Statistical Data Mining I
Homework 2

Due: Friday October 13th (11:59 pm)
30 points

Directions: See UB learns for Homework Guidelines.

1) (10 points) (Exercise 9 modified, ISL) In this exercise, we will predict the number of applications received using the other variables in the College data set in the ISLR package.
(a) Split the data set into a training set and a test set. Fit a linear model using least squares on the training set, and report the test error obtained.
(b) Fit a ridge regression model on the training set, with λ chosen by cross-validation. Report the test error obtained.
(d) 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) Fit a PCR model on the training set, with k chosen by cross-validation. Report the test error obtained, along with the value of k selected by cross-validation.
(f) Fit a PLS model on the training set, with k chosen by crossvalidation.
Report the test error obtained, along with the value of k selected by cross-validation.
(g) 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?

2) (10 points) The insurance company benchmark data set gives information on customers. Specifically, it contains 86 variables on product-usage data and socio-demographic data derived from zip area codes. There are 5,822 customers in the training set and another 4,000 in the test set. The data were collected to answer the following questions: Can you predict who will be interested in buying a caravan insurance policy and give an explanation why? Compute the OLS estimates and compare them with those obtained from the following variable-selection algorithms: Forwards Selection, Backwards Selection, Lasso regression, and Ridge regression. Support your answer.
(The data can be downloaded from https://kdd.ics.uci.edu/databases/tic/tic.html.)

3) (10 points) (Exercise 9 modified, ISL) 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.
Generate a data set with p = 20 features, n = 1, 000 observations, and an associated quantitative response vector generated according to the model
$Y = X\beta + \varepsilon$
where β has some elements that are exactly equal to zero. Split your data set into a training set containing 100 observations and a test set containing 900 observations.

Perform best subset selection on the training set, and plot the training set MSE associated with the best model of each size. Plot the test set MSE associated with the best model of each size.

For which model size does the test set MSE take on its minimum value? Comment on your results. 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.

(*Note: 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 until you come up with a scenario in which the test set MSE is minimized for an intermediate model size.*)