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?  
  
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?

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.  
  
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.

train\_data <- read.csv("sonar\_train.csv",header = FALSE)

test\_data <- read.csv("sonar\_test.csv", header = FALSE)

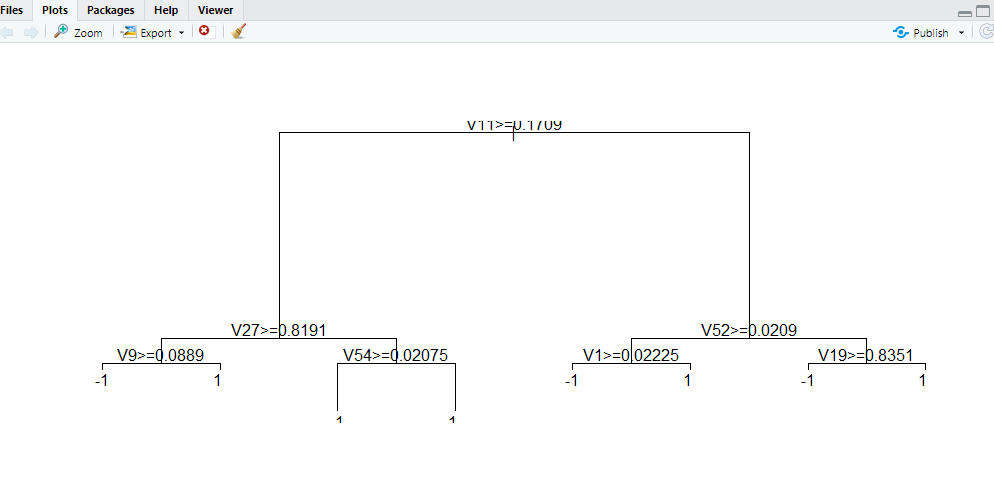
y <- as.factor(train\_data[,61])

x <- train\_data[, 1:60]

fit <- rpart(y ~ ., x, control = rpart.control(minsplit = 0, minbucket = 0,maxdepth = 3))

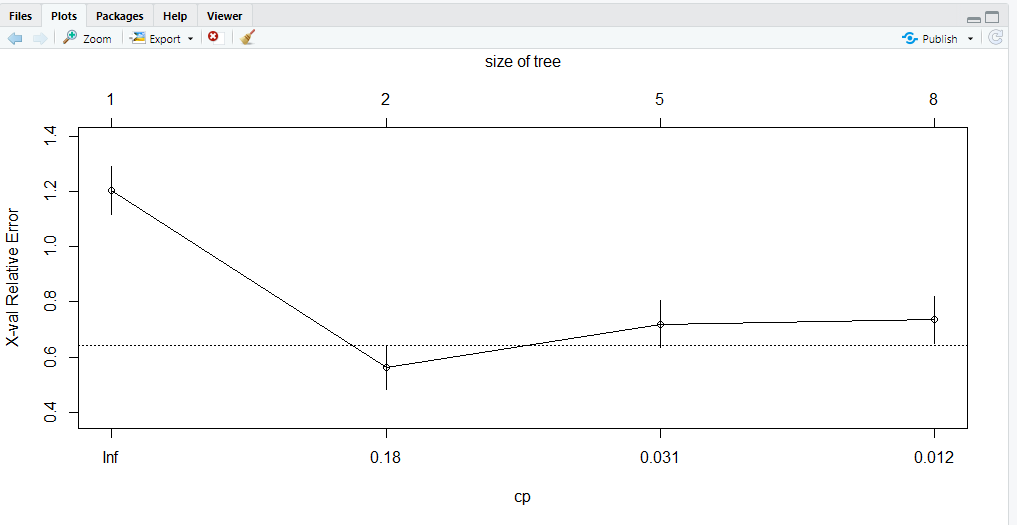
plot(fit)

text(fit)



print(fit)

plotcp(fit)



predictions <- as.numeric(predict(fit, test\_data, type = "class"))

predictions <- replace(predictions, predictions == 1, -1)

predictions <- replace(predictions, predictions == 2, 1)

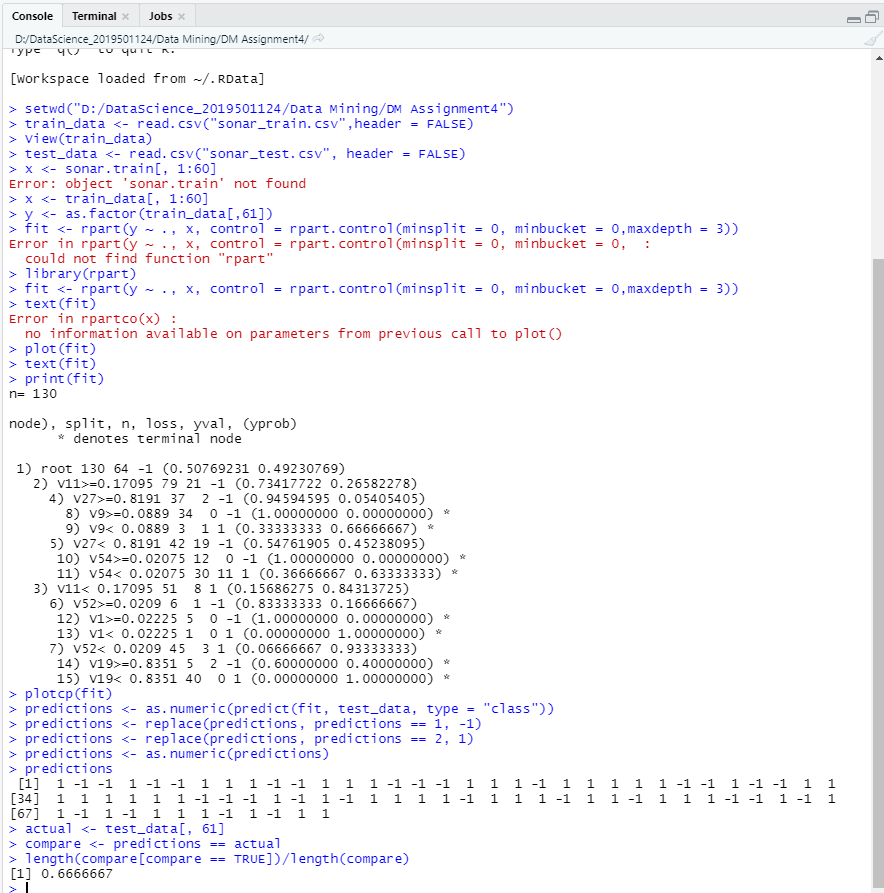
predictions <- as.numeric(predictions)

predictions

actual <- test\_data[, 61]

compare <- predictions == actual

length(compare[compare == TRUE])/length(compare)



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.

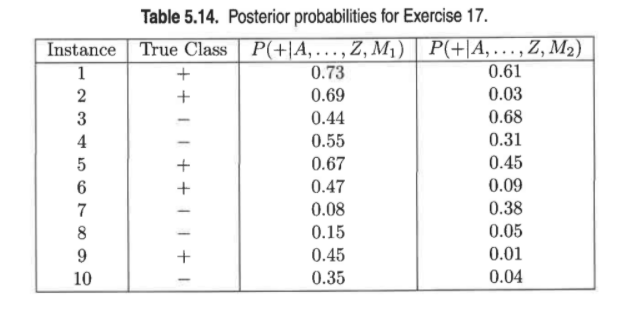


Table shows the posterior probabilities obtained by applying the models

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(-lA, . . ., Z) :I - P(+lA, . . ., Z). Assume that we are mostly interested in detecting instances 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.

The ROC curve for M1 and M2 are shown in the Figure 4.5. M1 is better, since its area under the ROC curve is larger than the area under ROC curve for M2.

(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?

Answer: When t = 0.5, the confusion matrix for M2 is shown below.

+ -

Actual + 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 still better than M2.

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.