CS 6220 - Fall 2015 Olga Vitek

Homework 8

Each part of the problems 5 points
Due on Blackboard by **5pm on Tuesday November 24**.
JWHT is 'An Introduction to Statistical Learning' by James, Witten, Hastie and Tibshirani.

A tax assessor recorded residential home sales prices in a midwestern city, along with various characteristics of the home and surrounding property. Data on 522 arms-length transactions were obtained for home sales during the year 2002, and is stored in the file `realEstate.RData` on the course website. Each line of the data set has an identification number and provides information on 12 other variables. The 13 variables are:

1. Identification number: 1-522
2. Sales price: Sales price of residence (dollars)
3. Finished square feet: Finished area of residence (square feet)
4. Number of bedrooms: Total number of bedrooms in residence
5. Number of bathrooms: Total number of bathrooms in residence
6. Air conditioning: Presence or absence of air conditioning: 1 if yes; 0 otherwise
7. Garage size: Number of cars that garage will hold
8. Pool: Presence or absence of swimming pool: 1 if yes; 0 otherwise
9. Year built: Year property was originally constructed
10. Quality: Index for quality of construction: 1 indicates high quality; 2 indicates medium quality; 3 indicates low quality
11. Style: Qualitative indicator of architectural style
12. Lot size: Lot size (square feet)
13. Adjacent to highway: Presence or absence of adjacency to highway: 1 if yes; 0 otherwise

In this problem we are interested in predicting the quality of the construction from the transaction data. Create a new binary response, by letting Y = I if quality (variable 10) equals I, and Y =0 otherwise (i.e., if quality equals 2 or 3).

1. Answer the questions in JWHT Problem 9, page 334, using this dataset. Randomly select 350 observations for the training set, and the remainder for the validation set.
2. Perform bagging on the training set, and compare the % of correct predictions on the training and the validation set to those of the best single-tree classification.
3. Perform random forest on the training set, and compare the % of correct predictions on the the training and the validation set to those of the bagging, and of the best single-tree classification.

4. Perform boosting on the training set with 1,000 trees for a range of values of the shrinkage parameter λ.

   (a) Produce a plot with different shrinkage values on the x-axis and the corresponding training set % of correct predictions on the y-axis.

   (b) Produce a plot with different shrinkage values on the x-axis and the corresponding test set % of correct predictions on the y-axis.

   (c) Compare the training and the test set % of correct predictions to those of the approaches above.

   (d) Which variables appear to be the most important predictors in the boosted model? Are these the same variables chosen in the best single-tree approach?