Applied Machine Learning

Feature Selection: Further Techniques

Learning goals

- Embedded Feature Selection
- Domain Knowledge
- Multi-objective Optimization
- Boruta Algorithm

EMBEDDED FEATURE SELECTION

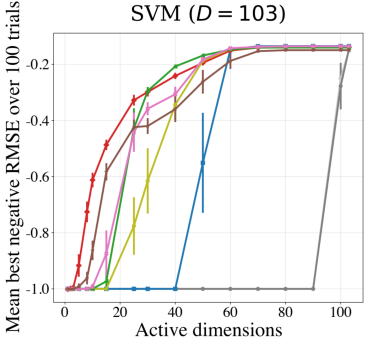
- Select features during the learning process
- Explicitly – via regularization
  - LASSO/L1
  - Requires a solver that can set weights to zero
- Implicitly – by construction of the algorithm
  - Decision-tree based algorithms
  - Can ignore irrelevant features by not selecting them (no guarantee, though)

DOMAIN KNOWLEDGE

- ID columns
  - Often ordered, can leak time information
  - If they identify a person etc, replace by features describing that person
  - Contains no information that can generalize
- Duplicate features
  - Cause issues for (linear) models
  - Slow down learning
  - Reduce model interpretability

MULTI-OBJECTIVE OPTIMIZATION OF FEATURES

- Optimize two competing Objectives:
  1. goodness-of-fit
  2. number of variable
- Use global optimization algorithm
  - Evolutionary Algorithm
  - Sparse Bayesian optimization



BORUTA (1)

- All-relevant feature selection
  Goal: find all features that have affect the prediction
  $\rightarrow$ Can include redundant features
- Idea: use *shadow variables* to contrast existing features against
- Method: Extend a feature scoring to an iterative testing mechanism

Boruta for those in a Hurry by Miron B. Krusa

BORUTA (2)

1. Create a copy of each feature, shuffle these copies
2. Fit a random forest on the new dataset
3. An attribute is deemed important, if its feature importance is higher than the maximal importance of all randomised attributes
4. Repeat steps 1-3 for N iterations
5. Execute SHT with the null hypothesis that the importance of a feature is equal to the maximal importance of all randomised attributes
6. Repeat steps 1-5 until the feature set is stable

Wrap-up

TAKEAWAYS

- There can be multiple goals in feature selection:
  - Single Feature Selection
  - Multiple Feature Selection
  - All-relevant Feature Selection
- Important step for good prediction quality and fast models
- Different feature selection methods have different goals and use different mechanisms $\rightarrow$ model selection problem