Bayesian-Neural Methods for Missing Data Imputation with Applications in Bioinformatics

Main Supervisor: Jesse Read – https://jmread.github.io/

LIX Laboratory, Ecole Polytechnique, Institute Polytechnique de Paris, France

Topic Description

Missing data is a universal problem in data science and machine learning, and impacts many domains. For example, genetic and genomic studies can involve SNP (Single Nucleotide Polymorphism) datasets where associations between genetic variants and traits (usually, diseases) are investigated; and such datasets are prone to a number of missing values caused by, e.g., inaccurate sequencing, an imbalance between dominant and recessive variants, and/or a lack of overlapping regions between datasets (when different studies are combined).

Imputation can be approached as a machine learning task itself, by building models and using them to predict the missing values. These models can predict single (one at a time) or multiple values (i.e., multi-output and structured prediction), either within a common row-instance or across the dataset within a common feature-column. Deep neural networks are a popular and successful class of model in general, particularly when multiple outputs are involved, yet usually these methods only provide point estimates. When imputing values there are many scenarios where we wish to measure our uncertainty of the value and other aspects of interpretation surrounding the prediction, to be taken into account by a human domain expert or a separate prediction algorithm.The same mechanism can be used to check the 'correctness' of existing values.

This proposed topic balances between the areas of machine learning and Bayesian inference. There are a number of methods found on this boundary that can be of interest, and among them this project would target specifically Bayesian Neural Networks. We want not only to test and further develop this approach for missing values imputation (and specifically, in the domain of interest of SNP data), but also explore ways to leverage uncertainty information from the imputation task in a separate classification/regression task (using the imputed data). Although the main application domain is SNP datasets, methods can be tested in other domains including medical data (which we have available) and other other tasks related to imputation, such as anomaly detection and recommendation systems.

Keywords: Machine learning, Missing value imputation, Deep neural networks, Bayesian inference, Bioinformatics, SNP data

Necessary Skills

Knowledge of and experience in machine learning, including at least one deep-learning framework (e.g., TensorFlow or PyTorch) and scientific programming in Python (including use of libraries such as Numpy), and specifically awareness of probabilistic views of inference, including Bayesian methods. Although not a requirement, some relevant bioinformatics knowledge could make the topic more appreciable – however, as missing data is relevant to many areas, it is possible to also pursue other domain interests.

Practical Details

Supervision from Jesse Read (Professor at LIX) and with involvement from Ekaterina Antonenko (PhD candidate at LIX). The project can be adapted to either Bachelor or Master level (please contact to discuss). The mode of work will be initially online, but travel and stay at the LIX laboratory will be a possibility to explore, depending on the local restrictions and recommendations respective of the global pandemic, and funding availability.

For questions or discussion, contact: jread@lix.polytechnique.fr

References

[1] Nicole S. Erler. *Bayesian Imputation of Missing Covariates*. 2019.

[2] Yang Guo, Zhengyuan Liu, Pavitra Krishnswamy, and Savitha Ramasamy. Bayesian recurrent framework for missing data imputation and prediction with clinical time series. January 2020.

[3] Laurent Valentin Jospin, Wray Buntine, Farid Boussaid, Hamid Laga, and Mohammed Bennamoun. Hands-on bayesian neural networks – a tutorial for deep learning users. September 2021.

[4] Meng Song, Jonathan Greenbaum, Joseph Luttrell IV, Weihua Zhou, Chong Wu, Hui Shen, Ping Gong, Chaoyang Zhang, and Hong-Wen Deng. A review of integrative imputation for multi-omics datasets. *Frontiers in genetics*, 11, 2020.