20 BME 7082/26 BE 7082/26 PH 7028

Autumn 2020

MB Rao

Homework Sheet No. 2 Due Date: September 10, 2020 Maximum Points: 30

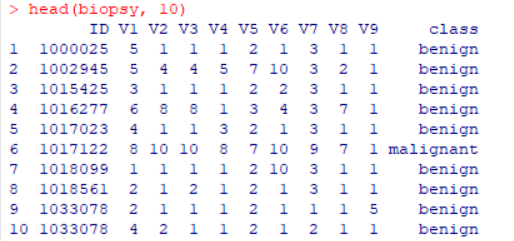
Wisconsin Breast Cancer Data: A gold standard procedure for detecting breast cancer is biopsy. This is a definitive procedure. Women, fifty years or older, are advised get checked once every year for the presence or absence of cancer. Biopsy is painful, time-consuming, and expensive. It cannot be done every year. An alternative procedure is mammogram. This diagnostic test is not accurate. Its sensitivity is about 85% (True positives) and specificity 80% (True negatives). Wisconsin Medical Research Center proposed another diagnostic procedure (breast aspiration), which they touted more accurate than the mammogram. A needle is inserted into the breast and cells are extracted. Various properties (nine in all) of the cells are noted for each case (malignant) and control (benign). The determination of whether it is malignant or benign comes from the gold standard procedure. Download the data (biopsy) from the package (MASS). The response variable is ‘class,’ which is binary. We have nine predictors in all. They are cryptically labeled V1 through V9.

One needs to prepare the data before building a classification tree.

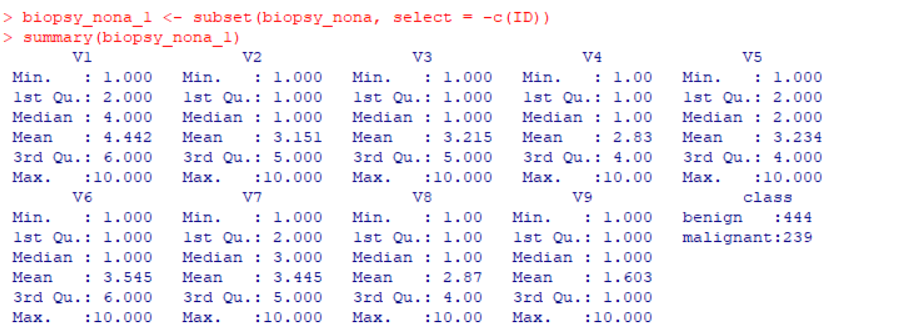
1. What is the dimension of the data? 1 point

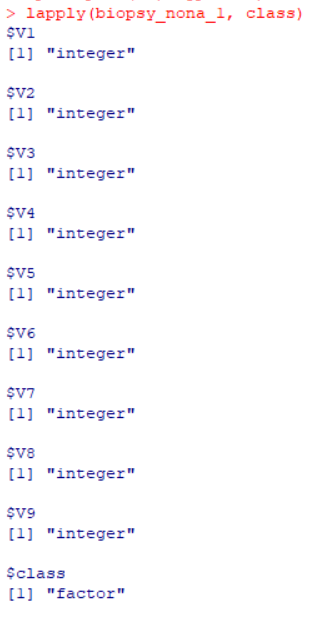


The data has 699 rows and 11 columns.

2. Show the top ten rows of the data. 1 point

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

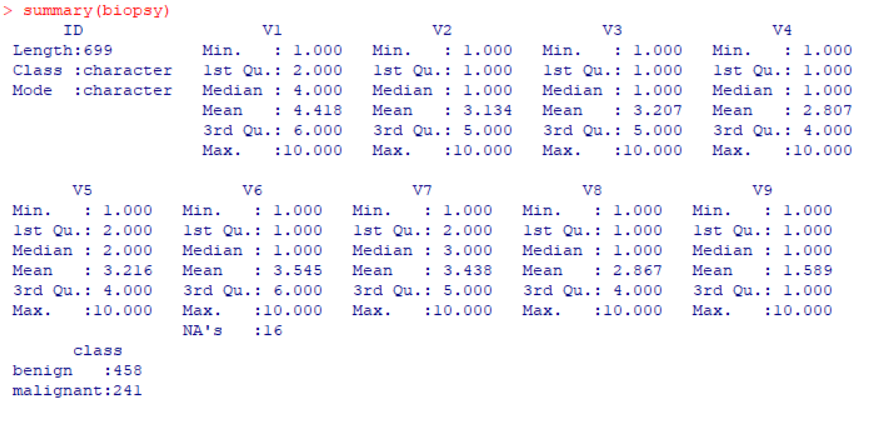


4. What is the class of each of the variables now? 1 point

5. Explain each variable. 4 points

Variables V1 through V9 are integer type meaning their numerical values will be used by R to classify (or predict) the result in the column entitled “Class”. Aforementioned command also enlists the last class column type as factor meaning the column will be used to categorize and store data in vectors or as in our case words “benign” or “malignant”.

6. Do summary statistics. Are there any missing observations? 1+1 points

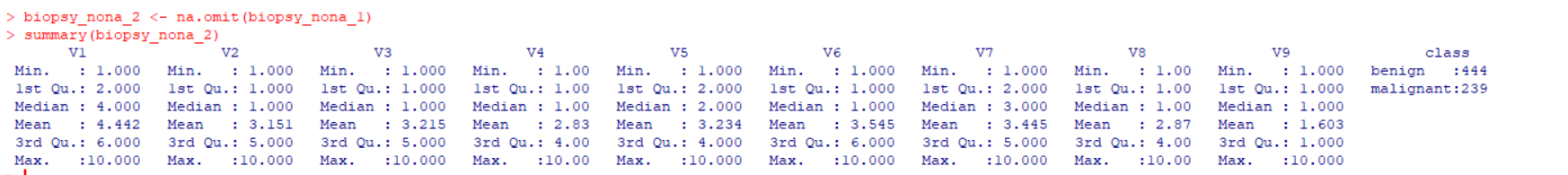


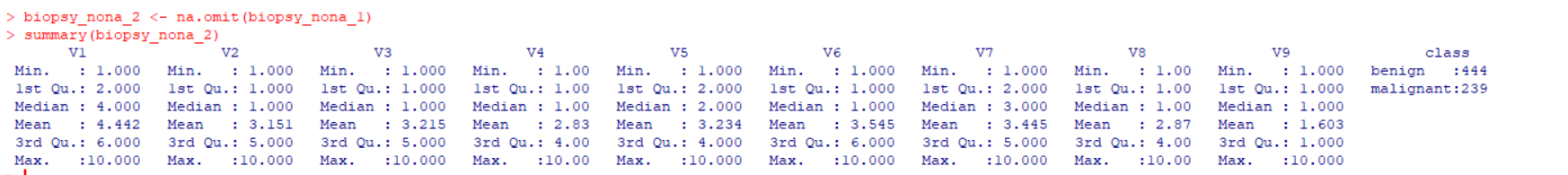
**Yes, there are 16 missing values spread in different rows.**

7. 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 this using na.omit.

> biopsy\_nona\_2 <- na.omit(biopsy\_nona\_1)





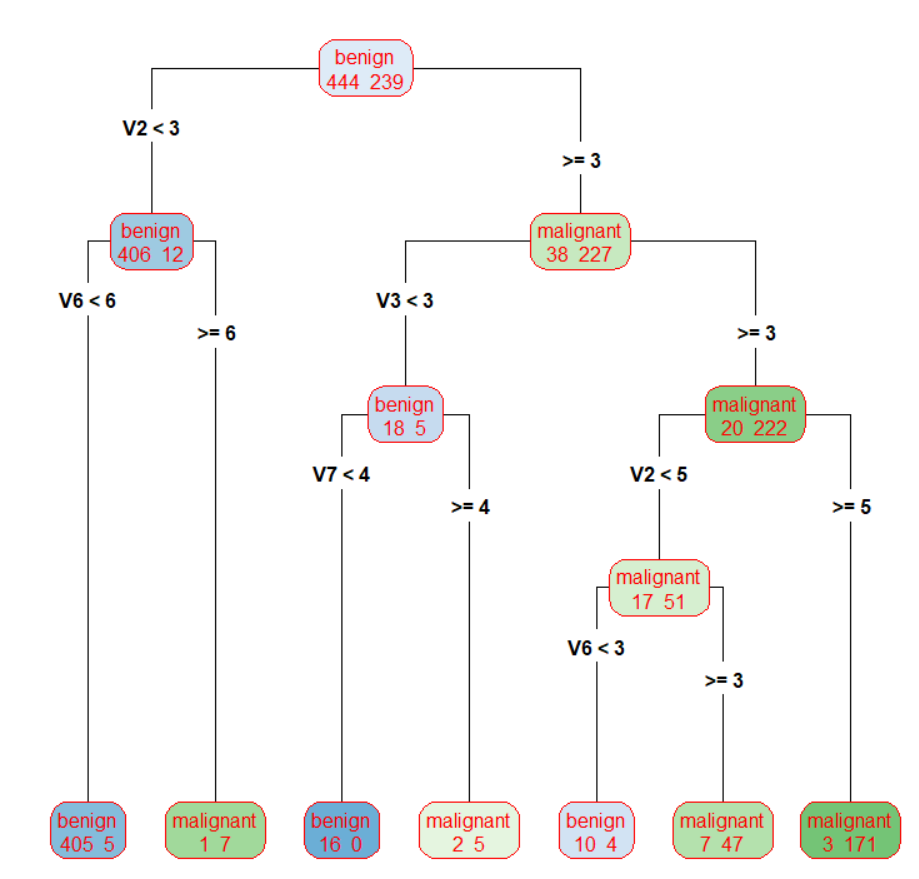
8. Build a classification tree. Show the tree. 3 points

**I built the classification tree using the following lines of code:**

*MB <- rpart(class~., data=biopsy\_nona\_1)*

*> rpart.plot(MB, type = 4, extra =1, digits=3, col="red")*

**Output**



9. Calculate the misclassification rate. 2 points

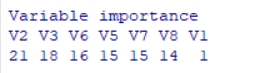
Miscalculation Rate = 5+1+0+2+4+7+3/683 = 0.0322 or 3.22%

10. Provide a verbal description of the classification protocol of the tree. 5 points

* If V2 >=5 the cancer is malignant.
* If V2 >=3 and V2<5 and V6 >=3 the cancer is malignant.
* If V2 >=3 and V2<5 and V6 <3 the cancer is benign.
* If V2 >=3 and V3 <3 and V7 >= the cancer is malignant
* If V2 >=3 and V3 <3 and V7 < 4 the cancer is benign.
* If V2<3 and V6>=6 the cancer is malignant.
* If V2<3 and V6<6 the cancer is benign.

11. Identify the predictors that made a mark in the tree. 2 points

The most important predictor turned out to be V2 while the least important predictor was V1. As evident from the classification tree and the >summary(MB) command. The order of variable importance is shown below:

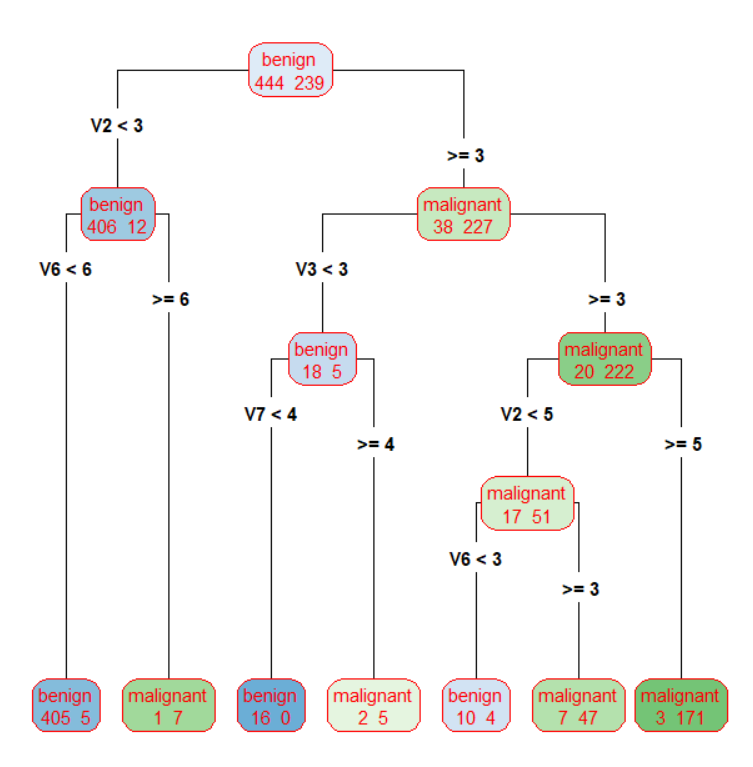


12. The default pruning principle stipulates that if the size of a node is 20 or less stop splitting the node. Suppose we change the size from 20 to 15. Explain how the tree changes and examine its impact on misclassification rate. 4 points

I used rpart.control to change the split size to 15 as shown below:

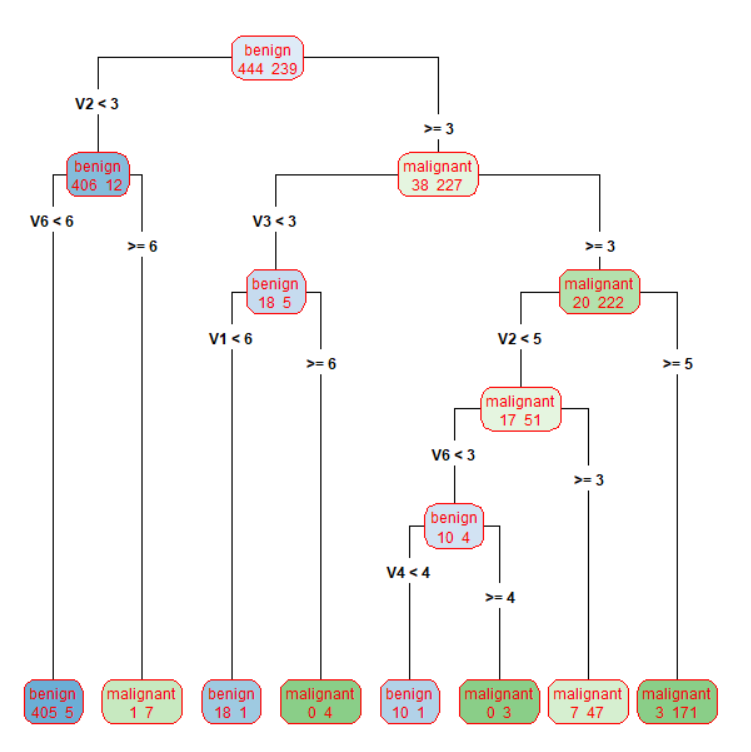


The classification tree obtained is shown below:



The misclassification rate remains the same as 22/683 = 3.22%

**However, changing the minsplit node size to 10 results in the following classification tree that yields a Misclassification rate of 18/683 = 0.0263 or 2.63%**



**Hence a conclusion can be drawn from the above results that reducing the minsplit size reduces the misclassification rate.** This leads to an improved prediction of the target with the same data and same variables.