DIPLÔME NATIONAL DE DOCTORAT

(Arrêté du 25 mai 2016)

Date de la soutenance : **8 janvier 2020**

Nom de famille et prénom de l'auteur : **BAUDRY Samuel**

Titre de la thèse : « *Quelques problèmes d'apprentissage statistique en présence de données incomplètes* ».

Résumé

La plupart des méthodes statistiques ne sont pas adaptées pour nativement tenir compte de l'incomplétude des données, et attendent en général en entrée un ensemble d'observations indépendantes et identiquement distribuées (i.i.d.). Le terme de données incomplètes réfère souvent dans la littérature aux données manquantes, néanmoins nous considérons dans ce document que cela fait référence à un ensemble de phénomènes plus large, dont les données manquantes font partie. Dans le panel de schémas d'incomplétude de données, on citera notamment les phénomènes de censure, très présents en assurance et dans les études cliniques, les données manquantes ou encore les données bruitées, pour n'en citer que quelques-uns. Par ailleurs, les travaux en deep learning se concentrent principalement sur l'étude de données non structurées, comme les images, le texte, les vidéos ou encore le son, mais très peu de travaux approfondis appliqués sur des données tabulaires ont été proposés, et encore moins sur des données tabulaires incomplètes.

Les problèmes d'incomplétude de données sont omniprésents en pratique, et peuvent conduire à un biais significatif dans l'étude statistique. Une attention particulière est donc à porter en amont de toute étude statistique afin de s'assurer de la cohérence du jeu de données observé, et éventuellement effectuer un traitement statistique en amont de l'étude (e.g. pour imputer des valeurs manquantes, ou correctement tenir compte des valeurs aberrantes). Pour autant, peu de travaux proposent de travailler sur des problèmes de modélisation en présence de données incomplètes.
Les données tabulaires et les séries temporelles n'ont pas reçu assez d'attention de la part de la communauté autour du deep learning, et nécessitent un traitement particulier. En effet, certaines métriques et fonctions de perte utilisées en analyse d'image ne sont pas applicables sur des données tabulaires. De même, vérifier qu'un modèle a bien estimé la distribution d'un ensemble d'images revient à vérifier "à la main" que le modèle génère des images (ou du texte) cohérentes, i.e. vérifier si le modèle dépasse les capacités humaines.
Cela n'est en revanche pas applicable à un vecteur multivarié, et nécessite des outils plus complexes pour comparer des distributions.
Cette thèse se concentre sur l'utilisation de méthodes de machine learning appliquées à des problèmes comportant des données incomplètes, dans un contexte assurantiel. Le secteur de l'assurance est en train de changer de paradigme avec d'une part la collecte de nouvelles données (e.g. les données GPS telematics, ou encore les photographies et les descriptions textuelles des sinistres), et d'autre part l'émergence de nouvelles méthodes d'analyse de données. Ce changement de paradigme représente un intérêt fort pour le secteur de l'assurance, car cela permet de mieux estimer le risque encouru par l'assureur de par son activité, et de mieux connaître ses clients.

Ce document regroupe trois contributions, regroupées dans trois chapitres distincts précédés d'un chapitre qui introduit les éléments techniques nécessaires à la compréhension des contributions. La première contribution propose une structure de modélisation permettant d'entraîner un modèle à prédire une variable cible fortement censurée. Nous appliquons notre méthode sur des problèmes de provisionnement en assurance, et comparons notre approche avec les outils standards de calcul de réserves en assurance, et montrons que nous donnons des prédictions plus précises en matière de variance.
La deuxième contribution propose une architecture permettant de combiner plusieurs jeux de données afin de construire un modèle prédictif sur chaque jeu de données qui tient compte de l'information prédictive présente dans chaque autre jeu. Les assureurs conçoivent souvent plusieurs produits différents couvrant le même risque. Cela est en partie causé par le fait que la population couverte est fondamentalement différente (e.g. pays différents, ou entités différentes de l'assurance). Cela conduit à des jeux de données distincts portant sur l'étude d'un risque commun, et les combiner permet potentiellement d'affiner l'étude du risque. Nous comparons notre approche proposée avec l'état de l'art, et montrons que nous donnons de meilleurs résultats.
La troisième contribution propose une architecture qui permet d'estimer la distribution conditionnelle d'une série temporelle. Cela permet à l'assureur de générer des scénarios économiques de type "monde réel", et d'anticiper des changements significatifs dans leurs placements financiers. Les méthodes usuelles employées pour répondre à ce type de problématique font en général une hypothèse structurelle sur le processus stochastique. Ici, nous proposons une méthode non-paramétrique pour estimer la distribution conditionnelle d'une série temporelle, afin d'en générer de nouvelles réalisations. Nous comparons notre approche avec l'état de l'art, et montrons qu'elle donne une meilleure estimation des auto-corrélations du processus stochastique, ainsi que de son support.