Homework Assignment 2

Due by Thursday Oct 8th*

SDS 383C Statistical Modeling I

1. (2+5+3 pts) Get the passenger car mileage data from `http://lib.stat.cmu.edu/DASL/Datafiles/carmpgdat.html`

   (a) Fit multiple linear regression model to predict MPG (miles per gallon) from the other variables. Summarize your analysis.

   (b) Use Mallow C_p to select a best sub-model. To search through the models try (i) forward stepwise, (ii) backward stepwise. Summarize your findings.

   (c) Use the Zheng-Loh model selection method and compare to (b). You will write your own code to do the steps.

2. (8+2) Get the Prostrate cancer data from `http://statweb.stanford.edu/~tibs/ElemStatLearn/datasets/prostate.data`. More information about this dataset can be found in `http://statweb.stanford.edu/~tibs/ElemStatLearn/datasets/prostate.info.txt`.

   (a) Carry out a best-subset linear regression analysis (as in Table 3.3 of the H-T-F book). Compute the AIC, BIC, five and tenfold cross-validation estimates of prediction error. Read section 7.10.2 first from the book.

   (b) Compare the results. You will write your own code for cross validation, but you can use builtin code for doing best subset selection.

3. (8 pts) Consider a linear regression with p parameters, fit by least squares to a set of training data $(x_1, y_1), \ldots, (x_N, y_N)$ drawn at random from a population. Let $\hat{\beta}$ be the least squares estimate. Suppose we have some test data $(\tilde{x}_1, \tilde{y}_1), \ldots, (\tilde{x}_M, \tilde{y}_M)$ drawn at random from the same population as the training data. If $R_{\rm tr}(\beta) = \frac{1}{N}\sum_{i=1}^{N}(y_i - \beta^{\rm T}x_i)^2$ and $R_{\rm te}(\beta) = \frac{1}{M}\sum_{i=1}^{M}(\tilde{y}_i - \beta^{\rm T}\tilde{x}_i)^2$, prove that

$$E\left[R_{\rm tr}(\hat{\beta})\right] \leq E\left[R_{\rm te}(\hat{\beta})\right]$$

   where expectations are over all that is random in each expression. Note that this setting is different from the setting we used for proving the optimism of training error in class, namely the in sample error. Here both x and y values are random, which makes the problem much easier.

*If you are late, you can use a grace period of 3 days, and turn it in via email.

4. (3+8 pts) Consider a design matrix $\boldsymbol{X}$ such that $\boldsymbol{X}$ has orthonormal columns, i.e. $\boldsymbol{X}^T\boldsymbol{X} = I$, where I is the $p \times p$ identity matrix. Consider the following regularization.

$$\min_{\beta} \frac{1}{2}(\boldsymbol{X}^T\boldsymbol{\beta} - \boldsymbol{y})^T(\boldsymbol{X}^T\boldsymbol{\beta} - \boldsymbol{y}) + \lambda\|\boldsymbol{\beta}\|_0, \tag{1}$$

where $\|\boldsymbol{\beta}\|_0 = \sum_{i=1}^{p} |\beta_i|$.

   (a) Show that the least squares estimate is $\hat{\boldsymbol{\beta}} = \boldsymbol{X}^T\boldsymbol{y}$.

   (b) Show that the solution to equation 1 is given by $\tilde{\boldsymbol{\beta}}$, where

$$\tilde{\beta}_i = \begin{cases} \boldsymbol{v}_i^T\boldsymbol{y} & \text{if } |\boldsymbol{v}_i^T\boldsymbol{y}| > \sqrt{2\lambda} \\ 0 & \text{if } |\boldsymbol{v}_i^T\boldsymbol{y}| \leq \sqrt{2\lambda} \end{cases} \tag{2}$$

   This is also called the hard thresholding estimator. $\boldsymbol{v}_i$ is the i^{th} column of $\boldsymbol{X}$.

5. (6 pts) In this problem we will look at how collinearity affects ridge regression and lasso.

   (a) Suppose we run a ridge regression with parameter λ on p variables $X_1, \ldots, X_p$. The coefficient I estimate for X_1 ($\hat{\beta}_{ridge}(1)$) is a. Now $m-1$ additional copies of variable X_1, i.e. $X_1^* = X_2^* = \cdots = X_{m-1}^* = X_1$ are included and the ridge regression is refit. How are the new coefficients of the identical copies related to a ? Prove your answer.

   (b) Repeat the above problem with lasso.