Exercise 2

Classification and Discriminant Analysis

October 22, 2014

Load the data `prostate` from the package `ElemStatLearn`. The data contain measurements about prostate cancer. The goal is to apply regression analysis to model the response `lpsa` with the other explanatory variables except of `train`. The variable `train` gives us the information, which observations are in the training set (TRUE) or in the test set (FALSE). All the mentioned methods should be applied only on the training set to fit the model, that is subsequently evaluated on the test set (use MSE as a criterion).

1. *Principal component regression*:

   (a) Apply the *principal component regression*, that is implemented in the `library(pls)` as the function `pcr()`, see help. Set the number of components to include in the model as 8. Perform cross-validation using 10 segments and scale the data (`scale=TRUE`). Interpret the results from `summary()`.

   (b) Plot the obtained prediction errors of cross-validation according to the lecture notes. How many components seem to be optimal?

   (c) Plot the measured y values against the predicted y values considering the optimal model.

2. *Partial least squares regression*:

   (a) Apply the *partial least squares regression*, that is implemented in the `library(pls)` as the function `plsr()`, see help. Set the number of components to include in the model as 8. Perform cross-validation using 10 segments and scale the data (`scale=TRUE`). Interpret the results from `summary()`.

   (b) Plot the obtained prediction errors of cross-validation according to lecture notes. How many components seem to be optimal?

   (c) Plot the measured y values against the predicted y values considering the optimal model.

Compare all obtained models by calculating the MSE for test data. Which model shows the best fit to the data?

Please, send your R scripts with the solution as a text file saved as "Surname2.R", via email to

`kynclova@statistik.tuwien.ac.at`

at latest until October 20.