INRIA Saclay,
INRIA Montpellier

https://project.inria.fr/dirtydata

Fast MICE: speeding up the most popular imputation method

Research theme: Machine learning, data science
Keywords: Missing values, stochastic optimization, gibbs sampling,
Duration & salary: 3 to 6 months, between 500 €and 800 € monthly
Research teams: INRIA Saclay (Parietal team) and INRIA Montpellier (J. Josse)
Adviser: Gaël Varoquaux & Julie Josse
Contact: gael.varoquaux@inria.fr, julie.josse@inria.fr
Application: Interested candidate should send CV and motivation letter

Context: Due to the difficulty of controlling the surveying, assembling, or measuring, data often come with missing values: some of the observations have only a fraction of the features measured. Standard statistical or machine-learning models can not longer be applied on such data. A common approach to circumvent the problem and recover valid statistical analysis is to use missing-values *imputation*: the predictive distribution of the unobserved values given the observed values and an (implicit) imputation model is computed and used to create a new dataset where missing values are replaced by plausible values. MICE [1] is probably the most popular imputation approach. This popularity is justified by its flexibility, and its success without much parameter tuning. MICE [2] works by using iteratively machine-learning models to predict missing values in one feature from the other features.
The drawback of MICE is its computational cost. It needs to fit a number of base machine-learning models scaling as $\mathcal{O}(p)$ where p is the number of features. As the cost of a machine-learning model is at least $\mathcal{O}(n \cdot p)$ –where n is the number of sample– typically more, the cost of fitting scales at least as $\mathcal{O}(n \cdot p^2)$. Using as a base model a ridge regression –cost of $\mathcal{O}(n \cdot p \min(n, p))$– leads to a total cost of $\mathcal{O}(n \cdot p^2 \min(n, p))$. The resulting costs are intractable in many modern settings where n is large (hundreds of thousands) and p is not small (hundreds).

Proposed work:
We propose to tackle the problem of fitting multiple base models on large datasets more efficiently. For this, we propose two alleys. The first one will take a stochastic approximation point of view, for instance fitting the models on subsamples of the total data. The second one will consider using multi-output machine learning models, to share the computational cost across several output features.

Required skills:

- Knowledge of statistics, machine learning, or applied maths background (mathematical optimization, algebra and statistics)
- Some skills in numerical programming

[1] S van Buuren and Karin Groothuis-Oudshoorn. mice: Multivariate imputation by chained equations in r. *Journal of statistical software*, pages 1–68, 2010.

[2] S. van Buuren. *Flexible Imputation of Missing Data. Second Edition.* CRC Press, Boca Raton, FL., 2018.