Machine Learning Notes

by Mattia Orlandi

5. Feature Selection

Problems with attributes:

- The significance of attributes for data mining can vary:
  - **irrelevant alteration**: alter results of mining algorithm in case of insufficient control of overfitting;
  - **redundancy**: attributes strongly related to other useful attributes;
    - **alteration**: some mining algorithms (e.g. Naive Bayes) are influenced by strong correlations between attributes;
  - **confounding**: some attributes can be misleading;
    - **hidden effect**: on the outcome variable;
- **mixed effect**: one attribute could be strongly related to the class in some cases, and random in the others.

Feature selection:

- enables machine learning algorithm to train faster;
- reduces complexity of a model and makes it easier to interpret;
- improves the accuracy of a model if the right subject is chosen;
- reduces overfitting.

Obs.: a specific selection action may obtain only some of the above effects.

- **Unsupervised learning**: several methods available
  - feature transformation techniques, such as PCA, can have the effect of reducing the number of features.
- **Supervised learning**: the relationship between each attribute and the *class* is considered
  - filter methods (i.e. Scheme-Independent Selection);
  - Scheme-Dependent Selection:
    - Wrapper methods;
    - Embedded methods.

Filter methods (Scheme-Independent Selection):

- Assessment based on general characteristics of data.
- Select the subset of attributes independently from mining model used.

- Types:
  - **Pearson's Correlation**: measure for quantifying linear dependence between two continuous variables (range: $[-1, 1]$);
  - **LDA**: Linear Discriminant Analysis is used to find a linear combination of features that characterizes or separates classes.
  - **ANOVA**: Analysis of Variance is similar to LDA, but it's operated using one or more categorical independent features and one continuous dependent feature.
  - **Chi-Square**: statistical test applied to groups of categorical features to evaluate likelihood of correlation/association between them using their frequency distribution.

Feature Response	Continuous	Categorical
Continuous	Pearson's Correlation	LDA
Categorical	ANOVA	Chi-Square

Set of all features $\rightarrow$ Selecting best subset $\rightarrow$ Learning algorithm $\rightarrow$ Performance

Wrapper methods:

- try to use a subset of features to train a model;
- search problem of what features to add/remove, with a test of the performance;
- computationally intensive.

Set of all features $\rightarrow$ [Generate subset $\leftrightarrow$ Learning algorithm] $\rightarrow$ Performance

Filter vs Wrapper:

- Filter methods measure the relevance of features by their correlation with dependent variables, while wrapper methods measure the usefulness of a subset of features by actually training a model on it.
- Filter methods are much faster, since they do not provide for training the model.
- Filter methods use statistics for evaluation of a subset of features while wrapper methods use cross-validation.
- Wrapper methods always provide the best subset of features, while filter methods might fail.
- The subset of features produced bt wrapper methods makes the model more prone to overfitting.

5.1. Dimensionality Reduction

Instead of considering which subset of attributes is to be ignored, it is possible to map the dataset into a new space with fewer attributes $\Rightarrow$ Principal Component Analysis (PCA).

PCA

- PCA makes use of covariance matrix and eigenvalues analysis.
- It finds a new ordered set of dimensions that better captures the variability of data.
- The fraction of variance captured by each new variable is measured.
- $\Rightarrow$ Few variables capture most of the variability.

Multi-Dimensional Scaling - MDS: a presentation technique which fits the projection of the elements into an m dimensional space s.t. distances among elements are preserved.

5.2. *Scikit-learn* solution for feature selection

- Main methods:
  - `.fit` learns empirical data from X;
  - `.fit_transform` fits to data, then transforms it;
  - `.transform` reduces X to the selected features.
- Main argument:
  - X, the dataset.

Baseline estimator:

- `VarianceThreshold` removes features with low variance, without taking into account the *class*.

Univariate feature selection: select the best set of features based on univariate statistical tests

- `SelectKBest` removes all but the k highest scoring features;
- `SelectPercentile` removes all but a user-specified highest scoring percentage of features;
- `GenericUnivariateSelect`selects the best univariate selection strategy with *hyper-parameter search estimator*.

Recursive Feature Elimination - RFE:

- Uses an external estimator to assign weights to features.
- Considers smaller and smaller sets of features.
- Estimator trained on initial set of features $\Rightarrow$ importance of each feature is obtained.
- Least important features are pruned.
- It stops when the desired number of features is reached.