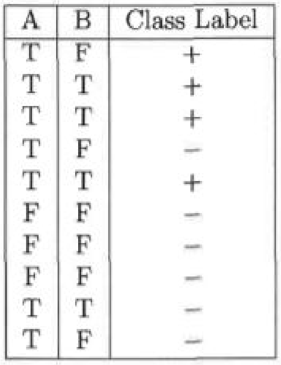
Assignment 4

Question – 2:

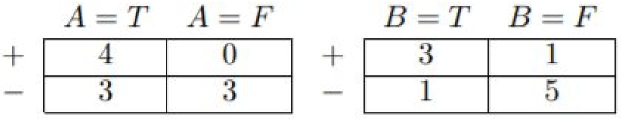
Repeat In-Class Exercise #38 using the misclassification error rate instead of information gain to determine the best split. Which of these splits considered is the best according to the misclassification error rate?



A	B	Class Label
T	F	+
T	T	+
T	T	+
T	F	−
T	T	+
F	F	−
F	F	−
F	F	−
T	T	−
T	F	−

Answer:

Contingency tables after splitting A and B are:



	$A = T$	$A = F$
+	4	0
−	3	3

	$B = T$	$B = F$
+	3	1
−	1	5

Misclassification error rate after splitting on A:

Misclassification error rate = (false positive + false negative)/total population

= (0+3)/ (4+3+0+3)

= 0.3

Misclassification error rate after splitting on B:

Misclassification error rate = (false positive + false negative)/total population

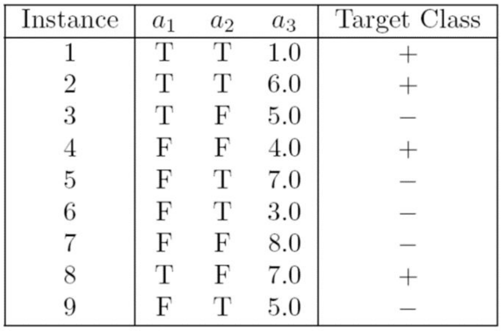
= (1+1)/ (3+1+1+5)

= 0.2

Therefore, from the above two misclassification errors, we can conclude that the misclassification error rate after splitting on B is best, as it has less error rate than A.

Question — 3:
Repeat In-Class Exercise #39 using the misclassification error rate instead of information gain to determine the best split. Which of these splits considered is the best according to the misclassification error rate?

Table 4.2. Data set for Exercise 3.



Instance	a_1	a_2	a_3	Target Class
1	T	T	1.0	+
2	T	T	6.0	+
3	T	F	5.0	−
4	F	F	4.0	+
5	F	T	7.0	−
6	F	T	3.0	−
7	F	F	8.0	−
8	T	F	7.0	+
9	F	T	5.0	−

Solution:

Contingency tables after splitting A1 and A2 are:

	A1 = T	A1 = F
+	3	1
-	1	4

	A2 = T	A2 = F
+	2	2
-	3	2

Misclassification error rate after splitting on A1:

Misclassification error rate = (false positive + false negative)/total population

= (1+1)/ (3+1+1+4)

= 0.22

Misclassification error rate after splitting on A2:

Misclassification error rate = (false positive + false negative)/total population

= (3+2)/ (3+2+2+2)

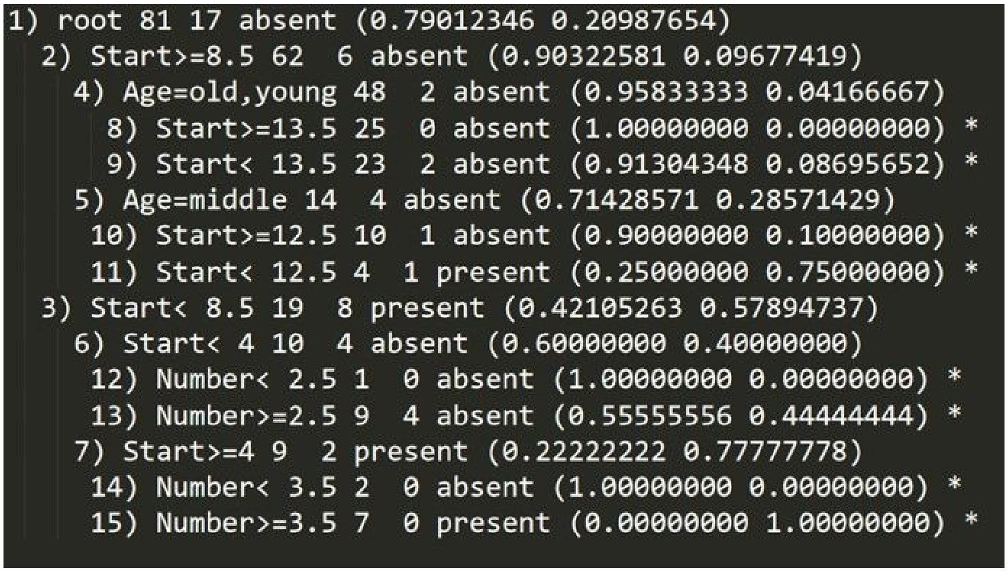
= 0.55

Therefore, from the above two misclassification errors, we can conclude that the misclassification error rate after splitting on A1 is best, as it has less error rate than A2.

Question – 4:

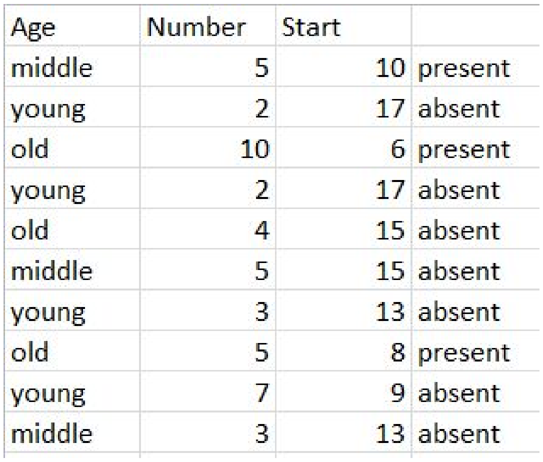
The file http://www-stat.wharton.upenn.edu/~dmease/rpart_text_example.txt gives an example of text output for a tree fit using the rpart() function in R from the library rpart. Use this tree to predict the class labels for the 10 observations in the test data http://www-stat.wharton.upenn.edu/~dmease/test_data.csv linked here. Do this manually - do not use R or any software.

rpart_text_example.txt



```
1) root 81 17 absent (0.79012346 0.20987654)
  2) Start>=8.5 62  6 absent (0.90322581 0.09677419)
    4) Age=old,young 48  2 absent (0.95833333 0.04166667)
      8) Start>=13.5 25  0 absent (1.00000000 0.00000000) *
      9) Start< 13.5 23  2 absent (0.91304348 0.08695652) *
    5) Age=middle 14  4 absent (0.71428571 0.28571429)
     10) Start>=12.5 10  1 absent (0.90000000 0.10000000) *
     11) Start< 12.5 4  1 present (0.25000000 0.75000000) *
  3) Start< 8.5 19  8 present (0.42105263 0.57894737)
    6) Start< 4 10  4 absent (0.60000000 0.40000000)
     12) Number< 2.5 1  0 absent (1.00000000 0.00000000) *
     13) Number>=2.5 9  4 absent (0.55555556 0.44444444) *
    7) Start>=4 9  2 present (0.22222222 0.77777778)
     14) Number< 3.5 2  0 absent (1.00000000 0.00000000) *
     15) Number>=3.5 7  0 present (0.00000000 1.00000000) *
```

Test_data.csv



Age	Number	Start	
middle	5	10	present
young	2	17	absent
old	10	6	present
young	2	17	absent
old	4	15	absent
middle	5	15	absent
young	3	13	absent
old	5	8	present
young	7	9	absent
middle	3	13	absent

Answer:

Observation-1: Age = middle, Number = 5, Start = 10

Path : 1 -> 2 -> 5 -> 11 -> **present**

Observation-2: Age = young, Number = 2, Start = 17

Path : 1 -> 2 -> 4 -> 8 -> **absent**

Observation-3: Age = old, Number = 10, Start = 6

Path : 1 -> 3 -> 7 -> 15 -> **present**

Observation-4: Age = young, Number = 2, Start = 17

Path : 1 -> 2 -> 4 -> 8 -> **absent**

Observation-5: Age = old, Number = 4, Start = 15

Path : 1 -> 2 -> 4 -> 8 -> **absent**

Observation-6: Age = middle, Number = 5, Start = 15

Path : 1 -> 2 -> 5 -> 10 -> **absent**

Observation-7: Age = young, Number = 3, Start = 13

Path : 1 -> 2 -> 4 -> 9 -> **absent**

Observation-8: Age = old, Number = 5, Start = 8

Path : 1 -> 3 -> 7 -> 15 -> **present**

Observation-9: Age = young, Number = 7, Start = 9

Path : 1 -> 2 -> 4 -> 9 -> **absent**

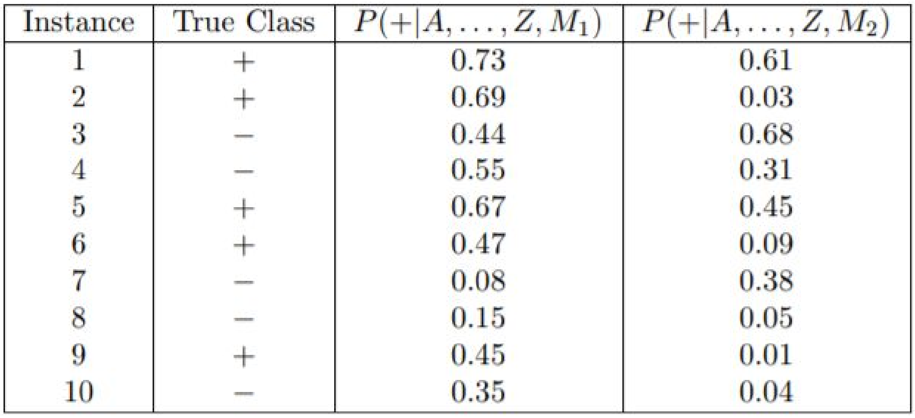
Observation-10: Age = middle, Number = 3, Start = 13

Path : 1 -> 2 -> 5 -> 10 -> **absent**

Question – 6
You are asked to evaluate the performance of two classification models, M1 and M2. The test set you have chosen contains 26 binary attributes, labeled as A through Z.

Table 5.5 shows the posterior probabilities obtained by applying the models to the test set. (Only the posterior probabilities for the positive class are shown). As this is a two-class problem, P(−)=1 – P(+) and P(−|A, . . . , Z)=1 – P(+|A, . . . , Z). Assume that we are mostly interested in detecting instances from the positive class.

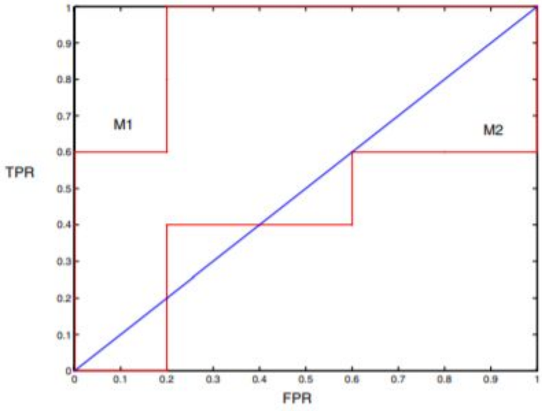
Table 5.5. Posterior probabilities for Exercise 17.



Instance	True Class	$P(+\|A,\ldots,Z,M_1)$	$P(+\|A,\ldots,Z,M_2)$
1	+	0.73	0.61
2	+	0.69	0.03
3	−	0.44	0.68
4	−	0.55	0.31
5	+	0.67	0.45
6	+	0.47	0.09
7	−	0.08	0.38
8	−	0.15	0.05
9	+	0.45	0.01
10	−	0.35	0.04

(a) Plot the ROC curve for both M1 and M2. (You should plot them on the same graph.) Which model do you think is better? Explain your reasons.

Answer:



The above image is the ROC curve for M1 and M2. From the figure, it is clear that the area under the curve for M1 is more than the area under the curve of M2.

(b) For model M1, suppose you choose the cut off threshold to be t = 0.5. In other words, any test instances whose posterior probability is greater than C will be classified as a positive example. Compute the precision, recall, and F-measure for the model at this threshold value.

When t = 0.5, confusion matrix for M1 is:

	+	-
+	3	2
-	1	4

Precision = ¾ = 75%

Recall = 3/5 = 60%.

F-measure = (2 × .75 × .6)/(.75 + .6) = 0.667.

(c) Repeat the analysis for part (b) using the same cut off threshold on model M2. Compare the F-measure results for both models. Which model is better? Are the results consistent with what you expect from the ROC curve?

When t = 0.5, confusion matrix for M2 is:

	+	-
+	1	4
-	1	4

Precision = 1/2 = 50%.

Recall = 1/5 = 20%.

F-measure = (2 × .5 × .2)/(.5 + .2) = 0.2857.

Based on F-measure, M1 is better than M2. And this result is consistent with the ROC plot.