Exercise 5

Classification and Discriminant Analysis

November 19, 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). This time use the variable *famhist* in the exercise.

1. *Quadratic Discriminant Analysis (QDA)*: function `qda` from `library(MASS)`

   (a) Randomly (set the same seed!) select a training data set of size 300 and apply QDA. Predict the group membership for the test data and compute the misclassification rate.

   (b) Repeat this procedure (without seed) 100 times and visualize misclassification rates with a boxplot. Compare the boxplot with the previous ones (LS, LDA). Which method works best?

2. *Regularized Discriminant Analysis (RDA)*: function `rda` from `library(klaR)`

   (a) Randomly (set the same seed!) select a training data set of size 300 and apply RDA. Predict the group membership for the test data and compute the misclassification rate.

   (b) Repeat this procedure (without seed) 20 times (slow!) and visualize misclassification rates with a boxplot. Compare the boxplot with the previous ones. Which method works best?

   (c) This is rather a remark than a task: RDA internally carries out a cross-validation, and as an output we obtain the misclassification error due to cross-validation. However, this error turns out to be much smaller than that from (b). Why? Even when using the parameters `gamma=0` and `lambda=1` in the RDA function (i.e. in the case of LDA, see help) we obtain different results. Why?

3. *Logistic Regression (LR)*: function `glm(...,family=binomial)`

   (a) Randomly (set the same seed!) select a training data set of size 300 and apply LR. Predict the group membership for the test data and compute the misclassification rate.

   (b) Repeat this procedure (without seed) 100 times and visualize misclassification rates with a boxplot. Compare the boxplot with the previous ones. Which method works best?

Please, send your R scripts with the solution as a text file saved as "Surname5.R", via email to

`kynclova@statistik.tuwien.ac.at`

at latest until November 17.