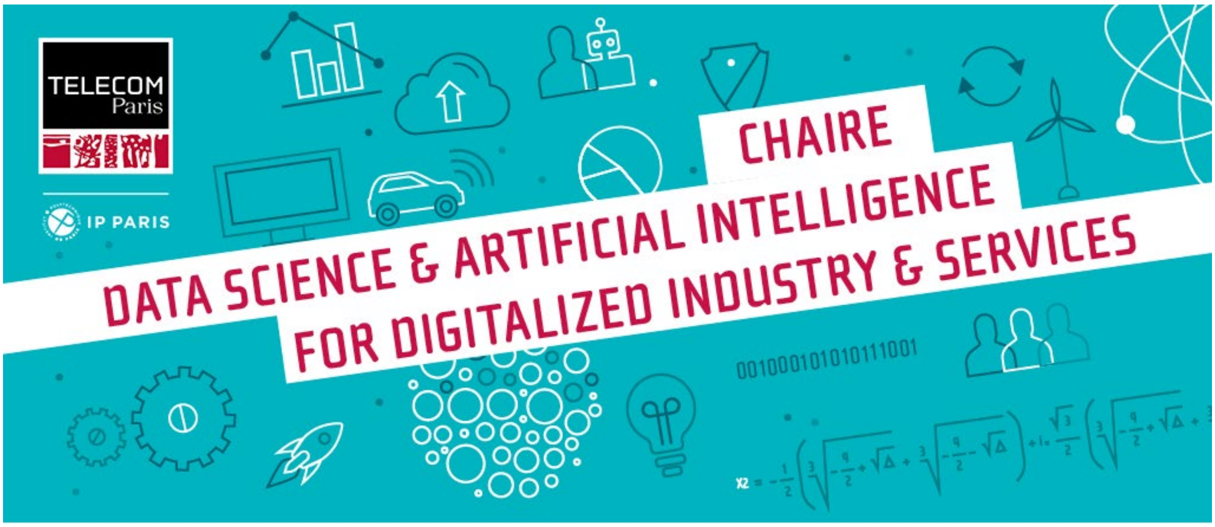
TELECOM Paris
IP PARIS
CHAIRE
DATA SCIENCE & ARTIFICIAL INTELLIGENCE
FOR DIGITALIZED INDUSTRY & SERVICES
0010001010101111001

AIRBUS engie IDEMIA augmented identity SAFRAN Valeo FONDATION Mines-Télécom

Offre de stage

Reliable Machine Learning: Learning to Weight Data

Possibilité de poursuivre sur une thèse

A propos de la chaire

La Chaire Data Science and Artificial Intelligence for Digitalized Industry and Services, portée par Florence d’Alché-Buc, enseignante-chercheuse dans le département Image, Données, Signal de Télécom Paris, la chaire DSAIDIS réunit cinq partenaires industriels : Airbus Defence & Space, Engie, Idemia, Safran et Valeo. Son objectif général est de développer, en liaison étroite avec les partenaires, une formation et une recherche de niveau international.

Ses quatre principaux axes de recherche sont :

1. Analyse et prévision de séries temporelles (Predictive Analytics on Time Series) ;

2. Exploitation de données hétérogènes, massives et partiellement étiquetées (Exploiting Large Scale and Heterogeneous, Partially Labelled Data) ;

3. Apprentissage pour une prise de décision robuste et fiable (Learning for Trusted and Robust Decision) ;

4. Apprentissage dans un environnement dynamique (Learning through Interactions with a Changing Environment).

Plus d’information sur la chaire : **www.telecom-paris.fr/dsaidis**

Description du stage

Encadrement

Stephan Clémençon

Lieu et dates du stage

Adresse : Télécom Paris, 19 Place Marguerite Perey, 91120 Palaiseau

Date de début du stage : début 2020

Équipe(s) d'accueil de la thèse

Département IDS, équipe 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.

Profil du candidat

Etudiant titulaire d'un master 2 recherche

- Apprentissage statistique / reconnaissance des formes
- Bon niveau en programmation (Java, C/C++, Python)
- Bon niveau d'anglais

Candidatures

A envoyer à stephan.clemencon@telecom-paris.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)

[5] Learning from Biased Training Samples. S. Clémençon & P. Laforgue, (2019)

[6] Weighted Empirical Risk Minimization: an Importance Sampling Approach to Transfer Learning. M. Achab, S. Clémençon, C. Tillier & R. Vogel, (2019)

[7] Empirical Risk Minimization under Random Censorship: Theory and Practice. G. Ausset, S. Clémençon & F. Portier