Assignment #5: Advanced Regression and Tree-based Methods

Problem 1

In this question, we will predict the number of applications received (Apps) using the other variables in the College data set (ISLR package).

(a) Perform best subset selection to the data. 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.

(b) Repeat (a) using forward stepwise selection and backwards stepwise selection. How does your answer compare to the results in (a)?

(c) Fit a lasso model on the data. 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.

(d) Fit a ridge regression model on the data. 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.

(e) Now split the data set into a training set and a test set.

i. Fit the best models obtained in the best subset selection (according to C_p, BIC or adjusted R^2) to the training set, and report the test error obtained.

ii. Fit a lasso model to the training set, with λ chosen by cross validation. Report the test error obtained.

iii. Fit a ridge regression model to the training set, with λ chosen by cross validation. Report the test error obtained.

iv. Compare the test errors obtained in the above analysis (i-iii) and determine the optimal model.

Problem 2

In the lab, a classification tree was applied to the Carseats data set after converting Sales into a binary response variable. This question will seek to predict Sales using regression trees and related approaches, treating the response as a quantitative variable (that is, without the conversion).

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

(b) Fit a regression tree to the training set. Plot the tree, and interpret the results. Then compute the test MSE.

(c) Prune the tree obtained in (b). Use cross validation to determine the optimal level of tree complexity. Plot the pruned tree and interpret the results. Compute the test MSE of the pruned tree. Does pruning improve the test error?

(d) Use the bagging approach to analyze the data. What test MSE do you obtain? Determine which variables are most important.

(e) Use random forests to analyze the data. What test MSE do you obtain? Determine which variables are most important.

Problem 3

In the lab, we applied random forests to the Boston data using mtry=6 and ntree=100.

(a) Consider a more comprehensive range of values for mtry: 1, 2,…,13. Given each value of mtry, find the test error resulting from random forests on the Boston data (using ntree=100). Create a plot displaying the test error rate vs. the value of mtry. Comment on the results in the plot.

(b) Similarly, consider a range of values for ntree (between 5 to 200). Given each value of ntree, find the test error resulting from random forests (using mtry=6). Create a plot displaying the test error vs. the value of ntree. Comment on the results in the plot.