N° d'ordre NNT : 2020LYSE2106

THESE de DOCTORAT DE L'UNIVERSITÉ DE LYON

Opérée au sein de

L'UNIVERSITÉ LUMIÈRE LYON 2

École Doctorale : ED 512

Informatique et Mathématiques

Discipline : Mathématiques appliquées

Soutenue publiquement le 13 novembre 2020, par :

Margot SELOSSE

Introducing parsimony to analyse complex data with model-based clustering.

Devant le jury composé de :

Stéphane CHRÉTIEN, Professeur des Universités, Université Lumière Lyon 2, Président

Pierre LATOUCHE, Professeur des Universités, Université de Paris, Rapporteur

Bettina GRÜN, Associate professor, Vienna University of Economics and Business, Rapporteuse

Charles BOUVEYRON, Professeur des Universités, INRIA et Université Côte d'Azur, Examinateur

Claire GORMLEY, Associate professor, University College Dublin, Examinatrice

Charlotte LACLAU, Maître de conférences, Université de Lyon, Examinatrice

Julien JACQUES, Professeur des Universités, Université Lumière Lyon 2, Directeur de thèse

Christophe BIERNACKI, Professeur des Universités, INRIA, Co-Directeur de thèse

Contrat de diffusion

Résumé long : Parcimonie dans les modèles probabilistes pour l'analyse de données complexes

Apprentissage automatique

Ces dernières années, l'apprentissage automatique (aussi appelé "machine learning" en anglais) a reçu beaucoup d'intérêt de la part de la communauté scientifique et du grand publique. "Intelligence artificielle", "apprentissage statistique", "science des données", sont tous des termes qui représentent une branche de l'apprentissage automatique ou qui y sont fortement liés.

On considère que les modèles d'apprentissage automatique apprennent à partir des données puisque leur comportement dépend des échantillons de données qui ont été introduits dans le programme en entrée. En outre, ces algorithmes peuvent être utilisés sur différents ensembles de données pour résoudre différents problèmes et c'est la raison pour laquelle nous les considérons comme intelligents. L'intérêt croissant pour l'apprentissage automatique est dû à deux facteurs. Premièrement, la production d'informations numériques a fortement augmenté ces dernières années, les entreprises et institutions privées ont désormais davantage accès à des flux de données massifs via les réseaux sociaux, les smartphones, les sites web et les plateformes d'achat. Deuxièmement, ces données n'auraient pas pu être stockées, prétraitées ou analysées sans l'énorme croissance de la puissance de calcul, qui permet de concevoir des modèles plus complexes et plus puissants.

Trois familles de paradigmes

Il existe plusieurs paradigmes en apprentissage automatique, qui utilisent les données de manière différentes, et qui réalisent différents type de tâches. Nous décrivons maintenant les trois grandes familles de paradigmes de l'apprentissage automatique.

Apprentissage supervisé

En apprentissage supervisé, nous avons deux ensembles de variables. Les variables d'entrée $\boldsymbol{x}_i$, et les variables labellisées $\boldsymbol{y}_i$. Le but est d'apprendre une application f de $\boldsymbol{x}_i$ vers $\boldsymbol{y}_i$, avec le jeu de données fait de paires $(\boldsymbol{x}_i, \boldsymbol{y}_i)_{i\in\{1,\dots,N\}}$. En notant $\hat{\boldsymbol{y}}_i = f(\boldsymbol{x}_i;\boldsymbol{\theta})$ la prédiction du modèle pour $\boldsymbol{y}_i$ sachant les paramètres $\boldsymbol{\theta}$, alors la fonction de perte $\mathcal{L}(\boldsymbol{y}_i, \hat{\boldsymbol{y}}_i)$ définit à quel point les prédicitions du modèle sont précises. Les paramètres $\boldsymbol{\theta}$ sont choisis pour minimiser cette fonction de perte sur un jeu de données avec les échantillons $(\boldsymbol{x}_i, \boldsymbol{y}_i)_{i\in\{1,\dots,N\}}$ donnés :

$$\sum_i^N \mathcal{L}(\boldsymbol{y}_i, \hat{\boldsymbol{y}}_i) = \sum_i^N \mathcal{L}(\boldsymbol{y}_i, f(\boldsymbol{x}_i;\boldsymbol{\theta})).$$

Le choix de fonction de perte dépend du problème à résoudre et de la nature de $\boldsymbol{x}_i$ et $\boldsymbol{y}_i$. de nombreux algorithmes supervisés existent déjà, tels que la régression linéaire, la régression logistique, les arbres de décisions, les machines à vecteur de support. Dernièrement, les algorithmes impliquant les réseaux de neurones atteignent les performances de l'état de l'art pour différentes tâches telles que la vision par ordinateur et la reconnaissance vocale.

Apprentissage non-supervisé

L'apprentissage non-supervisé a un sens plus large et moins bien défini que l'apprentissage supervisé, car il peut servir dans plusieurs cas. Globalement, le rôle de l'apprentissage non-

supervisé est de trouver des structures au sein des données $\boldsymbol{x}$. En général, les variables $\boldsymbol{y}$ ne sont pas fournies, seulement les variables d'entrées $\boldsymbol{x}$ le sont. Les principaux exemples de tâches non supervisées sont l'estimation de densité, la réduction de dimension, l'extraction de caractéristques, la modélisation générative et l'analyse de clusters.

- L'estimation de densité est la construction d'un estimateur de la desnité de probabilité, basé sur les données observées.
- La réduction de dimension consiste à trouver une représentation pour un jeu de données, et ce dans un espace de plus petite dimension.
- Les modèles génératifs consistent à considérer que les observations d'un jeu de données ont été échantillonés selon un processus par la fonction de densité $p(\boldsymbol{x}; \boldsymbol{\theta})$, dont les paramètres $\boldsymbol{\theta}$ doivent être estimés.
- L'analyse de clusters est la tâche qui consiste à regrouper les observations $\boldsymbol{x}_i$ dans différents groupes (ou clusters). Les observations qui se trouvent dans le même groupe sont alors supposées appartenir à une même catégorie.

Apprentissage semi-supervisé

L'apprentissage semi-supervisé est à mi-chemin entre l'apprentissage supervisé et l'apprentissage non-supervisé, dans le sens où tous les labels $\boldsymbol{y}_i$ ne sont pas nécessairement disponibles pour toutes les observations $\boldsymbol{x}_i$. Dans ce cas, l'objectif d'un algorithme semi-supervisé peut être de concevoir un modèle qui utilise les échantillons non-labellisés pour obtenir de meilleures performances en prédictions, en comparaison avec un algorithme qui n'utiliserait que les échantillons labellisés.

Données complexes

La plupart des techniques d'apprentissage automatiques sont très efficaces lorsque le jeu de données est dit "facile", ce qui signifie qu'il est structuré, en petite dimension, et qu'il ne contient pas de valeurs manquantes. Cependant, les jeux de données qui représentent la réalité sont souvent plus compliqués. Les propriétés qui nous font considérer un jeu de donénes comme "complexe" sont :

- La haute dimension, souvent associée à "la malédiction de la dimension", qui concerne les phénomènes qui apparaissent lors de la manipulation de jeux de données avec de nombreuses variables. Le principal problème est que lorsque le nombre de variables augmente, le volume de l'espace augmente si rapidement que les données deviennent sparses. De plus, de nombreux algorithmes ne peuvent pas estimer leurs paramètres lorsque le nombre d'observations N est supéreur au nombre de variables J.
- L'hétérogénéité des données, ou la mixité des données, concerne les données qui ne sont pas de même nature. Par exemple, un simple jeu de données sur les clients d'une entreprise pourrait contenir le statut social (une variable catégorielle), l'âge (une variable de comptage), la taille et le poids du client (des variables continues). Une telle diversité de type de données peut être difficile à modéliser mathématiquement car les valeurs des variables ne font pas parti du même espace. C'est donc difficile de choisir une distribution commune à toutes ces variables.

- La sparsité se réfère aux jeux de données avec peu d'information. Souvent, cela concerne les données qui contiennent une majorité de valeurs nulles. Par exemple, lorsque nous modélisons les intéractions entre utilisateurs d'un réseau social en comptant le nombre de messages qu'ils s'envoient, la matrice résultante est généralement sparses (beaucoup d'utilisateurs ne s'envoient jamais de messages).
- Les valeurs manquantes se réfèrent au fait que, parfois, certains éléments d'un jeu de données n'ont pas de valeur. Par exemple, lorsque l'on analyse un questionnaire auquel des personnes ont répondu, il est très courant d'observer que certaines questions n'ont pas été répondues par certains participants. Cela peut être modélisé de différentes manières, selon si on considère que le participant n'a pas répondu de manière intentionelle ou non.
- Les jeux de données en continu sont ceux dont les données arrivent en flux. L'exemple le plus courant concerne les données venant de capteurs dont les valeurs s'actualisent à différents instants. Ce genre données requièrent des algorithmes spéciaux capables de recevoir de nouvelles données au fil du temps, mais ce sujet ne sera pas abordé dans cette thèse.

Contenu de la thèse

Cette thèse se concentre sur l'apprentissage non-supervisé, et plus spécifiquement sur la parcimonie de modèles de clustering probabilistes dans le cadre de données complexes. Les modèles de clustering probabilistes marient les modèles génératifs et l'analyse de clusters. Ce type de modèle apporte de nombreux avantages tels que l'interprétabilité et la sélection de modèle. Grâce à leur flexibilité, ces techniques ont prouvé leur efficacité dans de nombreux domaines, et sont largement utilisées pour l'analyse de données. Un inconvénient des méthodes de clustering probabilistes classiques est le nombre élevé de paramètres à estimer, ce qui peut ralentir les algorithmes d'inférence et conduire à de mauvais résultats dans le cas de données complexes. Concevoir des modèles plus parcimonieux (c'est à dire avec moins de paramètres) est un moyen efficace de surmonter ce problème. Cette thèse a pour objectif de concevoir de nouvelles approches probabilstes adaptées aux données complexes. Nous nous intéressons à des données en grande dimension, mais aussi à des données hétérogènes, des données avec des valeurs manquantes et des données sparses telles que les données textuelles.

Le Chapitre 2 rappelle les notions nécessaires pour une bonne compréhension des contributions de la thèse. Premièrement, il détaille les aspects mathématiques des modèles de mélange finis, qui sont à la base des approches probabilistes de clustering. Ces notions seront utiles pour tous les autres chapitres de cette thèse. Deuxièmement, ce chapitre décrit l'anayse factorielle, et plus particulièrement le modèle de mélange d'analyse de facteurs, qui est la base du Chapitre 5. Finalement, ce chapitre définit le modèle des blocs latents (LBM), qui est une technique de co-clustering. Le co-clustering est une tâche qui consiste à réaliser le clustering simultané des lignes et des colonnes d'un jeu de données. Ces notions seront utiles pour le Chapitre 3 et pour le Chapitre 4.

Le Chapitre 3 présente une extension du modèle des blocs latents multiples (MLBM) (Robert, 2017) aux données hétérogènes. Ces données sont difficiles à modéliser avecune seule et même distribution car les valeurs des variables ne se trouvent pas dans le même espace. Dans le cas du co-clustering, c'est particulièrement compliqué, car l'agorithme doit regrouper les variables aussi. De plus, il peut sembler contre-intuitif de regrouper des variables de nature différente car l'objectif du clustering est de regrouper des éléments qui ont quelque chose

en commun. L'approche MLBM consiste à étendre le modèle des blocs latents (LBM) pour qu'il soit capable de prendre des données hétérogènes en compte.

Le Chapitre 4 présente le modèle SOCC (Self-Organised Co-Clustering model) pour les données textuelles et plus précisément pour les matrices document-terme. Ces matrices représentent des données textuelles tel que la cellule ($document_i$,$terme_j$) compte combien de fois le terme $terme_j$ a été utiisé dans le document $document_i$. Cette représentation a l'avantage d'être facile à construire et à lire. Cependant, les matrices qui en résultent sont de très haute dimension et extrêmement sparses, ce qui les rend difficiles à exploiter. Le modèle SOCC s'adapte à ces particularités et définit un modèle pour le clustering des termes et des documents qui offre des résultats simples à exploiter.

Le Chapitre 5 investigue le modèle de mélange Gaussian profond (DGMM) (Viroli and McLachlan, 2019) et ses propriétés. Ce modèle consiste à empiler des couches de MFA, ce qui résulte en une architecture imitant les réseaux de neurones. Cela est rendu possible en considérant les scores latents d'une couche comme étant l'entrée du MFA de la couche d'après. Dans ce chapitre, nous montrons empiriquement les difficultés pour estimer les paramètres du modèle, puis nous discutons les raisons possibles et les solutions à ces problèmes.

Mots-Clefs : modèles probabilistes – clustering – modèles de mélange – co-clustering – analyse de facteurs – parcimonie.