Assignment 3

Arthur Junges Schmidt

20 July 2020

Problem 1

(a) The file ps3-1.csv contains a data set with 34 features (x1, x2, ... x34) and 1 target variable (Y). Estimate a classifier model using Support Vector Machine and Random Forest algorithm in R, respectively, with the first 14628 rows of the data. Optimize your model so that a False Positive Rate is less than 10% for Y = 0 (actual Y = 1 cases falsely classified as Y = 0). Particularly, you are required to review relevant literature, use the k-fold cross-validation method to train the Random Forest model. Use grid search to find hyper-parameter setting: the best number of trees and features, maximum leaf nodes. Assess importance of each feature based on two criteria: Mean Decrease Accuracy and Mean Decrease Gini. Compare two estimation methods with confusion matrices

```
Crude_Data <- read.csv("ps3-1.csv");
Crude_Data <- Crude_Data[,-1];
```