Hacker Dojo Machine Learning
Homework 2
Mike Bowles, PhD & Patricia Hoffman, PhD.

Homework on Trees

1) For binary classification, consider the training examples in the table below (shown in Table 4.8 from the book on page 198). For column a_3, which is a continuous attribute, compute the misclassification rate for every possible split. Next compute the information gain for every possible split of that same variable. Which of these splits considered is the best according to misclassification error rate? Is that split also the best when considering information gain? (You can code this in r or you can do this by hand.)

Instance	a_1	a_2	a_3	Target Class
1	T	T	1.0	+
2	T	T	6.0	+
3	T	F	5.0	−
4	F	F	4.0	+
5	F	T	7.0	−
6	F	T	3.0	−
7	F	F	8.0	−
8	T	F	7.0	+
9	F	T	5.0	−

2) The following tree was created using rpart for the table given in this homework problem number one.

```
node), split, n, loss, yval, (yprob)
      * denotes terminal node

 1) root 9 4 0 (0.5555556 0.4444444)
   2) a1< 0.5 5 1 0 (0.8000000 0.2000000)
     4) a2>=0.5 3 0 0 (1.0000000 0.0000000) *
     5) a2< 0.5 2 1 0 (0.5000000 0.5000000)
      10) a3>=6 1 0 0 (1.0000000 0.0000000) *
      11) a3< 6 1 0 1 (0.0000000 1.0000000) *
   3) a1>=0.5 4 1 1 (0.2500000 0.7500000)
```

```
6) a2< 0.5 2 1 0 (0.5000000 0.5000000)
 12) a3< 6 1 0 0 (1.0000000 0.0000000) *
 13) a3>=6 1 0 1 (0.0000000 1.0000000) *
7) a2>=0.5 2 0 1 (0.0000000 1.0000000) *
```

Use this tree to predict the class labels (either a + or -) for the following test observations:

Observation	a1	a2	a3
1	T	T	2.5
2	T	F	5.5
3	F	T	2.5
4	F	F	8.5

3) Consider the table given in the text on page 200 in the book exercise number five (copied below). It is a binary class problem. Determine the best places to split the data for a decision tree. The first time use the Gini Index as the impurity measure. The second time use the classification error as the impurity measure. In both cases use the information gain rate to determine the best split. Note any changes relative to the tree built from the two methods.

A	B	Class Label
T	F	+
T	T	+
T	T	+
T	F	−
T	T	+
F	F	−
F	F	−
F	F	−
T	T	−
T	F	−

4) The UC Irvine web site has many interesting data sets. Sonar data is described at the web site: [http://archive.ics.uci.edu/ml/machine-learning-databases/undocumented/connectionist-bench/sonar/sonar.**names**](http://archive.ics.uci.edu/ml/machine-learning-databases/undocumented/connectionist-bench/sonar/sonar.names) The sonar data set can be found at http://archive.ics.uci.edu/ml/machine-learning-databases/undocumented/connectionist-bench/sonar/ Divide the sonar data

set into a training set (sonar_train.csv) and a test set (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. Documentation for the rpart package can be found at http://cran.r-project.org/web/packages/rpart/rpart.pdf

5) Check out the web page which describes a wine quality data set:

http://archive.ics.uci.edu/ml/machine-learning-databases/wine-quality/winequality.names

Use the Red Wine data set: winequality-red.csv This data set contains 1599 observations of 11 attributes. The median score of the wine tasters is given in the last column. Note also that the delimiter used in this file is a semi colon and not a comma. Use rpart on this data to create trees for a range of different tree depths. Use cross validation to generate training error and test error. Plot these errors as a function of tree depth. Which tree depth results in the best Test Error? What is that Test Error? Hint: look at the cross validation example given in the lecture.