Karlsruhe Institute of Technology

Institute for Program Structures and Data Organization (IPD)
Chair of Information Management Systems

# Leveraging Constraints for User-Centric Selection of Predictive Features

AI @ KIT, October 6, 2022

Jakob Bach (jakob.bach@kit.edu)

## Motivation

- Feature selection determines most important predictors in a dataset
  - Various benefits for predictions: Lower computational and memory requirements, better interpretability, etc.
  - But: Existing methods usually just optimize prediction quality
- Constraints can make feature selection more user-centric:
  - Express firm domain knowledge
  - Express hypotheses
  - Express preferences
  - Express alternatives

## Formalization: Constrained Feature Selection

- Given:
  - Dataset $X \in \mathbb{R}^{m \times n}$ (rows are instances, columns are features)
  - Prediction target $y \in \mathbb{R}^m$
- Goal:
  - Make a feature-selection decision $s \in \{0, 1\}^n$ ...
  - ... to optimize the feature-set quality $Q(s, X, y)$.
- Constraints induce conditions on decision variables $s$:
  - Example 1: $(s_1 \wedge s_2) \vee s_3 \leftrightarrow$ "Select Features 1 and 2, or select Feature 3, or select all of them."
  - Example 2: $\sum_{j=1}^{n} s_j \cdot c_j \leq C_{max} \leftrightarrow$ "Select features so that their summed cost is under some threshold $C_{max} \in \mathbb{R}$."
- Depending on quality function $Q(s, X, y)$ and constraint types, problem requires black-box optimization or white-box optimization

## Formalization: Alternative Feature Selection

- Special case of constrained feature selection
- Idea: Find multiple, differently composed feature sets with high quality
  - Optimization goal remains feature-set quality $Q(s, X, y)$
  - Constraints: Feature sets should be alternative, i.e., dissimilar to each other (dissimilarity threshold $\tau \in \mathbb{R}_{\geq 0}$)
  - E.g., Feature sets $F_1$, $F_2$ alternative if $d_{Dice}(F_1, F_2) = 1 - \frac{2 \cdot |F_1 \cap F_2|}{|F_1| + |F_2|} \geq \tau$
- Search for alternatives can progress:
  - Simultaneously: Find a fixed number of alternatives at once
  - Sequentially: Find alternatives one after the other

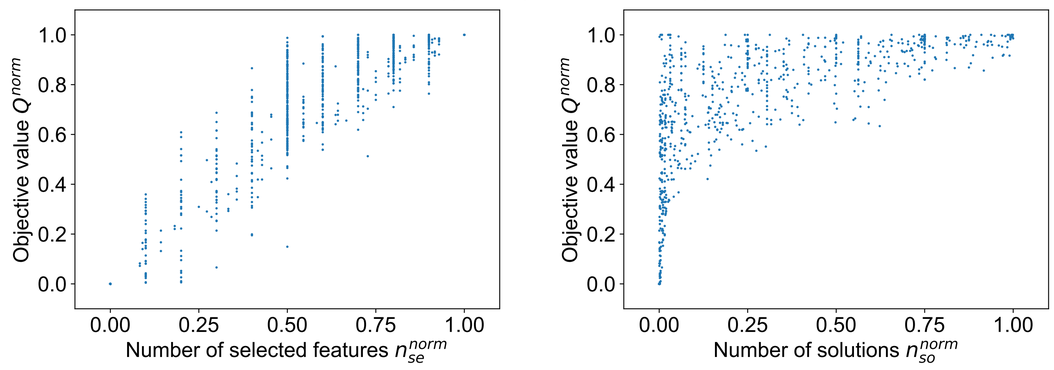
## References

J. Bach, K. Zoller, H. Trittenbach, *et al.*, "An empirical evaluation of constrained feature selection," *SN Computer Science*, vol. 3, no. 6, 2022. DOI: 10.1007/s42979-022-01338-z

J. Bach, "Finding optimal solutions for alternative feature selection," *Unpublished manuscript*, 2022

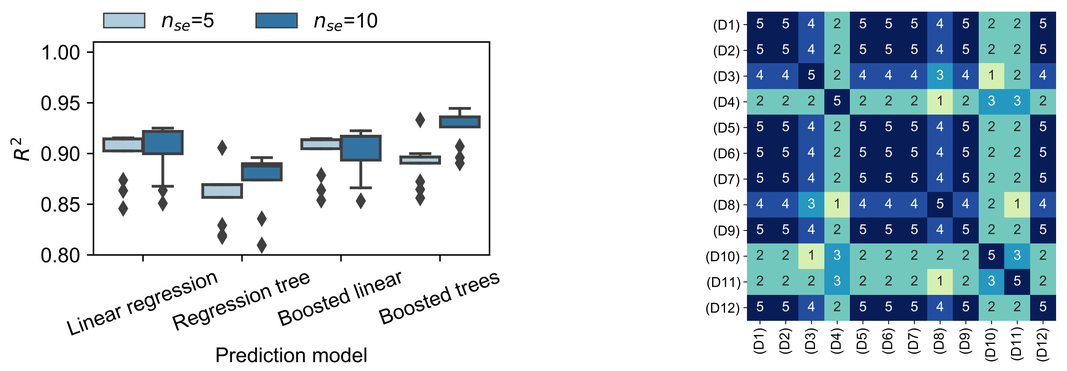
## Study: Evaluating the Impact of Constraints

- Experimental design:
  - 35 datasets from *OpenML* repository
  - Ten constraint types
- Key result: Stricter constraints (pruning more feature sets) can, but need not decrease predictive quality $Q$ of the optimal feature set:




## Study: Using Constraints to Express Domain-Specific Hypotheses

- Experimental design:
  - One materials-science dataset (evolution of a material's microstructure under load)
  - Twelve domain-specific constraint types
- Key result: Constraints may allow finding different feature sets adhering to domain constraints and yielding similar prediction performance:




## Study: Using Constraints to Find Alternative Feature Sets

- Experimental design:
  - 30 datasets from *PMLB* repository
  - Four feature-selection methods
  - Multiple search configurations for alternatives
- Key result: Predictive quality $Q$ decreases with the number of alternatives and the dissimilarity threshold $\tau$ for being alternative: