Statistical Learning Methods - Practical test

Exercise 1 *[20p] - Classification problem*

You are quantitative analyst in medical center. You were asked to help medical staff in decision making process to classify patients with breast cancer into two different therapies. For this purpose, build three concurrent classification models (using Conditional Inference Tree method, Recursive Partitioning and Regression Trees method and Random Forest method) and choose the best one with highest prediction accuracy.

Before You start:

1) Install and load package “mlbench” {install.packages(“mlbench”), library(mlbench)}.
2) Load dataset named “BreastCancer” {data(BreastCancer)}.
3) Set the seed according to the number of your index.
4) Keep informed that the endogenous variable is labeled “Class”.
5) Keep informed that if it is possible to show outcome for mini tasks, just copy it into MS Word. In case it is not possible, just copy the instruction to MS Word, which should lead to achieve results. Before you send the file to *artur.pluska.sgh@gmail.com*, save as PDF.

To get final results go through mini tasks:

1) *[2p]* How many exogenous variables and observations dataset have? What type of object this data set is?
2) *[2p]* Does the dataset contains missing values? If yes, impute missing values with method, that you are free to choose (e.g. average, median, regression of other variables, removing these observations, coding as a separate value, etc.).
3) *[2p]* Remove ID variable and check the percentage of observations for one chosen class.
4) *[2p]* Divide dataset into training subset and validation subset in relation 2:1.
5) *[2p]* Build classification tree using Conditional Inference Tree (ctree) method with parameters mincriterion at the level of 0.99 and minsplit equal to 20. Plot obtained model.
6) *[2p]* Build classification tree using Recursive Partitioning and Regression Trees (rpart) methods with parameters cp at the level of 0.00001 and minsplit equal to 3. Plot chart showing complexity parameter vs. misclassification error. Choose cp with lowest misclassification error.
7) *[2p]* Prune previous classification tree with optimal cp.
8) *[2p]* Build Random Forest model with number of trees at the level of 250.
9) *[2p]* For all classification trees make prediction on validation subset. Show confusion matrices for every model.
10) *[2p]* Print comparison table which show for every method misclassification error on validation subset. Which model is the best?

Statistical Learning Methods - Practical test

Exercise 2 [20p] - Regression problem

You are manager in baseball team. You are going to build dream team and win the league in the next season. The player transfer window is just opening and one determinant which can attract best players to your team is the salary. But you have budget limitations and you are going to adjust salaries according to player's skills. For these purpose, build regression model with three regression methods (Linear Model, Forward Selection and Ridge Regression) and compare the results. Point the model with highest prediction accuracy.

Before You start:

1) Install and load package "ISLR" {install.packages("ISLR"), library(ISLR)}.
2) Load dataset named "Hitters" {data(Hitters)}.
3) Set the seed according to the number of your index.
4) Keep informed that the endogenous variable is labeled "Salary".
5) Keep informed that if it is possible to show outcome for mini tasks, just copy it into MS Word. In case it is not possible, just copy the instruction to MS Word, which should lead to achieve results. Before you send the file to *artur.pluska.sgh@gmail.com*, save as PDF.

To get final results go through mini tasks:

1) *[2p]* Plot the chart showing the distribution of *Salary* variable? Does it have gaussian distribution (to determine the distribution use instruction: "shapiro.test")?
2) *[2p]* Does the dataset contains missing values? If yes, impute missing values with average value for the given variable. Remove qualitative variables from dataset.
3) *[2p]* Divide equally dataset into training, validation and testing subset.
4) *[2p]* Identify variables with correlation at least at |0.3| using correlation test (to determine the correlation and its significance use instruction: "cor.test").
5) *[2p]* Build regression with GLM on training subset with variables, which have correlation over |0.3|. Which variables and statistically significant?
6) *[2p]* Build regression with GAM method on training dataset with all variables.
7) *[2p]* Visualize (bar chart) the Mean Squared Error on training and validation subsets for obtained GLM and GAM regression.
8) *[2p]* Based on validation subset check, which model is the best. Show its summary.
9) *[2p]* Calculate Mean Squared Error on testing subset for two models. Which model is the best?
10) *[2p]* What the salary would you pay to average player (new exogenous variable is an average for every characteristics)? Use the best model, you have obtained.