Math 5365

Data Mining 1

Homework 15

Mary Barker

1. Compare the accuracies of bagging, boosting, and random forests on the wdbc data set, using 70% of the data as training data and 30% as testing data.

   The respective accuracies and runtimes of the three methods are shown in the table below.

Method	Accuracy	runtime
Bagging	98.24561%	21.354 seconds
Forest	99.4152%	0.268 seconds
Boosting	99.4152%	21.478 seconds

   All three methods displayed excellent accuracy rates. However, the Forest model and the boosting method had the highest accuracy. In addition, the forest model performed better with runtime. Therefore the best method is the fastest and tied for most accurate in forest. The boosting method is the second fastest and also the most accurate.

```
1  #Data Mining hw 15

2

3  # Compare the accuracies of bagging, boosting, and random forests on the wdbc

4  # data set, using 70% of the data as training data and 30% as testing data.

5  library(adabag)

6  library(randomForest)

7

8  source('~/Dropbox/Tarleton/data_mining/class_notes/extras.R')

9  source('~/Dropbox/Tarleton/data_mining/generic_functions/dataset_ops.R')

10  wdbc <- read.csv('~/Dropbox/Tarleton/data_mining/dfiles/wdbc.data',
```

```
11                   header=F, sep=',')
12   wdbc <- wdbc[,-1]
13   splitset <- splitdata(wdbc,0.7,F)
14   train <- splitset$train
15
16   bag1 <- proc.time()
17   baggingmodel <- bagging(V2~., wdbc[train,])
18   bagpred <- predict(baggingmodel,wdbc[-train,],type='class')
19   bagacc <- confmatrix(bagpred$class,wdbc$V2[-train])$accuracy
20   bag2 <- proc.time()
21
22   forest1 <- proc.time()
23   forestmodel <- randomForest(V2~.,wdbc[train,])
24   forestpred <- predict(forestmodel,wdbc[-train,])
25   forestacc <- confmatrix(forestpred,wdbc$V2[-train])$accuracy
26   forest2 <- proc.time()
27
28   boost1 <- proc.time()
29   boostingmodel <- boosting(V2~., wdbc[train,])
30   boostpred <- predict(boostingmodel,wdbc[-train,],type='class')
31   boostacc <- 1 - boostpred$error
32   boost2 <- proc.time()
```