Problem Set 4
CSCI 688

Problem 1 Let $X \in \mathbb{R}^{N \times P}$ be our design matrix and $Y \in \mathbb{R}^N$ be our response variables. Let $\widetilde{X}$ and $\widetilde{Y}$ be augmented versions of X and Y where

$$\widetilde{X} = \begin{bmatrix} X \\ \sqrt{\lambda} I_P \end{bmatrix}$$

adds P rows to X each with values $\sqrt{\lambda}$ and

$$\widetilde{Y} = \begin{bmatrix} Y \\ 0 \end{bmatrix}$$

adds P zeros on to the end of Y. Show that the coefficients $\hat{\beta}$ associated with regressing $\widetilde{Y}$ onto $\widetilde{X}$ is equivalent to the coefficients found from fitting a Ridge regression estimator of Y onto X. This can be interpreted as shrinking our estimate of $\hat{\beta}$ by adding hints into our data that, for many of our data points, the coefficient is zero. Hint: if

$$C_1 = \begin{bmatrix} A_1 \\ B_1 \end{bmatrix} \text{ and } C_2 = \begin{bmatrix} A_2 \\ B_2 \end{bmatrix}$$

are block matrices then

$$C_1^T C_2 = A_1^T A_2 + B_1^T B_2$$

Problem 2 Suppose we estimate the regression coefficients in a linear regression model by minimizing

$$\sum_{i=1}^{n} (y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij})^2$$

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

for a particular value of s. For parts (a) through (d), indicate which of i. through v. is correct.

(a) As we increase s from 0, the training RSS will:
   i. Increase initially, and then eventually start decreasing in an inverted U shape.
   ii. Decrease initially, and then eventually start increasing in a U shape.
   iii. Steadily increase.
   iv. Steadily decrease.
   v. Remain constant.

(b) Repeat (a) for test RSS.

(c) Repeat (a) for variance.

(d) Repeat (a) for (squared) bias.

Problem 3 In this exercise, we will look at gene expression data from prostate tumors. The dataset is called `prostate` and can be found in the `spls` package using the command `data(prostate)`.

(a) Get the matrix of gene expressions `prostate$x`. Split off the first column as a new variable `y = prostate$x[,1]` to predict from the remaining columns. Now split the data set into a training set and a test set.

(b) Fit a linear model using least squares on the training set. Does this model make sense?

(c) Suppose I wanted to do forward step-wise variable selection. If I wanted to compare all models including one covariate X how many linear models would I have to fit?

(d) Fit a ridge regression model on the training set, with λ chosen by cross-validation. Report the test error obtained.

(e) Fit a lasso model on the training set, with λ chosen by cross-validation. Report the test error obtained, along with the number of non-zero coefficient estimates.

(f) Fit an elastic-net model with $\alpha = 0.5$ on the training set, with λ chosen by cross-validation. Report the test error obtained, along with the number of non-zero coefficient estimates.

(g) Fit a PCR (PCA Regression) model on the training set. How many PCs do you choose? Report the test error obtained.

(h) Comment on the results obtained.