Exercise 6

Classification and Discriminant Analysis

November 26, 2014

Load the data `SAheart` from the package `ElemStatLearn`. The data contain information about males in a heart-disease high-risk region in South Africa (see help). The goal is to apply logistic regression with natural cubic splines to split the data into groups according to the variable *chd* (coronary heart disease). Do not use the variable *famhist* in the exercise.

1. *Logistic regression with natural cubic splines*: Use the function `ns()` from the `library(splines)` and the function `glm(...,family=binomial)`

   The form of the model is

   $$\text{logit}[P(\texttt{chd}|\boldsymbol{x})] = \log\left(\frac{P(\texttt{chd}=0|\boldsymbol{x})}{P(\texttt{chd}=1|\boldsymbol{x})}\right) = \theta_0 + \boldsymbol{h}_1(x_1)^\top\boldsymbol{\theta}_1 + \boldsymbol{h}_2(x_2)^\top\boldsymbol{\theta}_2 + \ldots + \boldsymbol{h}_p(x_p)^\top\boldsymbol{\theta}_p,$$

   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).

   Every term in the model should be represented by 4 natural cubic splines. For example, the variable `sbp` is assigned as x_1 and $\boldsymbol{h}_1(x_1)$ describes 4 basis functions for x_1.

   (a) Randomly select a training data set of 300 observations and apply logistic regression with cubic splines. Which variables (basis functions) are significant? Calculate the misclassification rate for an independent test set.

   (b) Apply stepwise variable selection for all observations using `step(...,direction="both")`. Which variables (basis functions) are significant?

   (c) Use the reduced model from (b) and compute the missclassification rate as in (a).

   (d) Plot the variables from the reduced model (b) against their estimated values. How could we interpret this plot?

Please, send your R scripts with the solution as a text file saved as "Surname6.R", via email to

`kynclova@statistik.tuwien.ac.at`

at latest until November 24.