Exercise 8

Classification and Discriminant Analysis

December 10, 2014

Use the data from `http://archive.ics.uci.edu/ml/datasets/Bank+Marketing`, that are available on the website with homework. Load the smaller data set using `d <- read.csv2("bank.csv")`. The data contain information about direct marketing campaigns (phone calls) of a Portuguese banking institution. The classification goal is to predict if the client will subscribe a term deposit or not. This information is represented by the binary variable y (last one).

Random forests: function `randomForest()` from the `library(randomForest)`

Random forests use different bootstrap samples of the data to construct multiple classification trees. The final classification is obtained based on the major desicion resulting from all the trees.

(a) Set randomly a training set of a reasonable size (as for random trees) and apply `randomForest()`. Predict the group membership for the test set and compute the missclassification rate. Can we see an improvement considering results from random trees?

(b) Although the misclassification rate is relatively small, the proportion of persons signing a contract is significant. A lot of "no" responses have been predicted. This is a very unpleasant issue, because the bank never wants to loose potential customers. Consider a strategy, how to reduce the number of missclassified customers who actually signed the contract. Apply the strategy also on the whole data set *bank-full.csv*.

Please, send your R scripts with the solution as a text file saved as "Surname8.R", via email to

`kynclova@statistik.tuwien.ac.at`

at latest until December 8.