Data Mining Assignment 4

1) Consider the following data set for a binary class problem.

Calculate the misclassification error rate when splitting on A and B to determine the best split. Which of these splits considered is the best according to misclassification error rate?



**Solution:**

If splitting is done on A, the misclassification error rate will be 3/10 as in the rows 4,9,10 we can see the three records are misclassified and the total number of records is 10 with respect to A. If splitting is done on B, the misclassification error rate will be 2/10 as the rows 1 and 9 are misclassified with respect to B.

2) Consider the training examples shown below for a binary classification problem. For a3, which is a continuous attribute compute misclassification error rate for every possible split to determine the best split. Which of these splits considered is the best according to misclassification error rate?



**Solution:**

Misclassification error rate while splitting on a1 is 2/9 and while splitting on a2 is 5/9. But while splitting on a3, as a3 is not a categorical value, splitting will not be straight. Here a3 has discrete values and we can split based on one condition i.e., a3 < 5.0 as + and a3 >= 5.0 as -. Then the misclassification error rate will be 3/9.  
  
3) 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.

Solution:

**1. Age = Middle, Number = 5 and Start = 10, the class label is present, as we traverse from 1 -> 2 -> 5 -> 11**

**2. Age = young, Number = 2, Start = 17, the class label is absent, as we traverse from 1 -> 2 -> 4 -> 8**

**3. Age = old, Number = 10, Start = 6, the class label is present, as we traverse from 1 -> 3 -> 7 -> 15**

**4. Age = young, Number = 2, Start = 17, the class label is absent, as we traverse from 1 -> 2 -> 4 -> 8**

**5. Age = old, Number = 4, Start = 15, the class label is absent, as we traverse from 1 -> 2 -> 4 -> 8**

**6. Age = middle, Number = 5, Start = 15, the class label is absent, as we traverse from 1 -> 2 -> 5 -> 10**

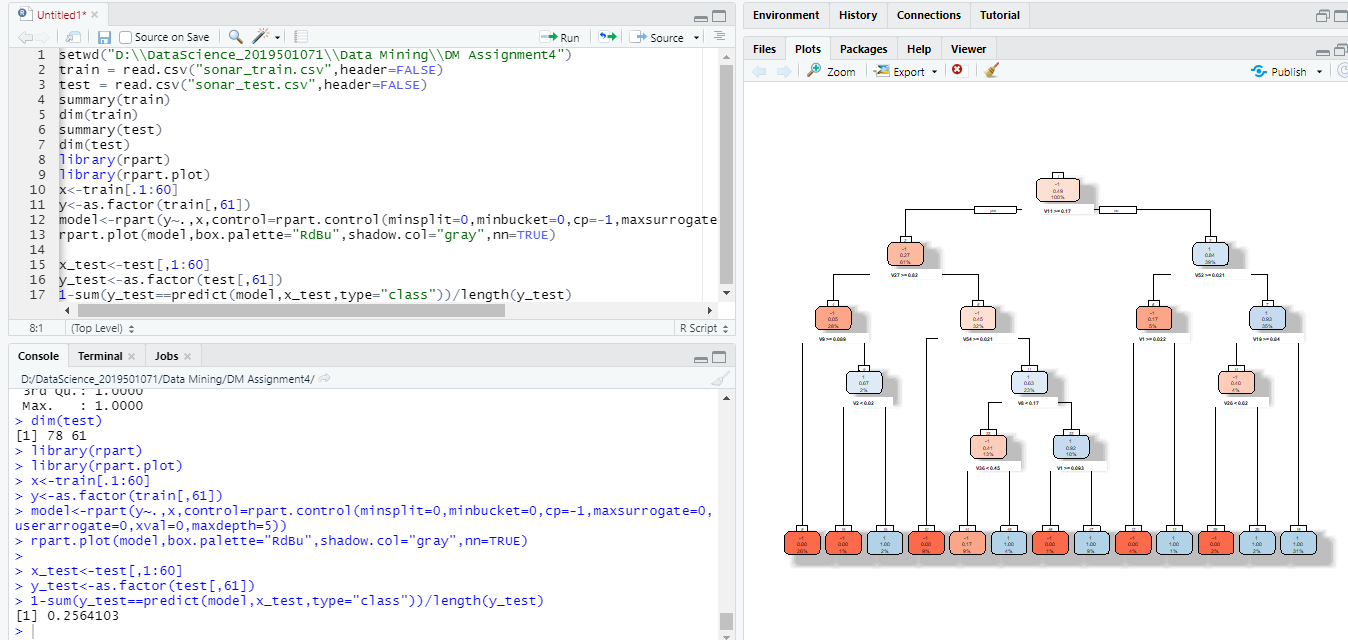
**7. Age = young, Number = 3, Start = 13, the class label is absent, as we traverse from 1 -> 2 -> 4 -> 9**

**8. Age = old, Number = 5, Start = 8, the class label is present, as we traverse from 1 -> 3 -> 7 -> 15**

**9. Age = young, Number = 7, Start = 9, the class label is absent, as we traverse from 1 - > 2 -> 4 -> 9**

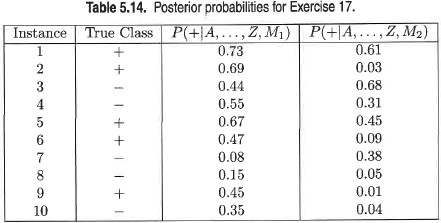
**10. Age = middle, Number = 3, Start = 13, the class label is absent, as we traverse from 1 -> 2 -> 5 -> 10**

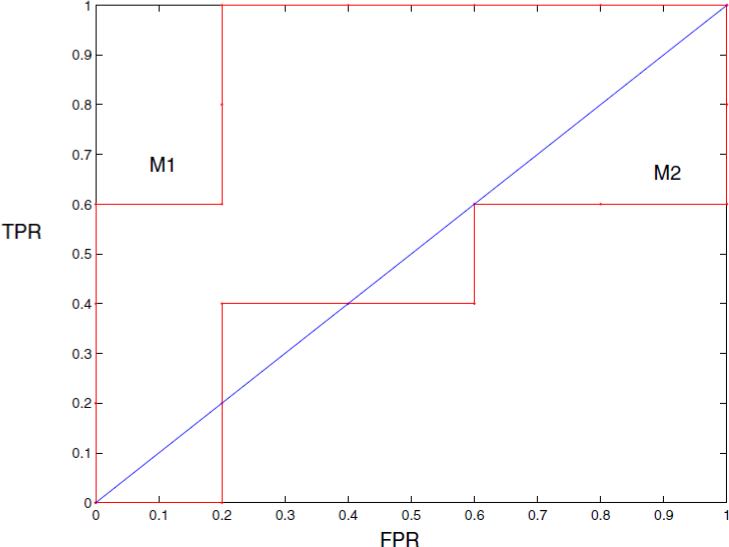
4) I split the popular sonar data set into a training set (<http://www-stat.wharton.upenn.edu/~dmease/sonar_train.csv>) and a test set (<http://www-stat.wharton.upenn.edu/~dmease/sonar_test.csv>). Use R to compute the misclassification error rate on the test set when training on the training set for a tree of depth 5 using all the default values except control=rpart.control(minsplit=0,minbucket=0,cp=-1, maxcompete=0, maxsurrogate=0, usesurrogate=0, xval=0,maxdepth=5). Remember that the 61st column is the response and the other 60 columns are the predictors.



5) Do Chapter 5 textbook problem #17 (parts a and c only) on pages 322-323. Note that there is a typo in part c - it should read "Repeat the analysis for part (b)". We will do part b in class.

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 5.14 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.

1. 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.

Area under ROC curve of M1 is larger than are under ROC curve of M2, so M1 is better.

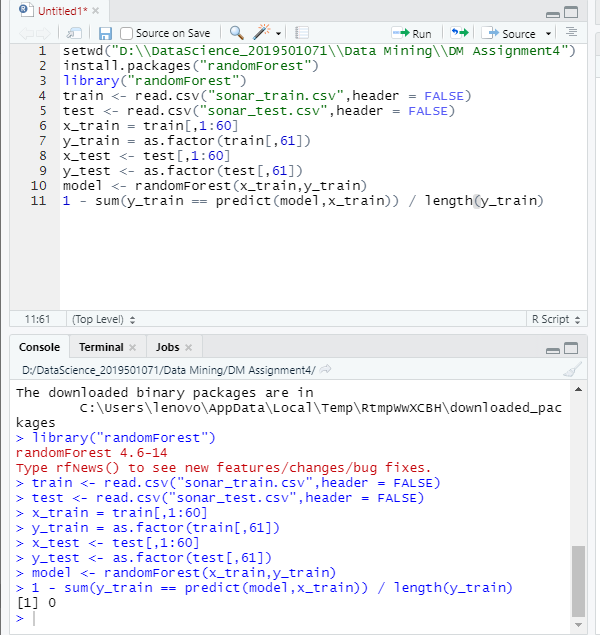
c) Repeat the analysis for part (c) using the same cutoff threshold on model M2. Compare the F-measure results for both models. Which model is better? Are the results consistent with what you expect from the ROC curve?

# For model M2 when t = 0.5: Precision = 1/2 = 50%

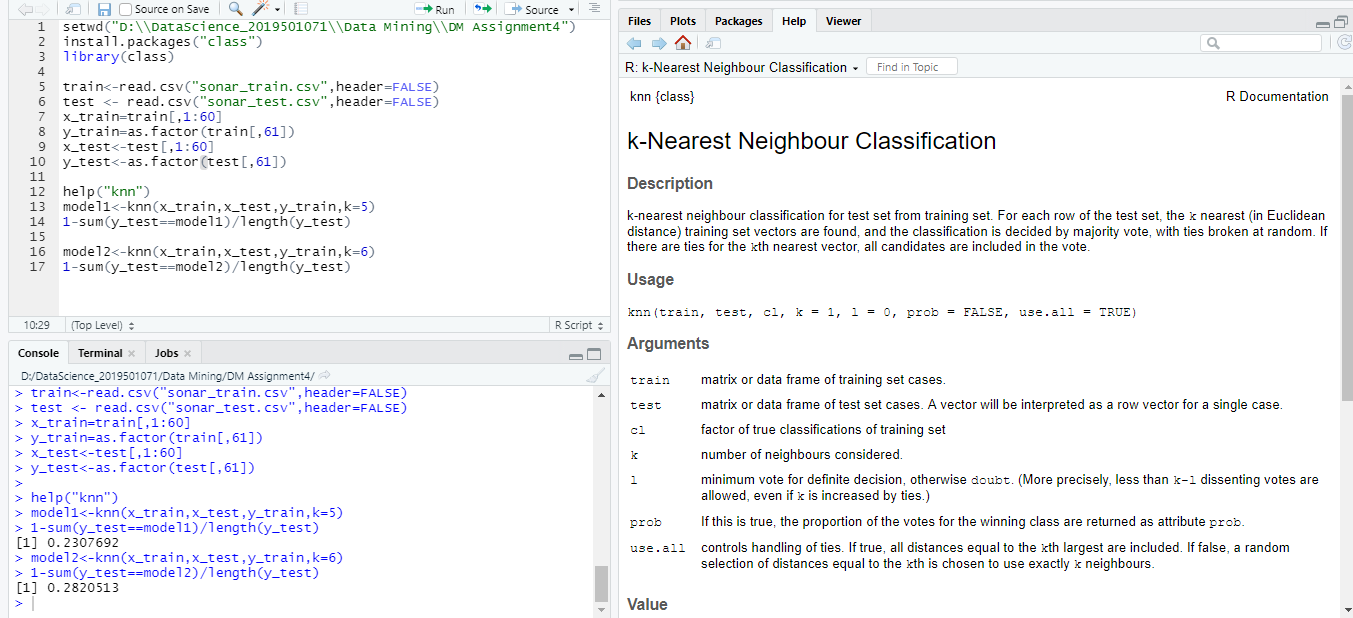
# Recall = 1/5 = 20%

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

6) Compute the misclassification error on the training data for the Random Forest classifier from In Class Exercise #47. Show your R code for doing this.



7) This question deals with sonar data   
  
a) Use knn() for the k-nearest neighbor classifier for k=5 and k=6 to the last column of the sonar training data. Compute the misclassification error on the training data and also on the test data.   
  
b) Repeat part a using the exact same R code a few times. Explain why both the training errors and the test errors often change for k=6 but not for k=5. Hint: Read the help on the knn function if you do not know.



# We can choose any value for k either even or odd but it is better to choose an odd value for k when there are two classes because if k value is even there might be a risk of a tie in the decision of which class to choose. So, the misclassification error rate often changes for k=6.