Supervised learning with missing values.

Supervision:

- Marine Le Morvan, Soda, Inria (marine.le-morvan@inria.fr)
- Gaël Varoquaux, Soda, Inria (gael.varoquaux@inria.fr)

Keywords: missing values, representation learning, neural networks, supervised learning, statistics

Context In applications –health, business, social sciences, . . . – the pervasiveness of missing values hinder the use of machine learning. In health for instance, two patients rarely undergo the exact same series of exams, doctors do not always have time to record the information, . . . Surveys suffer from non-responses. . .

Missing data has been heavily studied by the statistical literature, mostly focused on the estimation of model parameters and their variances in the presence of missing values, as well as imputation techniques, where imputation is concerned with replacing the missing entries with likely values. When the missingness occurs at random, imputation leads to the same parameter inference as on fully-observed data [5]. However, supervised learning with missing values leads to different tradeoffs which have been much less studied.

To date, the standard practice for learning in the presence of missing values consists in first imputing the missing values, and then learning on the completed data. Research on imputation techniques has been prolific in recent times, ranging from optimal transport-based techniques [7] to Generative Adversarial Networks [9] or Variational Auto-encoders [6]. However, many of these methods are either hard to train or computationally expensive. It is thus of interest to understand whether the computational efforts for imputing are worthwhile in terms of downstream performances. For supervised learning, imputation should be seen through the angle of **representation learning in the presence of missing values**. This internship will attempt to answer the following questions: **Is there a correlation between imputation quality and downstream performance on supervised tasks? What is the cost-benefit tradeoff of conditional imputation techniques for supervised learning?** Theoretically, it has been shown recently that low-quality imputations, such as the mean, are enough for learning the best possible predictors given enough data and a sufficiently rich prediction prediction model [3]. Here we would like to explore finite-sample regimes behaviors.

Aside from impute-then-regress approaches, recent attempts have proposed to learn the imputation and prediction functions jointly. This is for example the case of NeuMiss [2, 3], a **neural network architecture designed for learning with missing values**. Originally, NeuMiss was designed under the hypothesis of a linear-Gaussian model. In this internship, we will explore how this approach can be adapted to real data settings, and whether its superiority over two-steps approaches holds beyond linear-Gaussian models.

Environment This internship will take place at Inria Saclay, in the Soda team. Soda is doing computational and statistical research, both fundamental and applied, to harness large databases on health and society. Soda is also developing core software tools such as scikit-learn. The team now has a lot of experience with missing values [1, 4, 2, 3, 8]. This internship topic can be pursued by a PhD.

Requirements We are seeking a highly motivated student with a strong interest in statistical aspects of machine learning and an affinity for numerical experimentation. Factors of success include:

- Proficiency in Python.
- Knowledge of PyTorch is a plus.
- Good mathematical background.
- Curious mindset.

References

[1] Julie Josse, Nicolas Prost, Erwan Scornet, and Gaël Varoquaux. On the consistency of supervised learning with missing values, 2020.

[2] Marine Le Morvan, Julie Josse, Thomas Moreau, Erwan Scornet, and Gaël Varoquaux. Neumiss networks: differentiable programming for supervised learning with missing values, 2020.

[3] Marine Le Morvan, Julie Josse, Erwan Scornet, and Gaël Varoquaux. What's a good imputation to predict with missing values?, 2021.

[4] Marine Le Morvan, Nicolas Prost, Julie Josse, Erwan Scornet, and Gaël Varoquaux. Linear predictor on linearly-generated data with missing values: non consistency and solutions. In *International Conference on Artificial Intelligence and Statistics*, pages 3165–3174. PMLR, 2020.

[5] Roderick JA Little and Donald B Rubin. *Statistical analysis with missing data*, volume 793. John Wiley & Sons, 2019.

[6] Pierre-Alexandre Mattei and Jes Frellsen. Miwae: Deep generative modelling and imputation of incomplete data sets. In *International conference on machine learning*, pages 4413–4423. PMLR, 2019.

[7] Boris Muzellec, Julie Josse, Claire Boyer, and Marco Cuturi. Missing data imputation using optimal transport. In *International Conference on Machine Learning*, pages 7130–7140. PMLR, 2020.

[8] Alexandre Perez-Lebel, Gaël Varoquaux, Marine Le Morvan, Julie Josse, and Jean-Baptiste Poline. Benchmarking missing-values approaches for predictive models on health databases. *GigaScience*, 11, 2022.

[9] Jinsung Yoon, James Jordon, and Mihaela Schaar. Gain: Missing data imputation using generative adversarial nets. In *International conference on machine learning*, pages 5689–5698. PMLR, 2018.