20 BME 7082 + 26 BE 7082 + 26 PH 7028
Introduction to Data Science
Autumn 2020
MB Rao
Homework Sheet No. 3 Due Date: September 17, 2020 Maximum Points: 30

Theme: Classification Trees and Multi-level Responses Variables. Classification Trees in Industry

A preamble: A company specializes in glass that fall into six different types (1, 2, 3, 5, 6, 7). (Watch the levels.) A trained expert can distinguish them. The company wants to institute a mechanical way to identify the type of a piece of a glass based on physical and chemical properties of the glass. Data are obtained on the following nine properties: RF (Refractive Index); Na; Mg; Al; Si; K; Ca; Ba; Fe. Your task is to build a classification tree to meet the objective of the company. Download the data 'Glass' from the package 'mlbench.'

1. What is the dimension of the data? Show the top six rows of the data.

1 + 1 points

The data has 214 rows and 10 columns. *Following lines of code were executed to find the dimension and top six rows of the data.*

```
> dim(Glass)
[1] 214  10
```

```
+ > head(Glass)
       RI    Na   Mg   Al    Si    K   Ca Ba   Fe Type
1 1.52101 13.64 4.49 1.10 71.78 0.06 8.75  0 0.00    1
2 1.51761 13.89 3.60 1.36 72.73 0.48 7.83  0 0.00    1
3 1.51618 13.53 3.55 1.54 72.99 0.39 7.78  0 0.00    1
4 1.51766 13.21 3.69 1.29 72.61 0.57 8.22  0 0.00    1
5 1.51742 13.27 3.62 1.24 73.08 0.55 8.07  0 0.00    1
6 1.51596 12.79 3.61 1.62 72.97 0.64 8.07  0 0.26    1
```

2. Identify the nature of the last column of the data (numeric, integer, or factor).
   2 points

   The Last column is of class "factor." *Attribute query was executed using lapply statement.*

```
> lapply(Glass, class)
$RI
[1] "numeric"

$Na
[1] "numeric"

$Mg
[1] "numeric"

$Al
[1] "numeric"

$Si
[1] "numeric"

$K
[1] "numeric"

$Ca
[1] "numeric"

$Ba
[1] "numeric"

$Fe
[1] "numeric"

$Type
[1] "factor"
```

3. Obtain the summary statistics of the data including standard deviations.

3 + 3 points

Following lines of codes were executed to determine summary statistics and Standard Variation

```
> summary(Glass)
       RI              Na              Mg              Al       
 Min.   :1.511   Min.   :10.73   Min.   :0.000   Min.   :0.290  
 1st Qu.:1.517   1st Qu.:12.91   1st Qu.:2.115   1st Qu.:1.190  
 Median :1.518   Median :13.30   Median :3.480   Median :1.360  
 Mean   :1.518   Mean   :13.41   Mean   :2.685   Mean   :1.445  
 3rd Qu.:1.519   3rd Qu.:13.82   3rd Qu.:3.600   3rd Qu.:1.630  
 Max.   :1.534   Max.   :17.38   Max.   :4.490   Max.   :3.500  
       Si              K                Ca               Ba       
 Min.   :69.81   Min.   :0.0000   Min.   : 5.430   Min.   :0.000  
 1st Qu.:72.28   1st Qu.:0.1225   1st Qu.: 8.240   1st Qu.:0.000  
 Median :72.79   Median :0.5550   Median : 8.600   Median :0.000  
 Mean   :72.65   Mean   :0.4971   Mean   : 8.957   Mean   :0.175  
 3rd Qu.:73.09   3rd Qu.:0.6100   3rd Qu.: 9.172   3rd Qu.:0.000  
 Max.   :75.41   Max.   :6.2100   Max.   :16.190   Max.   :3.150  
       Fe          Type  
 Min.   :0.00000   1:70  
 1st Qu.:0.00000   2:76  
 Median :0.00000   3:17  
 Mean   :0.05701   5:13  
 3rd Qu.:0.10000   6: 9  
 Max.   :0.51000   7:29  
```

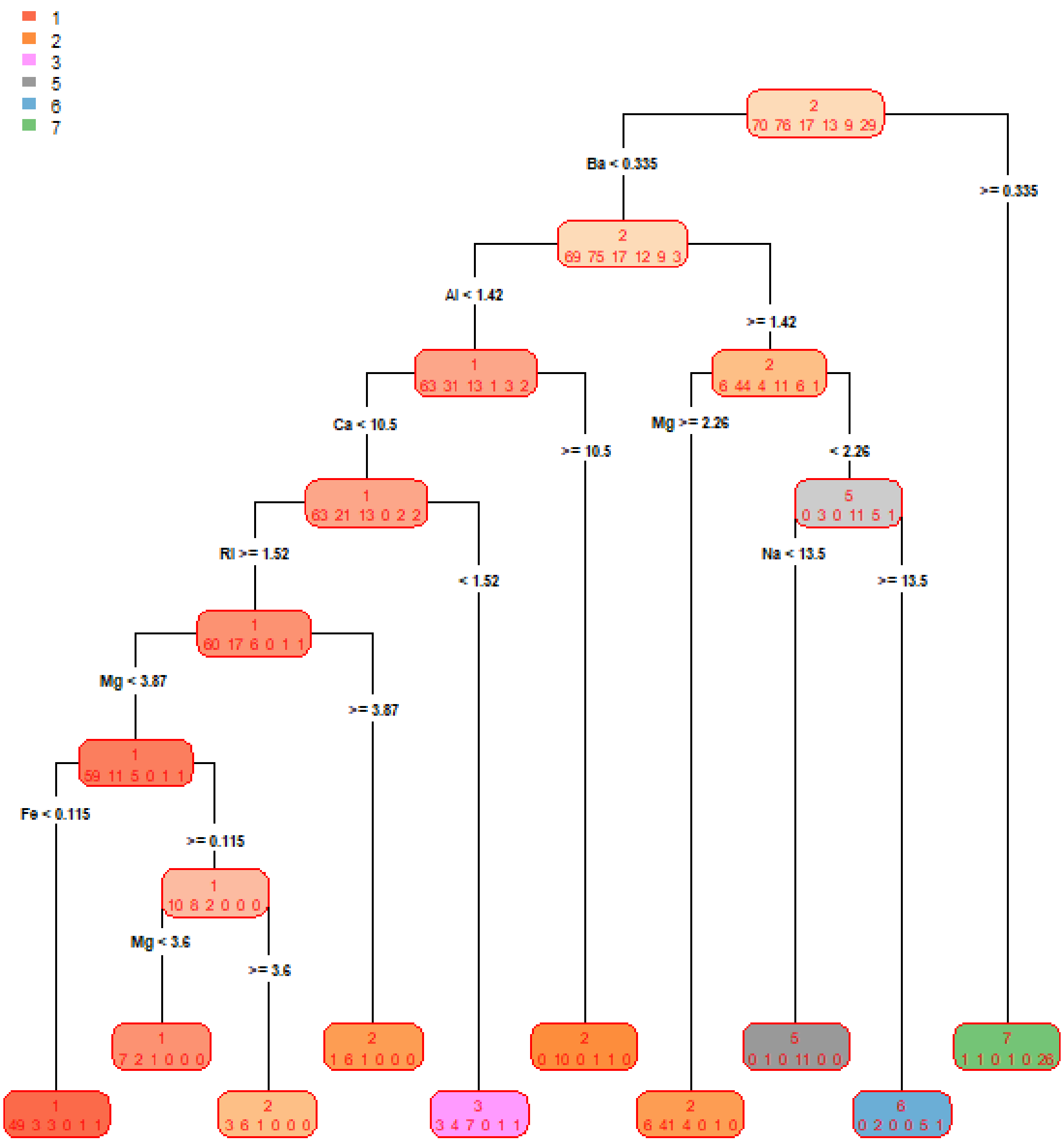
```
> sd(Glass$RI)
[1] 0.003036864
> sd(Glass$Na)
[1] 0.8166036
> sd(Glass$Mg)
[1] 1.442408
> sd(Glass$Al)
[1] 0.4992696
> sd(Glass$Si)
[1] 0.7745458
> sd(Glass$K)
[1] 0.6521918
> sd(Glass$Ca)
[1] 1.423153
> sd(Glass$Ba)
[1] 0.4972193
> sd(Glass$Fe)
[1] 0.0974387
```

4. Build a classification tree. 7 points

Following lines of code were executed to build the classification tree.

MB<-rpart(Type~., data=Glass)

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



5. Count the number of terminal nodes. Identify the missing chemical properties in the tree. 2 + 1 points

*There are **10 terminal nodes. "Si" and "K"** are missing in the classification tree.*

6. Provide a physical description when a glass falls into the first terminal node (first on the left side of the tree). 3 points

*The glass should **have Ba<0.335, AL<1.42, Ca<10.5, RI>=1.52, Mg<3.87 and Fe<0.115** to fall into the first terminal node on the left.*

The terminal node also suggests there were 49 classifications made correctly and 3 were mis-classified.

7. Calculate the accuracy rate of the tree. Is the tree worth? 2 points

***The accuracy rate of the tree is 78.5%.** Yes, the tree is worth it because any further pruning may not result in appreciable change in classification accuracy and over-fitting issues may arise. The accuracy rate was calculated using the following lines of code:*

```
> SF<-length(MB$y)
> print(AccRate<-100*(49+7+6+6+7+10+41+11+5+26)/SF)
[1] 78.50467
```

8. Predict the type of two pieces of glass with the following properties. (2 pts)

RF = 1.517; Na = 12.91; Mg = 2.115; Al = 1.190; Si = 72.28; K = 0.1225; Ca = 8.240; Ba = 0.000; Fe = 0.000

RF = 1.519; Na = 13.82; Mg = 3.600; Al = 1.630; Si = 73.09; K = 0.6100; Ca = 9.172; Ba = 0.000; Fe = 0.100 (Use R)

First, the given chemical and physical data was copied from excel and fed in R using read.table and then a predict statement was executed as shown below:

```
> data_new<-read.table(file="clipboard", sep="\t", header=TRUE)
> head(data_new)
     RI    Na    Mg   Al    Si      K    Ca Ba  Fe Type
1 1.517 12.91 2.115 1.19 72.28 0.1225 8.240  0 0.0   NA
2 1.519 13.82 3.600 1.63 73.09 0.6100 9.172  0 0.1   NA
> data_1<-subset(data_new, select=-c(Type))
```

```
> predict(MB, data_1, type="class")
1 2
3 2
```

The First piece of Glass is of Type 3 and Second piece of Glass is predicted as Type 2.

9. In splitting nodes in a classification tree, recall that entropy or Gini Index of distributions play a prominent role. I want you to understand these measures. I have 6 binary probability distributions on X:

X:	A	B	X:	A	B	X:	A	B
Pr:	0.3	0.7	Pr:	0.4	0.6	Pr:	0.5	0.5

X:	A	B	X:	A	B	X:	A	B
Pr:	0.0	1.0	Pr:	0.9	0.1	Pr:	0.2	0.8

(a) Arrange the distributions from the least chaotic to the most chaotic using your personal judgement. 1 point

	PR:		
X:	A	B	
	0.0	1.0	Least Chaotic
	0.9	0.1	
	0.2	0.8	
	0.3	0.7	
	0.4	0.6	
	0.5	0.5	Most Chaotic

(b) Calculate the entropy of each distribution and arrange the distributions according to increasing level of entropy. 1 point

Entropy was calculated using the following formula: $\sum_{i=1}^{m} -p_i \, ln \, p_i$

```
> 0.9 * log2(0.9) + 0.1 * log2(0.1)
[1] -0.4689956
> 0.2 * log2(0.2) + 0.8 * log2(0.8)
[1] -0.7219281
> 0.3 * log2(0.3) + 0.7 * log2(0.7)
[1] -0.8812909
> 0.4 * log2(0.4) + 0.6 * log2(0.6)
[1] -0.9709506
> 0.5 * log2(0.5) + 0.5 * log2(0.5)
[1] -1
```

	PR:			
X:	**A**	**B**	**Entropy**	
	0.0	1.0	0	Least Entropy
	0.9	0.1	0.4689956	
	0.2	0.8	0.7219281	
	0.3	0.7	0.8812909	
	0.4	0.6	0.9709506	
	0.5	0.5	1	Most Entropy

(c) Calculate the Gini's index of uncertainty for each distribution and arrange the distributions according to increasing level of uncertainty. 1 pt

Gini's Measure of Uncertainty was calculated using $\sum_{i \neq j} p_i p_j$ *and arranged in order of increasing uncertainty.*

	PR:			
X:	**A**	**B**	**Gini's Index**	
	0.0	1.0	=2*0*1 = **0**	Least Uncertainty
	0.9	0.1	=2*0.9*0.1=**0.18**	
	0.2	0.8	=2*0.2*0.8=**0.32**	
	0.3	0.7	=2*0.3*0.7=**0.42**	
	0.4	0.6	=2*0.4*0.6=**0.48**	
	0.5	0.5	=2*0.5*0.5=**0.50**	Most Uncertainty