CS 5565, LAB5(Subset Selection, Ridge and Lasso, PCR and PLS) 100 pts.

Name ______________________________________

1. View the videos at the following URLs
   https://www.youtube.com/watch?v=3kwdDGnV8MM
   https://www.youtube.com/watch?v=mv-vdysZIb4
   https://www.youtube.com/watch?v=F8MMHCCoALU
   https://www.youtube.com/watch?v=1REe3qSotx8
   You may download the R Code for Labs and the Data Sets to use from the textbook website.
   http://www-bcf.usc.edu/˜gareth/ISL/

2. (30 points total) In this exercise, we will generate simulated data, and will then use this data to perform best subset selection.

   (a) (5 points) Use the `rnorm()` function to generate a predictor X of length $n = 100$, as well as a noise vector ϵ of length $n = 100$.

   (b) (5 points) Generate a response vector Y of length $n = 100$ according to the model

   $$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3 + \epsilon,$$

   where $\beta_0, \beta_1, \beta_2$, and β_3 are constants of your choice.

   (c) (5 points) Use the `regsubsets()` function to perform best subset selection in order to choose the best model containing the predictors $X, X^2, \ldots, X^{10}$. What is the best model obtained according to C_p, BIC, and adjusted R^2? Show some plots to provide evidence for your answer, and report the coefficients of the best model obtained. Note you will need to use the `data.frame()` function to create a single data set containing both X and Y.

   (d) (5 points) Repeat (c), using forward stepwise selection and also using backwards stepwise selection. How does your answer compare to the results in (c)?

   (e) (5 points) Now fit a lasso model to the simulated data, again using $X, X^2, \ldots, X^{10}$ as predictors. Use cross-validation to select the optimal value of λ. Create plots of the cross-validation error as a function of λ. Report the resulting coefficient estimates, and discuss the results obtained.

   (f) (5 points)Now generate a response vector Y according to the model

   $$Y = \beta_0 + \beta_7 X^7 + \epsilon,$$

   and perform best subset selection and the lasso. Discuss the results obtained.

3. (35 points total) In this exercise, we will predict the number of applications received using the other variables in the `College` data set.

   (a) (5 points) Split the data set into a training set and a test set.

   (b) (5 points) Fit a linear model using least squares on the training set, and report the test error obtained.

(c) (5 points) Fit a ridge regression model on the training set, with λ chosen by cross-validation. Report the test error obtained.

(d) (5 points) Fit a lasso model on the training set, with λ chosen by crossvalidation. Report the test error obtained, along with the number of non-zero coefficient estimates.

(e) (5 points) Fit a PCR model on the training set, with M chosen by crossvalidation. Report the test error obtained, along with the value of M selected by cross-validation.

(f) (5 points) Fit a PLS model on the training set, with M chosen by crossvalidation. Report the test error obtained, along with the value of M selected by cross-validation.

(g) (5 points) Comment on the results obtained. How accurately can we predict the number of college applications received? Is there much difference among the test errors resulting from these five approaches?

4. (35 points total) We have seen that as the number of features used in a model increases, the training error will necessarily decrease, but the test error may not. We will now explore this in a simulated data set.

   (a) (5 points) Generate a data set with $p = 20$ features, $n = 1,000$ observations, and an associated quantitative response vector generated according to the model

   $$Y = X\beta + \epsilon,$$

   where β has some elements that are exactly equal to zero.

   (b) (5 points) Split your data set into a training set containing 100 observations and a test set containing 900 observations.

   (c) (5 points) Perform best subset selection on the training set, and plot the training set MSE associated with the best model of each size.

   (d) (5 points) Plot the test set MSE associated with the best model of each size.

   (e) (5 points) For which model size does the test set MSE take on its minimum value? Comment on your results. If it takes on its minimum value for a model containing only an intercept or a model containing all of the features, then play around with the way that you are generating the data in (a) until you come up with a scenario in which the test set MSE is minimized for an intermediate model size.

   (f) (5 points) How does the model at which the test set MSE is minimized compare to the true model used to generate the data? Comment on the coefficient values.

   (g) (5 points) Create a plot displaying $\sqrt{\sum_{j=1}^{p}(\beta_j - \hat{\beta}_j^r)^2}$ for a range of values of r, where $\hat{\beta}_j^r$ is the jth coefficient estimate for the best model containing r coefficients. Comment on what you observe. How does this compare to the test MSE plot from (d)?