The Elements of Statistical Learning, WS 2019/2020
Jilles Vreeken and Tobias Marschall
Exercise Sheet #5: *Trees and Beyond Linearity*

Deadline: Thursday, January 9th. 2020, 10:00 a.m.
Please read and adhere to the following requirements to generate a valid submission. This problem set is worth 50 points. You may submit your solutions in groups of two students. The solutions to the theoretical problems should be submitted either digitally (in PDF format) to esl-ta@mpi-inf.mpg.de or as a hard copy before the lecture. **Label your hard copy submissions with your name(s), Matriculation number, and Exercise group.**
Solutions to programming problems, including resulting plots and answers to the questions, need to be submitted in digital format (PDF). For the programming problems you have to submit an R script (that means it needs to run with `Rscript your_file.R`). Write comments within the script to explain what your code is supposed to do.
For digital submissions the subject line of your email should have the following format:

[SL][problem set 5][*tutorial group*]_lastname1_firstname1_lastname2_firstname2

where *tutorial group* is either "Mo" or "We" depending on your assigned time slot. Please include the numbers of the problems you submitted solutions to (both digitally and analogously) in the email's body. **Please make sure that all the files are attached to the email.** The attached files should only include an executable version of your code as .R file and one .pdf file with all the other solutions.

Problem 1 (T, 10 Points).

1. (7P) **Principal Components Analysis**
   The first principal component is the direction of maximum variance in the data. Show that this first principal component also minimizes the residual sum of squares, which is here the squared distance between the projected data point and the original data point.

2. (3P) **Partial Least Squares**
   Show that the first partial least squares direction solves:

$$\max_{\alpha} \ \mathrm{Cor}^2(y, X\alpha)\mathrm{Var}(X\alpha)$$

$$\text{subject to } ||\alpha|| = 1,$$

   i.e., the PLS direction is a compromise between the least squares regression coefficient and the principal component directions.

Problem 2 (T, 5 Points). **MARS: Multivariate Adaptive Regression Splines**
MARS is an adaptive procedure for regression, which uses pairs of piecewise linear basis functions (a sort of very simple splines that is also called *reflected pairs*) of the form $(x-t)_+$ and $(t-x)_+$ with "+" denoting the positive part, e.g.,

$$(x-t)_+ = \begin{cases} x-t & \text{if } x > t, \\ 0, & \text{otherwise.} \end{cases}$$

At the value t, the function has a knot and for each predictor X_j, we generate basis function pairs with knots at every observed value x_{ij}, such that we have a collection of basis function pairs

$$C = \{(X_j - t)_+, (t - X_j)_+\}, \qquad t \in \{x_{1j}, x_{2j}, ..., x_{Nj}\}, \quad j = 1, 2, ..., p.$$

The trained model uses products of the basis functions from C instead of the original inputs and has the form

$$f(X) = \beta_0 + \sum_{m=1}^{M} \beta_m h_m(X),$$

where $h_m(X)$ is one function from C, or a product of two or more such functions. All h_m form the model set $\mathcal{M}$. Model training is done similarly to forward stepwise linear regression, such that iteratively new terms are added to the model set $\mathcal{M}$. Initially, we start with $\mathcal{M}$ containing only the constant function $h_0(X) = 1$ and add in each step a product of a function $h_l \in \mathcal{M}$ with one reflected pair from C. This product is chosen such that adding the term

$$\hat{\beta}_{M+1} h_l(X) \cdot (X_j - t)_+ + \hat{\beta}_{M+2} h_l(X) \cdot (t - X_j)_+$$

to our model decreases the training error most. All coefficients $\hat{\beta}_0, .., \hat{\beta}_{M+2}$ are estimated together using least squares. This process ends when some preset maximum number of terms are contained in $\mathcal{M}$. In order to avoid overfitting, the final model is cropped, i.e., one iteratively removes that term whose removal leads to a minimum increase in training error. This gives us for each possible number of terms an estimated best model. Among all these models, the best one is chosen using generalized cross-validation.

1. How do you have to change the procedure of generating a MARS model to make a decision tree?
2. Can you argue on the basis of the relationship between MARS and decision trees revealed in (a) what is an advantage of MARS over decision trees and what is an advantage of decision trees over MARS?

Problem 3 (T, 15 Points). (Equation 5.4, p. 145 in ESL)
Consider the truncated power series representation for cubic splines with K interior knots. Let

$$f(X) = \sum_{j=0}^{3} \beta_j X^j + \sum_{k=1}^{K} \theta_k (X - \xi_k)_+^3.$$

Prove that the natural boundary conditions for natural cubic splines imply the following linear constraints on the coefficients

$$\beta_2 = 0, \qquad \sum_{k=1}^{K} \theta_k = 0,$$
$$\beta_3 = 0, \qquad \sum_{k=1}^{K} \xi_k \theta_k = 0.$$

Hence, derive the basis

$$N_1(X) = 1, \quad N_2(X) = X, \quad N_{k+2}(X) = d_k(X) - d_{K-1}(X), \quad k = \{1, ..., K-2\}$$

where

$$d_k(X) = \frac{(X - \xi_k)_+^3 - (X - \xi_K)_+^3}{\xi_K - \xi_k}.$$

Problem 4 (P, 20 Points). Go through **6.7 Lab: PCR and PLS Regression** (ISLR p.256–259) and **10.4 Lab 1: Principal Components Analysis** (ISLR p.401–404). We continue the analysis of the prostate dataset from the previous problem set. Download the normalized data set provided in PROSTATE.RDATA. The objective is to predict LPSA from the other features.

1. (4P) Apply best subset selection to the training set. Generate plots for R^2, adjusted R^2, C_p, and BIC in dependence of the number of features. What can you observe? Which model would you chose and why? Which features are used in this model? Calculate training and test error measured in MSE for this model.

2. (4P) Fit principal components regression models for $M = 1, ..., 8$. Plot the train and test error against the number of principal components M. What can you observe?

3. (4P) Fit partial least squares models for $M = 1, ..., 8$. Plot the train and test error against the number of directions M. What can you observe? Compare to the results you obtained when using PCA.

4. (3P) Visualize the whole data set (combining training and test data) and the training data only projected on the first four principal components (using the scores obtained by PCA). Color the data points according to their LPSA value: Set a threshold at 2.5, all samples with an LPSA below should be colored in one color, all other samples in a different color. What can you observe?

5. (3P) Perform the same visualization task using the first four PLS directions. Compare the resulting plots to the PCA plots.

6. (2P) Explain the role of M in the bias-variance tradeoff. Which model would you choose for PCR and PLS, respectively?