Feature selection is an important step in machine learning pipelines. Using fewer features speeds up training of prediction models and makes the models more comprehensible. Furthermore, performance of simple prediction models might even improve if unnecessary features are excluded. Many feature selection algorithms exist. An important category is wrapper feature selection, which searches the space of feature subsets and evaluates each subset with a prediction model. An appropriate search strategy is vital to obtain good results, as the search space grows exponentially: For $n$ features, there are $2^n$ combinations. Existing search strategies include simple approaches like greedy forward selection, but also general-purpose optimization algorithms like genetic algorithms or simulated annealing.

Many existing search strategies only keep track of a small set of candidate feature sets. They iteratively modify this candidate set with a mixture of simple modification rules and randomization. As an alternative, one could also model the influence of features on prediction quality and use this to sample new candidates. For example, Bayesian optimization techniques have already been used successfully to model the influence of hyper-parameters on prediction quality and to tune prediction models. While Bayesian optimization already is an iterative approach, more systematic one-off sampling, e.g., using principles from experimental design, might also allow modeling the influence of features and guide feature selection.

The goal of this thesis is to study the use of meta-modeling approaches for wrapper feature selection. The following questions are particularly interesting:

- How can we use meta-modeling approaches for wrapper feature selection, and which adaptations might be necessary for this purpose?
- How does this meta-modeling-aided feature selection compare to existing wrapper feature selection methods?
- Is there a benefit compared to prediction models with built-in feature selection?
- Can we provide guidelines in which scenarios one should use our new approaches?

The following steps are part of your thesis:

- Review literature about (wrapper) feature selection and meta-modeling approaches.
- Design and implement one or more approaches using meta-modeling approaches for wrapper feature selection.
- Design and implement experiments to compare feature selection approaches. You can use existing machine learning libraries as well as datasets from public repositories.
- 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-modeling approaches as well as feature selection and gain an understanding of their usefulness as well as limitations for different kinds of datasets. You will gain experience in running and evaluating large-scale 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 machine learning in Python is beneficial, but not necessary.

Contact
Jakob Bach  
jakob.bach@kit.edu  
+49 721 608-43911  
Room: 363
Am Fasanengarten 5  
76131 Karlsruhe  
Building: 50.34