Quantifying the uncertainty of any algorithm handling missing values with a conformal procedure

Keywords : confidence; conformal prediction; missing data; application on a real medical dataset.

Objective The increasing quantity of available data, coming from multiple sources, is a real opportunity to better understand and anticipate many phenomena. However, it comes hand to hand with the multiplication of missing data. There is a rich literature on how to impute missing values and how to conduct a statistical inference in presence of missing values [Little and Rubin, 2019], for example considering the EM algorithm [Dempster et al., 1977], low rank models [Sportisse et al., 2020], random forests [Stekhoven and Bühlmann, 2012] or deep learning techniques with variational autoencoders [Mattei and Frellsen, 2019]. Most methods propose a single imputation or provide a single estimation, which does not account for the uncertainty of the algorithm. On the contrary, multiple imputation [Rubin, 2004] allows to get confidence intervals, but its main drawback is that for each algorithm, specific rules must be defined to combine the results.

Introduced by [Vovk et al., 1999], conformal prediction is a very promising technique for building predictive intervals for arbitrary machine learning models, which have the great advantage to be valid in finite sample and without strong assumptions on the data distribution. In recent years, conformal prediction has emerged as a key framework for quantifying uncertainty in machine learning algorithms, in particular with the development of split conformal prediction [Lei et al., 2018], which considerably reduces the computational cost, and with recent works that allow to go beyond the classical assumption of exchangeable data [Tibshirani et al., 2019, Zaffran et al., 2022, Barber et al., 2022].

The objective of this internship will be to propose a conformal procedure to quantify the uncertainty of any algorithm handling missing values. It will be illustrated on the Traumabase dataset (Assistance Publique - Hôpitaux de Paris), a public clinical dataset on the management of polytraumatised patients who have suffered a major trauma (30,000 individuals, 250 clinical variables, containing many missing values). The goal is to assist doctors in making decisions in emergency situations (e.g. administration of an active substance); in this medical context, it is essential to be able to quantify the uncertainty of the results given by the algorithm. This intership will involve both theoretical work (the candidate should have a strong background in statistics/machine learning) and to implement the proposed methods in Python and/or R (depending on the candidat's skills and will). The code might be also integrated to the R-miss-tastic platform, a resource website on missing values.

Context of the internship The intern will join the Maasai team of Inria Sophia-Antipolis and Université Côte d'Azur, which is composed of 25 researchers in statistical and machine learning (web: `https://team.inria.fr/maasai/`). The team is part of the Institut 3IA Côte d'Azur `https://3ia.univ-cotedazur.eu/`, which offers a lot of opportunities (thesis offers, seminars & meetings with PhD students/postdoc in machine learning).
A collaboration with Claire Boyer (LPSM, Sorbonne University) will also be considered.
Duration: 6 months
Salary: approx. 550€ / month
PhD opportunities within the Maasai team may be pursued after the intership, to continue this work.

Contact To apply, please contact Aude Sportisse (aude.sportisse@inria.fr) and Pierre-Alexandre Mattei (pierre-alexandre.mattei@inria.fr).

References

Rina Foygel Barber, Emmanuel J Candes, Aaditya Ramdas, and Ryan J Tibshirani. Conformal prediction beyond exchangeability. *arXiv preprint arXiv:2202.13415*, 2022.

Arthur P Dempster, Nan M Laird, and Donald B Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, 39(1): 1–22, 1977.

Jing Lei, Max G'Sell, Alessandro Rinaldo, Ryan J Tibshirani, and Larry Wasserman. Distribution-free predictive inference for regression. *Journal of the American Statistical Association*, 113(523): 1094–1111, 2018.

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

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.

Donald B Rubin. *Multiple imputation for nonresponse in surveys*, volume 81. John Wiley & Sons, 2004.

Aude Sportisse, Claire Boyer, and Julie Josse. Estimation and imputation in probabilistic principal component analysis with missing not at random data. *Advances in Neural Information Processing Systems*, 33:7067–7077, 2020.

Daniel J Stekhoven and Peter Bühlmann. Missforest—non-parametric missing value imputation for mixed-type data. *Bioinformatics*, 28(1):112–118, 2012.

Ryan J Tibshirani, Rina Foygel Barber, Emmanuel Candes, and Aaditya Ramdas. Conformal prediction under covariate shift. *Advances in neural information processing systems*, 32, 2019.

Volodya Vovk, Alexander Gammerman, and Craig Saunders. Machine-learning applications of algorithmic randomness. 1999.

Margaux Zaffran, Olivier Féron, Yannig Goude, Julie Josse, and Aymeric Dieuleveut. Adaptive conformal predictions for time series. In *International Conference on Machine Learning*, pages 25834–25866. PMLR, 2022.