Homework I

Statistics 640 / 444 Fall 2014

Assigned: September 4

Due: September 18

Stat 640:

1. ESL Textbook Problem 3.27 parts (a), (b), (c).
   Hint: More information on the KKT conditions can be found in the book Convex Optimization, `http://www.stanford.edu/~boyd/cvxbook/`*, chapter 5.5.*

2. Suppose $p >> n$ and you would like to compute the ridge regression estimator without inverting a $p \times p$ matrix. Devise a strategy to do so. If you use any matrix algebra results or tricks, prove that these give the desired solution.

3. Prove that the least squares solution has zero training error (RSS = 0) when $\text{rank}(\mathbf{X}) > n$. Interpret this result.

4. Gene Expression Data:
   Download the SRBCT microarray data from the course website. This is a gene expression data set from a childhood cancer study with $n = 83$ patients and $p = 2308$ genes. Your response is the expression profile of p53, a major oncogene that acts as a tumor suppressor. Your goal is to select other genes whose expression profiles are associated with p53.

   (a) Visualize regularization paths for the following methods:
      i. Elastic net.
      ii. Lasso.
      iii. SCAD.
      iv. MC+.

   (b) Reflection. Interpret the results. What are the top genes selected by each method? Are they different? If so, why? Which regularization paths look most variable? Why is this the case? If you had to report to a scientist the top 10 genes associated with p53, which ones would you report? Why?

Stat 444:

1. Show empirically that least squares has zero training error (RSS = 0) when $p > n$. You may do this by creating your own simulation or by using one of the data sets below. Interpret this result.

2. Gene Expression Data:
   Download the SRBCT microarray data from the course website. This is a gene expression data set from a childhood cancer study with $n = 83$ patients and $p = 2308$ genes. Your response is the expression profile of p53, a major oncogene that acts as a tumor suppressor. Your goal is to select other genes whose expression profiles are associated with p53.

   (a) Visualize regularization paths for the following methods:
      i. Lasso.
      ii. Elastic Net.

(b) Reflection. Interpret the results. What are the top genes selected by each method? Are they different? If so, why? Which regularization paths look most variable? Why is this the case? If you had to report to a scientist the top 10 genes associated with p53, which ones would you report? Why?

3. Prostate Cancer Data:
   Download the prostate cancer data from `http://www-stat.stanford.edu/~tibs/ElemStatLearn/`. For this problem, you will need to randomly split the data set into training and test sets of equal sizes, fit the models, and repeat this procedure 10 times.

   (a) Compare the training and test errors for the following methods using the best 5 predictors:
   - Least squares (No selection necessary).
   - Best Subsets.
   - Forward Step-wise.
   - Backwards Step-wise.
   - LASSO.
   - Principal Components Regression (Use top 5 PCs).
   - Partial Least Squares Regression (Use top 5 PLS components).
   - Ridge Regression (Instead, fix a λ value).
   - Elastic Net.
   - Adaptive Lasso.
   - SCAD.

   (b) Present a graphic or table summarizing the results.

   (c) Reflection. Which method gives the best test error? Why do these methods perform well? Do any methods seem to overfit to the training set? If so, why? Which variables are in the model? Do all the methods choose the same subset of variables? Is variable selection necessary for this data? Explain and expand on your responses.

Hint: The `R` *packages* `leaps`, `MASS`, `ncvreg` *and* `glmnet` *may be helpful.*