HW6: Linear Model Selection and Regularization

**Directions.** Show and explain all work to receive full credit. Homework is due on **Thursday, May 18th** at the beginning of class by 12:00 pm.

*Note:* Throughout the assignment, you are allowed to use the `LinearRegression`, `RidgeCV`, `LassoCV` classes from scikit-learn.

**Problem 1.** 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. *Show your computations by displaying the function calls.*

(a) Use the the NumPy function `random.randn` to generate a data set with \( p = 20 \) features and \( n = 1,000 \) observations, and an associated quantitative response vector generated according to the model

\[ Y = \beta_0 + \beta_1 X_1 + \ldots + \beta_{20} X_{20} + \varepsilon, \]

and set 5 coefficients (besides the intercept) to zero exactly.

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

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

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

(e) 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) 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) Create a plot displaying \[ \sqrt{\sum_{j=1}^{20} (\beta_j - \hat{\beta}_j^{(d)})^2} \] for \( d = 0, 1, \ldots, 20 \), where \( \hat{\beta}_j^{(d)} \) is the \( j \)-th coefficient estimate for the best model containing exactly \( d \) predictors. Comment on what you observe. How does this compare to the test MSE plot from (d)?
Problem 2. In this problem, we will predict the number of applications received using the other variables in the College data set. Show your computations by displaying the function calls.

(a) Split the data into a training set and a test set. The training set will contain 80% of the data points and the test set will contain the remaining observations.

(b) Standardize the training data using the function standardize coded in HW3. Then standardize the test data using the empirical means and standard deviations computed from the training data.

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

From now on, you will use the standardized data.

(d) Fit a ridge regression model on the training set, where the tuning parameter \( \lambda \) is chosen by ten-fold cross-validation. Report the test error obtained.

(e) Fit a lasso model on the training set, where the tuning parameter \( \lambda \) is chosen by ten-fold cross-validation. Report the test error obtained.

(f) 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?