Multiple Imputation for supervised learning.

Supervision:

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

Context In applications –health, business, social sciences, . . . – the pervasiveness of missing values hinder the use of machine learning. Electronic Health records are typically plagued with missing values. The reasons are multiple: two patients rarely undergo the exact same series of exams, doctors do not always have time to record the information, . . . Surveys are also subject to missingness issues, due to non-responses.

Since the 70s, an abundant statistical literature on missing data has flourished. This literature has been mostly focused on 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 [6]. However, missing values in a supervised-learning setting has been much less studied.

For inference in the presence of missing values, it is widely known that single imputation distorts distributions and leads to biased estimates. Rather Multiple Imputation techniques are commonly used, giving Monte-Carlo approximations of the distribution of missing values. In supervised learning, imputation is also widely used as a preprocessing step. It would be tempting to think that as in the case of inference, single imputation should be avoided and Multiple Imputation should be preferred. However, recent results have shown that single imputation, even by a constant, allows optimal prediction results given a sufficiently powerful learner [5, 2]. As for Multiple Imputation, it is not even clear as of today what it means in a supervised learning setting. For example, [3] presents many different ways to implement it using bagging. There is thus a need to study what aspect of the distribution of missing values is important in a supervised learning setting.

In terms of estimation procedures in the presence of missing values, a number of deep learning architectures have recently emerged. It includes NeuMiss [4, 5], a ResNet-based architecture with a new kind of non-linearities that can approximate good imputations in a differentiable way; supMIWAE [1], a Variational Autoencoder-based architecture that mimicks Multiple Imputation; as well as permutation-invariant input layers [7].

Methods A first goal of this internship will be to clarify what Multiple Imputation means in a supervised learning setting, and whether it benefits theoretical guarantees. We will consider several simple procedures combining sampling with supervised learning. We will study them both in asymptotic settings (consistency), as well as finite-sample considerations such as the complexity of the function to learn on the imputed data or applying results from ensemble prediction. A second goal of this internship will be to establish which practical procedures provide the best results on large tabular databases among (Multiple)Impute-then-learn procedures. This will require benchmarking the procedures on real data, including in the benchmark the new deep learning architectures which mimick such procedures [1].

Environment The internship will take place in Inria Saclay, in the Parietal team. This is a large team focused on mathematical methods for statistical modeling of brain function and health, developing core software tools such as scikit-learn. The internship will be set in the DirtyData project.

Requirements

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

References

[1] Niels Bruun Ipsen, Pierre-Alexandre Mattei, and Jes Frellsen. How to deal with missing data in supervised deep learning? In *Artemiss - ICML Workshop on the Art of Learning with Missing Values*, Vienne, Austria, July 2020.

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

[3] Shehroz S. Khan, Amir Ahmad, and Alex Mihailidis. Bootstrapping and multiple imputation ensemble approaches for classification problems. *J. Intell. Fuzzy Syst.*, 37:7769–7783, 2019.

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

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

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

[7] Chao Ma, Sebastian Tschiatschek, Konstantina Palla, Jose Miguel Hernandez Lobato, Sebastian Nowozin, and Cheng Zhang. EDDI: Efficient dynamic discovery of high-value information with partial VAE, 2019.