The Elements of Statistical Learning, WS 2017/18
Prof. Dr. Dr. Thomas Lengauer
Michael Scherer
D3: Computational Biology & Applied Algorithmics
Max Planck Institute for Informatics

2017-12-13

Problem Set 5

Deadline: Wednesday, January 3. 2018, 10:00 a.m.

Please read and follow 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 `mscherer@mpi-inf.mpg.de` or as a hard copy before the lecture. **Label your hard copy submissions with your name(s).**
Solutions to programming problems and resulting plots need to be submitted in digital format (.pdf). For the programming problems you have to submit an executable version of your code (R script).

For digital submissions the subject line of your email should have the following format:

`[SL][problem set 5] lastname1,firstname1;lastname2,firstname2`

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)

(a) (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.

(b) (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)

Splines will be discussed in the lecture on December, 20th.

Show that regression splines of degree d with K knots form a vector space of dimension $d + K + 1$ by providing a balance of the degrees of freedom in every region of the input data range and the lost degrees of freedom due to the smoothness constraints at the knots. Do not use bases of the spline vector space for your argument.

Problem 3 (T, 15 Points)

(Exercise 5.4 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

$$\beta_2 = 0, \quad \sum_{k=1}^{K} \theta_k = 0,$$

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

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.

(a) (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?

(b) (5P) 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.

(c) (4P) 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?

(d) (4P) Perform the same visualization task using the first four PLS directions. Compare the resulting plots to the PCA plots.

(e) (3P) Explain the role of M in the bias-variance tradeoff. Which model would you choose for PCR and PLS, respectively?