Introduction to Machine Learning – NPFL 054

Barbora Hladká and Martin Holub
Charles University in Prague
2016/17

Overview of the course and summary of examination requirements

I. Data analysis

- **Elementary data analysis**
  – empirical and standard probability distributions, joint and conditional distributions
  – categorical data distribution
  – binomial distribution, normal distribution, z-transformation
  – density function, distribution function, quantile function
- **Inter-annotator agreement**
  – gold standard data and manual annotation
  – confusion matrix, probability of disagreement, Cohen's kappa
- **Analysis of data associations**
  – methods for data exploration – plotting and summarizing
  – Q-Q plots
  – Pearson's correlation coefficient
  – Pearson's contingency coefficient
- **Statistical tests**
  – statistical hypotheses, significance level, critical values, p-value, confidence intervals
  – meaning of t-distribution and chi-square distribution
  – t-tests, chi-square tests, practical applications
- **Entropy and information gain**
  – entropy and conditional entropy – motivation, definition, meaning, main properties
  – joint entropy, mutual information, and relationships
  – information gain and its application in feature selection
- **Clustering**
  – division of clustering methods
  – k-means algorithm
  – hierarchical clustering, dendrogram, agglomerative and divisive clustering
  – measures of cluster similarity
- **Principal Component Analysis**
  – feature covariance
  – PCA algorithm and geometric interpretation, biplot
  – Proportion of Variance Explained, scree plot

II. Supervised learning methods

- **Formal foundations of machine learning**
  – supervised and unsupervised learning, classification and regression
  – training examples, feature vectors, discrete and continuous features, scaling
  – target variable and prediction function
  – loss/cost function, squared loss, zero-one loss, logistic loss, Hinge loss
  – learning methods, model, hypothesis, predictor
  – machine learning process and development cycle
  – PAC learning (not required at exam)
- **Decision Trees**
  – prediction function and learning algorithm
  – impurity measures / splitting criteria
  – misclassification error, information gain, Gini index
  – classification and regression trees
- **Linear Regression and Logistic Regression**
  – prediction function and learning algorithm
  – simple and polynomial linear regression
  – multivariate linear regression and gradient descent algorithm
  – linear regression on binary classification
  – linear and non-linear decision boundary
  – logistic regression for binary and multi-class classification
  – interpretation of hypothesis parameters
  – linear/logistic regression with a categorical feature
- **Instance-Based Learning**
  – k-NN and distance weighted k-NN algorithm for classification and regression
  – locally weighted linear regression
- **Naive Bayes algorithm and Bayesian belief networks**
  – generative and discriminative classifiers
  – naïve assumption of feature conditional independence
  – derivation of Naïve Bayes prediction function
  – relation to linear classifiers
  – Bayesian belief networks – motivation, structure, and prediction
- **Support Vector Machines**
  – linear separability of training examples, separating hyperplane
  – large margin and soft margin classifier, support vectors
  – primal and dual optimization problem, idea of solution (details are not required at exam)
  – kernel tricks, common kernel functions
  – SVM for multi-class classification
- **Perceptron**
  – error-driven learning
  – perceptron learning algorithm and geometric interpretation
- **Learning parameters tuning**
  – hypothesis parameters and learning parameters
  – grid search
  – gradient descent algorithm

- **Ensemble learning**
  - general scheme, ideas, algorithms and advantages
  - bootstrapping, bagging, boosting
  - Random Forests
  - AdaBoost
- **Regularization**
  - Ridge regression regularization (L2)
  - Lasso regularization (L1)
  - regularized linear and logistic regression
  - soft margin classifier as a regularization problem
- **Maximum Likelihood Estimation**
  - likelihood function and the idea of MLE
  - relation to MSE and logistic regression
- **Feature selection heuristics**
  - curse of dimensionality
  - feature frequency
  - filters, wrappers, embedded methods
  - greedy forward selection and backward elimination
- **Neural Networks** (not required at exam)

III. Evaluation

- **Generalization error estimation**
  - sample error and generalization error, MSE, accuracy and classification error
  - confusion matrix, different types of classification errors
  - division of development data and test data
  - model complexity and overfitting
  - bias and variance error decomposition
  - cross validation process, aggregated confusion matrix
  - error estimation by bootstrapping
  - using statistical tests for evaluation
  - irreducible Bayes error rate and Bayes classifier – meaning and definition
- **Binary classifier evaluation**
  - true/false positives/negatives
  - accuracy, precision, recall, specificity, F-measure
  - ROC, AUC measure