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

2018-11-29

Problem Set 4

Deadline: Thursday, December 13. 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 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)

LDA and QDA

- (5P) (Exercise 4.7.2 in ISLR):
  This problem relates to the LDA model. It was stated in the text that classifying an observation to the class for which (4.12, ISLR) is largest is equivalent to classifying an observation to the class for which Equation (4.13, ISLR) is largest. Prove that this is the case. In other words, under the assumption that the observations in the kth class are drawn from a $N(\mu_k, \ \sigma^2)$ distribution, the Bayes' classifier assigns an observation to the class for which the discriminant function is maximized.
- (5P) (Exercise 4.7.3 in ISLR):
  This problem relates to the QDA model, in which the observations within each class are drawn from a normal distribution with a class-specific mean vector and a class-specific covariance matrix. We consider the simple case where p = 1; i.e. there is only one feature. Suppose that we have k classes and that if an observation belongs to the kth class then X comes from a one-dimensional normal distribution, $X \sim N(\mu_k, \sigma_k^2)$. Recall that the density function for the one-dimensional normal distribution is given in Equation (4.11, ISLR). Show that in this case, the Bayes' classifier is not linear. Argue that it is in fact quadratic.

Problem 2 (T, 8 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 3 (T, 12 Points)

- (6P) Cross-Validation (Exercise 5.4.3 in ISLR):

  (a) (3P) Explain how k-fold cross-validation is implemented.

  (b) (3P) Argue about the advantages and disadvantages of k-fold cross-validation relative to the validation set approach.

- (6P) The Bootstrap (Exercise 5.4.1 in ISLR):
  Show that formula 5.6 in ISLR:

$$\alpha = \frac{\sigma_X^2 - \sigma_{XY}}{\sigma_X^2 + \sigma_Y^2 - 2\sigma_{XY}}$$

  indeed minimizes the “risk of investment" $Var(\alpha X + (1 - \alpha)Y)$.

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) Use LOOCV, 5- and 10-fold cross-validation on the training data set to estimate the test error of using linear regression to predict lpsa from all other features. Use the full training data set to train a linear regression model and compute the test error. Compare your estimates obtained from cross validation to the error obtained from the test set and argue about your findings. Which of the methods is (theoretically) fastest?

(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 10-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 10-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) Compare the models generated in (d) and (f) to the model generated in (a). Which model would you choose? What alternative model could have been used?