Exercise 4

Classification and Discriminant Analysis

November 12, 2014

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 methods of discriminant analysis to split the data into groups according to the variable *chd* (coronary heart disease). Do not use the variable *famhist* in this exercise.

1. Plot the (scaled) data into the space of principal components (function `princomp`) to distinguish the groups given by the variable *chd*.

2. *Linear regression with the indicator matrix (LS)*:

   (a) Construct the indicator matrix consisting of two columns for the variable *chd*. Select the training set of 300 observations (set the random seed!) and apply the LS-regression with the indicator matrix. Predict the group membership for the test data and compute the rate of missclassification.

   (b) Repeat the procedure 100 times (without seed) and plot the rates of missclassification using a boxplot.

3. *Linear Discrimant Analysis (LDA)*: function `lda` from `library(MASS)`

   (a) Select the training set of 300 observations (set the random seed!) and apply the LDA. Predict the group membership for the test data and compute the rate of missclassification.

   (b) Repeat the procedure 100 times (without seed) and plot the rates of missclassification using a boxplot. Compare the boxplot with the previous one. Which method works better in this case?

Please, send your R scripts with the solution as a text file saved as "Surname4.R", via email to

`kynclova@statistik.tuwien.ac.at`

at latest until November 10.