Introduction to Statistical Learning, Spring 2016

Homework exercise 4

Due date: 9 June 2016 to my mailbox on floor 1 of Schreiber building

1. In class we showed how the regular Boosting implementation as described in the book, which uses the current residuals as the response for building each tree, is in fact performing gradient descent for regression with a squared error loss function. In this problem, we will discuss applying the same idea to regression with an absolute loss function: $L(y, \hat{y}) = |y - \hat{y}|$.

   (a) For the b^{th} iteration, write the loss function in terms of the current residual (which we denoted $R_b = y_i - \epsilon \sum_{k=1}^{b} \hat{y}_{k,i}$ in class).

   (b) Calculate its derivative, and explain what would need to change in the boosting algorithm so it would implement gradient descent in this loss function.

   (c) The code `boost.r` available from the class homepage, implements regular gradient boosting with squared error loss on the Boston dataset. Implement the absolute loss version as well. Using appropriate choices of depth, ϵ and B (you can try several), compare prediction performance of both boosting models on the test set. Use both squared error loss and absolute loss in comparing the test performance. Which method does better on which performance task? Is this as you would expect?
   Hints: You need to change only one line in the main loop to implement the change of loss function, think carefully which one... You may find the R function `sign()` useful.

2. We said in class that in Bootstrap sampling, the probability that an observation does not appear in the sample is $(1 - \frac{1}{n})^n \approx \frac{1}{e}$ (when n is large). Write down the exact probability that an observation appears once and the probability that it appears twice, and find similar approximations to them, using the same formula.
   Hint: Try to make use of the Binomial distribution.

3. Suppose we produce ten bootstrapped samples from a data set containing two classes, denoted Yes and No. We then apply a classification tree to each bootstrapped sample and, for a specific value of X, produce 10 estimates of $P(Class\ is\ Yes|X)$:
   0.1, 0.15, 0.2, 0.2, 0.55, 0.6, 0.6, 0.65, 0.7, and 0.75.
   There are two common ways to combine these results together into a single class prediction. One is the majority vote approach discussed in this chapter. The second approach is to classify based on the average probability. In this example, what is the final classification under each of these two approaches?

4. Fitting and comparing different classification models to the Caravan dataset (available from the ISLR package). The dependent variable is "purchase". Use the first 1000 observations as a training set and the remaining as test set.

   (a) Fit Random Forests to this dataset. Try at least ten different parameter settings and compare the results on the test set. You may vary the size of the trees, the number of variables sampled at each node, or the number of trees. Comment on the results.

(b) Fit Boosting to this dataset. Try at least ten different parameter settings and compare the results on the test set. You may vary the size of the trees, the step size ϵ, or the number of trees. Comment on the results and compare them to Random Forest results.
Hint: If you use gbm for this, make sure you apply it correctly to the classification setting. Notice specifically the distribution argument in the function.

(c) Fit logistic regression to this dataset. Evaluate the model on the test set and compare to the Random Forest and Boosting results.