# Assignment 4

Freya D Mello

November 2018

## 1 Question 1

For the prostate data of Chapter 3, carry out a best subset linear regression analysis, as in Table 3.3 (third column from the left). Compute the AIC, BIC, five- and tenfold cross-validation, and bootstrap .632 estimates of prediction error

### 1.1 Best Subset Selection

Best subset selection showed a min of 3 predictors were sufficient to and suitable for feature selection with the minimum CP and BIC. The best three variable model includes the attributes - 'lcavol', 'lweight', 'svi' for the prediction of 'lspa'.

### 1.2 Bootstrap

The bootstrap model shows the minimum error when number of predictors are 3 with an error value of .632. 1

### 1.3 5 fold Cross Validation

The 5 fold cross validation was performed over the dataset to learn that a minimum of 7 predictors gave the best model.

### 1.4 10 fold Cross Validation

As compared to 5 fold cross validation, the 10 fold cross validation gives us a better model as the best set of predictors out of all the predictors is 3 which suits the same answer we got from AIC and BIC.

## 2 Question 2

A access the wine data from the UCI machine learning repository. These data are the results of a chemical analysis of 178 wines grown over the decade 1970-1979 in the same region of Italy, but derived from three different cultivars (Barolo, Grignolino, Barbera). The Babera wines were predominately from a period that was much later than that of the Barolo and Grignolino wines. The analysis determined the quantities MalicAcid, Ash, AlcAsh, Mg, Phenols, Proa, Color, Hue, OD, and Proline. There are 50 Barolo wines, 71 Grignolino wines, and 48 Barbera wines. Construct the appropriate-size classification tree for this dataset. How many training and testing samples fall into each node? Describe the resulting tree and your approach.

### 2.1 Preparing the Data

The wine dataset is learnt to identify classification feature - Wine type. There are 3 main types of wines - 1-Barolo 2-Grignolino 3-Barbera. Using the appropriate seed number, the data is split into test and train datasets. The columns names were given having the response variable and the predictors as "WineType", "MalicAcid","Ash", "AlcaAsh", "Mg", "Phenols", "Proa", "Colour", "Hue", "OD", "Proline".

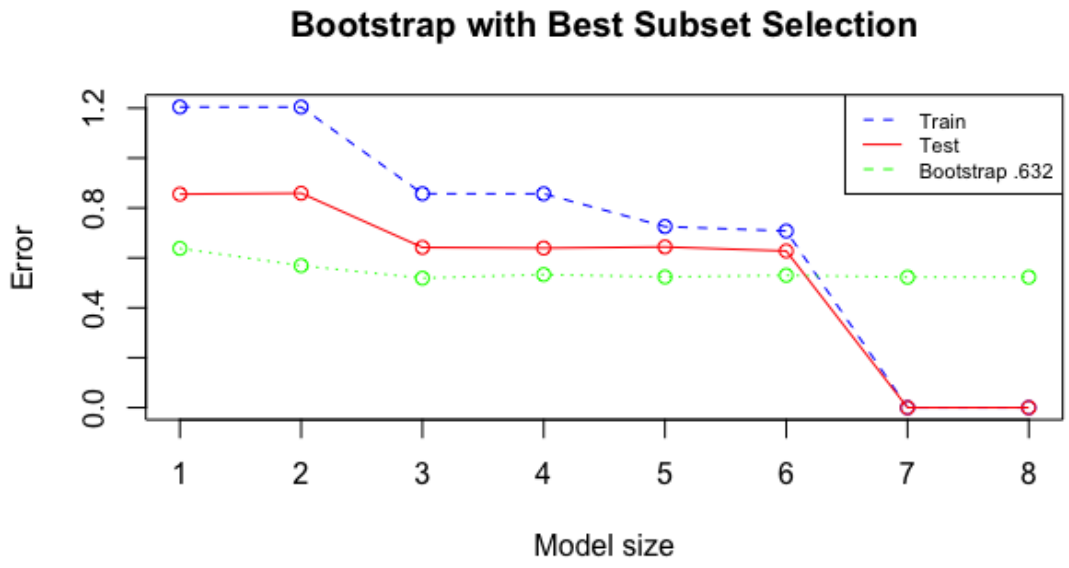


Figure 1: Bootstrap

| Nodes | Sample Percent |
|---|---|
| 5 | 4 |
| 6 | 32 |
| 8 | 33.6 |
| 9 | 2.4 |
| 14 | 3.2 |
| 15 | 24.8 |

Table 1: Percentage of train samples falling into each node

## 2.2 Inference

Using the train set, the random forest model was used with minimum split value of 5. The tree before pruning showed 2

After pruning, the misclassification train error was recorded as 0.048 and for test data is was 0.01886792. 2

The pruned fit has its frame values changed. The response values at every node were replaced with the value of the node. The frame of the fit was then fed with train data and then test data. The values at every node were noted and its percentage was computed so as to learn the ratio of how the number samples present at every node differ among same node numbers.

The tables 1 and 2 contain the percentage of train and test data samples present at every node and they are found to be more or less of the same ratio.

| Nodes | Sample Percent |
|---|---|
| 5 | 1.88 |
| 6 | 41.50 |
| 8 | 30.188 |
| 14 | 3.77 |
| 15 | 22.64 |

Table 2: Percentage of test samples falling into each node

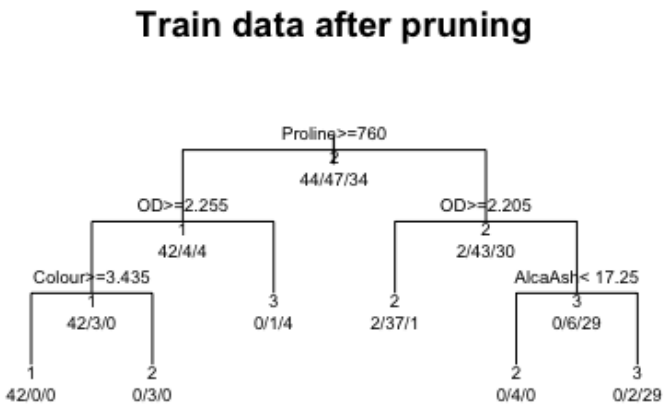


Figure 2: Train Data after Pruning

# 3 Question 3

Apply bagging, boosting, and random forests to a data set of your choice (not one used in the committee machines labs). Fit the models on a training set, and evaluate them on a test set. How accurate are these results compared to more simplistic (non-ensemble) methods (e.g., logistic regression, kNN, etc)? What are some advantages (and disadvantages) do committee machines have related to the data set that you selected?

## 3.1 Bagging

The data set considered is extracted from Wisconsin Breast Cancer Database. The response variable is classifying whether the patient is benign or malignant of cancer. With the value of 10 for randomly sampling the variables and and number of trees as 1000, the bagging is performed on the random forest model. The variable importance across all predictors can be shown in 3. The misclassification error rate is shown as 0.03846154.

## 3.2 Boosting

For two shrinkage values, 0.1 and 0.6 boosting models are fitted on the train data and tested using test data to obtain misclassification rates 0.02980456 and 0.03057541 respectively and number of trees 1000.

## 3.3 Random forest

The random forest is implemented for number of tree values 1000 and the variable importance is noted from 4 The misclassification rate is computed to be 0.02884615.

## 3.4 Logistic Regression

Logistic Regression is performed over the data set taking the train data and the model is tested on the test data. The misclassification rate is computed to be 0.0384615.

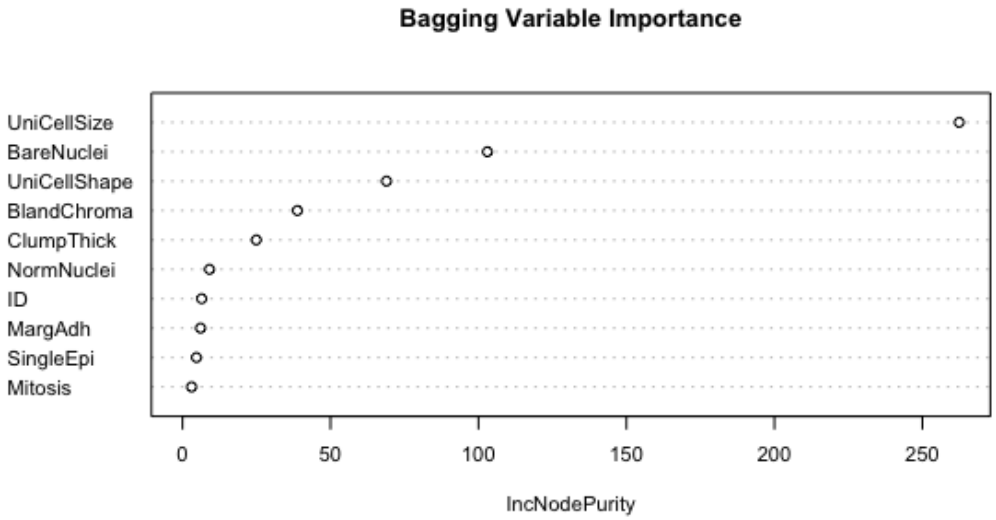


Figure 3: Variable Importance

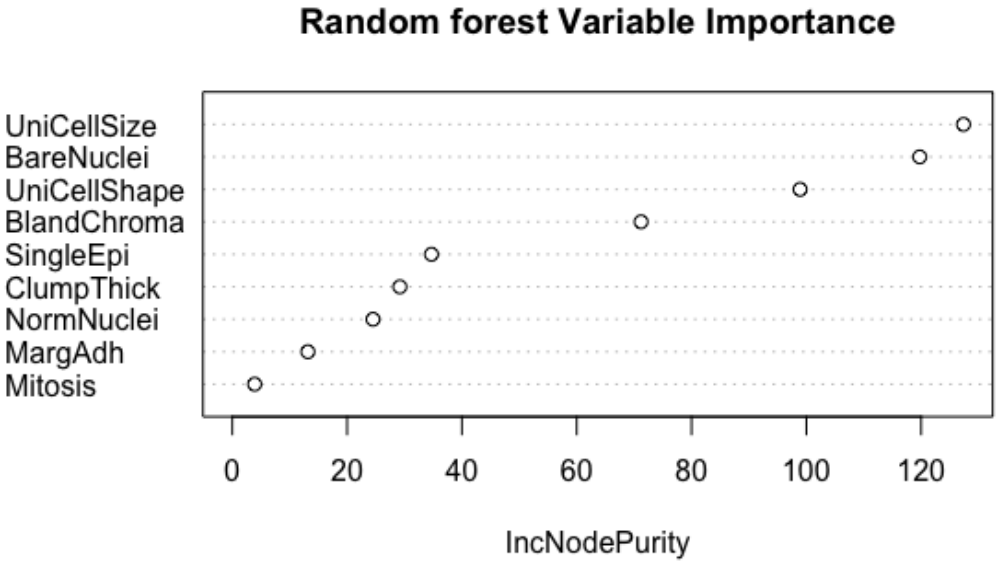


Figure 4: Variable Importance

| Method | Error |
|---|---|
| Bagging | 0.03846154 |
| Logistic Regression | 0.03846154 |
| Boosting | 0.03057541 |
| Random Forest | 0.02884615 |

Table 3: Misclassification Error

## 3.5 Comparison and Inference

The misclassification error is noted most in Bagging with a value of 0.03846154 followed by Logistic Regression with a value of 0.03846154 which was almost the same, followed by Boosting with a error rate of 0.03057541 and lastly random forest with a value of 0.02884615.

### 3.5.1 Advantages

The concept of splitting the data into modules on several iterations and considering the average of them is fostered in Committee machines. Here, the response variable does not depend on the model chosen, but also on several other models and therefore, giving reliable stable predictions.

### 3.5.2 Disadvantages

The data set considered had the response variable as 2 for benign and 4 for malignant of breast cancer. The method adaboost used was by taking response variables to be binary. i.e. only 0 or 1 so it had the over head of data transformation.

Boosting adds weight to observations that are incorrectly classified and if the dataset contained outliers, it would wrongly apply weights on them so that it gets classified under the right category, which infact is wrong.

The random forest model needs the number of distinct values in under each predictor to be less than or equal to 53.