Exercise 3

Classification and Discriminant Analysis

October 29, 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. *Ridge Regression*:

   (a) Use the function `lm.ridge()` from the `library(MASS)`. Consider different values for the ridge parameter λ to find an optimal parameter. Plot the resulting GCV against the examined λ. You can also use a function `plotRidge` from the `library(chemometrics)`.

   (b) Use the optimal ridge parameter for the final estimation. Which values do you obtain as regression coefficients? Calculate the predicted y values.

   (c) Use `ridgeCV()` from the `library(chemometrics)` to perform cross validation with the optimal ridge parameter to estimate the prediction error.

   (d) Calculate the MSE for the test data.

2. *Lasso Regression:*

   (a) Use the function `lars()` from the `library(lars)`. Visualize and interpret the results of Lasso regression.

   (b) Apply the function `lassoCV()` from the `library(chemometrics)` to perform cross validation with 10 segments and investigate the prediction error.

   (d) Calculate the MSE for the test 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 27.