26 BE 7082 + 26 PH 7028 + 20 BME 7082
Introduction to Data Science
Autumn 2020

MB Rao

Homework No. 5 Due date: October 01, 2020 Maximum points: 30

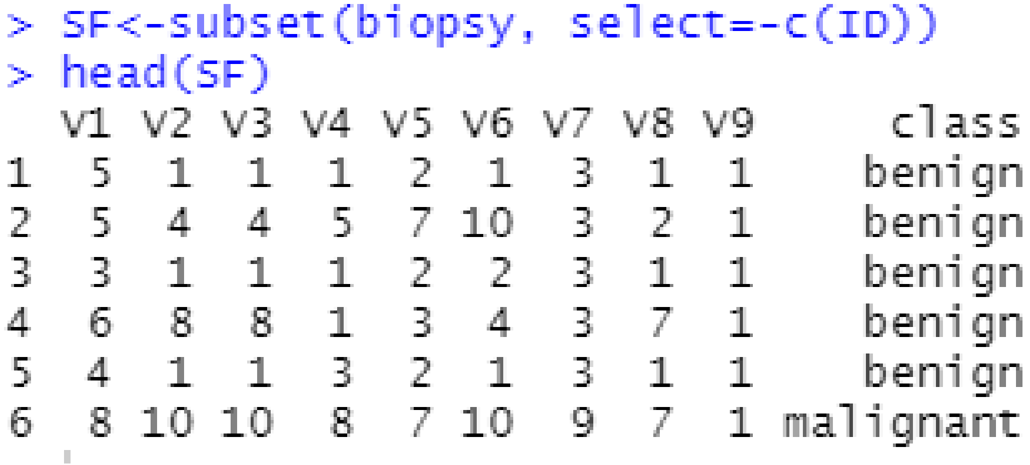
Theme: Random forests for binary response variables.

In Lecture 6, we have seen the evolution of random forests for feature selection, especially, when the data has a large number of predictors. You have worked in the past on Wisconsin Breast Cancer data (HW 2). I want you to work on the same data from the random forest angle.

Download the Wisconsin Breast Cancer data. Carry out the following steps.

1. What is the dimension of the data? 1 point
> dim(biopsy)
[1] 699 11
The data has **699 rows and 11 columns.**

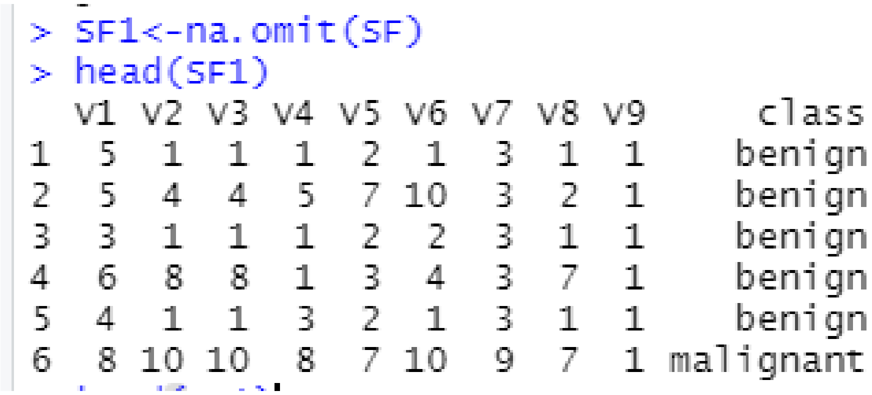
2. The first column is id. This is a variable. It is useless. Create a new folder eliminating the first column. 1 point



```
> SF<-subset(biopsy, select=-c(ID))
> head(SF)
  V1 V2 V3 V4 V5 V6 V7 V8 V9     class
1  5  1  1  1  2  1  3  1  1    benign
2  5  4  4  5  7 10  3  2  1    benign
3  3  1  1  1  2  2  3  1  1    benign
4  6  8  8  1  3  4  3  7  1    benign
5  4  1  1  3  2  1  3  1  1    benign
6  8 10 10  8  7 10  9  7  1 malignant
```

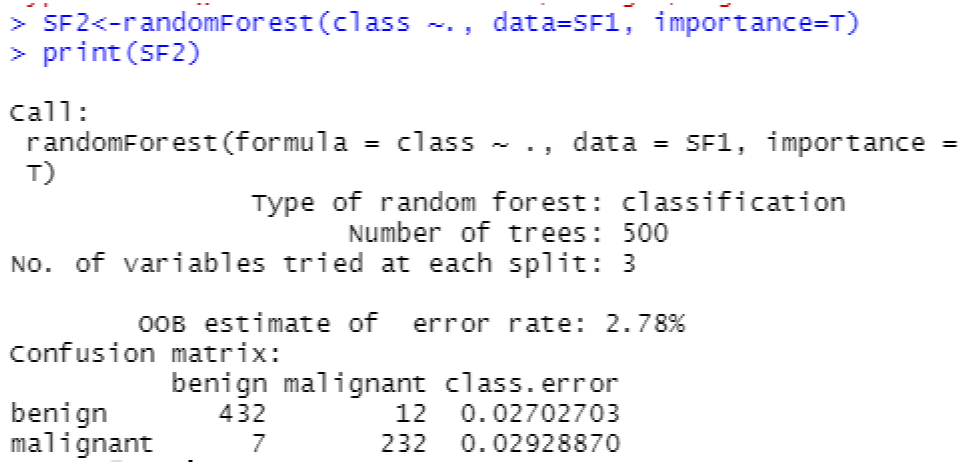
3. Eliminate the missing observations. (If any observation is missing in a row, the entire row is deemed missing.) (The R function complete.cases(biopsy) should help.) Or, find your own way to eliminate the missing observations. 4 points

I did it using na.omit and stored the new data in a new folder SF1 as shown below:



```
> SF1<-na.omit(SF)
> head(SF1)
  V1 V2 V3 V4 V5 V6 V7 V8 V9     class
1  5  1  1  1  2  1  3  1  1    benign
2  5  4  4  5  7 10  3  2  1    benign
3  3  1  1  1  2  2  3  1  1    benign
4  6  8  8  1  3  4  3  7  1    benign
5  4  1  1  3  2  1  3  1  1    benign
6  8 10 10  8  7 10  9  7  1 malignant
```

4. Build a random forest. Explain each entry in the output. 6 points



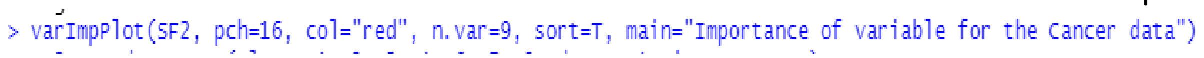
```
> SF2<-randomForest(class ~., data=SF1, importance=T)
> print(SF2)

Call:
 randomForest(formula = class ~ ., data = SF1, importance = 
 T) 
               Type of random forest: classification
                     Number of trees: 500
No. of variables tried at each split: 3

        OOB estimate of  error rate: 2.78%
Confusion matrix:
          benign malignant class.error
benign       432        12  0.02702703
malignant      7       232  0.02928870
```

The output contains the following:

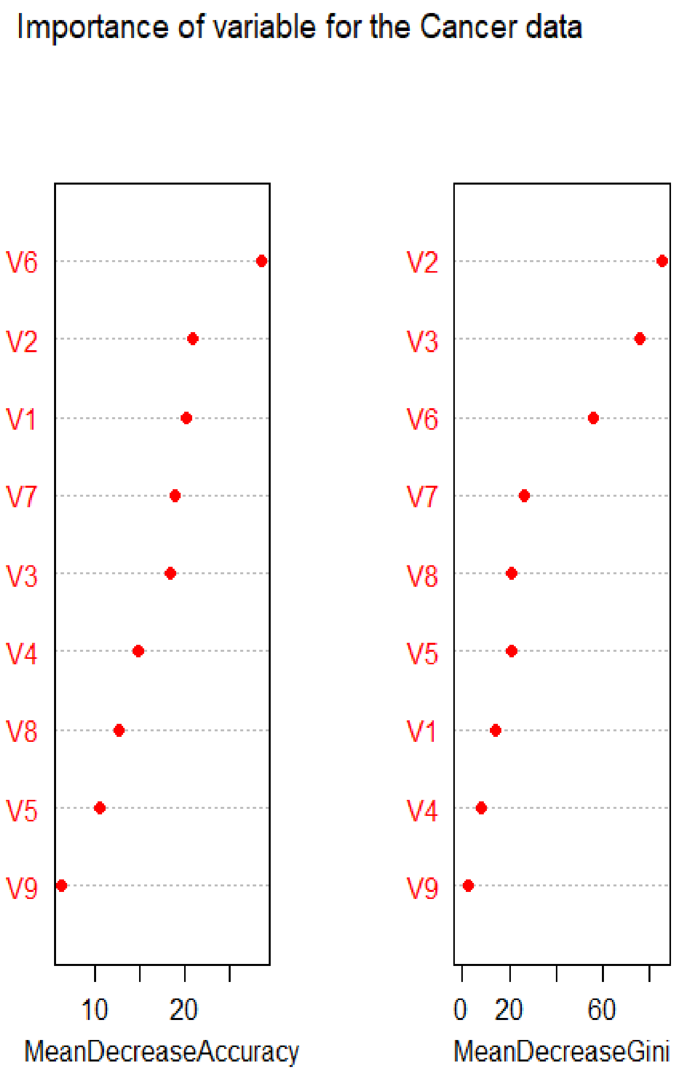
1. Type of random forest is classified as a **classification problem**.
2. Total number of forests produced are **500**.
3. Total number of predictors is 9, however, it uses **3 variables for splitting each node**.
4. The out-of-bag **(OOB)** error is estimated at **2.78%.**
5. The confusion matrix depicts classification error in predicting each of the two-cancer type separately; **2.70% error in classifying benign cancer** and **2.92% error in classifying malignant cancer**.

5. Get the variable importance graph. Explain the meaning of the first graph.

4 + 4 points



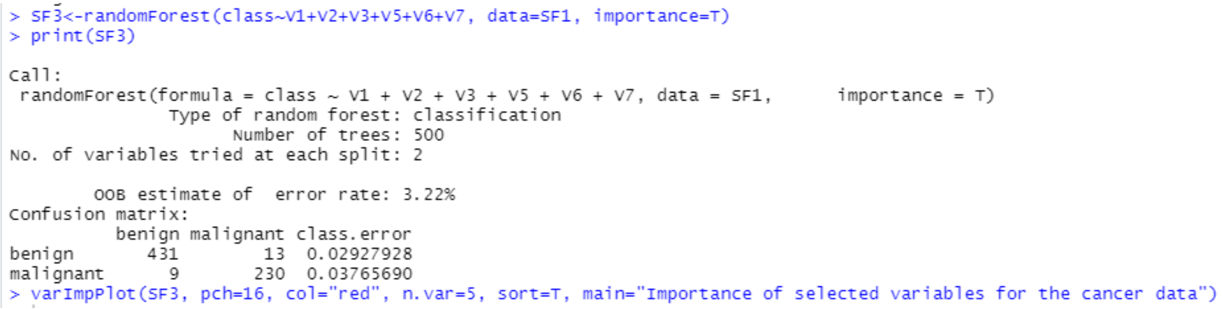
```
> varImpPlot(SF2, pch=16, col="red", n.var=9, sort=T, main="Importance of variable for the Cancer data")
```



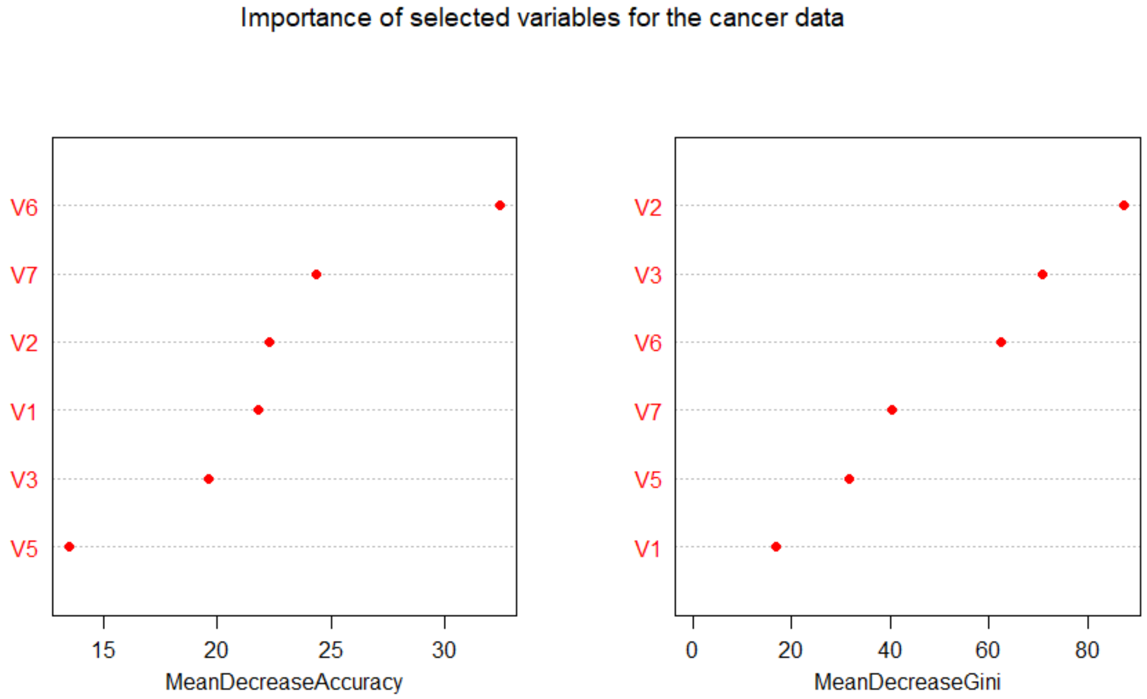
The first graph is showing mean decrease (on x-axis) in accuracy if we were to mess up the variable (on y-axis) data shown in the graph. In other words, the OOB error of the overall data set will increase by this mean decrease in accuracy for each messed up variable by the amount as indicated in the graph for the said variable. For eg. If we omit or mess up V6 and create a new random forest classifier, we will end up with an OOB error which will be 29% higher than original OOB error of 2.78%. This indicates that V6 is an important variable. Also, variable V9 will have least effect on overall OOB error if were to mess up V9.

Variable names	Importance of variables
V1	20.95
V2	20.34
V3	18.89
V4	13.77
V5	12.58
V6	27.70
V7	19.45
V8	13.82
V9	6.55

6. Redo the random forest by using what you think are the important predictors. Do the corresponding variable importance graph. Contrast the output in 6 with that of 4 and 5. 4 + 4 + 4 points



```
> SF3<-randomForest(class~V1+V2+V3+V5+V6+V7, data=SF1, importance=T)
> print(SF3)

Call:
 randomForest(formula = class ~ V1 + V2 + V3 + V5 + V6 + V7, data = SF1,      importance = T) 
               Type of random forest: classification
                     Number of trees: 500
No. of variables tried at each split: 2

        OOB estimate of  error rate: 3.22%
Confusion matrix:
          benign malignant class.error
benign       431        13  0.02927928
malignant      9       230  0.03765690
> varImpPlot(SF3, pch=16, col="red", n.var=5, sort=T, main="Importance of selected variables for the cancer data")
```



In my opinion, V6, V7, V2, V1, V3 and V5 are important predictors in determining whether the cancer is benign or malignant. I used the varImpPlot to identify these variables and then using created a new randomForest with just these 6 variables. The difference between output obtained in 6 with 4 and 5 are highlighted below:

Q4 and Q5	**Q6**
Number of variable in building random forest = 9	Number of variable in building random forest = 6
Number of variables for splitting each node=3	Number of variables for splitting each node=2
OOB error = 2.78%	OOB error = 3.22%

Misclassification rate for benign cancer=2.7%	Misclassification rate for benign cancer=2.92%
Misclassification rate for malignant cancer=2.92%	Misclassification rate for malignant cancer=3.7%

Another aspect of difference between random forest obtained in 4, 5 and 6 is highlighted below:

Using

```
> SF4<-varImpPlot(SF3, col="red", n.var=6, sort=T, main="Importance of selected variables for the cancer data")
> SF4
```

We can obtain exact values of Mean Decrease in Accuracy and Mean Decrease in Gini Index values for Q6 (with only 6 variables).

Q6	Mean Decrease Accuracy	Mean Decrease Gini
V1	21.81	16.78
V2	22.32	87.37
V3	19.62	70.91
V5	13.49	31.56
V6	32.42	62.37
V7	24.36	40.44

Also, using,

```
> SF5<-varImpPlot(SF2, pch=16, col="red", n.var=, sort=T, main="Importance of variable for the Cancer data")
> SF5
```

We can obtain exact values of Mean Decrease in Accuracy and Mean Decrease in Gini Index values for Q4 and Q5 (with all 9 variables).

Q4 & Q5	Mean Decrease Accuracy	Mean Decrease Gini
V1	20.95	16.16
V2	20.34	81.13
V3	18.89	70.09
V4	13.77	8.79
V5	12.58	26.06
V6	27.70	50.03
V7	19.45	33.50
V8	13.82	22.06
V9	6.55	2.43

7. Are we getting anything special from here over the simple classification tree?
2 points

A random forest tells us about the role each variable plays in a classification problem by generating multiple decision trees and generating variable importance. In random forest we can obtain mean decrease in accuracy (importance values) and Gini index for removal of each variable. Random forest chooses variables randomly to generate training set. **This randomization makes Random forest much more accurate compared to classification tree.**