ID de Contribution: **1** Type: **Non spécifié**

Supervised Learning with Missing Values

vendredi 5 février 2021 10:30 (40 minutes)

Some data come with missing values. For instance, a survey's participant may ignore some questions. There is an abundant statistical literature on this topic, establishing for instance how to fit model without biases due to the missingness, and imputation strategies to provide practical solutions to the analyst. In machine learning, to build models that minimize a prediction risk, most work default to these practices. As we will see, these different settings lead to different theoretical and practical solutions.

I will outline some conditions under which machine-learning models yield the best-possible predictions in the presence of missing values. A striking result is that naive imputation strategies can be optimal, as the supervised-learning model does the hard work [1]. A challenge to fitting a machine-learning model is that there is a combinatorial explosion of possible missing-values patterns such that even when the output is a linear function of the fully-observed data, the optimal predictor is complex [2]. I will show how the same dedicated neural architecture can approximate well the optimal predictor for multiple missing-values mechanisms, including difficult missing-not-at-random settings [3].

[1] Josse, J., Prost, N., Scornet, E., & Varoquaux, G. (2019). On the consistency of supervised learning with missing values. arXiv preprint arXiv:1902.06931.

[2] Le Morvan, M., Prost, N., Josse, J., Scornet, E., & Varoquaux, G. (2020). Linear predictor on linearly-generated data with missing values: non consistency and solutions. AISTATS 2020.

Orateur: VAROQUAUX, Gaël (INRIA Parietal)