Advances in Positive Unlabelled data modeling

Jan Mielniczuk

Institute of Computer Science, Polish Academy of Sciences

miel@ipipan.waw.pl, https://home.ipipan.waw.pl/j.mielniczuk/

Partially observed data such as Positive Unlabeled (PU) data is frequently encountered and analysed in e.g. text mining, medicine and ecology (Bekker, Davis (2020)). However, the properties of modelling approaches are frequently not thoroughly investigated leading to problems with valid inference for such data (Łazęcka et al (2021)). The present project aims to answer some questions in this field. Firstly, properties of Maximum Likelihood estimators for parametric PU models and its regularised versions will be investigated under SCAR (Selected Completely at Random) assumption when optimisation is performed using Maximisation-Minimisation (MM) or Convex-Concave (CC) approach. Secondly, the consequences of misspecification will be investigated when the the model for aposteriori probabability of success is a single index model with the unknown response function. Thirdly, the situation when SCAR assumption is not valid will be analysed and modelled using parametric models of propensity score.
Successful candidate needs to have MSc in Computer Science or Mathematics, exhibit proficiency in programming, have some hands-on experience in Machine Learning/Statistics and embedded motivation to solve computational and analytical problems.

References

1. J. Bekker, J. Davis (2020) Learning from positive and unlabeled data: a survey. *Machine Learning*

2. P. Teisseyre, J. Mielniczuk, M. Łazęcka (2020) Different strategies of fitting logistic regression for positive and unlabelled data, *Proceedings of the International Conference on Computational Science ICCS'20*

3. M. Łazęcka, J. Mielniczuk, P. Teisseyre (2021) Estimating the class prior for positive and unlabelled data via logistic regression, *Advances in Data Analysis and Classification, to appear*