Research internship: Classification with missing entities / EM algorithm

1 Introduction

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 devices, 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 [3].

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 [4]. 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 - Classification 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. Under various assumption on the missing mechanisms and the distribution of the full input vector, we know that the Bayes predictor on each missing pattern remains linear. Unfortunately, this property is not valid for logistic regression, even in the simple case of Gaussian data [6] and missing completely at random (MCAR) mechanisms. A natural question arises: can we identify settings in which the Bayes predictor on each missing pattern takes the form of a logistic regression? Other methods like linear Discrimant analysis [6] and variants of the EM algorithms [2] are promising but come at the price of strong assumption on the input distribution or the missingness mechanisms. More practically, we would

like to propose guidelines to handle binary classification with missing values in all sorts of settings (input/missing distributions).

Second axis - EM algorithm EM algorithm has been successfully applied to missing data in the context of linear or logistic regression. Statistical guarantees on EM algorithms are notoriously difficult to obtain. Recently, several works focus on providing asymptotic/non-asymptotic bounds on the parameters resulting from EM strategies [see 1]. Extending these results to the case of linear regression with missing values would allow us to compare rate of convergence of different procedures in a regression setting, and thus to identify the more efficient procedures.

Third axis - MissForest. 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 [7] consists in iteratively predicting one input variable based on the other ones. This method shows excellent performance in practice, but there is very little theoretical ground to justify these performances [5]. 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 axes 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 one to establish nice statistical results. Both aspects are important and can be modulated according to the candidate's affinities.

Supervisors : Claire Boyer (LMO, Université Paris-Saclay), Rafaël Pinot (LPSM, Sorbonne Université), Erwan Scornet (LPSM, Sorbonne Université)

Required skills: M1 or M2 level trainee in statistics/machine learning. Applicants should send CV, transcripts of the last two years, and the name of a referee to

- claire.boyer@sorbonne-universite.fr
- rafael.pinot@sorbonne-universite.fr
- 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 in early April. This internship can be followed by a Ph.D. thesis.

References

[1] Raaz Dwivedi, Koulik Khamaru, Martin J Wainwright, Michael I Jordan, et al. Theoretical guarantees for em under misspecified gaussian mixture models. *Advances in Neural Information Processing Systems*, 31, 2018.

[2] Wei Jiang, Julie Josse, Marc Lavielle, TraumaBase Group, et al. Logistic regression with missing covariates—parameter estimation, model selection and prediction within a joint-modeling framework. *Computational Statistics & Data Analysis*, 145:106907, 2020.

[3] Julie Josse, Nicolas Prost, Erwan Scornet, and Gaël Varoquaux. On the consistency of supervised learning with missing values. *arXiv preprint arXiv:1902.06931*, 2019.

[4] Marine Le Morvan, Julie Josse, Erwan Scornet, and Gaël Varoquaux. What'sa good imputation to predict with missing values? *Advances in Neural Information Processing Systems*, 34:11530–11540, 2021.

[5] Jingchen Liu, Andrew Gelman, Jennifer Hill, Yu-Sung Su, and Jonathan Kropko. On the stationary distribution of iterative imputations. *Biometrika*, 101(1):155–173, 2014.

[6] Angel D Reyero Lobo, Alexis Ayme, Claire Boyer, and Erwan Scornet. Harnessing pattern-by-pattern linear classifiers for prediction with missing data. *arXiv preprint arXiv:2405.09196*, 2024.

[7] Daniel J Stekhoven and Peter Bühlmann. Missforest—non-parametric missing value imputation for mixed-type data. *Bioinformatics*, 28(1):112–118, 2012.