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

Homework on Trees

1) Repeat In Class Exercise #38 (from David Mease Lecture 5 - our lecture 4) twice. The first time use the information gain rate instead of misclassification error to determine the best split. The second time use the change in the Gini Index. Note any changes relative to the tree built from the miss classification error rate.

2) Repeat In Class Exercise #39 (from David Mease Lecture 5 - our lecture 4) using the misclassification error rate instead of information gain to determine the best split. Which of these splits considered is the best according to misclassification error rate? This is also the text book question 3(c) on page 199.

3) The file 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 test_data.csv linked here. Do this manually - do not use R or any software.

4) David Mease split the popular 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.

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. Hint: look at the cross validation example given in the lecture.