Classification with missing entities / Retrieving the missing ones in a forest

Keywords: missing values, supervised learning, nonparametric methods, linear models.

Missing data are present in most, if not all, real-world data sets. This is due for instance to "forgot to fill in the form" entry, failure of the measuring device, no time to measure in an emergency situation, aggregating data sets from multiple sources. Numerous works focused on inferring parameters (for example, that of a linear model) in presence of missing values, trying to uncover the true signal even in the absence of several entries of the design matrix. A new and promising avenue for research in the field of missing values is to consider them in a supervised learning framework, in which the aim is to predict the best value for an output, and not to recover the parameter of the true underlying distribution [1].

Recent works have shown that imputing missing data, that is replacing the missing values by some quantities (either deterministic, for example zero, or random, for example the mean) is asymptotically efficient in a prediction purpose [2]. Quite surprisingly, the story is different in an inferential framework in which one must not impute data as it distorts the data distribution (lowering the variance). Therefore, supervised learning with missing values offers a new point of view and exciting questions.

First axis. Recent works have shown that even linear regression with missing values is quite challenging, because of the exponential number of missing value patterns that may appear in a data set. In this context, a new simple method consisting in performing Ordinary Least Squares on the most frequent missing data patterns was proved to be optimal [3]. However, nothing has been done yet in the context of classification. The aim of this internship is to study what challenges arise in the problem of classification with missing values and to try to extend the framework of linear regression to logistic regression. Comparison of various existing methods (single/multiple imputation, non-parametric methods, EM, see [4]) will provide practical guidance on the procedure to use when facing classification with missing data.

Second axis. Approximation properties of linear models remain weak, as they depend on a very small number of parameters. In supervised learning, non-parametric methods as tree-based methods (random forests, for example) are often preferred for their versatility to handle various types of data. A popular approach based on imputation with random forests, called quite logically MissForest [5] consists in iteratively predicting one input variable based on the other ones. This method shows superb performances in practice, but there is very little theoretical ground justifying these performances [6]. The aim of the internship will be to first understand in detail the method and propose a setting in which its performance can be understood from a theoretical perspective. Experiments will illustrate the theoretical findings.

The choice between the first and second axis will be discussed with the successful candidate. For each axis, the subject combines two aspects of a scientific work: on the one hand, a more methodological development could lead to efficient algorithms; on the other hand, a more thorough theoretical study of this issue will allow establishing nice statistical results. Both aspects are important, and can be modulated according to the candidate's affinities.

Supervisors: Claire Boyer (Sorbonne Université), Aymeric Dieuleveut (Ecole Polytechnique), Erwan Scornet (Ecole Polytechnique)

Required skills: M1 or M2 level trainee in statistics/machine learning/optimization. Applicants should send CV, transcripts of the last two years and the name of a referee to
claire.boyer@sorbonne-universite.fr,
aymeric.dieuleveut@polytechnique.edu,
erwan.scornet@polytechnique.edu

Practical information: the internship will take place at LPSM (Sorbonne Université) in the statistical team. This is a 6-month internship that can start at the beginning of April.

[1] On the consistency of supervised learning with missing values. Josse, Prost,, Scornet & Varoquaux (2019).
[2] What's a good imputation to predict with missing values?. Le Morvan, M., Josse, J., Scornet, E., & Varoquaux, G. (2021). Published in NeurIPS
[3] Near-optimal rate of consistency for linear models with missing values. Ayme, A., Boyer, C., Dieuleveut, A., & Scornet, E. (2022, June). Published in ICML.
[4] Stochastic approximation em for logistic regression with missing values. Jiang, W., Josse, J., Lavielle, M., & Gauss, T. (2018). *arXiv preprint arXiv:1805.04602*.
[5] MissForest—non-parametric missing value imputation for mixed-type data. Stekhoven, D. J., & Bühlmann, P. (2012). *Bioinformatics*, *28*(1), 112-118.
[6] On the stationary distribution of iterative imputations. Liu, J., Gelman, A., Hill, J., Su, Y. S., & Kropko, J. (2014) *Biometrika*, *101*(1), 155-173.