The Elements of Statistical Learning, WS 2018/19

Jilles Vreeken and Tobias Marschall
Michael Scherer, Fawaz Dabbaghie, Aryan Kamal
Center for Bioinformatics & Max Planck Institute for Informatics
CISPA Helmholtz Center i.G. & Cluster of Excellence MMCI

2019-01-03

Problem Set 6

Deadline: Thursday, January 17. 2019, 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 6] 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, 5 Points)

Loss functions
In the regression setting, two different loss functions are quite popular: squared error loss $L(y, f(x)) = (y - f(x))^2$ and absolute loss $L(y, f(x)) = |y - f(x)|$. Squared error loss is differentiable at zero, but the values increase very strongly in regions far away from the origin, which makes the loss function quite sensitive to outliers. Absolute loss does not have the latter drawback but it is not differentiable at zero. Define a loss function which combines the two advantages, i.e., which is differentiable everywhere, mimics squared error loss for values y such that $|y - f(x)| < \delta$ for some threshold δ and is linear for all other x.

Problem 2 (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 3 (T, 5 Points)

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 4 (T, 5 Points)

(Modified version of exercise 7.9.1, ISLR)
We discussed that regression splines with a single knot have the form:

$$f(x) = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \beta_4 (x - \xi)_+^3$$

Where $(x-\xi)_+^3 = (x-\xi)^3$, if $x > \xi$ and 0 otherwise. We will now show that this formula indeed represents a cubic regression spline, irrespective of the βs.

(a) (1P) Express the cubic polynomial

$$f_1(x) = a_1 + b_1 x + c_1 x^2 + d_1 x^3$$

such that $f(x) = f_1(x)$ for all $x \leq \xi$.

(b) (2P) Express the cubic polynomial

$$f_2(x) = a_2 + b_2 x + c_2 x^2 + d_2 x^3$$

such that $f(x) = f_2(x)$ for all $x > \xi$.

(c) (2P) Show that $f(x)$ is continuous, and differentiable up to degree 2 by showing

$$f_1(\xi) = f_2(\xi)$$
$$f_1'(\xi) = f_2'(\xi)$$
$$f_1''(\xi) = f_2''(\xi)$$

Problem 5 (P, 20 Points)

Go through **7.8 Lab: Non-Linear Modeling** (ISLR p.287–297) and **8.3 Lab: Decision Trees** (ISLR p.324–331).

(Adapted from Exercise 8.4.10 in ISLR) We use boosting to predict **Salary** in the **Hitters** data set. Load the **ISLR** library, which contains the data.

(a) (2P) Remove the observations for whom the salary information is unknown, and then log-transform the salaries and remove the qualitative variables *Division*, *League* and *NewLeague*. Create a training set consisting of the last 200 observations and a test set consisting of the remaining observations.

(b) (2P) Perform Best Subset Selection to select the three predictors that appear to be most highly associated with the salary. Which variables do you select?

(c) (5P) Create three polynomial regression models, each containing all the variables selected in b) and one cubic transformation. For which variables do you observe a positive effect of using the polynomial?

(d) (4P) Use cubic splines to describe the relationship between the three variables and the output. Is there a linear relationship for one of the variables? Are the results in line with c)?

(e) (5P) Perform boosting on the training set with 1000 trees for a range of values of the shrinkage parameter λ. Produce a plot showing training and test set MSE for the different shrinkage values. Comment on your observations, choose a value for λ and justify your choice.

(f) (2P) Compare the test MSE of boosting to the MSE that results from applying least squares regression and polynomial regression.