Exercise 1

Classification and Discriminant Analysis

October 15, 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. *Full model*: Apply the full regression model and interpret the results.
2. *Stepwise regression*: Use the function `step()`. Find the optimal model using *forward selection*, *backward selection* and selection in both directions. Compare all the obtained models using ANOVA (see the lecture notes).
3. *Best subset regression*:
   - (a) Use the *best subset regression*, that is implemented in `library(leaps)` as the function `regsubsets()`, see help. To find the models set the maximum size of subsets to 8 variables and examine the best 3 models of each size.
   - (b) Plot the results. Which model seems to be the best?
   - (c) Save the resulting `summary` as another object. Display the structure `str()` of this object and plot the size of models against BIC values. Which is the best model? Apply `lm()` on the final best model and interpret the results `summary()`.

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 ”Surname1.R”, via email to

`kynclova@statistik.tuwien.ac.at`

at latest until October 13.