# Data Mining Lectures

Piotr Wąsiewicz

Institute of Computer Science

*pwasiewi@elka.pw.edu.pl*

15 maja 2017

# Literature

## Books

- P. Cichosz, Data Mining algorithms: explained using R, Wiley 2015

## Internet sites

- http://www.wiley.com/go/data_mining_algorithms
- http://www.stat.wisc.edu/~larget/stat302/chap3.pdf

# Data Sets

## Tables

- A table (a set) consist of several columns, variables, input attributes or rows as vectors of attribute values.
- A limited dataset contains also a depended on input attributes variable called the target variable, the target attribute, the class label, the class concept, classes, the target concept, the group identifier, the response variable (regression).
- Each row can be called an instance, an object descibed by attribute values.
- Attributes can be nominal with a finite number of discrete values, ordinal - like nominal with a total order relation, continouos (numerical, linear) having numerical values.
- The target attribute classifies instances (objects) into classes or clusters them into groups.

# Data Sets

## Baskets

- Baskets consists e.g. of different length sets of product ids (ordinal numbers).

# Statistical elements

- If the p-value is below the significance level, then the null hypothesis is rejected (no relationship in the given domain).
- False positive (type I error) rejected null hypothesis is actually true. This risk increases with a larger significance level.
- False negative (type II error) rejected alternative hypothesis is actually true. This risk increases with a smaller significance level.
- Pearson's linear correlation coefficient measures the strength of linear relationship between two continuous attributes, null hypothesis - 0 value.
- Spearman's rank correlation also has values from -1 to 1 for increasing relationship.
- For discrete attributes Chi2 and the loglikelihood ratio (G-test) tests can be taken.
- For mixed attributes t-test (with Mann-Whitney-Wilcoxon nonparametric alternative) and f-test - one-way ANOVA (with Kruskal-Wallis alternative) can be used.
- The population is to the sample as the sample is to the bootstrap samples (samples with replacement).

# The model e.g. the decision tree

**Create the model to return a prediction after taking an instance as an argument**

- Made a classifier which is trained on a training set with a target attribute (supervised learning tasks using some expert knowledge put in the target variable) and predicts class label values for rows without this target classes.
- Find a regression target values based on input data.
- Without any expert knowledge cluster groups using the similarity structure discovered in the input data and write groups identifiers to the target attribute.

# Model performance

## Quality of predictions

- The expected quality of predictions on the whole domain including previously unseen instances

## Overfitting

- A model perfoms worse on the whole domain than on the training dataset, where e.g. it has excellent performance.

## Model ensembles

- For three binary classification models with independent mistakes voting reduces error if base model error below $1/2$ ($\epsilon^3 + 3\epsilon^2(1-\epsilon) < \epsilon$)

## Base model generation

- Different sets of training instances: use a different set of training instances to create each base model.
- Different sets of attributes: use a different set of attributes to create each base model.
- Different algorithms: use a different algorithm to create each base model.
- Different parameter setups: use a different algorithm parameter setup to create each base model.
- Algorithm randomization: use independent runs of a non-deterministic algorithm to create each base model.

# Ensemble prediction

- Voting: combine base model predictions by (possibly weighted) voting.
- Probability averaging: combine base model class probability predictions by (possibly weighted) averaging.
- Using as attributes: create a combined model using base model predictions as attributes.

## Bagging

- Base models created using a single algorithm applied to multiple bootstrap samples of the training set samples drawn at random with replacement - 63.8% bag, 37.2% out-of-bag (OOB).
- Ensemble prediction by (unweighted) voting.
- Requires an unstable algorithm for sufficient base model diversity e.g., decision trees, but after some number of base models it is stabilized.
- Moderate prediction improvement, but a useful stabilization effect.

## Stacking

- Different base model algorithms.
- Very difficult to design.

# Boosting

- Increase base model diversity by shifting focus during model creation.
- Force subsequent base model to compensate deficiencies of the preceding ones.
- Typically used with weight-sensitive classification algorithms increase weights of instances incorrectly classified by the previous model.
- Ensemble prediction by weighted voting.
- Substantial prediction improvement.
- Simple base models work well e.g., small decision trees or decision stumps.
- Can not be executed in parallel.
- AdaBoost - weighted base models and attributes.

# Random forest

- Bagging combined with algorithm randomization for increased base model diversity.
- Randomized decision trees sample attributes at each node prior to split selection.
- Ensemble prediction by voting.
- Maximum or near-maximum fit of base models individual overfitting gets compensated by combining many randomized trees.
- Substantial prediction improvement.
- Estimate prediction quality using OOB (Out-of-bag) instances.
- For each training instance: generate and combine predictions of those and only those trees for which the instance is OOB
- Use these predictions to calculate the misclassification error (or another prediction quality indicator).
- Estimate the predictive utility of attributes by measuring error on the mutated data.