Exercise 8

Advanced Methods for Regression and Classification

December 13, 2018

Load the data Auto from the package ISLR. The data contain different characteristics, information about the origin, and the name of the car. The goal is to predict hte variable mpg, so the miles per gallon, using the remaining variables. Select those variables which could be useful for the prediction. Randomly select a training set of about 2/3 of the observations, by using the random seed "123". Build the following models always based on the training set, and evaluate them for the test set data consisting of the remaining observations.

1. *Linear model with natural cubic splines*: Use the function `ns()` from the `library(splines)` and the function `lm()`

   The form of the model is

   $$y = \theta_0 + \boldsymbol{h}_1(x_1)^\top \boldsymbol{\theta}_1 + \boldsymbol{h}_2(x_2)^\top \boldsymbol{\theta}_2 + \ldots + \boldsymbol{h}_p(x_p)^\top \boldsymbol{\theta}_p + \varepsilon,$$

   where each $\boldsymbol{\theta}_j$ $(j = 1, \ldots, p)$ is a vector of coefficients that is multiplied by the basis function $\boldsymbol{h}_j$ (natural cubic splines) for the jth input variable.

   Every term in the model should be represented by 4 natural cubic splines. However, for some input variables (binary, categorical) this might not make sense, and they should enter the model in the usual way without splines.

   (a) Which variables (basis functions) are significant? Calculate the RMSE (root mean squared error) for the test set.

   (b) Apply stepwise variable selection using `step(...,direction="both")`. Which variables (basis functions) are significant? Compute the RMSE for the test set.

   (c) Plot the variables from the reduced model (b) against their estimated values, so e.g. x_j against $\hat{f}_j(x_j) = \boldsymbol{h}_j(x_j)^\top \hat{\boldsymbol{\theta}}_j$. How can you interpret these plots?

2. *Generalized Additive Models (GAMs)*: function `gam()` from the `library(mgcv)`

   (a) Apply regression with GAMs, and let the function select the optimal tuning parameters (degrees of freedom). You might have to be careful which of the input variables are appropriate for smooth functions in the model.

   (b) Which variables are significant in the model? How complex are the smooth functions?

   (c) Plot the explanatory variables against their smoothed values as they are used in the model. You can simply use:
   `plot(gam.object,page=1,shade=TRUE,shade.col="yellow")`
   How can you interpret this plot?

   (d) Compute the RMSE for the test set.

   (e) Try to tune the model to get a value of the RMSE below 2.9. Some ideas can be found in the help of `step.gam`.

Save your (successful) R code together with short documentations and interpretations of results in a text file (= R script file), named as *Matrikelnummer_8.R* (no word document, no plots). Submit this file to Exercise 8 of our tuwel course (deadline December 12).