STATS 415 Homework 6

Due **Tuesday** March 6, 2018

Please include your name, uniqname, and lab section (number or time or GSI). A point will be taken off homework without the section info. Turn in a printout of your homework in the lecture or in your GSI's mailbox across room 305A West Hall, no later than 5pm on the due date.

1. Suppose we estimate the regression coefficients in a linear regression model by minimizing

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

$$\text{subject to} \quad \sum_{j=1}^{p}|\beta|_j \leq s$$

for a particular value of s. Complete each sentence (a)-(e) below by choosing the best option among (1)-(5). All choices below refer to the overall trends, which may be subject to noise variations.

(a) As we increase s from 0, the number of variables included in the model ...

(b) As we increase s from 0, the training RSS ...

(c) As we increase s from 0, the test RSS ...

(d) As we increase s from 0, the variance of $\hat{\beta}$ ...

(e) As we increase s from 0, the squared bias of $\hat{\beta}$ ...

Answer options:

(1) Increases initially, and then eventually starts decreasing.

(2) Decreases initially, and then eventually starts increasing.

(3) Steadily increases.

(4) Steadily decreases.

(5) Remains constant.

2. In this exercise, we will predict the number of applications received using the other variables in the `College` data set.

   (a) Split the data set into a training set and a test set. Fix the random seed to the value 23456 and choose 30% (rounded down to the nearest integer) of the data at random for testing, and use the rest for training.

   (b) Fit a linear model using least squares on the training set, and report the training and test error obtained.

   (c) Perform forward and backward selection on the previous model with the threshold $\alpha = 0.05$, and report which variables they recommend including in the final model. Report training and test errors for their final models.

   (d) Use AIC and BIC to select a potentially smaller model instead of the model in question (b). Report which variables they recommend including, and the training and test errors for their final models.

   (e) Fit a ridge regression model on the training set, with λ chosen by cross-validation. Report the training and test errors.

   (f) Fit a lasso model on the training set, with λ chosen by cross-validation. Report which variables are included in the model, and the training and test errors obtained.

   (g) Comment on the results obtained. How accurately can we predict the number of college applications received? How much difference is there among the test errors resulting from different approaches? Which approach would you recommend for this dataset and why?

Please limit your solution to at most 6 pages.