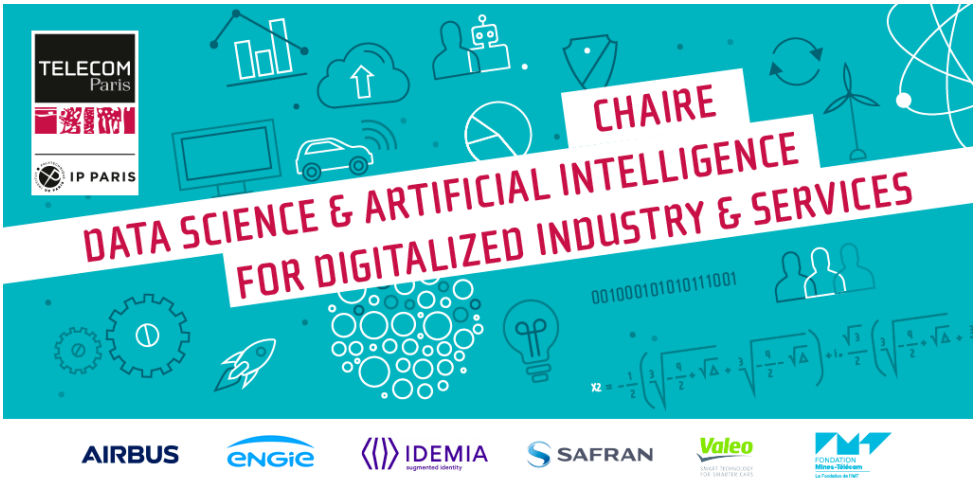

Paris, le 12/01/2022

Offre de stage

Sujet : Definition of conditional quantiles based on data depth

Possibility to continue as a PhD candidate YES

La Chaire Data Science and Artificial Intelligence for Digitized Industry and Services

Portée par Florence d'Alché-Buc, enseignante-chercheur dans le département Image, Données, Signal de Télécom ParisTech, la chaire DSAI réunit cinq partenaires industriels : Airbus Defence & Space, Engie, Idemia, Safran et Valeo Finance. Son objectif général est de développer, en liaison étroite entre les Parties, 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).

Description du stage

Encadrement

Stephan Clémençon (https://perso.telecom-paristech.fr/clemenco/)

Pavlo Mozharovskyi (https://perso.telecom-paristech.fr/mozharovskyi/)

Lieu et dates du stage

Laboratory/Institution: LTCI, Télécom Paris, Institut Polytechnique de Paris

Address: Télécom Paris, 19 Place Marguerite Perey, 91120 Palaiseau, France

Date de début du stage : Beginning 2022

Équipe(s) d'accueil de la thèse

Department IDS; Team Signal, Statistique et Apprentissage (S2A)

Mots clés

Machine learning, data depth, quantiles, local estimators, non-parametric statistics, robustness.

Sujet détaillé

Providing meaningful orderings of data is an important task in many scientific domains, which is reflected in prevailing techniques of statistics and machine learning. Being independent from assumptions on the data generating process and interpretable by its statistical nature, the concept of data depth [1] comes across as a versatile tool for the analysis of current large-scale data applications (*e.g.*, [2,3]). As different notions of data depth exist [4], varying in their statistical [5] and computational [6] properties as well as in their suitability for different applications, they constitute a comprehensive methodology capable to process data of different types from various sources. While data depth has already undergone theoretical and computational developments, most of the proposed depth notions treat the generality of the data (distribution) in a descriptive, unsupervised manner, which makes its direct application in regression problems rather complicated, with supervised classification [7] being among very few examples.

To overcome this limitation, a conditional version of data depth shall be developed. The goal of the current internship is to investigate further in this direction. As the first and very important aim, an overview of the literature is expected at the beginning, starting with such sources as, *e.g.*, [8,9]. This shall allow for a proposal of a localization-based methodology for defining the conditional depth. Further, based on these ideas – for well-manageable distributional assumptions (these can include, *e.g.*, restriction on the probability density shape or even limiting to a specific parametric family) – possible properties of the novel notion of data depth shall be formulated and studied. Existing in the literature freely-accessible data sets shall be used for construction of simple illustrative examples.

Profil du candidat

Student having master 2 research

- Statistical learning, bases of probability
- Knowing of a data-science language (Python)
- Good command of English

Candidatures

To send on stephan.clemencon@telecom-paris.fr and pavlo.mozharovskyi@telecom-paris.fr:

- Curriculum Vitae

- Personalized motivation letter that explains interest of the candidate in the subject (can be directly in the body of the email)

- Grade reports for recent years

- Contact of a person willing to give recommendation

Incomplete applications will not be considered.

Références

[1] Tukey, J.W. (1975). Mathematics and the picturing of data. In: *Proceedings of the International Congress of Mathematicians*, Volume 2, James, R.D. (Ed.), *Canadian Mathematical Congress*, 523–531.

[2] Dutta, S., Sarkar, S. and Ghosh, A. K. (2016). Multi-scale classification using localized spatial depth. *Journal of Machine Learning Research* 17(217), 1–30.

[3] Kleindessner, M. and Von Luxburg, U. (2017). Lens depth function and k-relative neighborhood graph: versatile tools for ordinal data analysis. *Journal of Machine Learning Research* 18(58), 1−52.

[4] Mosler, K. and Mozharovskyi, P. (2021). Choosing among notions of multivariate depth statistics. *Statistical Science*, in press.

[5] Zuo, Y. and Serfling, R. (2000). General notions of statistical depth function. *The Annals of Statistics* 28(2), 461–482.

[6] Pokotylo, O., Mozharovskyi, P., and Dyckerhoff, R. (2019). Depth and depth-based classification with `R`-package `ddalpha`. *Journal of Statistical Software* 91(5), 1–46.

[7] Lange, T., Mosler, K., and Mozharovskyi, P. (2014). Fast nonparametric classification based on data depth. *Statistical Papers* 55, 49–69.

[8] Hallin, M., Paindaveine, D., and Šiman, M. (2010). Multivariate quantiles and multiple-output regression quantiles: from L_1 optimization to halfspace depth. *The Annals of Statistics* 38(2), 635–669.

[9] Carlier, G., Chernozhukov, V. and Galichon, A. (2016). Vector quantile regression: An optimal transport approach. *The Annals of Statistics* 44(3), 1165–1192.