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

2017-11-29

Problem Set 4

Deadline: Wednesday, December 13. 2017, 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 4] 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)

Prove that for linear and polynomial least squares regression, the LOOCV estimate for the test MSE can be calculated using the following formula

$$\mathrm{CV}_{(n)} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{y_i - \hat{y}_i}{1 - h_i}\right)^2,$$

where h_i is the leverage. Note, that the leverage h_i for point i is given by the diagonal element pertaining to data point i of the hat matrix H.

Problem 2 (T, 12 Points)

Ridge Regression

(a) Ridge regression is done by minimizing the RSS with a quadratic penalty term:

$$\underset{\beta}{\text{minimize}} \quad (y - X\beta)^T(y - X\beta) + \lambda\beta^T\beta$$

Show that the solutions take the form:

$$\hat{\beta}^{ridge} = (X^TX + \lambda \mathrm{I})^{-1}X^Ty,$$

where I is the $p \times p$ identity matrix.

(b) Ridge regression can be expressed as an unconstrained optimization problem:

$$\underset{\beta}{\text{minimize}} \sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij}\right)^2 + \lambda\sum_{j=1}^{p}\beta_j^2$$

Show that this is equivalent to the constrained optimization problem:

$$\underset{\beta}{\text{minimize}} \quad \sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij}\right)^2$$

$$\text{subject to} \quad \sum_{j=1}^{p}\beta_j^2 \le s$$

Comment on the relationship between s and λ.
Hint: Use Lagrange multipliers.

Problem 3 (T, 8 Points)

Assume a scenario in which the number of observations equals the number of features (n=p) and X is the $n \times n$ identity matrix. Furthermore, assume that we perform regression without an intercept. In this setting, lasso simplifies to

$$\underset{\beta}{\text{minimize}} \sum_{j=1}^{p} (y_j - \beta_j)^2 + \lambda \sum_{j=1}^{p} \left|\beta_j\right|.$$

Show that the lasso estimates take the form:

$$\hat{\beta}_j^{\text{lasso}} = \begin{cases} y_j - \frac{\lambda}{2}, & \text{if } y_j > \frac{\lambda}{2}; \\ y_j + \frac{\lambda}{2}, & \text{if } y_j < -\frac{\lambda}{2}; \\ 0, & \text{if } |y_j| \leq \frac{\lambda}{2}; \end{cases}$$

Problem 4 (P, 20 Points)

Go through **5.3 Lab: Cross-Validation and the Bootstrap** (ISLR p.190–197), **6.5 Lab 1: Subset Selection Methods** (ISLR p. 244–251), and **6.6 Lab 2: Ridge Regression and the Lasso** (ISLR p. 251–255). The objective of this programming exercise is to predict the logarithm of the prostate specific antigen (PSA) level based on the other predictors. You find the dataset *prostate.txt* on the course website.

(a) (2P) Read and normalize the data: use *read.table()* to load the data; column 9 is the output lpsa for the regression and column 10 determines whether this data entry belongs to the training set. Column 1 is just an index and should not be used for prediction. Normalize each input feature to a mean of 0 and a variance of 1. Split up the data set into training and test set respectively. Useful functions: mean(), sd(), and the MASS library

(b) (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.

(c) (3P) Use the training set to fit ridge regression models and generate a plot showing the values of the coefficients in relation to the parameter λ (cf. Figure 6.4, p. 216, ISLR). What can you observe?

(d) (3P) Perform 5-fold cross-validation on the training set to determine the optimal value for λ for the ridge regression model. Report train and test error measured in MSE for this λ.

(e) (3P) Use the training set to fit lasso models and generate a plot showing the values of the coefficients in relation to the parameter λ (cf. Figure 6.6, p. 220, ISLR). What can you observe in comparison to the plot generated in (c)?

(f) (3P) Perform 5-fold cross-validation on the training set to determine the optimal value for λ in the lasso. Report train and test error measured in MSE for this λ. How many and which features are used? Compare this to the coefficients determined for ridge regression in (d).

(g) (2P) Train a linear regression model using all the features on the training data. Report train and test error measured in MSE. Compare the models generated in (d) and (f) to this one.