Data Mining Assignment 4

1) Read Chapter 4 (all sections) and Chapter 5 (Sections 5.2, 5.5, 5.6 and 5.7).  
  
2) 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?

split on A

Error(A=T) = 1 - max(4/7, 3/7) = 3/7 = 0.43

Error(A=F) = 1- max(0/3, 3/3) = 0

weighted average error = 7/10 \* 0.43 + 3/10 \* 0 = 0.30

split on B

Error(B=T) = 1 - max(3/4, 1/4) = 1/4 = 0.25

Error(B=F) = 1 - max(1/6, 5/6) = 1/6 = 0.16

weighted average error = 4/10 \* 0.25 + 6/10 \* 0.16 = 0.20

According to misclassification error rate, splitting on B is the best split, because the weighted average error for B is less than A.  
  
  
  
3) 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?

|  |  |  |  |
| --- | --- | --- | --- |
| a3 | Class Label | Split point | weighted Error |
| 1.0 | + | 2.0 | 0.33 |
| 3.0 | - | 3.5 | 0.42 |
| 4.0 | + | 4.5 | 0.33 |
| 5.0 | - | 5.5 | 0.44 |
| 5.0 | + | 0.44 |
| 6.0 | + | 6.5 | 0.44 |
| 7.0 | - | 7.5 | 0.44 |
| 7.0 | + | 0.44 |

split point 2.0 or 4.5 is the best split according to weighted error.

1. 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.

|  |  |  |  |
| --- | --- | --- | --- |
| Age | Number | Start | Prediction |
| 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 |

Predictions on the above test data are:

i)Age = middle, Number = 5, Start = 10

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

ii)Age = young, Number = 2, Start = 17

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

iii)Age = old, Number = 10, Start = 6

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

iv)Age = young, Number = 2, Start = 17

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

v)Age = old, Number = 4, Start = 15

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

vi)Age = middle, Number = 5, Start = 15

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

vii)Age = young, Number = 3, Start = 13

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

viii)Age = old, Number = 5, Start = 8

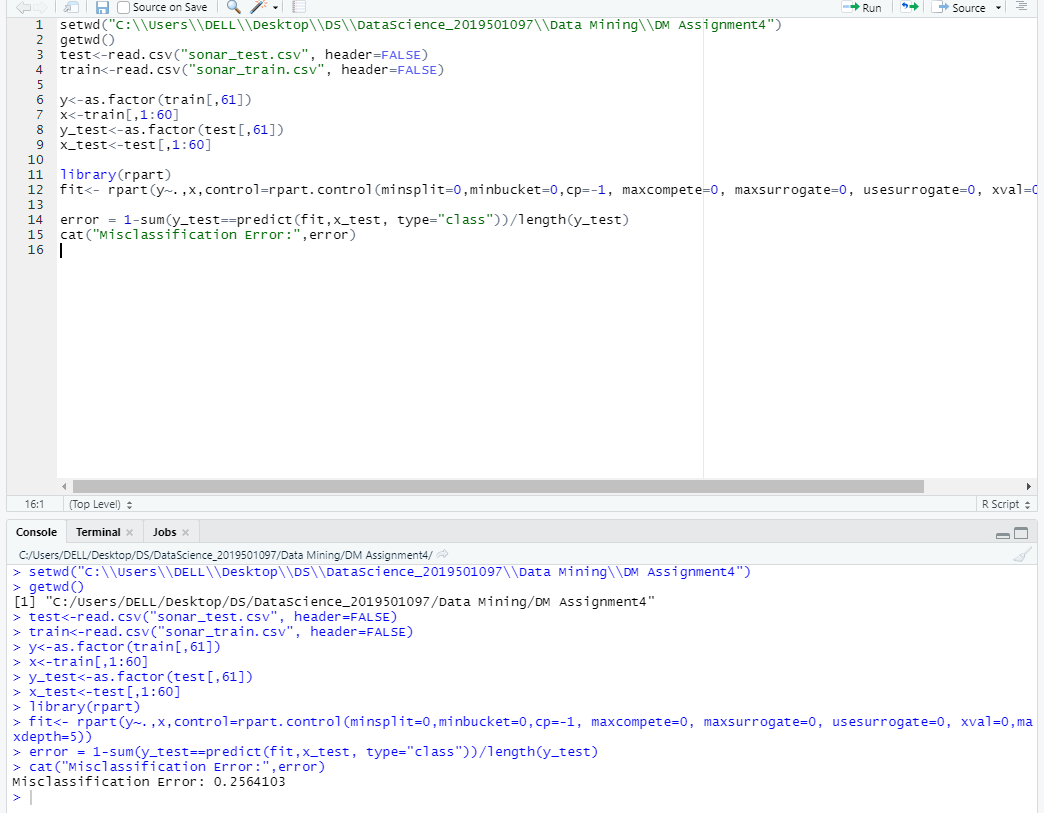
Path:1->3->7->15->present

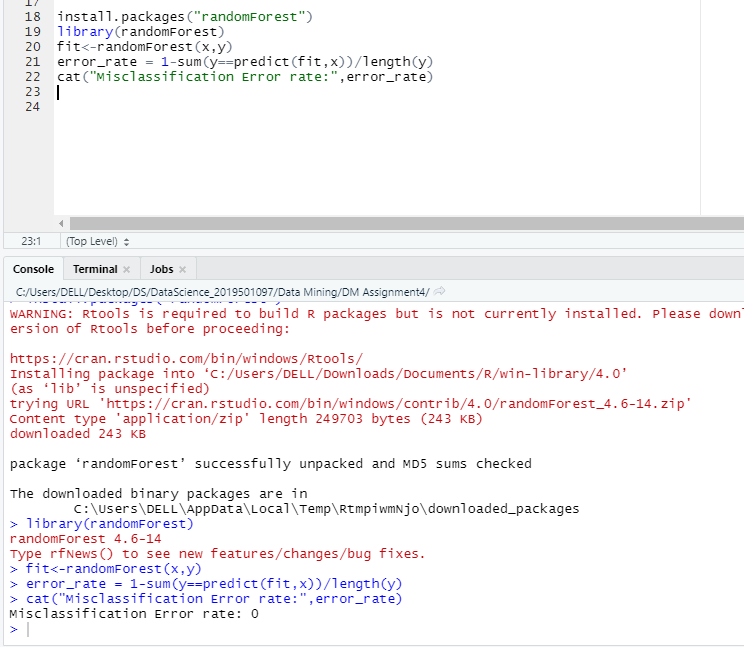
ix)Age = young, Number = 7, Start = 9

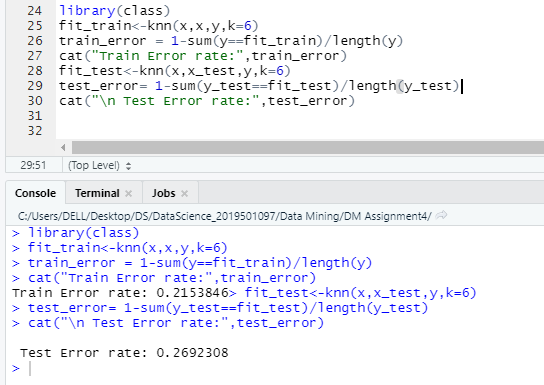
Path:1->2->4->9->absent

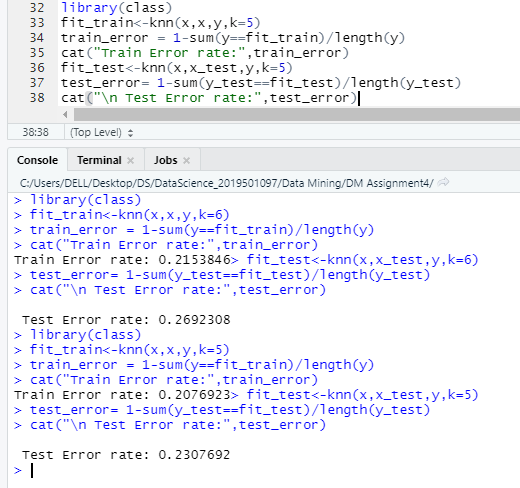
x)Age = middle, Number = 3, Start = 13

Path:1->2->5->10->absent  
  
5) 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.

  
  
6) 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.   
  
7) Compute the misclassification error on the training data for the Random Forest classifier to the last column of the sonar training data. Show your R code for doing this.

  
  
8) 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.

Knn function states “ties broken at random”. For each row of the test set, the k nearest (in Euclidean distance) training set vectors are found, and the classification is decided by majority vote, with ties broken at random. If there are ties for the kth nearest vector, all candidates are included in the vote.

For odd k, there will never be ties, while for even k, there are frequently ties.