Multivariate Statistics: Exercise 9

December 11, 2014

Discriminant analysis:

1. (a) We consider the data `hemophilia` from the `library(rrcov)`. The first two variables provide information about potential hemophilia carriers, while the third variable informs about the group membership to carriers or non-carriers.

    Use *all* available data to estimate the discriminant function for classical and robust versions of linear and quadratic discriminant analysis (functions `LdaClassic()`, `Linda()`, `QdaClassic()`, `QdaCov()` from the `library(rrcov)`). Predict the group memberships of *all* data with the estimated discriminant functions (command `predict()`). Compute the misclassification errors of the different discriminant analysis methods based on the true group memberships (use the command `table(prediction,truth)`). Which method works best?

   (b) The above misclassification error is usually too optimistic (why?), and we thus try a better form of evaluation of the discriminant function. The data set should be split into a "training data set" and a "test data set". Select randomly half of the observations as training data (command `sample()`), and the remaining observations as test data. Now the discriminant functions can be estimated based on the training data, and the "new" test data can be classified using these functions. Compute the misclassification errors for the test data. Repeat the procedure 100 times and show the results in boxplots. How do the results differ from above?

2. Consider the data `olives` from the package `classifly`, see Exercise 2. We will try to discriminate the data based on the fatty acid concentrations (do not use *eicosenoic*) into the three regions (variable *Region*). Perform the same analyses as in 1. Which difficulties are occurring?

Save your (successful) R code together with short documentations and interpretations of results in a text file, named as *Familyname09.R*. Send the file as an email attachment to *mehmet.mert@tuwien.ac.at*, at latest Tuesday (09.12)