Exercise 2

Advanced Methods for Regression and Classification

October 25, 2018

Consider the data `Hitters` from the package `ISLR`, see last exercise. Again, our goal is to predict the variable `Salary`.

For the following tasks, split the data randomly into training and test data (about equal halves), build the model with the training data, and evaluate for the test data (using the MSE as a criterion).

1. *Principal component regression (PCR)*:

   (a) Apply *principal component regression*, which is implemented in the `library(pls)` as the function `pcr()`, see help. Use the complete data set (without missings) for this purpose. Set the number of components to include in the model as 19. Perform cross-validation using 10 segments (see help of `pcr()`) and scale the data (`scale=TRUE`). Interpret the results from `summary()`.

   (b) Plot the obtained prediction errors of cross-validation, see lecture notes. How many components seem to be optimal? Compare with the MSEs from the last exercise.

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

2. *Partial least squares regression (PLS)*:

   (a) Apply *partial least squares regression*, implemented in the `library(pls)` as function `plsr()`, see help. 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, similar as shown in the lecture notes. How many components seem to be optimal? Compare with the MSE from PCR.

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

3. *PCR with variable selection*:

   - Perform principal component analysis (PCA) on the scaled predictor variables, using the function `princomp()`. There will be a problem with the non-numeric variables, which first need to be converted to numeric ones. This can be done “by hand”, or using e.g. the function `model.frame()`.
   - Apply `biplot()` to the result object of PCA. What can you see?
   - Use the PCA scores, available as list argument `$scores` from the PCA result object, for best subset regression on the variable `Salary`, only for the training data. Select the best model and compute the MSE for the test data. Can you beat the results from standard PCR?

Save your (successful) R code together with short documentations and interpretations of results in a text file (= R script file), named as *Matrikelnummer_2.R* (no word document, no plots). Submit this file to Exercise 2 of our tuwel course (deadline October 24).