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?

For A :

|  |  |  |
| --- | --- | --- |
|  | T | F |
| + | 4 | 0 |
| - | 3 | 3 |
|  | 7 | 3 |

T error = 1 - max[4/7, 3/7] = 1 - 4/7 = 3/7 = 0.43

F error = 1 - max[0, 3/3] = 1-1 = 0

Weighted avg. error = (7/10 \* 0.43) + (3/10 \* 0) = 0.301

For B:

|  |  |  |
| --- | --- | --- |
|  | T | F |
| + | 3 | 1 |
| - | 1 | 5 |
|  | 4 | 6 |

T error = 1 - max[3/4, 1/4] = 1 - 0.75 = 0.25

F error = 1 - max[1/6, 5/6] = 1 - 5/6 = 1/6 = 0.1667

Weighted avg. error = (4/10 \* 0.25) + (6/10 \* 0.1667) = 0.1+0.10002 = 0.20002

Best split is the lowest error B = 0.20002  
  
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?

For a1 :

|  |  |  |
| --- | --- | --- |
|  | T | F |
| + | 3 | 1 |
| - | 1 | 4 |
|  | 4 | 5 |

T error = 1 - max[1/4, 3/4] = 0.25

F error = 1 - max[1/5, 4/5] = 1 - 4/5 = 0.2

Weighted avg. error = (4/9 \* 0.25) + (5/9 \* 0.2) =0.333

For a2 :

|  |  |  |
| --- | --- | --- |
|  | T | F |
| + | 2 | 2 |
| - | 3 | 2 |
|  | 5 | 4 |

T error = 1 - max[2/5, 3/5] = 1 - 0.6 = 0.4

F error= 1 - max[2/4, 2/4] = 1 - 2/4 = 0.5

Weighted average error= (5/9 \* 0.4) + (4/9 \* 0.5) = 0.4444

Best split is the lowest error a1 = 0.333

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.

Observations 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.   
  
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.

You are asked to evaluate the performance of two classification models, *M*1

and *M*2. The test set you have chosen contains 26 binary attributes, labeled

as *A* through *Z*.

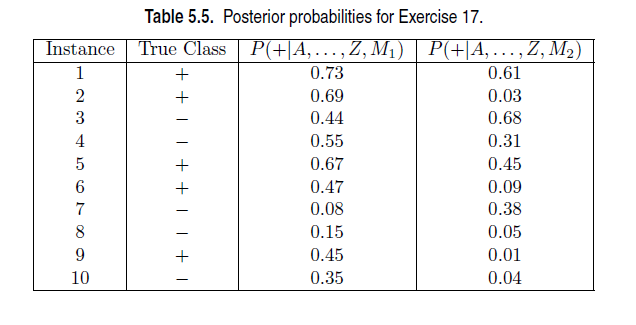
Table 5.5 shows the posterior probabilities obtained by applying the models to

the test set. (Only the posterior probabilities for the positive class are shown).

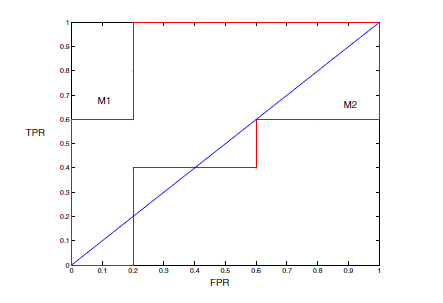
As this is a two-class problem, *P*(*−*) = 1 *− P*(+) and *P*(*−|A, . . . , Z*) = 1 *−*

*P*(+*|A, . . . , Z*). Assume that we are mostly interested in detecting instances

from the positive class.



(a) Plot the ROC curve for both *M*1 and *M*2. (You should plot them on the same graph.) Which model do you think is better? Explain your reasons.



*M*1 is better, since its area under the ROC curve is larger than the area under ROC curve for *M*2.

(c) Repeat the analysis for part (c) using the same cutoff threshold on model *M*2. 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, the confusion matrix for *M*2 is shown below.

+ -

Actual + 1 4

- 1 4

Precision = 1*/*2 = 50%.

Recall = 1*/*5 = 20%.

F-measure = (2 *×* 0*.*5 *×* 0*.*2)*/*(0*.*5 + 0*.*2) = 0*.*2857.

Based on F-measure, *M*1 is still better than *M*2. This result is consistent

with the ROC plot.  
  
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.