CHAIRE
MACHINE LEARNING
FOR BIG DATA

Paris, le 27/11/2017

Offre de stage

Sujet : «Reliable Machine-Learning: Learning to Weight Data»

Possibilité de poursuivre sur une thèse

Encadrement

Stephan Clémençon (Télécom ParisTech), Patrice Bertail (Université Paris Ouest & Télécom ParisTech)

Lieu et dates du stage

Telecom ParisTech, 46 rue Barrault, 75013 Paris

Date de début du stage : printemps 2018

Équipe(s) d'accueil de la thèse

Télécom ParisTech, Département IDS, équipe Statistique, Signal et Apprentissage (S2A)

Mots clés

Predictive learning, (importance) sampling, survey schemes, Horvitz-Thompson estimation, post-stratification, transfer learning

Sujet détaillé

In many situations, data are not the only materials that can be exploited by machine-learning algorithms. Sometimes, they can also make use of weights resulting from sampling stratification. Such weights correspond either to true inclusion probabilities or else to calibrated or post-stratification weights, minimizing some discrepancy under certain margin constraints for the inclusion probabilities. Asymptotic analysis of Horvitz-Thompson estimators (see Horvitz & Thompson, 1951}) based on survey data, in the context of mean estimation and regression in particular, has received a good deal of attention in the statistical literature and the last few years have witnessed significant progress towards a comprehensive functional limit theory for distribution function estimation. In parallel, the field of machine-learning has been the subject of a spectacular development, its practice has been revitalized in particular by various breakout algorithms (e.g. SVM, boosting methods) and is supported by a sound probabilistic theory based on recent (non asymptotic) results in the study of empirical processes. However, our increasing capacity to collect data, due to the ubiquity of sensors, has improved much faster than our ability to process and analyze Big Datasets, for predictive purpose in particular. Whereas massive information, which machine-learning procedures could theoretically now rely on, is available in the Big Data era, with the advent of IoT (Internet of Things) in particular, exploiting it may be challenging, insofar as data are now more and more rarely collected through controlled experimental designs, specified in advance, but much more frequently on the fly and using

them as training examples may yield strong biases in learning methods, and improper predictive models. The goal of this research internship is to investigate to which extent appropriate re-weighting of the data could be learnt in presence of auxiliary information, so as to correct possible biases. The idea is to formulate this objective as an optimization problem and develop computational methods, together with a theoretical validity framework, in order to solve it. Practical applications (IoT and smart cities, public transportations) will be also considered.

La Chaire Machine Learning for Big Data

Le traitement statistique des masses de données convoque à la fois mathématiques appliquées et informatique, à travers une discipline en plein essor : le Machine Learning ou apprentissage statistique.

Créée en septembre 2013 avec le soutien de la Fondation Télécom et financée à hauteur de près de 2 M€ par quatre entreprises partenaires : Criteo, PSA Peugeot Citroën, Safran et BNP Paribas, la Chaire Machine Learning for Big Data est portée par le mathématicien Stéphan Clémençon, Enseignant-Chercheur, Professeur au sein du Département du Traitement du Signal et des Images à Télécom ParisTech.

Proposant cinq axes de recherche méthodologiques, enrichis par des applications industrielles concrètes, cette Chaire a pour objectif d'animer, en interaction avec ses partenaires, une activité de recherche de pointe en Machine Learning, ainsi que de proposer des programmes de formation.

La variété des données aujourd'hui disponibles (nombres, images, textes, signaux), leur grande dimension et leur volumétrie rendent souvent inopérantes les méthodes statistiques traditionnelles reposant sur le prétraitement humain et un long travail de modélisation. Le Machine Learning vise donc à élaborer et étudier des algorithmes, à vocation prédictive le plus souvent, permettant à des machines d'apprendre automatiquement à partir des données et à effectuer des tâches de façon performante.

Les avancées technologiques, l'omniprésence des capteurs (systèmes embarqués, objets connectés, Internet…) et l'explosion des réseaux sociaux s'accompagnent d'un véritable déluge de données, propulsant les sciences de l'information au centre du processus de valorisation des masses de donnés. En plus de la collecte et du stockage, l'enjeu est de pouvoir analyser ces données afin d'optimiser les décisions et mettre au point de nouvelles applications.

Au-delà du buzz médiatique dont il fait l'objet, le Big Data est donc un sujet stratégique majeur, au cœur d'enjeux économiques et sociétaux considérables. Son impact est désormais perçu dans presque tous les secteurs de l'activité humaine : de la recherche scientifique à la médecine en passant, entre autres, par la finance, le bâtiment, l'e-commerce, la défense ou les transports.

En savoir plus sur la Chaire, ses axes de recherche, ses activités, ses publications :

http://machinelearningforbigdata.telecom-paristech.fr

Profil du candidat

Etudiant titulaire d'un master 2 recherche

- Apprentissage statistique, Probabilité, Optimisation

- Bon niveau en programmation (e.g. Java, C/C++, Python)

- Bon niveau d'anglais

Candidatures

à envoyer à stephan.clemencon@telecom-paristech.fr:

- Curriculum Vitae

- Lettre de motivation personnalisée expliquant l'intérêt du candidat sur le sujet (directement dans le corps du mail)

- Relevés de notes des années précédentes

- Contact d'une personne de référence

Les candidatures incomplètes ne seront pas examinées.

Références

[1] Horvitz, D., Thompson, D.: A generalization of sampling without replacement from a finite universe. JASA 47, 663–685 (1951)
[2] Sampling and Empirical Risk Minimization. S. Clémençon, P. Bertail & E. Chautru (2016). In Statistics.
[3] Learning from Survey Training Samples: Rate Bounds for Horvitz-Thompson Risk Minimizers. S. Clémençon, G. Papa & P. Bertail (2016). In the Proceedings of ACML (2016).
[4] Empirical processes in survey sampling. Bertail, P., Chautru, E., Clémençon, S. Scandinavian Journal of Statistics, (2016)