Bachelor/Master Thesis

Karlsruher Institut für Technologie

Institut für Programmstrukturen und Datenorganisation (IPD)
Lehrstuhl für Systeme der Informationsverwaltung, Prof. Böhm

Bachelor/Master Thesis

Meta-Learning Feature Importance

Feature selection is an important step in machine learning pipelines. In the context of this thesis, we consider feature selection for classification tasks. Training classifiers with fewer features is faster, and the models are more comprehensible. Furthermore, classifier performance tends to improve when excluding unnecessary features. Many classification models perform implicit feature selection by assigning unimportant features low weights or even excluding them completely. Thus, one can assess the importance of individual features easily *after* training such a model. However, it would even be better to estimate feature importance beforehand, to use this knowledge for feature selection prior to training.

Meta-learning is a field of machine learning which aims to predict the performance of models by training meta-models. In a meta-data set, one data object (row) usually represents the performance of a particular base model on a particular base data set. Common meta-features are dataset characteristics and hyper-parameters of the base model. The meta-target is often a classification performance measure or the name of the algorithm that performs best on the particular base data set. This form of meta-learning has gained momentum in the machine learning community over the last few years. However, it is unclear how well one can predict the importance of individual features of a data set with meta-learning.

The goal of this thesis is to develop and evaluate a meta-learning approach to predict the importance of features. The following questions are particularly interesting:

- What meta-features are useful to describe individual features in a dataset?
- How useful are existing filter-feature-selection scores as meta-features?
- How does our approach compare against existing feature-selection methods?
- How well can we predict the feature importance for different types of classification models?

The following steps are part of your thesis:

- Review literature about meta-learning and feature selection/importance.
- Design and implement meta-features describing base features.
- Design and implement experiments to predict feature importance with meta-learning. You can use existing machine learning libraries as well as data sets from public repositories like the UCI Machine Learning Repository.
- Evaluate your approach experimentally. You can use the server infrastructure of our chair.

During your work on this thesis, you will acquire practical knowledge about state-of-the-art machine learning libraries. You will get familiar with meta-learning as well as feature selection and gain an understanding of their usefulness as well as limitations for different kinds of data sets. You will gain experience in running and evaluating large scientific experiments.

The scope of the topic can be adapted to a Bachelor as well as a Master thesis. You can write the thesis in English or German. Prior experience with a programming language suitable for data science (e.g. Python, R) is beneficial, but not necessary if you are motivated to learn.

Contact

Jakob Bach | jakob.bach@kit.edu | +49 721 608-47339 | Room: 351
Am Fasanengarten 5 | 76131 Karlsruhe | Building: 50.34