CS 5525 Assignment 2

Due 1st Oct, 2020

Problem 1. Decision Tree based on Gini Index **15 points**

Consider the training examples shown in the Table below for a binary classification problem.

Customer ID	Gender	Car Type	Shirt Size	Class
1	M	Family	Small	C0
2	M	Sports	Medium	C0
3	M	Sports	Medium	C0
4	M	Sports	Large	C0
5	M	Sports	Extra Large	C0
6	M	Sports	Extra Large	C0
7	F	Sports	Small	C0
8	F	Sports	Small	C0
9	F	Sports	Medium	C0
10	F	Luxury	Large	C0
11	M	Family	Large	C1
12	M	Family	Extra Large	C1
13	M	Family	Medium	C1
14	M	Luxury	Extra Large	C1
15	F	Luxury	Small	C1
16	F	Luxury	Small	C1
17	F	Luxury	Medium	C1
18	F	Luxury	Medium	C1
19	F	Luxury	Medium	C1
20	F	Luxury	Large	C1

Fig. 1. The data set for applying the Decision Tree approach.

Compute the Gini Index value of the parent and the child nodes obtained when the split was made on the following attributes (i) Gender and (ii) Car Type. Show all your calculations including the Gini of the parent, contingency tables, Gini of the individual children.

Problem 2. Greedy Heuristic for Decision Trees **15 points**

Consider the following data set that contains 100 training examples (50 labeled as positive class while the remainder labeled as negative class).

X	Y	Z	No. of + Examples	No. of − Examples
1	1	1	5	0
1	1	0	0	20
1	0	1	20	0
1	0	0	0	5
0	1	1	10	0
0	1	0	15	0
0	0	1	0	10
0	0	0	0	15

1) Build a *two-level* decision tree using gini index as the criterion for splitting. You need to show your computations for each candidate splitting attribute at each level clearly to obtain full credit. What is the overall training error rate of the induced tree? Note: we consider a tree with only 1 internal node and two leaf nodes as a *one-level* decision tree. [8]

2) Use variable X as the first splitting attribute, then choose the best available splitting attribute at each of the two successor nodes. What is the training error rate of the induced tree? [5]

3) Discuss the results obtained in parts (a) and (b) above. Comment on the suitability of the greedy heuristic used as the splitting attribute selection. [2]

Problem 3. Properties of Entropy **25 points**

1) Show that the entropy measure $-p(x)\log p(x)$ is non-negative. [3]

2) The following dataset consists of 7 examples, each with 3 attributes, (A, B, C), and a label, Y. [22]

A	B	C	Y
1	1	0	0
1	1	2	1
1	0	0	1
1	1	2	1
0	0	2	0
0	1	1	0
0	0	0	0

Use the data above to answer the following questions. Please show the formula and steps.

a. What is the entropy of Y in bits, H(Y) ? In this and subsequent questions, when we request the **units in bits, this simply means that you need to use log base 2 in your calculations**. (Please include one number rounded to the fourth decimal place, e.g. 0.1234) [2]

b. What is the information gain of Y w.r.t. A in bits, I(Y ; A)? (Please include one number rounded to the fourth decimal place, e.g. 0.1234) [3]

c. What is the information gain of Y w.r.t. B in bits, I(Y ; B)? (Please include one number rounded to the fourth decimal place, e.g. 0.1234) [3]

d. What is the information gain of Y w.r.t. C in bits, I(Y ; C)? (Please include one number rounded to the fourth decimal place, e.g. 0.1234) [3]

e. If the same algorithm continues until the tree perfectly classifies the data, what would the depth of the tree be? [3]

f. Consider the dataset given above. Which attribute (A, B, or C) would a decision tree algorithm pick first to branch on, if its splitting criterion is mutual information? [3]

g. Consider the dataset given above. Which is the second attribute you would pick to branch on, if its splitting criterion is information gain? (Hint: Notice that this question correctly presupposes that there is exactly one second attribute.) [3]

h. Draw your completed Decision Tree. Label the non-leaf nodes with which attribute the tree will split on (e.g. B), the edges with the value of the attribute (e.g. 1 or 0), and the leaf nodes with the classification decision (e.g. Y = 0). [2]

Problem 4. Information Gain and Gain Ratio **20 points**

Consider the problem of predicting how well a baseball player will bat against a particular pitcher. The training set contains ten positive and ten negative examples. Assume there are two candidate attributes for splitting the data—ID (which is unique for every player) and Handedness (left or right). Among the left-handed players, nine of them are from the positive class and one from the negative class. On the other hand, among the right-handed players, only one of them is from the positive class, while the remaining nine are from the negative class.

1) Compute the information gain if we use ID as the splitting attribute. [4]

2) Repeat part (1) using Handedness as the splitting attribute. [4]

3) Based on your answers in parts (1) and (2), which attribute will be chosen according to information gain? [2]

4) Repeat part (1) using gain ratio (instead of information gain). [4]

5) Repeat part (2) using gain ratio (instead of information gain). [4]

6) Based on your answers in parts (4) and (5), which attribute will be chosen according to gain ratio? [2]

Problem 5. Model evaluation and statistics **25 points**

1. You have been asked to develop a classification model for diagnosing whether a patient is infected with a certain disease. To help you construct the models, your collaborator has provided you with a small training set (N = 10) with equal number of positive and negative examples. You tried several approaches and found two most promising models, C_1 and C_2. The outputs of the models in terms of predicting whether each of the training examples belong to the "positive" class are summarized in the table below. The first row shows the probability a training example belongs to the positive class according to classifier C1, while the second row shows the same information for classifier C2. The last row indicates the true class label of the 10 training examples.

$P(y = +/C_1)$	0.1	0.15	0.2	0.3	0.31	0.4	0.62	0.77	0.81	0.95
$P(y = +/C_2)$	0.25	0.49	0.05	0.35	0.66	0.6	0.7	0.65	0.55	0.99
y	-	+	-	-	+	-	+	+	-	+

a. Draw the corresponding ROC curves for both classifiers on the same plot. [6]

b. Compute the area under ROC curve for each classifier. Which classifier has a larger area under the curve? [4]

c. Compute the Wilcoxon Mann Whitney statistic for both classifiers. The statistic can be computed as follows:

$$WMW = \frac{\sum_{i=0}^{m-1} \sum_{j=0}^{n-1} I(xi, yj)}{mn}$$

Where

$$I(x,y) = \begin{cases} 1, & xi > yj \\ 0, & otherwise \end{cases}$$

Note that $\{x_0, x_1, \cdots, x_{m-1}\}$ correspond to the classifier outputs for the m positive examples while $\{y_0, y_1, \cdots, y_{n-1}\}$ correspond to the classifier outputs for the n negative examples (in this exercise, $m = n = 5$). Which classifier has a larger WMW value? [6]

Based on your answers, state the relationship between WMW and the ROC curve. [4]

2. Consider the following confusion matrix for a dataset of 1000 points with binary labels predicted using a classifier f(X).

	Predicted Class = Yes	Predicted Class = No
Actual Class = Yes	345	225
Actual Class = No	195	235

Calculate the following metrics. All values should be rounded upto 3 decimal digits. [5]

a. Precision.

b. Recall / True Positive Rate.

c. False Positive Rate.

d. F-score.

e. Accuracy.
