Bootstrap for local inference: a nearest neighbors approach

This is an announcement for an internship funded by the "chaire" **Data Science and Artificial Intelligence for Digitalized Industry and Services**. Doing a PhD after the internship is possible.

Location: Télécom Paris

Duration: 6 months

Supervisor: François Portier

Abstract: The bootstrap is a useful resampling method that first emerged as a statistical tool for making inference on some estimated parameter (Efron, 1979; Bickel et al., 1981). Recently, the bootstrap and more generally the use of resampling methods has been at the core of many modern applications withing the machine learning framework. For instance it has been used to stabilize prediction (Bühlmann et al., 2002), to save computing time in Monte-Carlo simulation (Delyon and Portier, 2020), to calibrate tests of conditional independence (Bianchi et al., 2020).

When facing a local estimation problem, i.e., when one is interesting in the conditional distribution given an event (rather than in the full distribution), the use of resampling methods is still subjected to many challenging questions. In particular, although the bootstrap of the empirical expectation has been well studied, only a few research works have been carried-out about the conditional expectation. A particularly relevant tool to deal with local estimation problem is the nearest neighbors estimate. However, the asymptotic variance of the nearest neighbor predictor is in practice hard to estimate and a bootstrap approach would be useful if not necessary. A first goal in this internship is to provide theoretical guarantees on bootstrap procedures based on the nearest neighbors estimate of the conditional distribution (see (Bianchi et al., 2020) for a definition of this estimate). A further avenue of research is the one of bagging nearest neighbors predictor in order to stabilize the estimation.

During the internship, the student will design and study new resampling methods to reproduce the behavior of certain local estimate such as nearest

neighbor. He or she will compare the proposed method to other state of the art methods in real problems.

For the internship, a strong background in ML and in mathemaical statistics is required. Especially some knowledge on weak convergence and kernel smoothing are appropriate. To apply send an email at **francois.portiergmail.com** including your cv, a description of the grades you obtain last year, and describing your motivations.

References

Bianchi, P., K. Elgui, and F. Portier (2020). Conditional independence testing via weighted partial copulas. *arXiv preprint arXiv:2006.12839*.

Bickel, P. J., D. A. Freedman, et al. (1981). Some asymptotic theory for the bootstrap. *The annals of statistics 9*(6), 1196–1217.

Bühlmann, P., B. Yu, et al. (2002). Analyzing bagging. *The Annals of Statistics 30*(4), 927–961.

Delyon, B. and F. Portier (2020). Safe adaptive importance sampling: a mixture approach. *Annals of statistics*.

Efron, B. (1979). Bootstrap methods: Another look at the jackknife. *The Annals of Statistics 7*(1), 1–26.