Master Thesis Proposal: Interaction-Aware Feature Selection with an Information-theoretic Approach

Lincen Yang

October 2022

1 Project Description

Feature selection, i.e., the task of selecting a subset of relevant features from all features, is a crucial task in mostly all kinds of machine learning tasks, as this leads to simpler models in general, which not only generalizes better to new data points but also gives better interpretability.

Intuitively, the optimal subset of features can be defined as the subset of features that preserves all information for the (functional) relationship between the target variable and all the features in the original dataset. This can be formally defined using information theory, and specifically characterized by the *mutual information (MI) and conditional mutual information (CMI)*. An implementation of this idea can be found in [1].

However, estimating MI and CMI is an extremely challenging task, especially when the dimensionality of features is high and/or when the feature types are *mixed* with both continuous and discrete variables. The state-of-the-art method for estimating MI and CMI with mixed type data is based on *MDL-based histogram models* [2], where MDL stands for the *minimum description length principle* [5, 3, 4].

The object of this project is to develop a high-dimensional CMI estimation method for interaction-aware feature selection for possibly mixed-type features. Specifically, the method will be built upon [2]. The core task for the student is to develop efficient algorithms to tackle this challenging problem, and empirically compare its performance against other existing methods.

2 Requirements

We are looking for a highly self-motivated student who wants to participate in the cutting-edge research, and who has very good programming skills in Python or R. Familiarity with one or more of the following is a plus: 1) information theory, 2) advanced statistical learning, 3) the MDL principle. The student

will be working with Lincen Yang and Dr. Matthijs van Leeuwen. If you are interested, please contact us by email: l.yang@liacs.leidenuniv.nl;

3 References

[1] Shishkin, Alexander, et al. "Efficient high-order interaction-aware feature selection based on conditional mutual information." Advances in neural information processing systems 29 (2016).
[2] Marx, Alexander, Lincen Yang, and Matthijs van Leeuwen. "Estimating conditional mutual information for discrete-continuous mixtures using multi-dimensional adaptive histograms." Proceedings of the 2021 SIAM International Conference on Data Mining (SDM). Society for Industrial and Applied Mathematics, 2021.
[3] Grünwald, Peter D. The minimum description length principle. MIT press, 2007.
[4] Kontkanen, Petri, and Petri Myllymäki. "MDL histogram density estimation." Artificial intelligence and statistics. PMLR, 2007.
[5] Grünwald, Peter, and Teemu Roos. "Minimum description length revisited." International journal of mathematics for industry 11.01 (2019): 1930001.