Exercise 4

Advanced Methods for Regression and Classification

November 14, 2018

1. Load the data `OJ` from the package `ISLR`. Our goal is to find a classification model that allows to predict the grouping variable `Purchase`, using the remaining variables in the data set.

   Select randomly a training set of about 2/3 of the observations, build the classification models, predict the group membership for the (remaining) test data and compute the misclassification rates. Consider the following methods:

   (a) *Linear regression with indicator variable (LS)*

   (b) *Linear Discrimant Analysis (LDA)*: function `lda` from `library(MASS)`
   How do you need to construct the matrix $\mathbf{X}$? Are there problems with binary (factor) or categorical variables? How to solve such problems?

   (c) *Quadratic Discrimant Analysis (QDA)*: function `qda` from `library(MASS)`
   How can problems with possible singularities be solved?

   (d) *Regularized Discrimant Analysis (RDA)*: function `rda` from `library(klaR)`
   Interpret the meaning of the resulting tuning parameters `gamma` and `lambda`.

2. Use the data from
   `http://archive.ics.uci.edu/ml/datasets/Bank+Marketing`,
   which are also available on our TUWEL course. 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 contained in the binary variable y (last one).

   (a) Select randomly a training set with 3000 observations, and use LDA to predict the group label of the remaining test set observations. Look at the confusion table and the resulting misclassification rate.

   (b) Although the misclassification rate is relatively small, the misclassifications of the two groups are heavily imbalanced. Many "yes" clients (those are the interesting ones) have been predicted incorrectly. This is very unpleasant, because the bank never wants to lose potential customers. Develop a strategy how to reduce the number of misclassified customers who actually signed the contract, and evaluate this strategy. One idea could be to use "balanced" samples from both groups (under-sampling, etc.).

Save your (successful) R code together with short documentations and interpretations of results in a text file (= R script file), named as *Matrikelnummer_4.R* (no word document, no plots). Submit this file to Exercise 4 of our tuwel course (deadline November 13).