INRIA Saclay DirtyData project https://project.inria.fr/dirtydata AppStat LAL, bât 200, Orsay

Machine learning for data-science with missing data

Research theme: machine learning, database, missing data
Keywords: Distributional semantic, graph embeddings.
Duration & salary: 24 months, 2500€ monthly or more depending on experience
Research team: AppStat (LAL, CNRS) & Parietal (INRIA Saclay), collaboration with Julie Josse (École Polytechnique)
Supervisor: Balazs Kegl, Gael Varoquaux
Contact: balazs.kegl@gmail.com, gael.varoquaux@inria.fr
Application: Interested candidate should send CV and motivation letter

Context: "data science" leverages data that is often observational and compound, rather than experimental and homogeneous. As such, it encounters many missing value problems: features can be unobserved for some data points. A wider problem is that of data cleaning: data scientist spend most of their time preparing the data for analysis, rather than doing the actual analysis. The "dirty data project" tackles this problem, to facilitate analysis of non-curated data.
While missing data is a classic problem is statistics, new settings and new tools bring new challenges and new opportunities. Challenges arise from the heterogeneity of data, which breaks classic missing-data assumptions such as "missing at random" [1] and creates structured missingness patterns. Opportunities arise from recent machine-learning models, that can both detect these patterns and provide good candidate replacements for data imputation. For an analysis, missing data imputation should be tackled as one step in a whole pipeline.

Research projects: The goal of this post-doc is to explore the specificities of modern missing-data challenges: the heterogeneity of the data and how a full analysis pipeline can best deal with incomplete data. There are several alleys that can be considered:

Empirical work : most empirical study have been conducted on fairly clean or synthetic data (as from the UCI machine learning repository). Such studies do not capture the structured patterns of missingness that arise, *eg* from merging multiple data sources. Via collaborators of the DirtyData project, we have access to unique sources of real complex data with missing data in healthcare, epidemiology, and business.

The questions are:

- How to characterize the structure of missingness empirically?
- What are the typical mechanisms (censoring, imperfect merge between databases, non-relevant attribute)?
- How to impute missing data best to improve predictive analytics?

Controling impact on conclusions When missing-data is structured, it creates biases in the conclusions drawn with standard analysis frameworks. Classic assumptions to control for these biases (missing at random) are too restrictive in real-world scenario.

- "pretext tasks" to probe for biases and correct them.

- Missing-data as a selection bias problem, and adapting solutions from the importance-weighting literature, combined with multiple imputation.
- Adapting causal frameworks (counter factual analysis, mediation analysis, and instrumental variables) for missing-data situations

Machine learning for imputation More powerful models, as developed in machine learning can benefit from the large sample sizes to impute missing values. There is already promising work in this direction [2]. The new developments that we want to tackle are:

- Making use of data type and structure (in databases, which have very heterogeneous and structured attributed)
- Using these models for efficient and statistically-controlled multiple imputation.

Candidate:
We are seeking a motivated individual, with a strong ability to learn and a desire to work in a group. Given that part of the work is empirical, we are open to **candidates with a scientific background outside machine learning and statistics** for instance in physics or any other natural science with a mathematical or a data-processing culture.
The ideal candidate will:

- be good at data manipulation or scientific programming with a high-level language
- have a good level of written English
- have a PhD with a data-processing, statistics, or mathematical aspect

Desired but non-essential skills:

- Good Python programming skills
- A statistical culture and mindset
- Knowledge in mathematical optimization
- Knowledge in non-parametric statistics
- Experience in designing machine-learning algorithms
- Experience in using machine-learning algorithms
- Practical data-science experience

Given that there are several possible research alleys (listed above), we are open to different profiles, that match the various aspects of our research agenda.

References

[1] Donald B Rubin. Inference and missing data. *Biometrika*, 63(3):581–592, 1976.

[2] Jinsung Yoon, James Jordon, and Mihaela van der Schaar. Gain: Missing data imputation using generative adversarial nets. *ICML*, 2018.