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

2018-12-13

Problem Set 5

Deadline: Thursday, January 3. 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 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.

There won't be a lecture on January, 3. Either scan your handwritten solutions and send them to mscherer@mpi-inf.mpg.de or submit them in Prof. Marschall's mailbox at the CBI (ground floor).

Problem 1 (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:

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

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

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

Problem 2 (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 3 (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:

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

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

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

Problem 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) 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.

(b) (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?

(c) (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.

(d) (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?

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

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