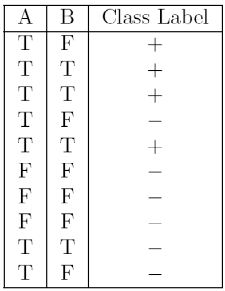
1) Read Chapter 4 (all sections) and Chapter 5 (Sections 5.2, 5.5, 5.6 and 5.7).  
  
2) Consider the following data set for a binary class problem.



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?

Ans.

splitting 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

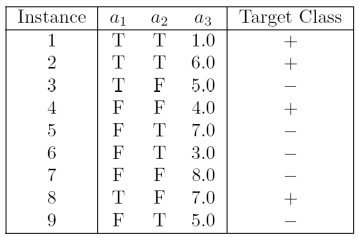
splitting on B

Error(B=T) = 1 - max(3/4, 1/4) = ¼ = 0.25

Error(B=F) = 1 - max(1/6, 5/6) = ⅙ = 0.16

weighted average error = 4/10 \* 0.25 + 6/10 \* 0.17 = 0.20

According to misclassification error rate, splitting on B is the best split, because the weighted average error for B is lesser 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?

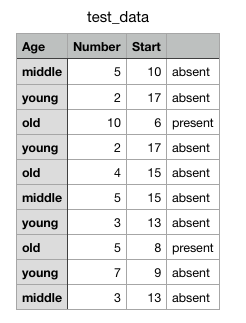
Ans.

|  |  |  |  |
| --- | --- | --- | --- |
| 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.

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.

Ans.



predictions on the above test data are:

1. Age = middle, Number = 5, Start = 10

Path : 1 → 2 → 5 → 11 → Present

* + 1. Age = young, Number = 2, Start = 17

Path : 1 → 2 → 4 → 8 → Absent

* + 1. Age = old, Number = 10, Start = 6

Path : 1 → 3 → 7 → 15 → Present

* + 1. Age = young, Number = 2, Start = 17

Path : 1 → 2 → 4 → 8 → Absent

* + 1. Age = old, Number = 4, Start = 15

Path : 1 → 2 → 4 → 8 → Absent

* + 1. Age = middle, Number = 5, Start = 15

Path : 1 → 2 → 5 → 10 → Absent

* + 1. Age = young, Number = 3, Start = 13

Path : 1 → 2 → 4 → 9 → Absent

* + 1. Age = old, Number = 5, Start = 8

Path : 1 → 3 → 7 → 15 → Present

* + 1. Age = young, Number = 7, Start = 9

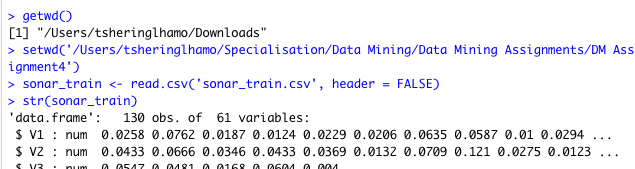
Path : 1 → 2 → 4 → 9 → Absent

* + 1. Age = middle, Number = 3, Start = 13

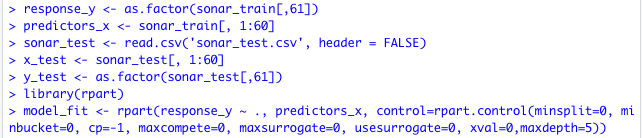
Path : 1 → 2 → 5 → 10 → Absent

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.

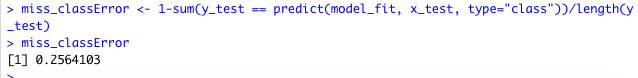
Ans.



make class variable a factor(a nominal variable)



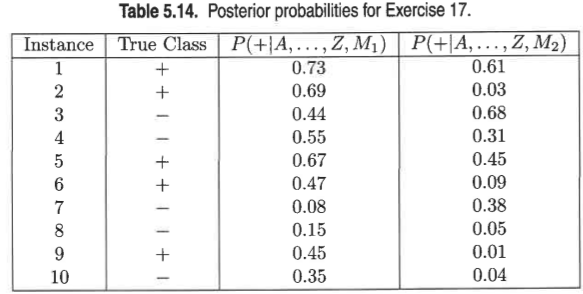
call library to load rpart into your r session before using it.



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.

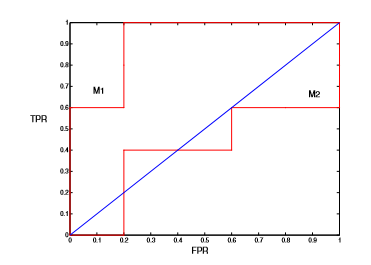
17) You are asked to evaluate the performance of two classification models, M1 and M2. The test set you have chosen contains 26 binary attributes, labeled as ,4 through Z.

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(+) andP(-|A, . . . , Z) = 1 - P(+|A, ...,Z). Assume that we are mostly interested in detecting instances I from the positive class.



(a) PIot 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.

Ans.

  
 M1 is better, since its area under the ROC curve is larger than the area under the ROC curve for M2.

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

Ans. confusion matrix when t = 0.5 is given below.

|  |  |  |
| --- | --- | --- |
| Actual | Predicted | |
|  |  |
|  | 3 | 2 |
|  | 1 | 4 |

Precision, p = TP/TP+FP = 3/3+1 = ¾ = 75%

Recall, r = TP/TP+FN = 3/3+2 = ⅗ = 60%

F-measure = 2rp/r+p = (2 \* .75 \* .60)/ .75 + .60 = 0.667

(c) Repeat the analysis for part (b) 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?

Ans. confusion matrix when t = 0.5 is given below.

|  |  |  |
| --- | --- | --- |
| Actual | Predicted | |
|  |  |
|  | 1 | 4 |
|  | 1 | 4 |

Precision, p = 1/1+1 = 1/2 = 50%

Recall, r = 1/1+4 = 1/5 = 20%

F-measure = 2rp/r+p = (2 \* .20 \* .50)/ .20 + .50 = 0.285

Based on F-measure, M1 is better than M2. The result is consistent with the ROC curve.

(d) Repeat part (c) for model M1 using the threshold t = 0.1. Which threshold do you prefer, t = 0.5 or t = 0.1? Are the results consistent with what you expect from the ROC curve?

Ans. confusion matrix when t = 0.1 is given below.

|  |  |  |
| --- | --- | --- |
| Actual | Predicted | |
|  |  |
|  | 5 | 0 |
|  | 4 | 1 |

Precision, p = 5/5+4 = 5/9 = 55.6%

Recall, r = 5/5+0 = 5/5 = 100%

F-measure = 2rp/r+p = (2 \* 1 \* .556)/ 1 + .556 = 0.715

Based on F-measure, t = 0.1 is better.

When t = 0.1, F P R = 0.8 and T P R = 1. On the other hand, when t = 0.5, F P R = 0.2 and T RP = 0.6. Since (0.2, 0.6) is closer to the point (0, 1), we favor t = 0.5. This result is inconsistent with the results using F-measure.

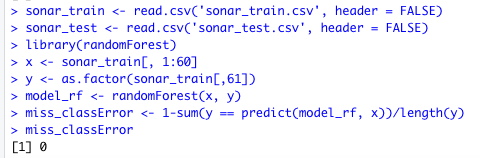
We can also show this by computing the area under the ROC curve.

For t = 0.5, area = 0.6 × (1 − 0.2) = 0.48.

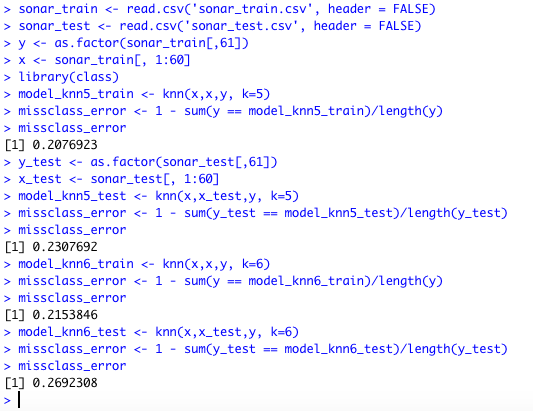
For t = 0.1, area = 1 × (1 − 0.8) = 0.2.

Since the area for t = 0.5 is larger than the area for t = 0.1, we prefer t = 0.5.

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.  
Ans. 



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.   
Ans. load class package to r session before using knn.



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.

Ans.

k-nearest neighbour classification for test set from training set. 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.

From the definition of the algorithm, I understood that as the value of k(neighbors) increases there can be more chances of ties, and using random tie breaking, the data points are classified into different classes. Therefore, the errors also keep varying for higher values of k.