STAT 5650

Statistical Learning and Data Mining I

Homework #4

Due: *Monday, March* 30.

1. In a typical bootstrap sample approximately 37% of the observations that are in the original dataset do not occur in the bootstrap sample. Here's a derivation of that result. Consider a dataset with n observations which we may label $1, 2, \cdots, k-1, k, k+1, \cdots, n-1, n$.

    a) Suppose I select n observations from the original dataset *with replacement*. What is the chance that observation k is not among the selected observations? (Another way to think about this question is the following. Suppose I have a fair n–sided die with sides labelled $1, 2, \cdots, k, \cdots, n$. I roll the die n times—and the rolls are all independent. What is the chance that the side labelled k does not occur in the n rolls?)

    b) Evaluate the expression you obtained in part a) for an increasing sequence of values of n.

    c) Do you recognize the limit as $n \to \infty$ of the expression in part a)? If so, identify and evaluate it. If not, compute the expression in part a) for some very large values of n.

    d) (Quite hard). What is the standard error of the observed *number* and *proportion* of observations in the original sample that are not is a bootstrap sample?

2. This problem continues the analysis of the *Forensic Glass* data.

    a) Apply random forests to the data and obtain the out-of-bag confusion matrix. How well can we classify these data, and where are the major misclassifications? How do your results compare to the classification tree you fitted in Homework #3.

    b) Use random forests to select a subset of the variables (which may be all the variables!) Refit random forests with only the important variables and obtain the out-of-bag confusion matrix. Did you observe any change in predictive accuracy?

    c) Summarize your results for your analyses of the forensic glass data using classification trees and random forests.

3. This problem continues your analyses of the Uintah Mountains cavity nesting birds' data.

    a) Apply random forests to all the data with Species as the response variable and obtain the out-of-bag confusion matrix. How well can we classify these data, and where are the major misclassifications? How do your results compare to the classification tree you fitted in Homework #3.
    b) Use random forests to select a subset of the variables (which may be all the variables!) Refit random forests with only the important variables and obtain the out-of-bag confusion matrix. Did you observe any change in predictive accuracy?
    c) Summarize your results for your analyses of the birds' nest data using classification trees and random forests.

4. This problem also continues your analyses of the Uintah Mountains cavity nesting birds' data.

    a) Apply random forests to all the data with nest as the response variable and obtain the out-of-bag confusion matrix. How well can we classify these data, and where are the major misclassifications? How do your results compare to the classification tree you fitted in Homework #3.
    b) Use random forests to select a subset of the variables (which may be all the variables!) Refit random forests with only the important variables and obtain the out-of-bag confusion matrix. Did you observe any change in predictive accuracy?
    c) Now apply *adaboost* to the data and add the classification accuracies to those you have previously obtained for classification trees and random forests.
    d) Fit untuned and tuned *gradient boosting machines* to the data and compare the results to those previously obtained.
    e) Fit untuned and tuned *support vector machines* to the data and compare the results to those previously obtained.
    f) Briefly discuss the results of all your analyses. Which method(s) did best on these data?