ITU

Homework 2
Patricia Hoffman, PhD.

The Entropy of a node is defined by equation(1), where $p(i|\text{node})$ denotes the fraction of records belonging to class i at the given node. The Gini index and Classification Error are defined in the equations (2) and (3) respectively.

$$\textbf{Entropy(node)} = -\sum_{i=0}^{c-1} p(i|\text{node}) \log_2 p(i|\text{node}) \tag{1}$$

$$\textbf{Gini(node)} = 1 - \sum_{i=0}^{c-1} [p(i|\text{node})]^2 \tag{2}$$

$$\textbf{Classification Error(node)} = 1 - \max_i [p(i|\text{node})] \tag{3}$$

If $I(\text{node})$ is any one of the three impurity measure given in one of the first three equations, the Purity Gain is defined by the following equation:

$$\textbf{Purity Gain} = I(\textbf{parent}) - \sum_{j=1}^{k} \frac{N(\nu_j)}{N} I(\nu_j) \tag{4}$$

where N is the total number of records at the parent node, k is the number of attribute values, and $N(\nu_j)$ is the number of records associated with each child node, ν_j.

$I(\text{node})$ can be any of the three impurity measures defined by the first three equations. The best split for a node is the one that maximizes the Purity Gain. When $I(\text{node})$ is Entropy the Purity Gain is called Information Gain

The data for this homework assignment is in the “Data Files for use in Homeworks” section under the course resources tab.

Homework on Trees

1) This question uses the following ages for a set of trees: 19, 23, 30, 30, 45, 25, 24, 20. Store them in R using the syntax
   ages<-c(19, 23, 30, 30, 45, 25, 24, 20).
   a) Compute the standard deviation in R using the sd() function. Also

compute the mean and median.
b) Compute the same value in R without the sd function.
c) Using R, how does the standard deviation from part a) change if you add 10 to all the values?
d) Using R, how does the standard deviation in part a) change if you multiply all the values by 100?
e) Next add another tree of age 70 to the sample. Compute the mean and median with this tree added to the sample. How have the mean and median changed?

2) Here is the data table for question 2.

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	−

The following tree was created using rpart for the data table given above.

```
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. Would it be possible to create a model which would correctly classify this training data? If it is possible create a tree which gives the correct answer (either + or -) for each training observation. Otherwise, give the reason that it is not possible to do so.

Observation	A	B	Class Label
1	T	F	+
2	T	T	+
3	T	T	+
4	T	F	-
5	T	T	+
6	F	F	-
7	F	F	-
8	F	F	-
9	T	T	-
10	T	F	-

4) The UC Irvine web site has many interesting data sets. The Sonar Data is described at the web site: http://archive.ics.uci.edu/ml/machine-learning-databases/undocumented/connectionist-bench/sonar/sonar.**names** . The sonar data has been divided into a training set (sonar_train.csv) and a test set (sonar_test.csv). The file sonar_test.csv should be used as the hold out set, while the file sonar_train.csv should be used to build the tree. Use R to compute the classification error on the test set when training on the training set for a tree of depth 5 using control=rpart.control(maxdepth=5). Remember that the 61st column is the response and the other 60 columns are the predictors.

What is the error on the training set? What is the error on the test set? What is the differences in these errors?

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 for each tree depth. 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. Are you considering the tasters score as a class, an ordered factor, or a numeric? How did you calculate the error?

Which attribute is at the root node? Make a scatter plot of the wine quality score vs. this root node attribute.

What is the correlation between each of the eleven attributes and the wine quality (hint: cor(wineData[, 1:11], wineData$quality). Which attribute has the highest (in absolution value) correlation with the wine quality? Make a scatter plot of the wine quality score vs. this attribute.