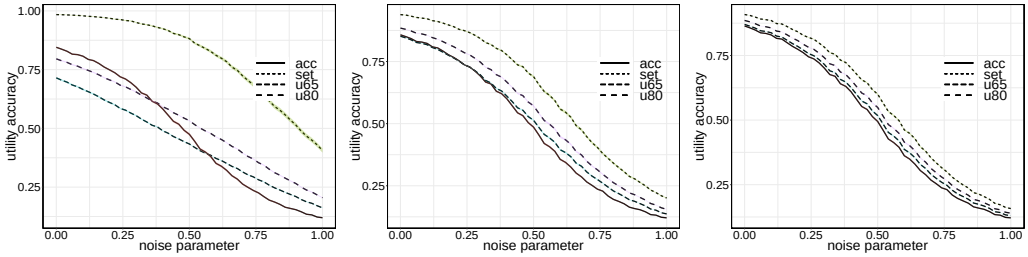


(c) (Imprecise) Naive discriminant analysis

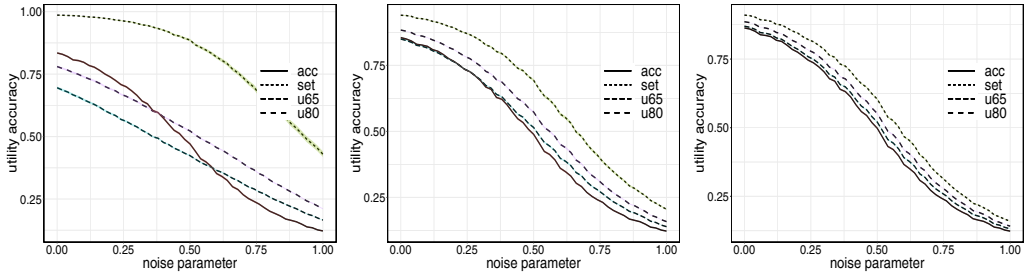


(d) (Imprecise) Quadratic discriminant analysis

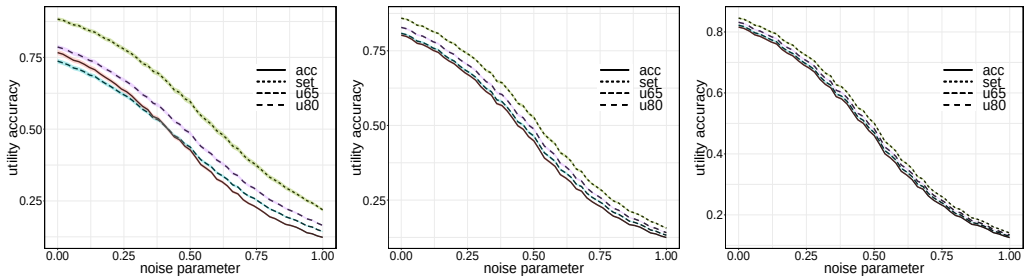
Figure A.6: Utility accuracies (%) with confidence intervals on corrupt test data sets $\mathbb{T}_3^\epsilon$. The first column ($\mathbb{D}_3^{10}$), the second column ($\mathbb{D}_3^{25}$) and the third column ($\mathbb{D}_3^{50}$). In each row a different Gaussian classifier model is fitted.



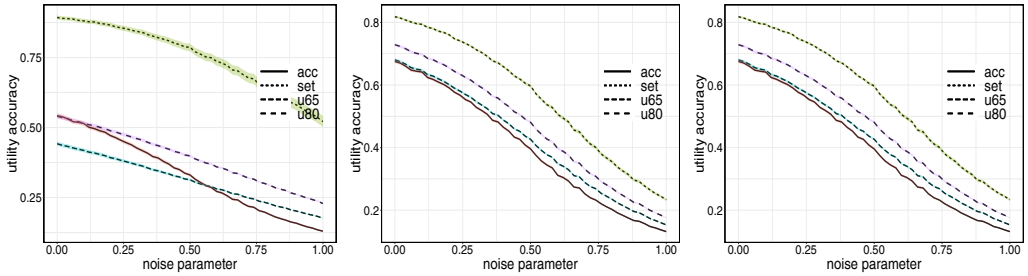
(a) (Imprecise) Euclidian discriminant analysis



(b) (Imprecise) Linear discriminant analysis



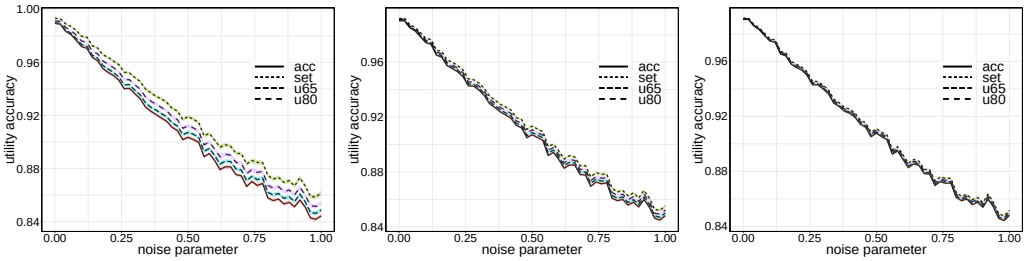
(c) (Imprecise) Naive discriminant analysis



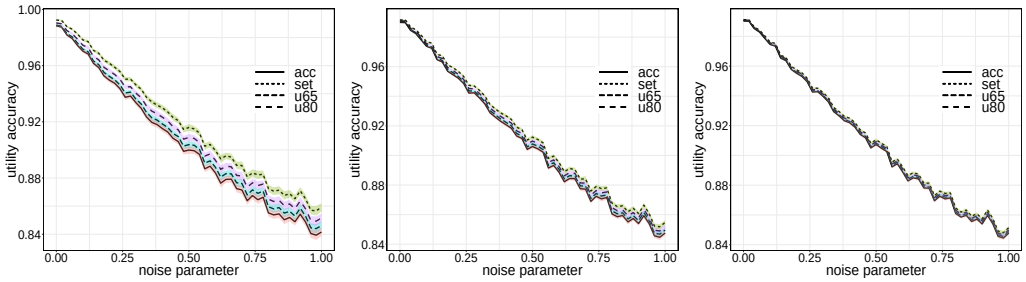
(d) (Imprecise) Quadratic discriminant analysis

Figure A.7: Utility accuracies (%) with confidence intervals on corrupt test data sets $\mathbb{T}_4^\epsilon$. The first column ($\mathbb{D}_4^{10}$), the second column ($\mathbb{D}_4^{25}$) and the third column ($\mathbb{D}_4^{50}$). In each row a different Gaussian classifier model is fitted.

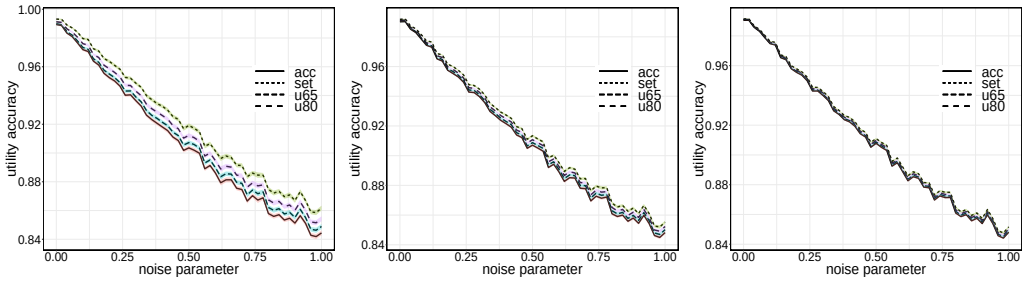
Complementary experimental results using the ψ noise parameter to corrupt test instances are shown in theFigure A.8, A.9, A.10 and A.11.



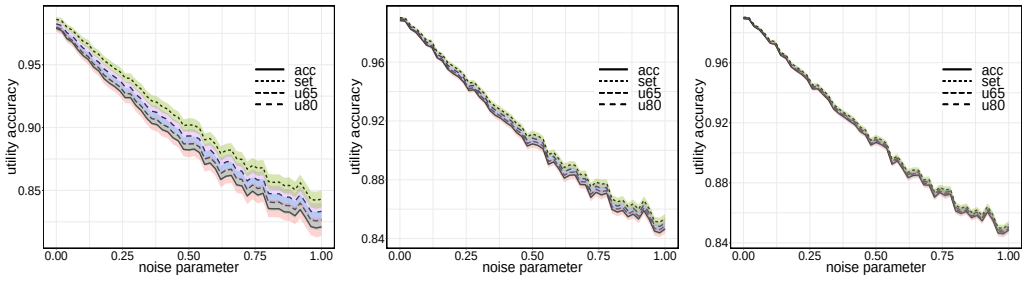
(a) (Imprecise) Euclidian discriminant analysis



(b) (Imprecise) Linear discriminant analysis

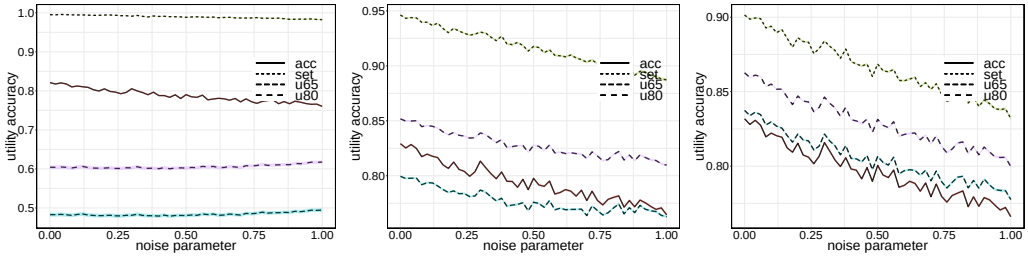


(c) (Imprecise) Naive discriminant analysis

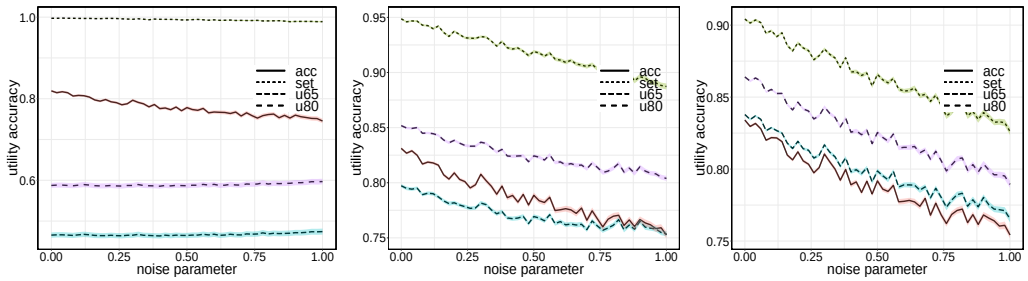


(d) (Imprecise) Quadratic discriminant analysis

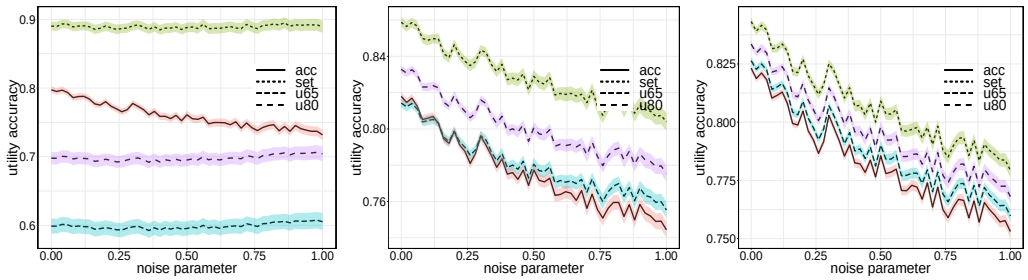
Figure A.8: Utility accuracies (%) with confidence intervals on corrupt test data sets $\mathbb{T}_1^{\psi}$. The first column ($\mathbb{D}_1^{10}$), the second column ($\mathbb{D}_1^{25}$) and the third column ($\mathbb{D}_1^{50}$). In each row a different Gaussian classifier model is fitted.



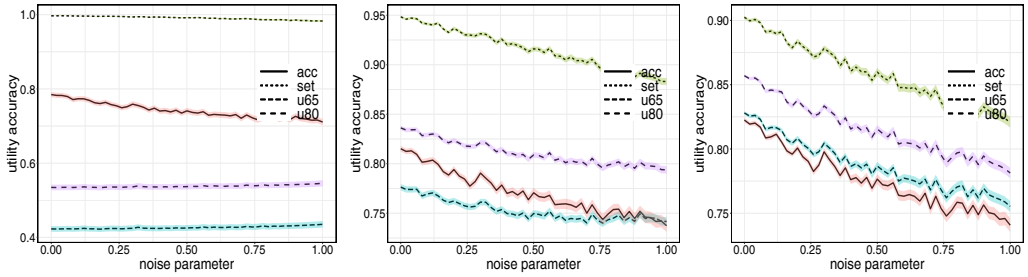
(a) (Imprecise) Euclidian discriminant analysis



(b) (Imprecise) Linear discriminant analysis

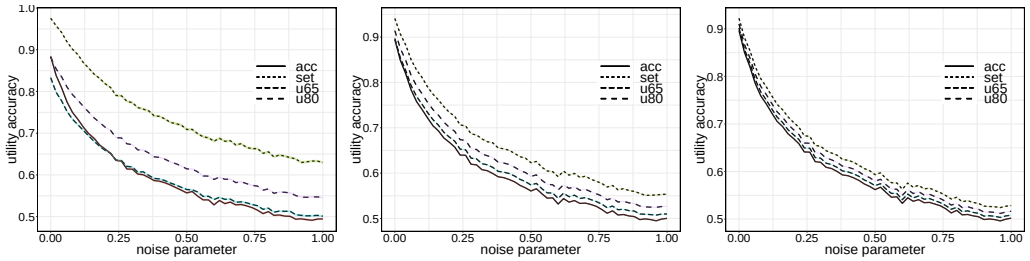


(c) (Imprecise) Naive discriminant analysis

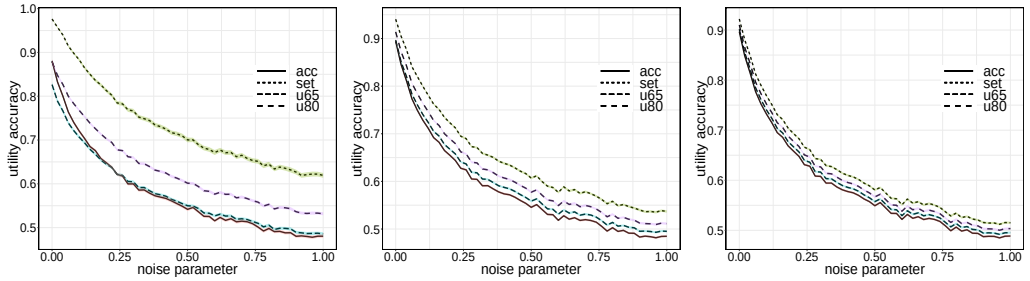


(d) (Imprecise) Quadratic discriminant analysis

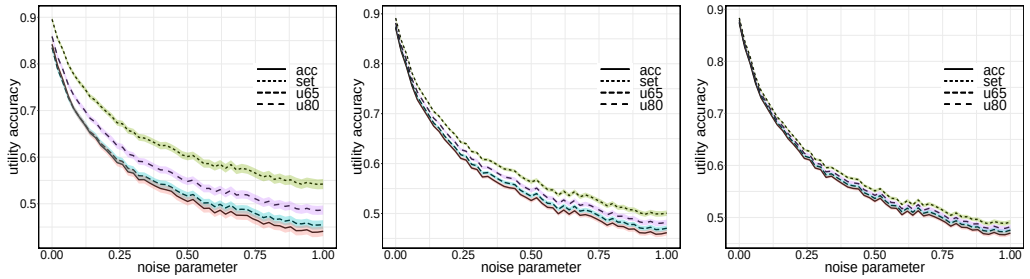
Figure A.9: Utility accuracies (%) with confidence intervals on corrupt test data sets $\mathbb{T}_2^{\psi}$. The first column ($\mathbb{D}_2^{10}$), the second column ($\mathbb{D}_2^{25}$) and the third column ($\mathbb{D}_2^{50}$). In each row a different Gaussian classifier model is fitted.



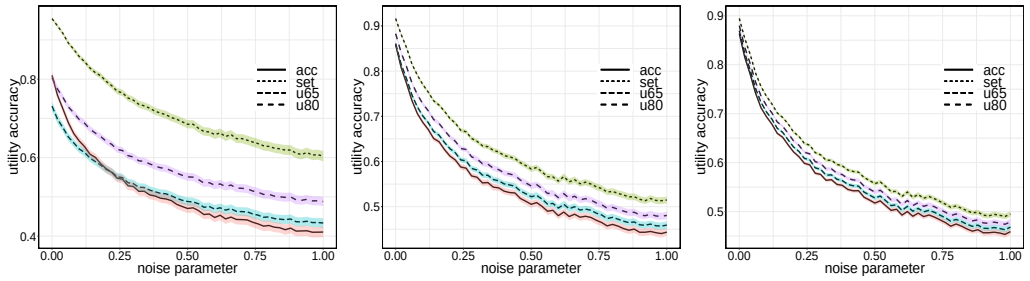
(a) (Imprecise) Euclidian discriminant analysis



(b) (Imprecise) Linear discriminant analysis

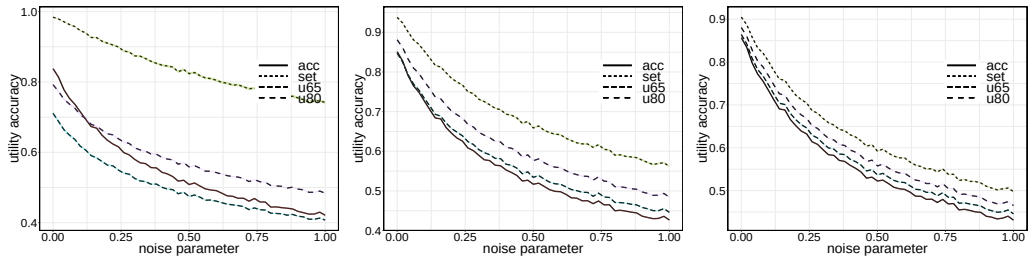


(c) (Imprecise) Naive discriminant analysis

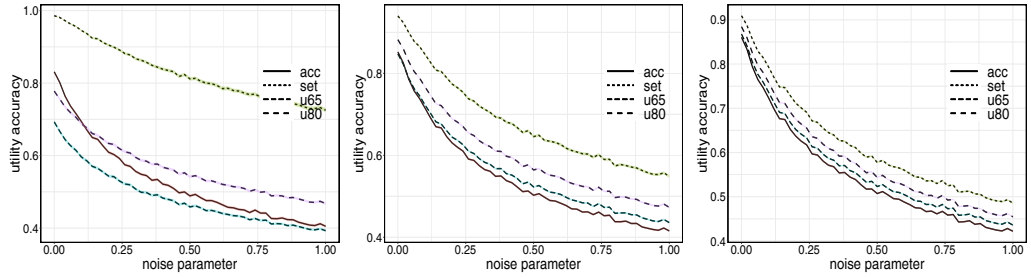


(d) (Imprecise) Quadratic discriminant analysis

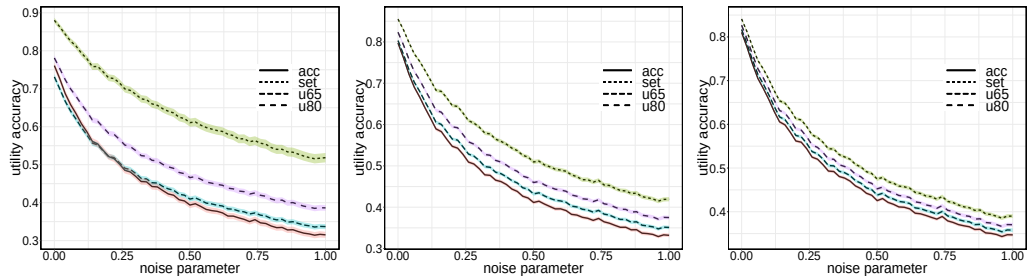
Figure A.10: Utility accuracies (%) with confidence intervals on corrupt test data sets $\mathbb{T}_3^{\psi}$. The first column ($\mathbb{D}_3^{10}$), the second column ($\mathbb{D}_3^{25}$) and the third column ($\mathbb{D}_3^{50}$). In each row a different Gaussian classifier model is fitted.



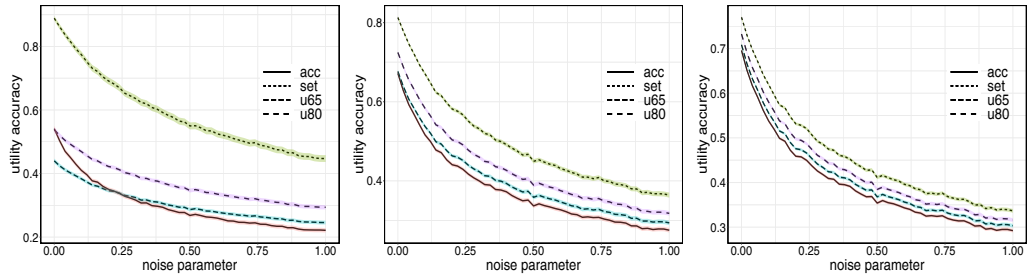
(a) (Imprecise) Euclidian discriminant analysis



(b) (Imprecise) Linear discriminant analysis



(c) (Imprecise) Naive discriminant analysis



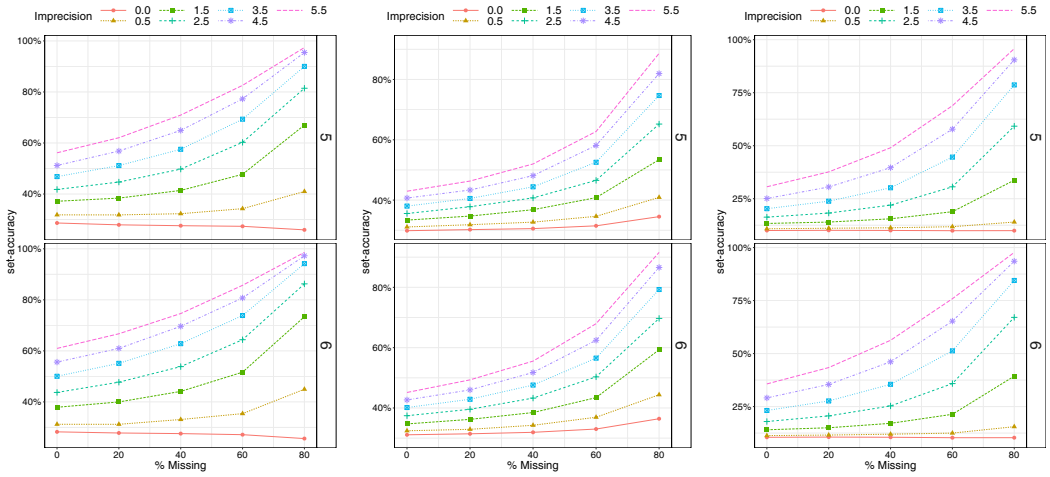
(d) (Imprecise) Quadratic discriminant analysis

Figure A.11: Utility accuracies (%) with confidence intervals on corrupt test data sets $\mathbb{T}_4^{\psi}$. The first column ($\mathbb{D}_4^{10}$), the second column ($\mathbb{D}_4^{25}$) and the third column ($\mathbb{D}_4^{50}$). In each row a different Gaussian classifier model is fitted.

Appendix B

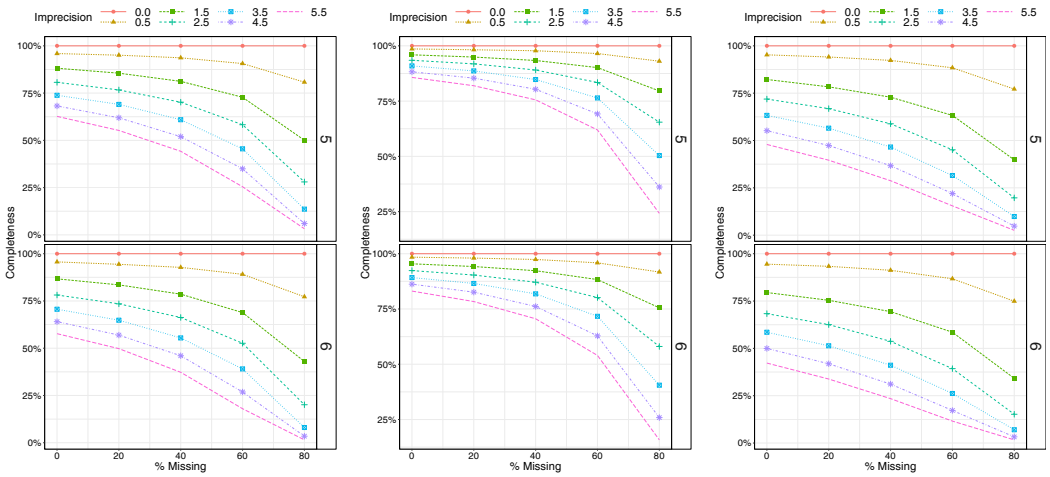
Complementary experimental

B.1 Missing precise



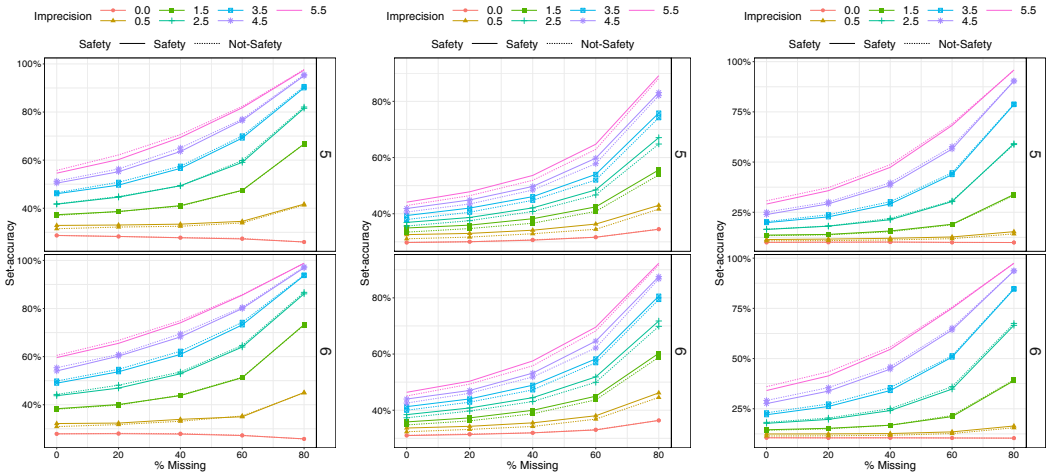
(a) Emotions **(b)** Scene **(c)** Yeast

Figure B.1: **Missing labels - Marginalization** Evolution of the average set-accuracy (%) for each level of imprecision (a curve for each one) and discretization $z = 5$ (top) and $z = 6$ (down), with respect the the percentage of missing labels.



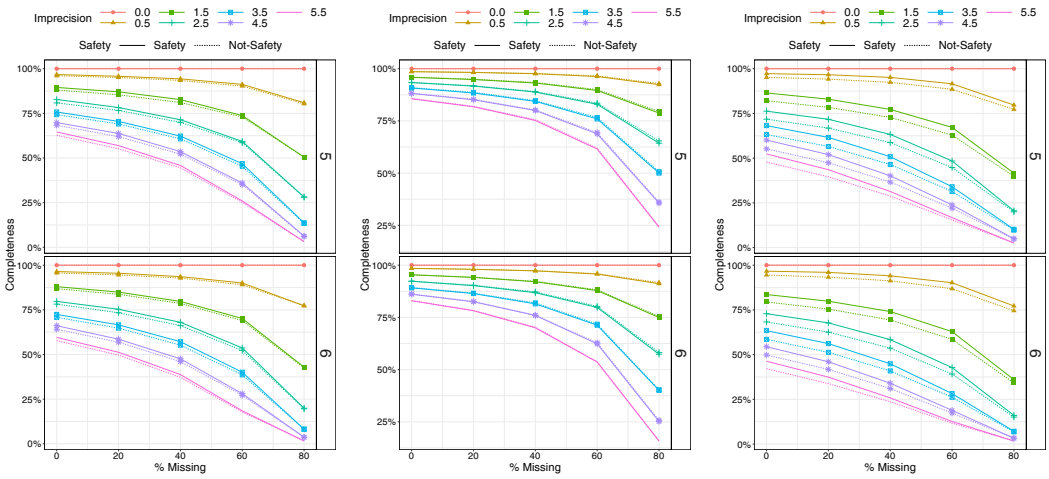
(a) Emotions **(b)** Scene **(c)** Yeast

Figure B.2: **Missing labels - Marginalization** Evolution of the average completeness (%) for each level of imprecision (a curve for each one) and discretization $z = 5$ (top) and $z = 6$ (down), with respect the the percentage of missing labels.



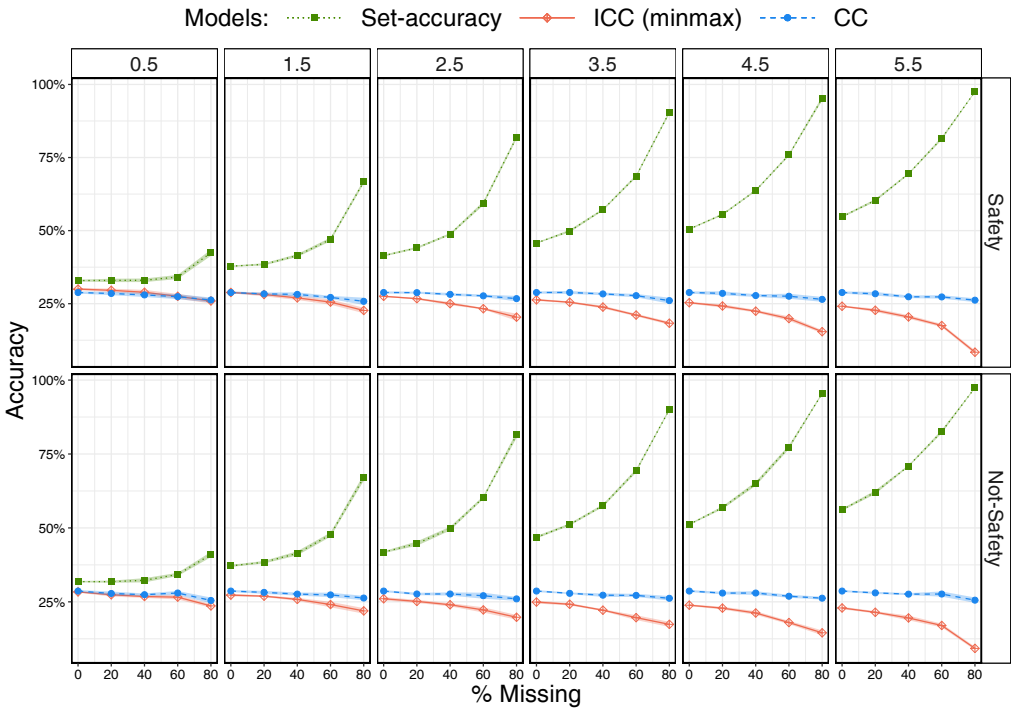
(a) EMOTIONS **(b)** SCENE **(c)** YEAST

Figure B.3: **Missing labels - Marginalization - Safety imprecise** Evolution of the average set-accuracy (%) for each level of imprecision (a different shape point and color for each one), and **safety imprecise chaining** in dotted line and **not-safety one** in solid line, and discretization $z = 5$ (top) and $z = 6$ (down), with respect the the percentage of missing labels.

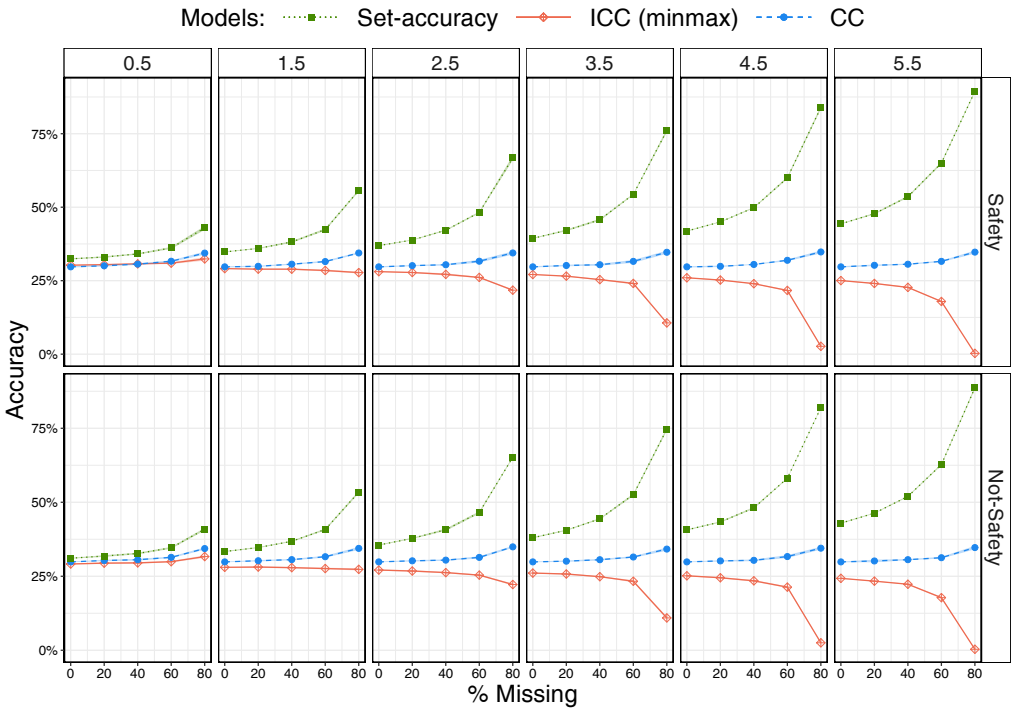


(a) EMOTIONS **(b)** SCENE **(c)** YEAST

Figure B.4: **Missing labels - Marginalization - Safety imprecise** Evolution of the average set-accuracy (%) for each level of imprecision (a different shape point and color for each one), and **safety imprecise chaining** in dotted line and **not-safety one** in solid line, and discretization $z = 5$ (top) and $z = 6$ (down), with respect the the percentage of missing labels.

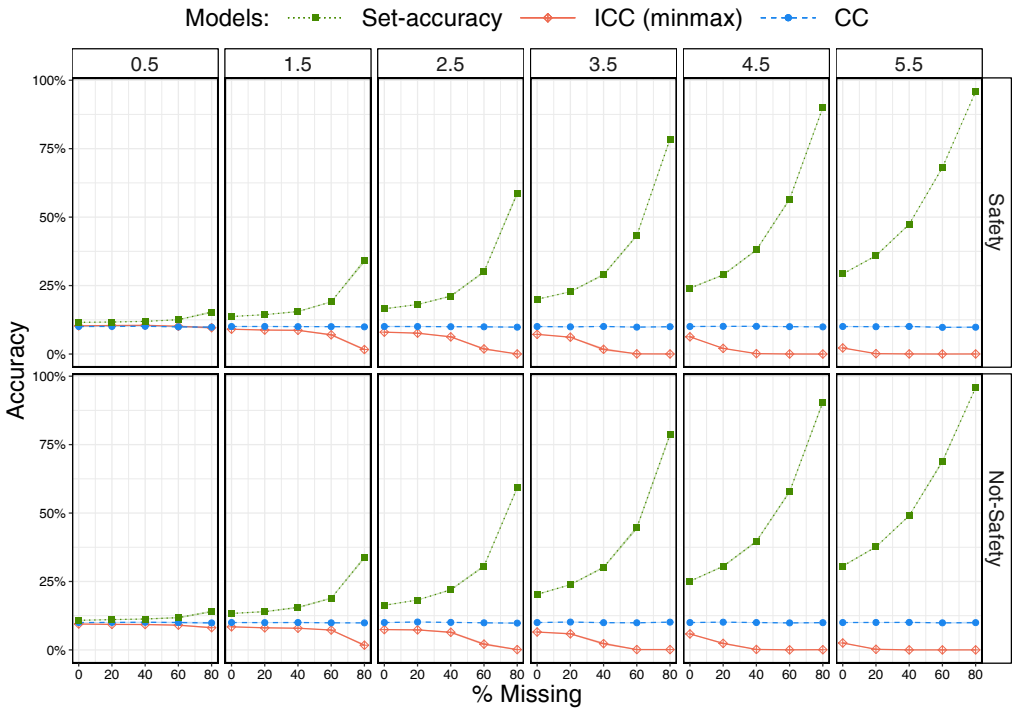


(a) EMOTIONS



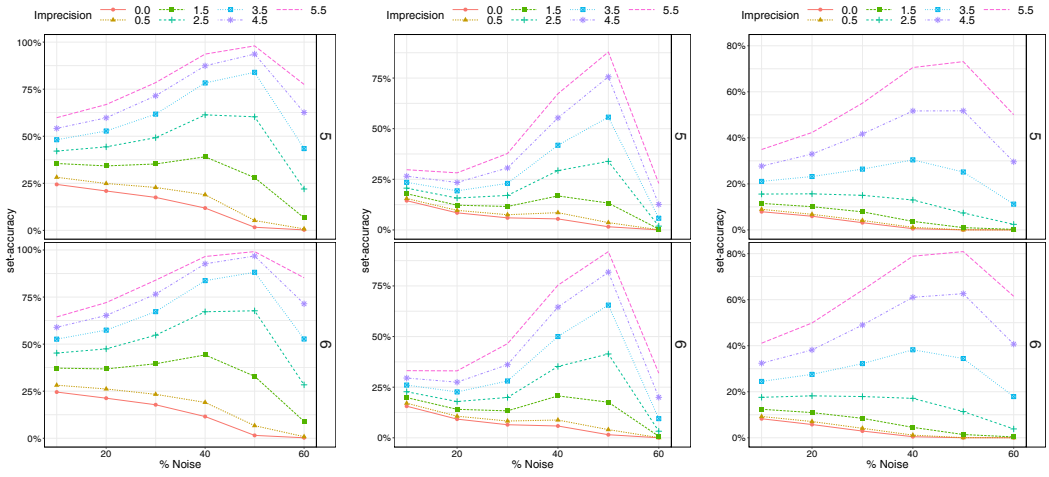
(b) SCENE

Figure B.5: **Missing - ICC versus CC - Marginalization.** Figures show performance evolution (%) of the imprecise and precise classifier-chains approaches for all levels of imprecision and using the **safety imprecise chaining** (top) and not (down), with respect to the percentage of missing (x-axis).



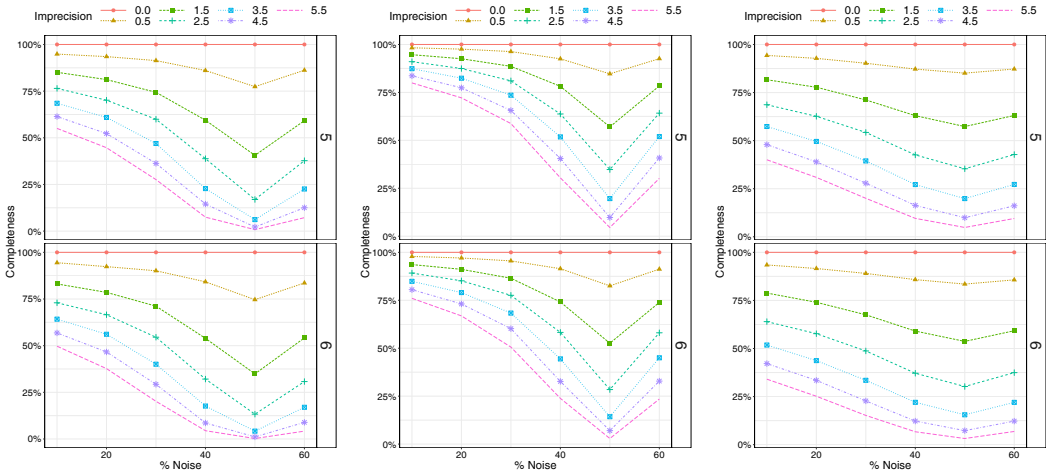
(a) YEAST

Figure B.6: **Missing - ICC versus CC - Marginalization.** continuation of Figure B.5



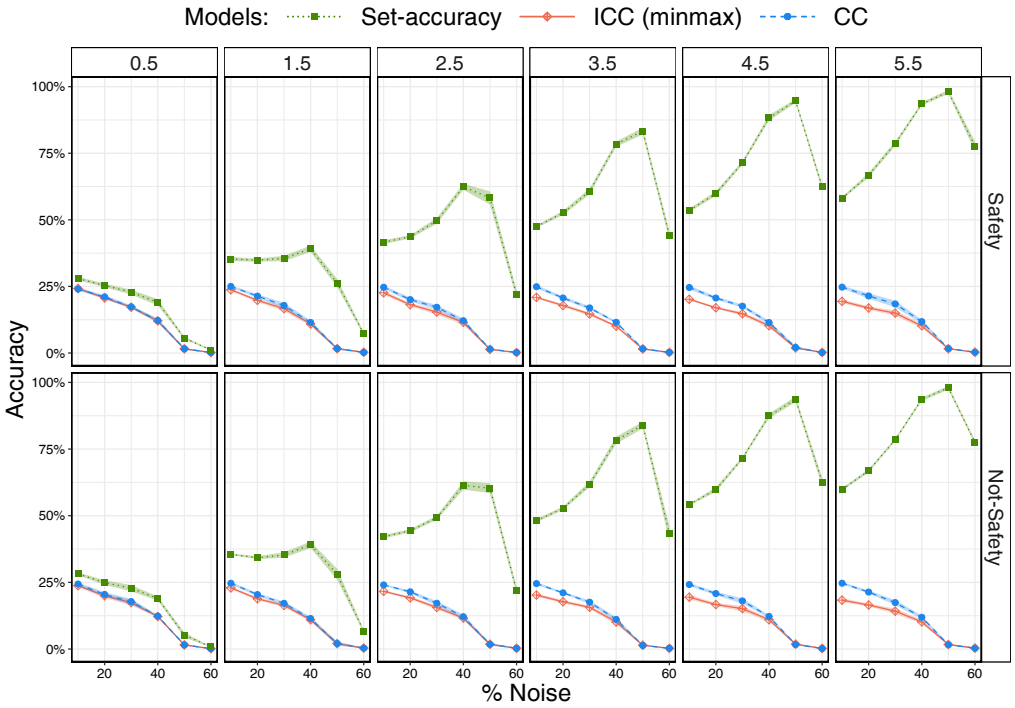
(a) EMOTIONS **(b)** SCENE **(c)** YEAST

Figure B.7: **Reversing - Marginalization** Evolution of the average set-accuracy (%) for each level of imprecision (a curve for each one) and two levels of discretization $z = 5$ (top) and $z = 6$ (down), with respect to the percentage of noise



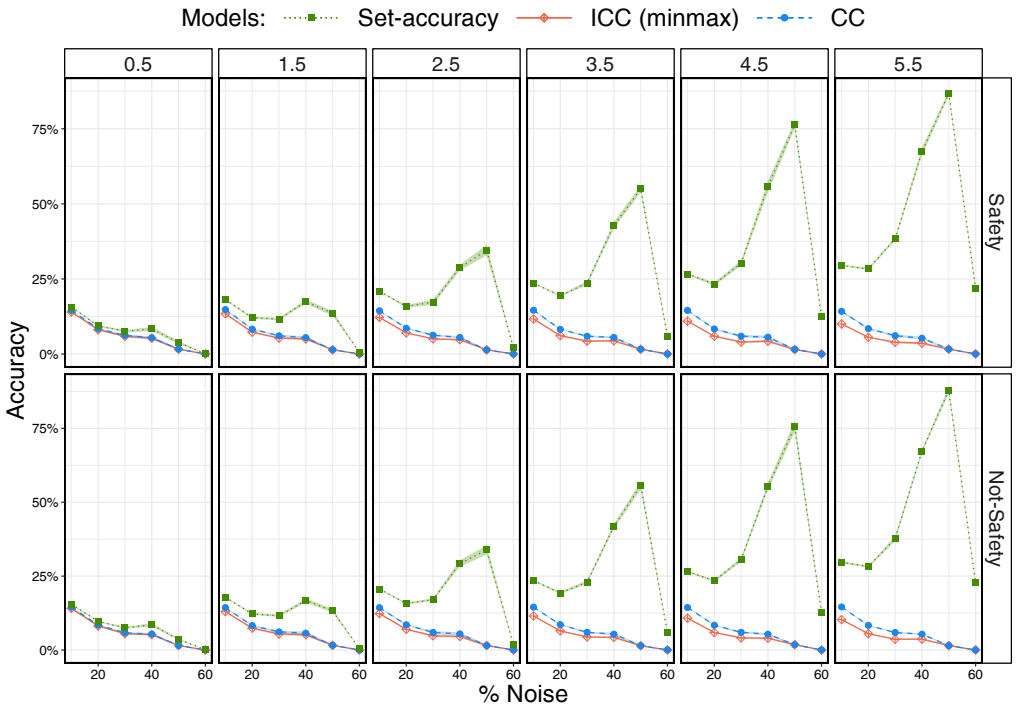
(a) EMOTIONS **(b)** SCENE **(c)** YEAST

Figure B.8: **Reversing - Marginalization** Evolution of the average completeness (%) for each level of imprecision (a curve for each one) and two levels of discretization $z = 5$ (top) and $z = 6$ (down), with respect to the percentage of noise

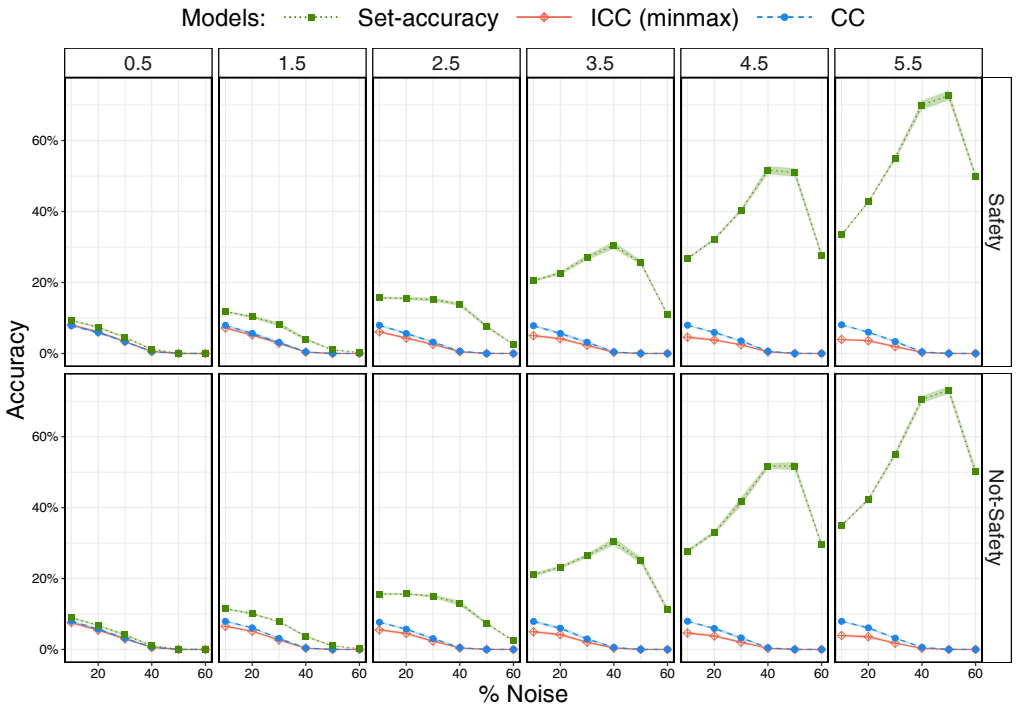


(a) EMOTIONS

Figure B.9: **Reversing - ICC versus CC - Marginalization.** Figures show performance evolution (%) of the imprecise and precise classifier-chains approaches for all levels of imprecision and using the **safety imprecise chaining** (top) and not (down), with respect to the percentage of noiseness (x-axis).



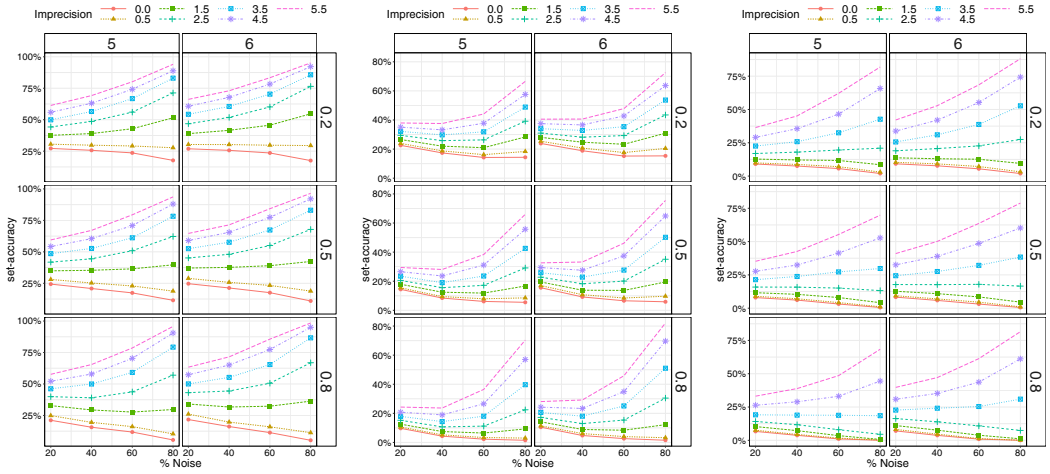
(a) SCENE



(b) YEAST

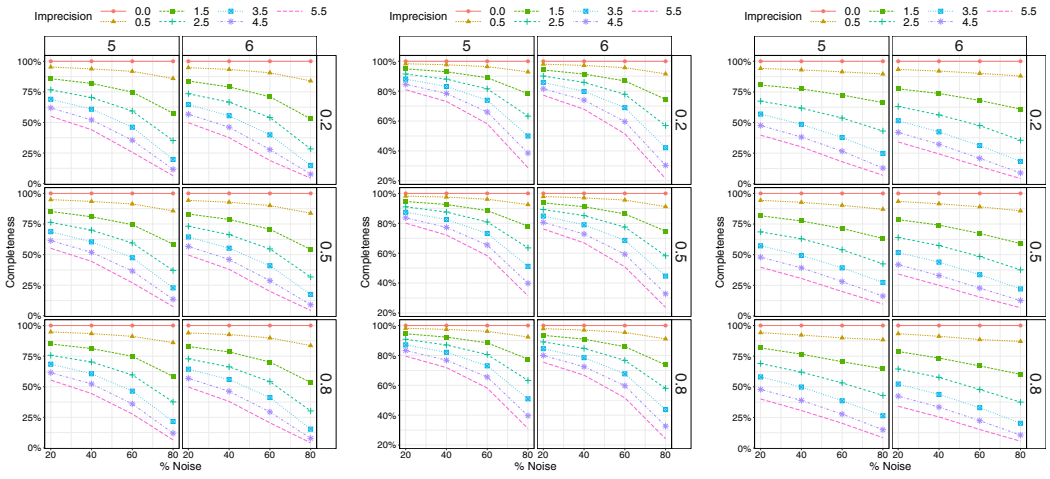
Figure B.10: **Missing - ICC versus CC - Marginalization.** continuation of Figure B.9

B.3 NOISY FLIPPING



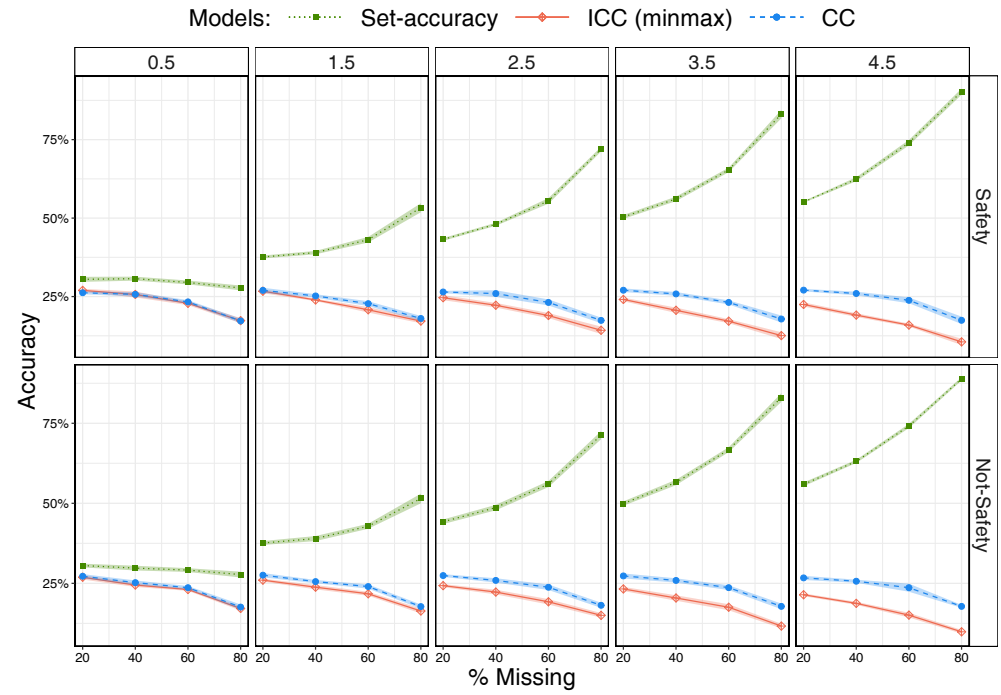
(a) EMOTIONS **(b)** SCENE **(c)** YEAST

Figure B.11: **Flipping - Marginalization** Evolution of the average set-accuracy (%) for each level of imprecision (a curve for each one) and two levels of discretization $z = 5$ (top) and $z = 6$ (down), with respect to the percentage of noise.

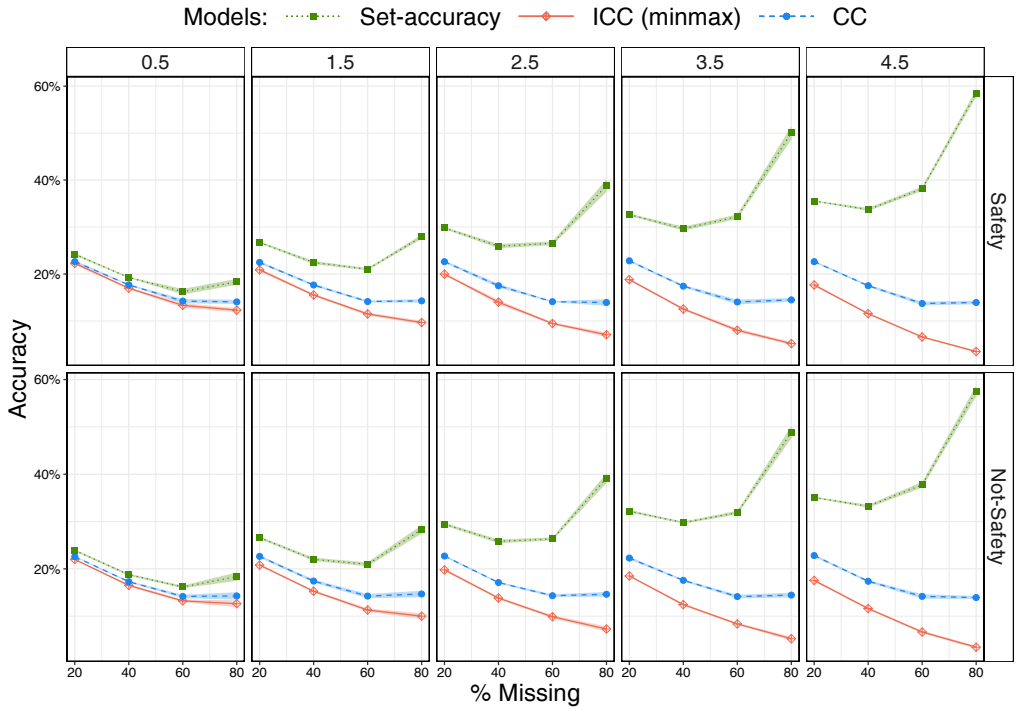


(a) EMOTIONS **(b)** SCENE **(c)** YEAST

Figure B.12: **Flipping - Marginalization** Evolution of the average completeness (%) for each level of imprecision (a curve for each one) and two levels of discretization $z = 5$ (top) and $z = 6$ (down), with respect to the percentage of noise.

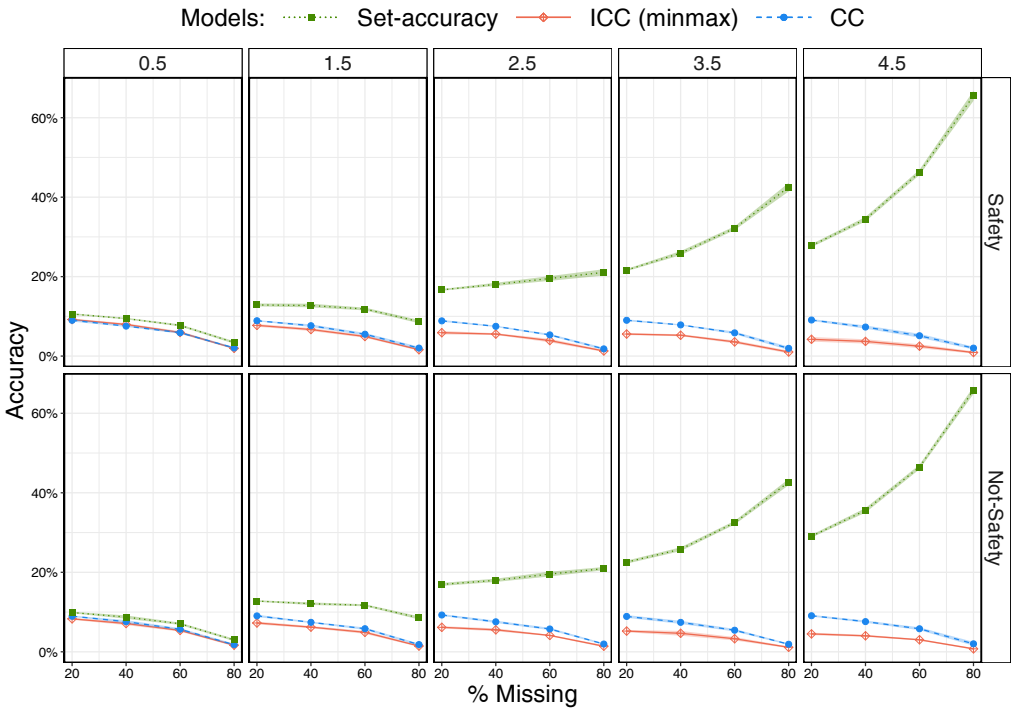


(a) EMOTIONS



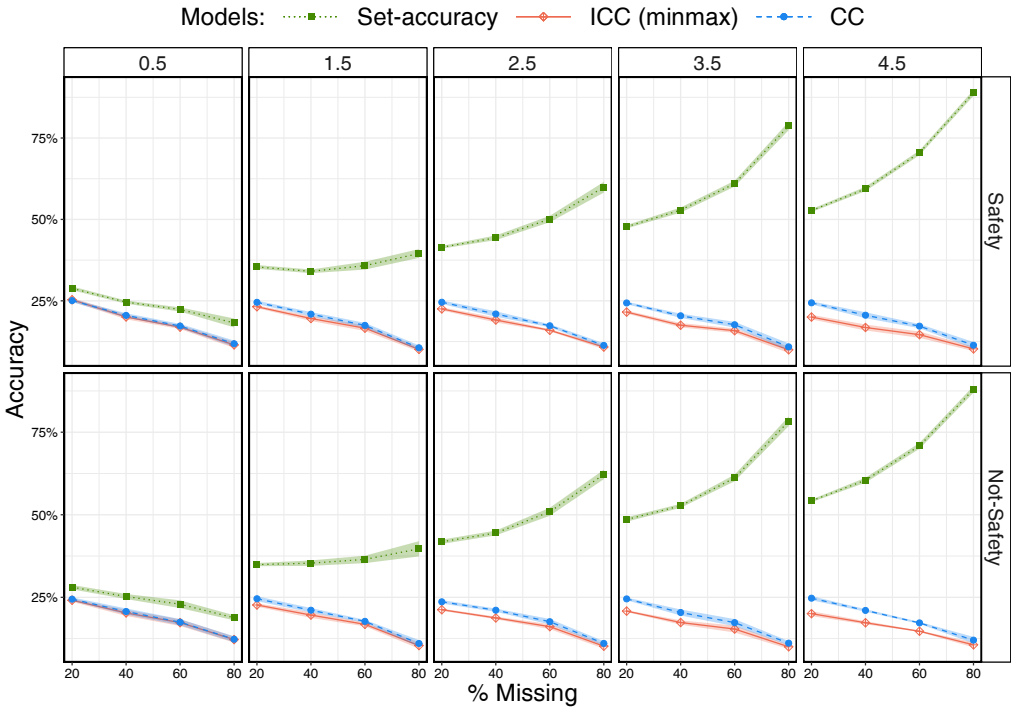
(b) SCENE

Figure B.13: **Flipping - ICC versus CC - Marginalization.** Figures show performance evolution (%) of the imprecise and precise classifier-chains approaches for 0.5 (left) and 1.5 (right) levels of imprecisions, and $\beta = 0.2$, and using the **safety imprecise chaining** (top) and not (down), with respect to the percentage of noisy (x-axis).



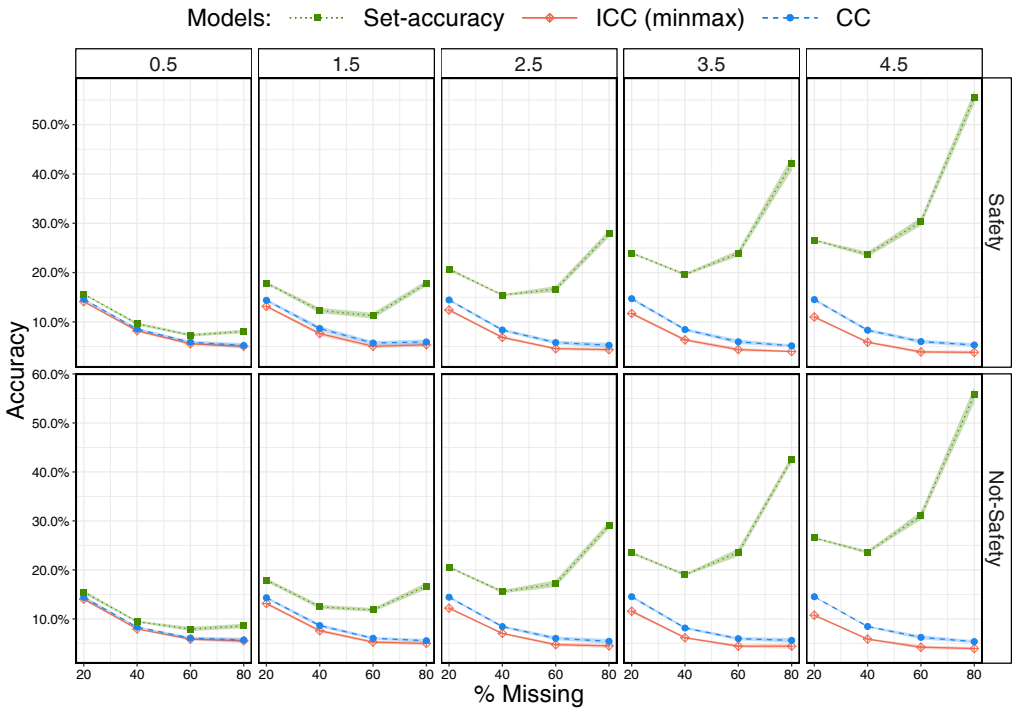
(a) YEAST

Figure B.14: **Flipping - ICC versus CC - Marginalization.** continuation of Figure B.13

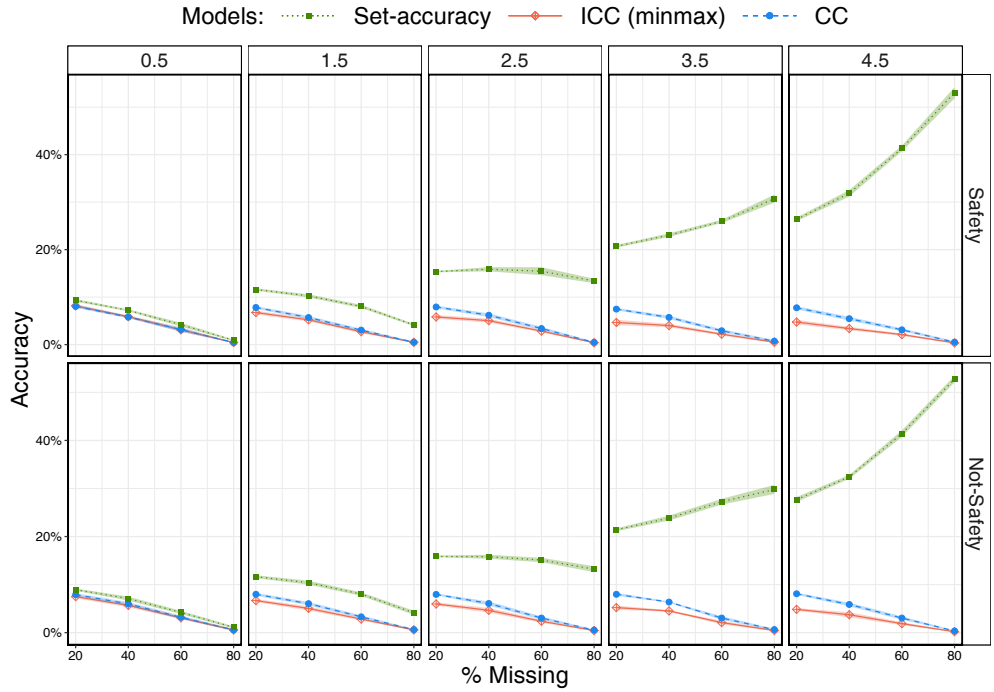


(a) EMOTIONS

Figure B.15: **Flipping - ICC versus CC - Marginalization** Figures show performance evolution (%) of the imprecise and precise classifier-chains approaches for 0.5 (left) and 1.5 (right) levels of imprecisions, and $\beta = 0.5$, and using the **safety imprecise chaining** (top) and not (down), with respect to the percentage of noisy (x-axis).

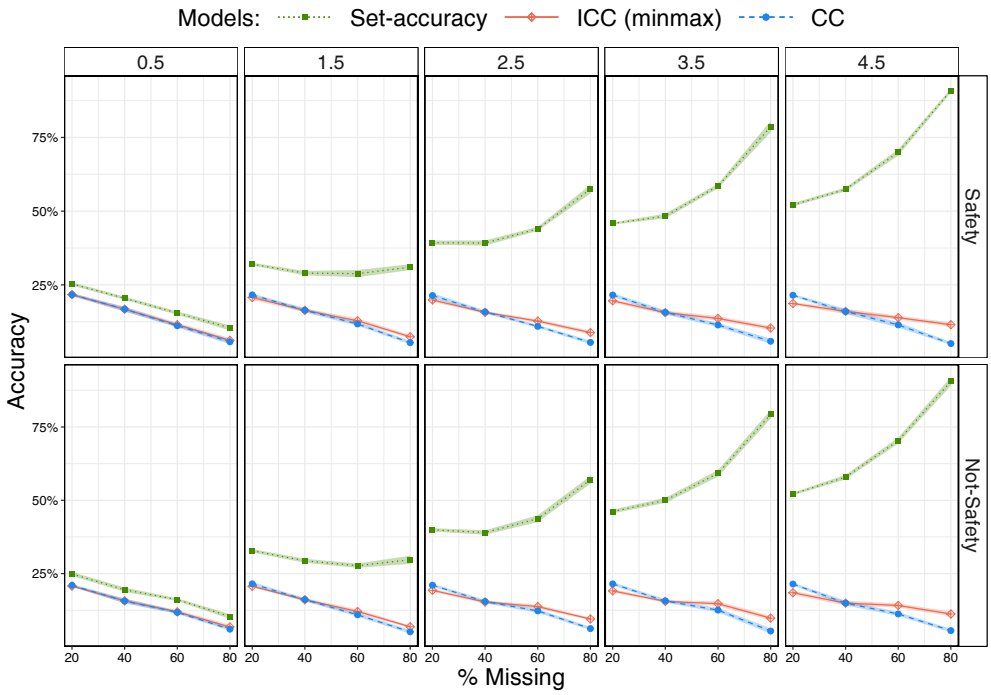


(a) SCENE

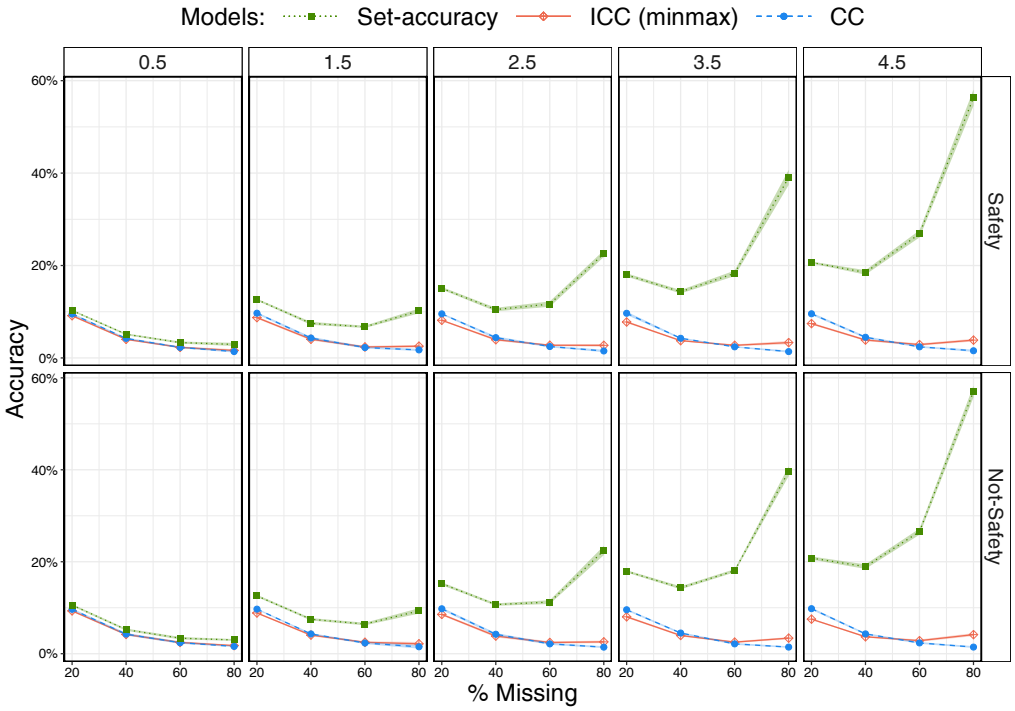


(b) YEAST

Figure B.16: **Flipping - ICC versus CC - Marginalization.** continuation of Figure B.15

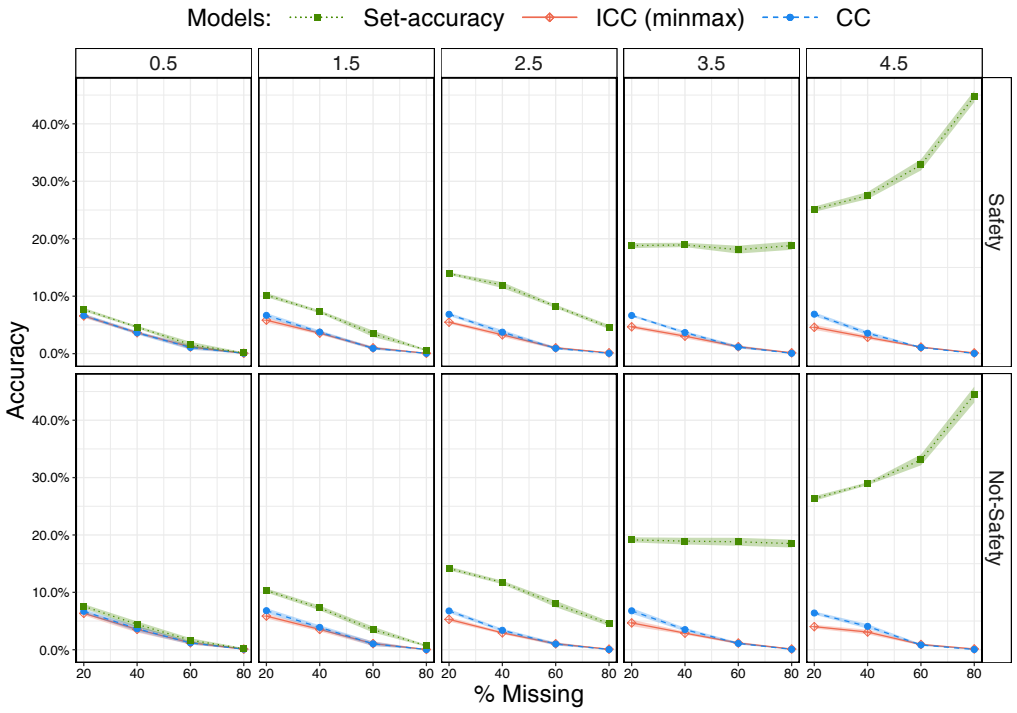


(a) EMOTIONS



(b) SCENE

Figure B.17: **Flipping - ICC versus CC - Marginalization.** Figures show performance evolution (%) of the imprecise and precise classifier-chains approaches for 0.5 (left) and 1.5 (right) levels of imprecisions, and $\beta = 0.8$, and using the **safety imprecise chaining** (top) and not (down), with respect to the percentage of noisy (x-axis).



(a) YEAST

Figure B.18: **Flipping - ICC versus CC - Marginalization.** continuation of Figure B.17

Bibliography

Abellan, Joaquin and Andrés R Masegosa (2012). "Imprecise classification with credal decision trees". In: *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems* 20.05, pp. 763–787 (cited on p. 28).

Abellán, Joaquín, Carlos J Mantas, and Javier G Castellano (2017). "A random forest approach using imprecise probabilities". In: *Knowledge-Based Systems* 134, pp. 72–84 (cited on p. 28).

Aitchison, John and Ian Robert Dunsmore (1975). *Statistical prediction analysis*. Cambridge University Press. (cited on p. 4).

Andersen, Martin S, Joachim Dahl, and Lieven Vandenberghe (2018). "CVXOPT: A Python package for convex optimization, version 1.2.2". In: *Available at cvxopt.org* (cited on p. 47).

Antonucci, Alessandro, Andrea Salvetti, and Marco Zaffalon (2007). "Credal networks for hazard assessment of debris flows". In: *Advanced Methods for Decision Making and Risk Management in Sustainability Science*, pp. 237–256 (cited on p. 29).

Antonucci, Alessandro and Giorgio Corani (2017). "The multilabel naive credal classifier". In: *International Journal of Approximate Reasoning* 83, pp. 320–336 (cited on pp. 78, 80).

Augustin, Thomas, Frank PA Coolen, Gert de Cooman, and Matthias CM Troffaes (2014). *Introduction to imprecise probabilities*. John Wiley & Sons (cited on pp. 8, 18, 20, 21, 23, 25, 28, 32, 89, 117, 122).

Basu, Tathagata, Matthias CM Troffaes, and Jochen Einbeck (2020). "Binary Credal Classification Under Sparsity Constraints". In: *International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems*. Springer, pp. 82–95 (cited on p. 28).

Benavoli, Alessio and Branko Ristic (2011). "Classification with imprecise likelihoods: A comparison of TBM, random set and imprecise probability approach". In: *Proceedings of the 14th International Conference on Information Fusion*. IEEE, pp. 1–8 (cited on p. 43).

Benavoli, Alessio and Marco Zaffalon (2014). "Prior near-ignorance for inferences in the k-parameter exponential family". In: *Statistics* 49.5, pp. 1104–1140 (cited on pp. 9, 38, 40–43, 49, 57, 58, 63).

Bensmail, Halima and Gilles Celeux (1996). "Regularized Gaussian discriminant analysis through eigenvalue decomposition". In: *Journal of the American statistical Association* 91.436, pp. 1743–1748 (cited on p. 70).

Berger, James O (1985). *Statistical decision theory and Bayesian analysis; 2nd ed.* Springer Series in Statistics. New York: Springer (cited on pp. 4, 16, 18).

Bernard, Jean-Marc (2005). "An introduction to the imprecise Dirichlet model for multinomial data". In: *International Journal of Approximate Reasoning* 39.2-3, pp. 123–150 (cited on p. 57).

Bernardo, José M and Adrian FM Smith (2000). *Bayesian Theory*. John Wiley & Sons Ltd. (cited on pp. 32, 38, 42).

Bertrand, Joseph (1889). *Calcul des probabilités*. Vol. 1. The name of the publisher (cited on p. 7).

Borel, Émile (1924). "A propos d un traité de probabilités". In: *Revue Philosophique de la France et de l'Étranger* 98, pp. 321–336. ISSN: 00353833, 2104385X. URL: http://www.jstor.org/stable/41082164 (cited on p. 7).

Boutell, Matthew R, Jiebo Luo, Xipeng Shen, and Christopher M Brown (2004). "Learning multi-label scene classification". In: *Pattern recognition* 37.9, pp. 1757–1771 (cited on pp. 73, 75).

Braga-Neto, Ulisses M. and Edward R. Dougherty (2004). "Is cross-validation valid for small-sample microarray classification?" In: *Bioinformatics* 20.3, pp. 374–380 (cited on pp. 7, 38).

Burer, Samuel and Dieter Vandenbussche (2009). "Globally solving box-constrained nonconvex quadratic programs with semidefinite-based finite branch-and-bound". In: *Computational Optimization and Applications* 43.2, pp. 181–195 (cited on p. 47).

Carranza Alarcón, Yonatan-Carlos and Sébastien Destercke (2018). "Analyse Discriminante Imprécise basée sur l'inférence Bayésienne robuste". In: *27èmes Rencontres francophones sur la Logique Floue et ses Applications*. Cépaduès, pp. 85–92 (cited on p. 11).

Carranza Alarcón, Yonatan Carlos and Sébastien Destercke (June 2019). "Imprecise Gaussian Discriminant Classification". In: *Proceedings of the Eleventh International Symposium on Imprecise Probabilities: Theories and Applications*. Ed. by Jasper De Bock, Cassio P. de Campos, Gert de Cooman, Erik Quaeghebeur, and Gregory Wheeler. Vol. 103. Proceedings of Machine Learning Research. Thagaste, Ghent, Belgium: PMLR, pp. 59–67. URL: http://proceedings.mlr.press/v103/carranza-alarcon19a.html (cited on p. 10).

Carranza Alarcón, Yonatan-Carlos and Sébastien Destercke (2020a). "A first glance at multi-label chaining using imprecise probabilities". In: *Workshop on Uncertainty in Machine Learning*, pp. – (cited on p. 10).

— (2020b). "Apprentissage de rangements prudent avec satisfaction de contraintes". In: *29èmes Rencontres francophones sur la Logique Floue et ses Applications*. Cépaduès (cited on p. 11).

Carranza Alarcón, Yonatan-Carlos, Soundouss Messoudi, and Sébastien Destercke (2020c). "Cautious Label-Wise Ranking with Constraint Satisfaction". In: *Information Processing and Management of Uncertainty in Knowledge-Based Systems*. Ed. by Marie-Jeanne Lesot, Susana Vieira, Marek Z. Reformat, João Paulo Carvalho, Anna Wilbik, Bernadette

Bouchon-Meunier, and Ronald R. Yager. Cham: Springer International Publishing, pp. 96–111. ISBN: 978-3-030-50143-3 (cited on p. 11).

Carranza Alarcón, Yonatan-Carlos and Sébastien Destercke (2020d). "Distributionally robust, skeptical binary inferences in multi-label problems". In: *submitted* -.-, pp. – (cited on p. 10).

— (2020e). "Imprecise Gaussian Discriminant Classification". In: *submitted to journal* -.-, pp. – (cited on pp. 10, 89).

— (2020f). "Multi-label chaining using naive credal classifier". In: *submitted* -.-, pp. – (cited on p. 11).

Cattaneo, Marco EGV (2007). "Statistical decisions based directly on the likelihood function". PhD thesis. ETH Zurich (cited on p. 26).

— (2008). "Fuzzy probabilities based on the likelihood function". In: *Soft Methods for Handling Variability and Imprecision*. Springer, pp. 43–50 (cited on p. 26).

Chen, Ruidi and Ioannis Ch Paschalidis (2018). "A robust learning approach for regression models based on distributionally robust optimization". In: *The Journal of Machine Learning Research* 19.1, pp. 517–564 (cited on pp. 4, 80).

Cheng, Weiwei, Eyke Hüllermeier, and Krzysztof J Dembczynski (2010). "Bayes optimal multilabel classification via probabilistic classifier chains". In: *Proceedings of the 27th international conference on machine learning (ICML-10)*, pp. 279–286 (cited on p. 75).

Cheng, Weiwei, Eyke Hüllermeier, Willem Waegeman, and Volkmar Welker (2012). "Label ranking with partial abstention based on thresholded probabilistic models". In: *Advances in neural information processing systems*, pp. 2501–2509 (cited on p. 26).

Clarke, Bertrand S and Jennifer L Clarke (2018). *Predictive Statistics: Analysis and Inference beyond Models*. Vol. 46. Cambridge University Press (cited on p. 4).

Coolen, FPA (1993). "Imprecise conjugate prior densities for the one-parameter exponential family of distributions". In: *Statistics & probability letters* 16.5, pp. 337–342 (cited on p. 27).

Corani, Giorgio and Marco Zaffalon (2008a). "Credal model averaging: an extension of Bayesian model averaging to imprecise probabilities". In: *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer, pp. 257–271 (cited on p. 27).

— (2008b). "Learning reliable classifiers from small or incomplete data sets: the naive credal classifier 2". In: *Journal of Machine Learning Research* 9.Apr, pp. 581–621 (cited on pp. 27, 101).

Corani, Giorgio and Alessandro Antonucci (2014). "Credal ensembles of classifiers". In: *Computational Statistics & Data Analysis* 71, pp. 818–831 (cited on p. 28).

Corani, Giorgio and Andrea Mignatti (2015). "Credal model averaging for classification: representing prior ignorance and expert opinions". In:

International Journal of Approximate Reasoning 56, pp. 264–277 (cited on p. 27).

Dalton, Lori A. and Mohammadmahdi R. Yousefi (2015). "On optimal Bayesian classification and risk estimation under multiple classes". In: *EURASIP Journal on Bioinformatics and Systems Biology* 2015.1, p. 8. ISSN: 1687-4153 (cited on pp. 7, 38).

De Angelis, Pasquale L, Panos M Pardalos, and Gerardo Toraldo (1997). "Quadratic programming with box constraints". In: *Developments in global optimization*. Springer US, pp. 73–93 (cited on p. 47).

De Cooman, Gert and Filip Hermans (2008). "Imprecise probability trees: Bridging two theories of imprecise probability". In: *Artificial Intelligence* 172.11, pp. 1400–1427 (cited on p. 115).

De Finetti, Bruno (1937). "La prévision: ses lois logiques, ses sources subjectives". In: *Annales de l'institut Henri Poincaré*. Vol. 7. 1, pp. 1–68 (cited on pp. 1, 4, 7, 78).

— (1951). *Recent suggestions for the reconciliation of theories of probability*. Tech. rep. UNIVERSITY OF TRIESTE TRIESTE Italy (cited on p. 13).

— (1970). *Teoria delle probabilità*. Vol. 1. Wiley (cited on p. 4).

— (2017). *Theory of probability: a critical introductory treatment*. Vol. 6. John Wiley & Sons (cited on p. 115).

Dembczyński, Krzysztof, Willem Waegeman, Weiwei Cheng, and Eyke Hüllermeier (2012). "On label dependence and loss minimization in multi-label classification". In: *Machine Learning* 88.1-2, pp. 5–45 (cited on pp. 74, 76, 84, 106).

Dembczyński, Krzysztof, Wojciech Kotłowski, Oluwasanmi Koyejo, and Nagarajan Natarajan (2017). "Consistency analysis for binary classification revisited". In: *International Conference on Machine Learning*, pp. 961–969 (cited on p. 5).

Dembczynski, Krzysztof J, Willem Waegeman, Weiwei Cheng, and Eyke Hüllermeier (2011). "An exact algorithm for F-measure maximization". In: *Advances in neural information processing systems*, pp. 1404–1412 (cited on p. 77).

Dempster, Arthur P (1968). "A generalization of Bayesian inference". In: *Journal of the Royal Statistical Society: Series B (Methodological)* 30.2, pp. 205–232 (cited on pp. 9, 17).

Dendievel, Guillaume, Sébastien Destercke, and Pierre Wachalski (2018). "Density estimation with imprecise kernels: application to classification". In: *International Conference Series on Soft Methods in Probability and Statistics*. Springer, pp. 59–67 (cited on p. 27).

Destercke, Sébastien (2010). "A decision rule for imprecise probabilities based on pair-wise comparison of expectation bounds". In: *Combining Soft Computing and Statistical Methods in Data Analysis*. Springer, pp. 189–197 (cited on p. 18).

Destercke, Sebastien (2014). "Multilabel prediction with probability sets: the Hamming loss case". In: *Information Processing and Management of Uncertainty in Knowledge-Based Systems*. Springer, pp. 496–505 (cited on pp. 78, 80, 84, 95, 97, 100, 110, 123).

Díez, Jorge, Oscar Luaces, Juan José del Coz, and Antonio Bahamonde (2015). "Optimizing different loss functions in multilabel classifications". In: *Progress in Artificial Intelligence* 3.2, pp. 107–118 (cited on p. 76).

Domingos, Pedro and Michael Pazzani (1997). "On the optimality of the simple Bayesian classifier under zero-one loss". In: *Machine learning* 29.2-3, pp. 103–130 (cited on p. 28).

Fisher, R. A. (1925). "Theory of Statistical Estimation". In: *Mathematical Proceedings of the Cambridge Philosophical Society* 22.5, 700–725. DOI: 10.1017/S0305004100009580 (cited on p. 2).

Frank, A. and A. Asuncion (2010). *UCI Machine Learning Repository*. URL: http://archive.ics.uci.edu/ml (cited on p. 50).

Friedman, Jerome, Trevor Hastie, and Robert Tibshirani (2001). *The elements of statistical learning*. Springer New York Inc. (cited on pp. 3, 6, 37, 40, 50).

Friedman, Jerome H (1989). "Regularized discriminant analysis". In: *Journal of the American statistical association* 84.405, pp. 165–175 (cited on p. 50).

— (1997). "On bias, variance, 0/1—loss, and the curse-of-dimensionality". In: *Data mining and knowledge discovery* 1.1, pp. 55–77 (cited on p. 28).

Fürnkranz, Johannes, Eyke Hüllermeier, Eneldo Loza Mencía, and Klaus Brinker (2008). "Multilabel classification via calibrated label ranking". In: *Machine Learning* 73.2, pp. 133–153 (cited on pp. 73, 107).

Gatterbauer, Wolfgang and Dan Suciu (2014). "Oblivious bounds on the probability of boolean functions". In: *ACM Transactions on Database Systems (TODS)* 39.1, pp. 1–34 (cited on p. 106).

Gauss, Carl Friedrich (1810). *Méthode des moindres carrés: Mémoires sur la combinaison des observations*. Mallet-Bachelier (cited on pp. 2, 3).

Gibaja, Eva and Sebastián Ventura (2014). "Multi-label learning: a review of the state of the art and ongoing research". In: *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 4.6, pp. 411–444 (cited on p. 76).

Ha, Thien M (1997). "The optimum class-selective rejection rule". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 19.6, pp. 608–615 (cited on pp. 17, 26).

Hadamard, Jacques (1902). "Sur les problèmes aux dérivées partielles et leur signification physique". In: *Princeton university bulletin*, pp. 49–52 (cited on p. 3).

Hand, David J and Keming Yu (2001). "Idiot's Bayes—not so stupid after all?" In: *International statistical review* 69.3, pp. 385–398 (cited on p. 28).

Herbei, Radu and Marten H Wegkamp (2006). "Classification with reject option". In: *Canadian Journal of Statistics* 34.4, pp. 709–721 (cited on p. 26).

Hermans, Filip, Erik Quaeghebeur, et al. (2009). "Imprecise Markov chains and their limit behavior". In: *Probability in the Engineering and Informational Sciences* 23.4, pp. 597–635 (cited on p. 96).

Hinkley, David et al. (1979). "Predictive likelihood". In: *The Annals of Statistics* 7.4, pp. 718–728 (cited on p. 4).

Hu, Weihua, Gang Niu, Issei Sato, and Masashi Sugiyama (2018). "Does distributionally robust supervised learning give robust classifiers?" In: *International Conference on Machine Learning*, pp. 2029–2037 (cited on p. 80).

Jain, Himanshu, Yashoteja Prabhu, and Manik Varma (2016). "Extreme multi-label loss functions for recommendation, tagging, ranking & other missing label applications". In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 935–944 (cited on p. 98).

Jasinska Kalina, Dembczyński Krzysztof (2018). "Bayes optimal prediction for NDCG@k in extreme multi-label classification". In: *Workshop on Multiple Criteria Decision Aid to Preference Learning (DA2PL)* (cited on p. 76).

Joachims, Thorsten (2005). "A support vector method for multivariate performance measures". In: *Proceedings of the 22nd international conference on Machine learning*, pp. 377–384 (cited on p. 5).

Johnson, CR (1970). "Positive definite matrices". In: *The American Mathematical Monthly* 77.3, pp. 259–264 (cited on p. 47).

Kitchin, Rob and Tracey P Lauriault (2015). "Small data in the era of big data". In: *GeoJournal* 80.4, pp. 463–475 (cited on p. 7).

Kotlowski, Wojciech and Krzysztof Dembczyński (2016). "Surrogate regret bounds for generalized classification performance metrics". In: *Asian Conference on Machine Learning*, pp. 301–316 (cited on p. 90).

Koyejo, Oluwasanmi O, Nagarajan Natarajan, Pradeep K Ravikumar, and Inderjit S Dhillon (2015). "Consistent multilabel classification". In: *Advances in Neural Information Processing Systems*, pp. 3321–3329 (cited on p. 90).

Kuhn, Daniel, Peyman Mohajerin Esfahani, Viet Anh Nguyen, and Soroosh Shafieezadeh-Abadeh (2019). "Wasserstein distributionally robust optimization: Theory and applications in machine learning". In: *Operations Research & Management Science in the Age of Analytics*. INFORMS, pp. 130–166 (cited on pp. 4, 60).

Kumar, Abhishek, Shankar Vembu, Aditya Krishna Menon, and Charles Elkan (2013). "Beam search algorithms for multilabel learning". In: *Machine learning* 92.1, pp. 65–89 (cited on p. 131).

Laplace, Pierre Simon (1773). "Mémoire sur la probabilité de causes par les évenements". In: *Mémoires de l'Académie Royale des Sciences présentés par divers savants*, 621–656 (cited on p. 2).

Levi, Isaac (1983). *The enterprise of knowledge: An essay on knowledge, credal probability, and chance*. MIT press (cited on pp. 8, 24, 38).

Lewis, David D (1995). "Evaluating and optimizing autonomous text classification systems". In: *Proceedings of the 18th annual international ACM SIGIR conference on Research and development in information retrieval*, pp. 246–254 (cited on p. 5).

Mantas, Carlos J and Joaquin Abellan (2014). "Credal-C4. 5: Decision tree based on imprecise probabilities to classify noisy data". In: *Expert Systems with Applications* 41.10, pp. 4625–4637 (cited on p. 4).

Marco, Virgil R, Dean M Young, and Danny W Turner (1987). "The Euclidean distance classifier: an alternative to the linear discriminant function". In: *Communications in Statistics-Simulation and Computation* 16.2, pp. 485–505 (cited on p. 40).

Mauá, Denis D, Fabio G Cozman, Diarmaid Conaty, and Cassio P Campos (2017). "Credal sum-product networks". In: *Proceedings of the Tenth International Symposium on Imprecise Probability: Theories and Applications*, pp. 205–216 (cited on p. 27).

Mena, Deiner, Elena Montañés, José Ramón Quevedo, and Juan José del Coz (2016). "An overview of inference methods in probabilistic classifier chains for multilabel classification". In: *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 6.6, pp. 215–230 (cited on pp. 75, 131).

Menon, Aditya K, Ankit Singh Rawat, Sashank Reddi, and Sanjiv Kumar (2019). "Multilabel reductions: what is my loss optimising?" In: *Advances in Neural Information Processing Systems*, pp. 10600–10611 (cited on p. 75).

Miranda, Enrique, Ignacio Montes, and Sébastien Destercke (2019). "A Unifying Frame for Neighbourhood and Distortion Models". In: *International Symposium on Imprecise Probability: Theories and Applications (ISIPTA)* (cited on p. 58).

Mortier, Thomas, Marek Wydmuch, Krzysztof Dembczyński, Eyke Hüllermeier, and Willem Waegeman (2019). "Efficient Set-Valued Prediction in Multi-Class Classification". In: *arXiv preprint arXiv:1906.08129* (cited on p. 26).

Mouhagir, Hafida, Véronique Cherfaoui, Reine Talj, François Aioun, and Franck Guillemard (2017). "Using evidential occupancy grid for vehicle trajectory planning under uncertainty with tentacles". In: *2017 IEEE 20th International Conference on Intelligent Transportation Systems (ITSC)*. IEEE, pp. 1–7 (cited on p. 106).

Moyano, Jose M, Eva L Gibaja, Krzysztof J Cios, and Sebastián Ventura (2018). "Review of ensembles of multi-label classifiers: models, experimental study and prospects". In: *Information Fusion* 44, pp. 33–45 (cited on p. 75).

Nakharutai, Nawapon, Matthias C.M. Troffaes, and Camila C.S. Caiado (2019). "Improving and benchmarking of algorithms for decision making with lower previsions". In: *International Journal of Approximate Reasoning* 113, pp. 91 –105. ISSN: 0888-613X. DOI: https://doi.org/10.1016/j.ijar.2019.06.008. URL: http://www.sciencedirect.com/science/article/pii/S0888613X18307084 (cited on pp. 39, 51, 66).

Neumann, J. von and O. Morgenstern (1947). *Theory of games and economic behavior*. Princeton University Press (cited on p. 2).

Neyman, Jerzy (1937). "Outline of a theory of statistical estimation based on the classical theory of probability". In: *Philosophical Transactions of the Royal Society of London. Series A, Mathematical and Physical Sciences* 236.767, pp. 333–380 (cited on p. 2).

Pardalos, Panos M and Stephen A Vavasis (1991). "Quadratic programming with one negative eigenvalue is NP-hard". In: *Journal of Global Optimization* 1.1, pp. 15–22 (cited on p. 47).

Paton, Lewis, Matthias Troffaes, Nigel Boatman, Mohamud Hussein, and Andy Hart (2015). "A robust Bayesian analysis of the impact of policy decisions on crop rotations." In: SIPTA (cited on p. 27).

Paton, Lewis et al. (2016). "A robust Bayesian land use model for crop rotations". PhD thesis. Durham University (cited on p. 27).

Pillai, Ignazio, Giorgio Fumera, and Fabio Roli (2013). "Multi-label classification with a reject option". In: *Pattern Recognition* 46.8, pp. 2256–2266 (cited on pp. 77, 79).

Qi, Feng, Xiao-Jing Zhang, and Wen-Hui Li (2017). "The harmonic and geometric means are Bernstein functions". In: *Boletín de la Sociedad Matemática Mexicana* 23.2, pp. 713–736 (cited on p. 77).

Quaeghebeur, Erik and Gert De Cooman (2005). "Imprecise probability models for inference in exponential families". In: *4th International Symposium on Imprecise Probabilities and Their Applications*. International Society for Imprecise Probability: Theories and Applications (SIPTA), pp. 287–296 (cited on pp. 26, 27).

Quaeghebeur, Erik, Chris Wesseling, Emma Beauxis-Aussalet, Teresa Piovesan, and Tom Sterkenburg (2017). "The CWI world cup competition: Eliciting sets of acceptable gambles". In: *Proceedings of the Tenth International Symposium on Imprecise Probability: Theories and Applications*, pp. 277–288 (cited on p. 29).

Ramón Quevedo, José, Oscar Luaces, and Antonio Bahamonde (2012). "Multilabel classifiers with a probabilistic thresholding strategy". In: *Pattern Recognition* 45.2, pp. 876–883 (cited on p. 77).

Read, Jesse, Bernhard Pfahringer, Geoff Holmes, and Eibe Frank (2011). "Classifier chains for multi-label classification". In: *Machine learning* 85.3, pp. 333–359 (cited on pp. 75, 107).

— (2019). "Classifier Chains: A Review and Perspectives". In: *arXiv preprint arXiv:1912.13405* (cited on p. 75).

Robert, Christian (2005). *Le choix bayésien: Principes et pratique*. Springer Paris (cited on p. 42).

Roeser, Sabine, Rafaela Hillerbrand, Per Sandin, and Martin Peterson (2012). *Essentials of risk theory*. Springer Science & Business Media (cited on p. 7).

Salmon, Wesley C (1957). "The predictive inference". In: *Philosophy of Science* 24.2, pp. 180–190 (cited on p. 4).

Senge, Robin, Stefan Bösner, Krzysztof Dembczyński, Jörg Haasenritter, Oliver Hirsch, Norbert Donner-Banzhoff, and Eyke Hüllermeier (2014). "Reliable classification: Learning classifiers that distinguish aleatoric and epistemic uncertainty". In: *Information Sciences* 255, pp. 16–29 (cited on pp. 7, 38).

Shafer, Glenn and Vladimir Vovk (2008). "A tutorial on conformal prediction". In: *Journal of Machine Learning Research* 9.Mar, pp. 371–421 (cited on p. 26).

— (2019). *Game-Theoretic Foundations for Probability and Finance*. Vol. 455. John Wiley & Sons (cited on p. 115).

Sucar, L Enrique, Concha Bielza, Eduardo F Morales, Pablo Hernandez-Leal, Julio H Zaragoza, and Pedro Larrañaga (2014). "Multi-label classification with Bayesian network-based chain classifiers". In: *Pattern Recognition Letters* 41, pp. 14–22 (cited on p. 115).

Taylor, Samuel James (1973). *Introduction to measure and integration*. CUP Archive (cited on p. 8).

Troffaes, Matthias CM (2007). "Decision making under uncertainty using imprecise probabilities". In: *International Journal of Approximate Reasoning* 45.1, pp. 17–29 (cited on pp. 18, 25).

Troffaes, Matthias CM and Gert De Cooman (2014). *Lower previsions*. John Wiley and Sons (cited on pp. 1, 8).

Tsoumakas, Grigorios and Ioannis Katakis (2007). "Multi-label classification: An overview". In: *International Journal of Data Warehousing and Mining (IJDWM)* 3.3, pp. 1–13 (cited on pp. 73, 75, 107).

Utkin, Lev V (2015). "The imprecise Dirichlet model as a basis for a new boosting classification algorithm". In: *Neurocomputing* 151, pp. 1374–1383 (cited on p. 28).

Vapnik, V. N. and A. Ya. Chervonenkis (1974). *Theory of Pattern Recognition [in Russian]*. USSR: Nauka (cited on p. 2).

Vapnik, Vladimir N. (1995). *The nature of statistical learning theory*. Springer-Verlag New York, Inc. ISBN: 0-387-94559-8 (cited on p. 3).

Vapnik, Vladimir N and A Ya Chervonenkis (1982). "Necessary and sufficient conditions for the uniform convergence of means to their expectations". In: *Theory of Probability & Its Applications* 26.3, pp. 532–553 (cited on p. 3).

Vu-Linh Nguyen, Eyke Hüllermeier (2019). "Reliable Multilabel Classification: Prediction with Partial Abstention". In: *Thirty-Fourth AAAI Conference on Artificial Intelligence* (cited on pp. 76–79).

Wald, A. (1950). *Statistical Decision Functions*. Wiley Publications in Statistics: Mathematical statistics. Wiley (cited on pp. 2, 3).

Wald, Abraham (1939). "Contributions to the theory of statistical estimation and testing hypotheses". In: *The Annals of Mathematical Statistics* 10.4, pp. 299–326 (cited on p. 2).

Walley, Peter (1991). *Statistical reasoning with imprecise Probabilities*. Chapman and Hall (cited on pp. 1, 7, 8, 17, 18, 20, 22, 24, 26, 38, 49, 78, 115).

— (1996). "Inferences from multinomial data: learning about a bag of marbles". In: *Journal of the Royal Statistical Society: Series B (Methodological)* 58.1, pp. 3–34 (cited on pp. 1, 27, 31–33, 118).

Walter, Gero (2013). "Generalized Bayesian inference under prior-data conflict". PhD thesis. lmu (cited on p. 26).

Xia, Wei, Juan Vera, and Luis F Zuluaga (2015). "Globally solving non-convex quadratic programs via linear integer programming techniques". In: *arXiv preprint arXiv:1511.02423* (cited on p. 47).

Yang, Gen, Sébastien Destercke, and Marie-Hélène Masson (2014). "Nested Dichotomies with probability sets for multi-class classification." In: *ECAI*, pp. 363–368 (cited on p. 96).

— (2017). "Cautious classification with nested dichotomies and imprecise probabilities". In: *Soft Computing* 21.24, pp. 7447–7462 (cited on pp. 18, 51).

Ye, Nan, Kian Ming A Chai, Wee Sun Lee, and Hai Leong Chieu (2012). "Optimizing F-measures: a tale of two approaches". In: *Proceedings of the 29th International Coference on International Conference on Machine Learning*, pp. 1555–1562 (cited on p. 5).

Zaffalon, Marco (1999). "A Credal Approach to Naive Classification." In: *ISIPTA*. Vol. 99, pp. 405–414 (cited on pp. 1, 27, 30, 108).

— (2001). "Statistical inference of the naive credal classifier." In: *ISIPTA*. Vol. 1, pp. 384–393 (cited on pp. 9, 32).

— (2002). "The naive credal classifier". In: *Journal of statistical planning and inference* 105.1, pp. 5–21 (cited on pp. 8, 18, 28, 44, 117, 124).

— (2005). "Credible classification for environmental problems". In: *Environmental Modelling & Software* 20.8, pp. 1003–1012 (cited on p. 29).

Zaffalon, Marco, Keith Wesnes, and Orlando Petrini (2003a). "Reliable diagnoses of dementia by the naive credal classifier inferred from incomplete cognitive data". In: *Artificial intelligence in medicine* 29.1-2, pp. 61–79 (cited on p. 29).

Zaffalon, Marco and Enrico Fagiuoli (2003b). "Tree-based credal networks for classification". In: *Reliable computing* 9.6, pp. 487–509 (cited on p. 28).

Zaffalon, Marco, Giorgio Corani, and Denis Mauá (2012). "Evaluating credal classifiers by utility-discounted predictive accuracy". In: *International Journal of Approximate Reasoning* 53.8, pp. 1282–1301 (cited on p. 51).

Zhang, Min-Ling and Zhi-Hua Zhou (2013). "A review on multi-label learning algorithms". In: *IEEE transactions on knowledge and data engineering* 26.8, pp. 1819–1837 (cited on p. 75).

Zhang, Min-Ling, Yu-Kun Li, Xu-Ying Liu, and Xin Geng (2018). "Binary relevance for multi-label learning: an overview". In: *Frontiers of Computer Science* 12.2, pp. 191–202 (cited on p. 75).

List of Tables

List of Figures

List of Abbreviations

DTA	**Decision Theoretic Approach**
ERM	**Empirical Risk Minimization**
GDA	**Gaussian Discriminant Analisis**
GBI	**Generalized Bayesian Inference**
ICC	**Imprecise Classifier Chains**
IP	**Imprecise Probabilities**
ML	**Machine Learning**
MLC	**Multi-Label Classification**
PMP	**Probablistic Modelling Paradigm**
SP	**Subjective Probability**