STAT 542, Homework 1

February 10, 2018
Due date: Feb 23, 5:00 pm

Requirements: This homework consists of two problems. You should submit your **report and `R` code** as separate files to Compass2g. The report should be in PDF/html/MS Word format, with no more than 6 pages, including plots. Your report should focus on describing the analysis strategy, demonstrating and interpreting the main result. No excessive `R` code should be included in the report. Font size should be 12pt and plots need to be clearly labeled. If you use R markdown and need a template, there is an example source file provided at our course website.

Question 1 [50 points]: This is a simple implementation of the Lasso. For this question, you cannot use any additional `R` package. Lets consider the objective function

$$f(\boldsymbol{\beta}, \beta_0) = \frac{1}{2n}\|\mathbf{y} - \beta_0 - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda \sum_{j=1}^{p} |\beta_j|$$

Please note that the intercept term β_0 is not subject to any penalty. We will use the coordinate descent algorithm to solve this optimization problem.

a. [10 points] Fixing all other parameters, write down the one-variable optimization problem of β_0 based on this objective function. Given the explicit form of the solution to this problem.

b. [10 points] Fixing all other parameters, write down the one-variable optimization problem of β_j for $j = 1, \ldots, p$, based on this objective function. Give the explicit form of the solution to this problem.

c. [10 points] Fixing all β_j's, $j = 1, \ldots, p$ to be 0. Use part a) to update β_0 to its optimal solution. After this update, find the smallest λ value such that none of the β_j's, $j = 1, \ldots, p$ can be updated out of zero in the next iteration. Denote this value as $\lambda_{\max}$.

d. [20 points] Implement the above procedures to get a path-wise coordinate descent solution of the lasso problem. Use the code in the `HW2.r` file to generate the data. You can either follow my structure of the code or write your own based on your understanding. Generate an independent set of 1000 observations under the same model as the test dataset. Use the test data to select the best tuning parameter λ. Report details of your tuning procedure and the results.

Question 2 [50 points]: This is a simulation study for the degrees of freedom. Using the `HW2.r` file to generate the training data ($N = 500$, $P = 200$). Use the `glmnet` package and a 10 fold cross-validation to select the penalty value. We want to then understand what is the degrees of freedom using this penalty.

a. [10 points] Run the `glmnet` package and a 10 fold cross-validation to select the best lambda. Record the number of nonzero parameters using this lambda. For the rest of this question, fix this lambda value.

b. [25 points] Recall that the degrees of freedom of a fit is defined as $\sum_{i=1}^{n} \text{Cov}(\widehat{y}_i, y_i)/\sigma^2$, we need to repeatedly generate new training data and refit the model to calculate this quantity. Consider the following steps:

   - Fix the X values throughout this question.
   - Using our model, generate a new set of response variables $\mathbf{y}$ for the training data. Fit the lasso model with the pre-selected lambda value. Obtain the fitted value $\widehat{y}_i$'s.
   - Perform this procedure 20 times so that you can estimate $\sum_{i=1}^{n} \text{Cov}(\widehat{y}_i, y_i)/\sigma^2$.
   - Report your estimated degrees of freedom and compare it to the number of non-zero parameter values you obtained from part a).

c. [15 points] Repeat part a) and b) for the ridge regression. Use the generalized cross validation criteria to select the best lambda. What is the theoretical value of the degrees of freedom for this lambda value? Does your estimation matches (or close to) the theoretical value?