Data Analysis and Knowledge Discovery

Feature selection

Antti Airola

University of Turku
Department of Computing

antti.airola@utu.fi

Feature selection

Feature selection refers to techniques to choose a subset of the features (attributes) that is as small as possible and sufficient for the data analysis.

Feature selection includes

- removing (more or less) irrelevant features and
- removing redundant features.

- For removing irrelevant features, a performance measure is needed that indicates how well a feature or subset of features performs w.r.t. the considered data analysis task.
- For removing redundant features, either a performance measure for subsets of features or a correlation measure is needed.

Three goals of feature selection

1. Cost savings: smallest or cheapest set of features needed to do well on a task
   - Example: cheapest set of laboratory tests required to diagnose prostate cancer
2. Understandability: which are the most important features given my task
   - Example: which genes associated with high risk of a disease
   - Warning: correlation does not imply causality, and different methods often select different feature sets
3. Accuracy: remove noisy features to get more accurate predictions
   - relevant for old statistical methods that do not have mechanisms for avoiding overfitting
   - dimensionality reduction methods like PCA viable alternative, if understandability / costs not important
   - modern machine learning methods incorporate regularization mechanisms to avoid overfitting to noisy features, for these feature selection often does not improve accuracy

Three approaches to (supervised) feature selection

1. Filter methods
   - Univariate filters: compute absolute correlation (or p-value of some statistical test) between each attribute and the target variable. Pick *k* best, or all over some threshold.
   - Pros: computationally efficient, simple to implement, results easy to intrerpret
   - Cons: does not consider interactions between features, resulting subset contains redundant features
   - Advanced multivariate filter methods (e.g Relief methods) aim to address these limitations
2. Embedded methods: machine learning methods that incorporate feature selection mechanisms
   - Lasso, Decision trees, Random forests...
   - Pros: fairly efficient, finds features best suited for the embedded method
   - Cons: selection not main goal of optimization so "optimal" feature set can be large, applicable to only a quite restricted set of methods

Three approaches to (supervised) feature selection

3. Wrapper methods: search algorithm scores feature sets according to how accurately classifier/regressor method trained on them predicts
    - Forward search, backwards elimination, floating search, genetic algorithms...
    - Pros: can be combined with any classifier/regression method, finds features best suited for the given method, avoids redundant features, may find smaller feature sets than embedded methods
    - Cons: complex search methods computationally heavy and prone to overfitting

Supervised vs unsupervised feature selection

Two types of feature selection / dimensionality reduction settings

1. Unsupervised: given data matrix $\mathbf{X}$ find a smaller subset of features (feature selection) or mapping to lower dimensions (dimensionality reduction) such that preserves structure of $\mathbf{X}$
   - Feature selection: corresponds to removing redundant features
   - Usually safe: not easy to overfit using standard methods
2. Supervised: given data matrix $\mathbf{X}$ and target variables $\mathbf{y}$, find feature subset or low-dimensional mapping that best allows predicting $\mathbf{y}$ from $\mathbf{X}$
   - Dangerous: can completely bias your analysis if you do this as part of pre-processing. Should be done at same stage of analysis where you fit the prediction model.
   - Need separate test data not used in selection, or running the selection algorithm inside each round of cross-validation
   - Feature selection: corresponds to removing irrelevant features

Feature selection techniques

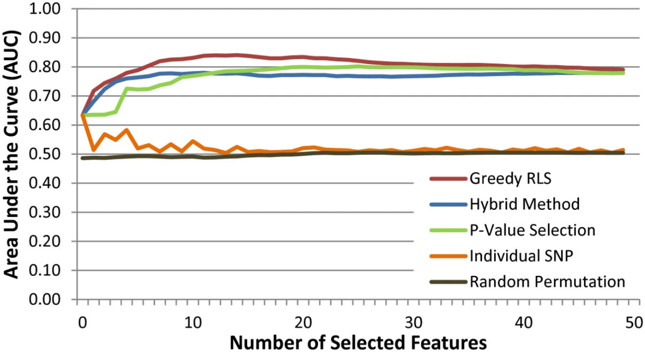
Selecting the top-ranked features. Choose the features with the best evaluation when single features are evaluated.

Selecting the top-ranked subset. Choose the subset of features with the best performance.
This requires exhaustive search and is impossible for larger numbers of features.
(For 20 features there are already more than one million possible subsets.)

Forward selection. Start with the empty set of features and add features one by one. In each step, add the feature that yields the best improvement of the performance.

Backward elimination. Start with the full set of features and remove features one by one. In each step, remove the feature that yields to the least decrease in performance.

Forward selection at work



- ”Wrapper-based selection of genetic features in genome-wide association studies through fast matrix operations”, Algorithms in Molecular Biology, 2012
- Few thousands of example, hundreds of thousands features to select from, AUC on independent test data