Training agenda

Predictive modeling with R

Instructor: Artur Suchwałko, Ph.D.

1. Introduction
   - applications of predictive models
   - data preparation
   - stages of learning and testing of effectiveness of models
   - selection of parameters of models

2. Data preparation
   - analysis of single features (characteristics)
     - distributions (contingency tables, histograms)
     - missing data and outliers
     - quality control and data cleaning
     - preliminary selection of features – analysis of discriminative power of features
   - classing (discretization) of continuous features
     - role of discretization
     - methods of of discretization
       - weight of evidence (WoE)
       - entropy maximization
       - classification trees
   - analysis of dependency between features and construction of derivative features (generated characteristics, cross characteristics)
   - standardization
   - sampling

3. Classification and regression methods
   - discriminative analysis
   - k-nearest neighbors method
   - neural networks
   - Support Vector Machines (SVM)
   - classification trees
   - regression trees
   - randomForest
• Bayes classifier
• linear regression
• logistic regression
• nonlinear regression

4. Tree-based models
  • specificity of tree-based models
  • overview of applications
  • visualization and interpretation of results
  • practical aspects related to tree-based models building:
    – feature selection criteria
    – split criteria
    – stop criteria
    – assessment of tree structure complexity
  • classification trees
  • regression trees
  • postprocessing of trees: simplifying and modification of structure (pruning), expert analysis
  • pros and cons of tree-based models
  • improvement of stability and effectiveness of trees (bagging algorithm, hybrid models)
  • randomForest

5. Classification quality assessment and tuning of classifiers parameters
  • classification error estimation
  • quality assessment strategies: train/test, cross-validation, leave-one-out, bootstrap
  • ROC curve, AUROC coefficient
  • cost-sensitive learning, cost-sensitive evaluation
  • selection of optimal cut-off point
  • selection of optimal model parameters (tuning)
  • comparison and selection of the best model

6. Feature selection for model building
  • criteria of application of features in models (statistical, business, operational)
  • graphical methods
  • complete search
  • one-step methods (filters)
  • stepwise methods (forward, backward, forward-backward)
  • methods built-in in classifiers (e.g. randomForest), committees of models, other methods

7. Very important practical aspects of modeling
  • building models for small data sets
  • building models for numerical features without categorization
  • dependency of features (numerical and categorical) – how to manage this issue
  • not equal proportions of groups and its consequences
  • comparison of popular approaches to model building: dummy variables, WoE encoding, models for continuous features

8. Additional practical topics related to predictive model building in R
  • working with different input data formats
  • working using MS Excel
  • export of models in PMML format