Assignment 2

May 15, 2018

In this assignment you will evaluate the use of non-linear models for predicting *Cscore* on the "prostate" dataset (prostate2.Rdata). In the end you will want to compare these models to what we observed in the first assignment, take that into account.

1. Study the relation between *Cscore* and *lpsa*. Make diagnostic plots for a linear regression of *lpsa* to *Cscore* to study the residuals. Do you see evidence for a non-linear trend? Explain.

2. Compare performance for predicting *Cscore* with *lpsa* using linear model with polynomial expansion, a cubic regression spline and a smoothing spline. Vary the degrees of freedom (until df=10) and evaluate their influence on the performance. Plot the fitted models.

3. Make a full generalized additive model (GAM) with smoothing curves with 5 degrees of freedom for all variables (except for the binary variable). Is this a good model? Does it improve performance over the minimal forward selected model of Assignment 1?

4. Start from a GAM with the variables *lpsa, svi, lweight, lcavol* and *lcp*. For all continuous variables use a smoothing spline with df=5. Optimize the formulation of the GAM, i.e. the variables and the df of the smoothing spline, using backward selection. Explain how you came to your final model and compare performance with the models of Assignment 1.

   - You can do the backward selection manually as in the labs of the chapter.
   - If you want to program the backward selection loop, the function `as.formula(textstring)` allows us to generate a formula for gam from a text string