# Exercise 7

## Classification and Discriminant Analysis

December 3, 2014

1. Load the data `SAheart` from the package `ElemStatLearn`. The data contain information about males in a heart-disease high-risk region in South Africa (see help). The goal is to apply logistic regression with natural cubic splines to split the data into groups according to the variable *chd* (coronary heart disease). Do not use the variable *famhist* in the exercise.

   *General Additive Models*: function `gam(..., family="binomial")` from the `library(mgcv)`

   (a) Randomly select a training data set of 300 observations and apply logistic regression with smoothing splines. Which variables are significant? Calculate the misclassification rate for an independent test set.

   (b) Plot the variables against their estimated values. You can simply use:
   `plot(gam.object,page=1,shade=TRUE,shade.col="yellow")`
   How could we interpret this plot?

2. For the following exercise 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).

   *Classification trees*: function `rpart()` from the `library(mvpart)`

   (a) Set randomly a training set of a reasonable size and apply a tree $T_0$ (see `help(rpart)` or lecture notes).

   (b) Visualize the tree with the function `plot()` and `text()`, and interpret the results.

   (c) Predict the group membership for the test set (see `help(predict.rpart)` or lecture notes) How high is the resulting missclassification rate?

   (d) Show and interpret results of cross-validation obtained by using `printcp()` und `plotcp()`. What is the optimal complexity?

   (e) Prune the tree $T_0$ of the optimal complexity using `prune()`. Visualize und interpret the results.

   (f) Predict the group membership for the test set and calculate the resulting missclassification rate. Do we observe any improvement?

Please, send your R scripts with the solution as a text file saved as "Surname7.R", via email to

`kynclova@statistik.tuwien.ac.at`

at latest until December 1.