Homework 2 – CS 5525

1. Consider the tabulated data for a binary classification problem. 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.

   Solution : Let us first compute the Gini index of the parent node before any splitting. Let us refer to this node as n_0. We have two classes $C0$ and $C1$ with relative frequencies $p(C0|n_0) = 0.5$ and $p(C1|n_0) = 0.5$. Thus the Gini index is

$$GINI(n_0) = 1 - (p(C0|n_0)^2 + p(C1|n_0)^2) = 0.5$$

   (i) Consider the first split on Gender that creates two children n_1 (male) and n_2 (female) with the following distribution.

	Gender	
	Male (n_1)	Female (n_2)
$C0$	6	4
$C1$	4	6

   Computing the Gini index of these two children yields

$$GINI(n_1) = 1 - (p(C0|n_1)^2 + p(C1|n_1)^2) = 1 - 0.6^2 - 0.4^2 = 0.48$$

   and

$$GINI(n_2) = 1 - (p(C0|n_2)^2 + p(C1|n_2)^2) = 1 - 0.4^2 - 0.6^2 = 0.48$$

   Therefore, we can find the Gini of split as $GINI_{split} = 0.5 * 0.48 + 0.5 * 0.48 = 0.48$.

   (ii) Next consider a three-way split on Car Type that creates three children n_3 (family) and n_4 (sports), and n_5 (luxury) with the following distribution.

	Car type		
	Family (n_3)	Sports(n_4)	Luxury (n_5)
$C0$	1	8	1
$C1$	3	0	7

Computing the Gini index of these three children yields

$$GINI(n_3) = 1 - 0.25^2 - 0.75^2 = 0.375,$$

$$GINI(n_4) = 1 - 1^2 - 0 = 0$$

$$GINI(n_5) = 1 - (1/8)^2 - (7/8)^2 = 0.2188$$

We can find the Gini of split as $GINI_{split} = 0.2 * 0.375 + 0.4 * 0 + 0.4 * 0.2188 = 0.1625$.

2. Consider the given data set that contains 100 training examples.
i) 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. What is the overall training error rate of the induced tree?

Solution : Let us refer to the root parent node as P, positive class as CP and negative class as CN. Thus, the Gini index of the parent node may be computed as

$$GINI(P) = 1 - (p(CP|P)^2 + p(CN|P)^2) = 1 - 0.5^2 - 0.5^2 = 0.5$$

Next, we construct the class distribution table for splitting attributes X, Y, and Z.

	$X = 0$	$X = 1$	$Y = 0$	$Y = 1$	$Z = 0$	$Z = 1$
CP(+)	25	25	20	30	15	35
CN(-)	25	25	30	20	40	10

Using the tabulated data, we next compute the $GINI_{split}$ for the three attributes as

$$GINI_{split}(X) = 0.5 * (1 - 0.5^2 - 0.5^2) + 0.5 * (1 - 0.5^2 - 0.5^2) = 0.5$$
$$GINI_{split}(Y) = 0.5 * (1 - 0.4^2 - 0.6^2) + 0.5 * (1 - 0.6^2 - 0.4^2) = 0.48$$
$$GINI_{split}(Z) = 0.55 * (1 - (15/55)^2 - (40/55)^2) + 0.45 * (1 - (35/45)^2 - (10/45)^2) = 0.3737$$

Since splitting at Z results in minimum GINI split, we do the first split at Z, and refer to the two child nodes as CZ_0 and CZ_1 for $Z = 0$ and $Z = 1$ respectively. We next prepare the count table for determining the splitting at node CZ_0

	CZ_0 $(Z = 0)$			
	$X = 0$	$X = 1$	$Y = 0$	$Y = 1$
CP(+)	15	0	0	15
CN(-)	15	25	20	20

Let us next compute the Gini split for attributes X and Y at node CZ_0

$$GINI_{split}(X;CZ_0) = (30/55) * (1 - 0.5^2 - 0.5^2) + (25/55) * (1 - 0^2 - 1^2) = 0.2727$$
$$GINI_{split}(Y;CZ_0) = (20/55) * (1 - 0^2 - 1^2) + (35/55) * (1 - (15/35)^2 - (20/35)^2) = 0.3117$$

Since splitting on X gives a better gain (lower Gini split) here, the node CZ_0 will be split based on attribute X into leaf nodes CX_{00} and CX_{01}. Repeating the process for node CZ_1,

	CZ_1 $(Z = 1)$			
	$X = 0$	$X = 1$	$Y = 0$	$Y = 1$
CP(+)	10	25	20	15
CN(-)	10	0	10	0

Let us next compute the Gini split for attributes X and Y at node CZ_0

$$GINI_{split}(X;CZ_1) = (20/45) * (1 - 0.5^2 - 0.5^2) + (25/45) * (1 - 0^2 - 1^2) = 0.2222$$
$$GINI_{split}(Y;CZ_1) = (30/45) * (1 - (2/3)^2 - (1/3)^2) + (15/45) * (1 - 1^2 - 0^2) = 0.2963$$

Thus, splitting on X again gives a better gain. The resulting decision tree is shown in Fig. 1. **Note that the naming the nodes may be done using any other convention, that may flip the results for node n0 and n1. The quiz has been graded to account for such changes in conventions.**

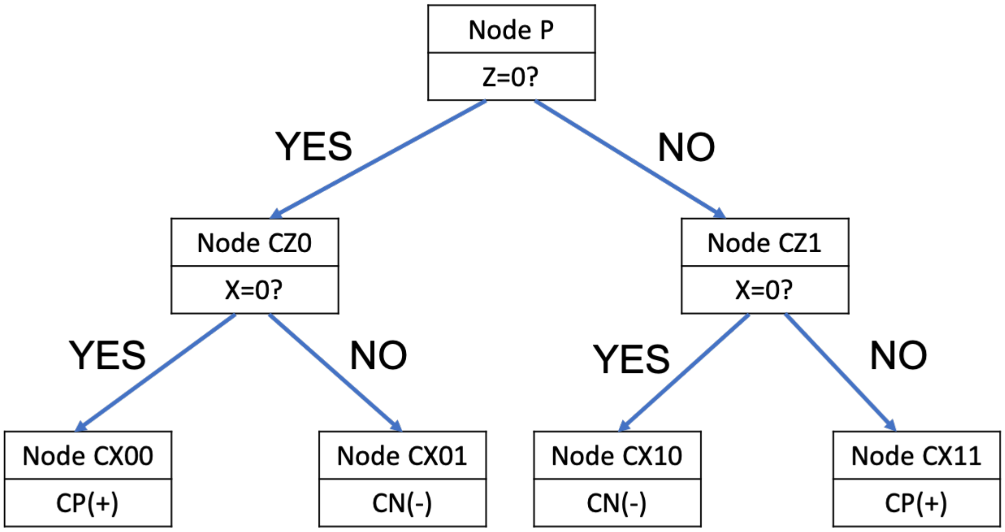


Figure 1: Two layer decision tree for binary classification.

Note that the nodes CX_{00} and CX_{10} have equal number of positive and negative cases. So they could have been classified as class CN and CP respectively without any extra error introduced. This would result in a single layer tree.

The decision tree in Fig. 1 misclassifies 15 samples at node CX_{00} and 10 samples at node CX_{10}, out of the total 100 samples. Thus the **training error rate is 0.25**.

ii) 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?

Solution : Using X as the first splitting attribute, let us refer to the two child nodes as CX_0 and CX_1 for $X = 0$ and $X = 1$ respectively. The count matrices and Gini split computation for node CX_0 provides

	CX_0 $(X = 0)$			
	$Y = 0$	$Y = 1$	$Z = 0$	$Z = 1$
CP(+)	0	25	15	10
CN(-)	25	0	15	10

Let us next compute the Gini split for attributes Y and Z at node CX_0

$$GINI_{split}(Y; CX_0) = 0.5 * (1 - 1^2 - 0^2) + 0.5 * (1 - 0^2 - 1^2) = 0$$
$$GINI_{split}(Z; CX_0) = (30/50) * (1 - 0.5^2 - 0.5^2) + (20/50) * (1 - 0.5^2 - 0.5^2) = 0.5$$

Since splitting at Y provides a better gain, node CX_0 shall be split into CY_{00} and CY_{01}. Proceeding to node CX_1, the count matrix can be found as

	CX_1 $(X = 1)$			
	$Y = 0$	$Y = 1$	$Z = 0$	$Z = 1$
CP(+)	20	5	0	25
CN(-)	5	20	25	0

The Gini split for attributes Y and Z at node CX_1 are

$$GINI_{split}(Y; CX_1) = 0.5 * (1 - 0.2^2 - 0.8^2) + 0.5 * (1 - 0.8^2 - 0.2^2) = 0.32$$
$$GINI_{split}(Z; CX_1) = 0.5 * (1 - 0^2 - 1^2) + 0.5 * (1 - 1^2 - 0^2) = 0$$

Since splitting on Z brings higher Gini gain, node CX_1 shall be split into CZ_{10} and CZ_{11} as shown in Fig. 2

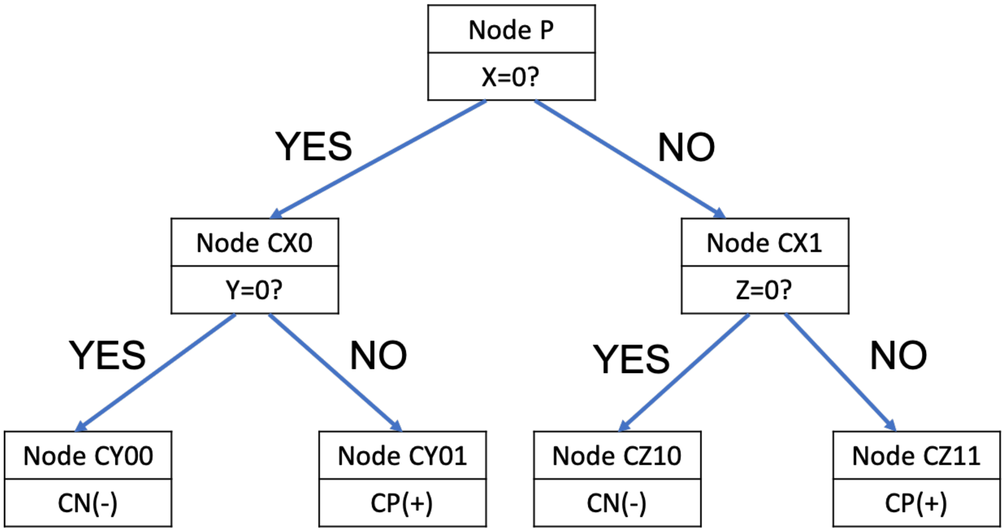


Figure 2: Two layer decision tree for binary classification.

Since all resulting leaf nodes are pure, the **training error rate is 0.**

iii) Discuss the results obtained in the two parts above. Comment on the suitability of the greedy heuristic used as the splitting attribute selection.

Solution : The decision tree obtained from the second part results in a significantly lower training error as compared to the first part, obtained via greedy heuristic. This serves as a demonstration that a greedy heuristic-based splitting approach may provide sub-optimal results. However, the combinatorial nature of optimal-tree search makes it difficult to find the globally best result, particularly for large set of attributes.

3. Properties of Entropy.
i) Show that the entropy measure $-p(x)\log p(x)$ is non-negative.

Solution : A probability $p(x)$ is known to exist in $[0, 1]$. Therefore, we get $\log p(x) \leq 0$. Using the fact $p(x) \geq 0$, we obtain $-p(x)\log p(x) \geq 0$.

ii) Consider the given data set and answer the following questions.
a.) What is the entropy of Y in bits, $H(Y)$?

Solution : The entropy of Y may be computed as

$$\begin{aligned} H(Y) &= -\sum_y p(Y=y)\log p(Y=y) \\ &= -p(Y=0)\log p(Y=0) - p(Y=1)\log p(Y=1) \\ &= -(4/7)\log(4/7) - (3/7)\log(3/7) = 0.9852 \text{ bits} \end{aligned}$$

b.) What is the information gain of Y w.r.t. A in bits, $I(Y;A)$?

Solution : The information gain of Y w.r.t A may be computed as

$$\begin{aligned} I(Y;A) &= H(Y) - \frac{n(A=0)}{n}H(Y|A=0) - \frac{n(A=1)}{n}H(Y|A=1) \\ &= H(Y) - 0 - \frac{4}{7}(-\frac{1}{4}\log\frac{1}{4} - \frac{3}{4}\log\frac{3}{4}) \\ &= 0.9852 - 0.4636 = 0.5216 \text{ bits} \end{aligned}$$

c.) What is the information gain of Y w.r.t. B in bits, $I(Y;B)$?

Solution : The information gain of Y w.r.t B may be computed as

$$\begin{aligned} I(Y;B) &= H(Y) - \frac{n(B=0)}{n}H(Y|B=0) - \frac{n(B=1)}{n}H(Y|B=1) \\ &= H(Y) - \frac{3}{7}(-\frac{1}{3}\log\frac{1}{3} - \frac{2}{3}\log\frac{2}{3}) - \frac{4}{7}(-\frac{1}{2}\log\frac{1}{2} - \frac{1}{2}\log\frac{1}{2}) \\ &= 0.9852 - 0.3936 - 0.5714 = 0.0202 \text{ bits} \end{aligned}$$

d.) What is the information gain of Y w.r.t. C in bits, $I(Y;C)$?

Solution : The information gain of Y w.r.t B may be computed as

$$\begin{aligned} I(Y;C) &= H(Y) - \frac{n(C=0)}{n}H(Y|C=0) - \frac{n(C=1)}{n}H(Y|C=1) - \frac{n(C=2)}{n}H(Y|C=2) \\ &= H(Y) - \frac{3}{7}(-\frac{1}{3}\log\frac{1}{3} - \frac{2}{3}\log\frac{2}{3}) - \frac{1}{7} * (0) - \frac{3}{7}(-\frac{1}{3}\log\frac{1}{3} - \frac{2}{3}\log\frac{2}{3}) \\ &= 0.9852 - 0.39355 - 0.39355 = 0.1981 \text{ bits} \end{aligned}$$

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

Solution : To get a perfect classification in this case, all the three attributes would be needed. Therefore, the depth of the tree would be three. To verify this, note that the information gains computed above motivate us to choose A as the first splitting attribute. Next, for $A = 1$ neither b nor C result in a pure node. Thus, both B and C attributes would be needed one after the another; hence resulting in a 3-level tree.

f.) Which attribute (A, B, orC) would a decision tree algorithm pick first to branch on, if its splitting criterion is mutual information?

Solution : Considering mutual information as the splitting criteria, attribute A would be picked first to branch as the mutual information or information gain for A is the highest.

g.) Consider the dataset given above. Which is the second attribute you would pick to branch on, if its splitting criterion is information gain?

Solution : Given the first splitting attribute is A from part *f*, we get two children, namely $A = 0$ and $A = 1$. The node $A = 0$ is pure and requires no further splitting. Thus we need to split $A = 1$. Following from part *b*, we know that the entropy

$$H(Y|A=1) = -\frac{1}{4}\log\frac{1}{4} - \frac{3}{4}\log\frac{3}{4} = 0.8113 \text{ bits}$$

. If we next split on B, the information gain is

$$\begin{aligned} I(Y;B|A=1) &= H(Y|A=1) - \frac{1}{4}H(Y|B=0,A=1) - \frac{3}{4}H(Y|B=A=1) \\ &= H(Y|A=1) - \frac{1}{4} * 0 - \frac{3}{4}(-\frac{1}{3}\log\frac{1}{3} - \frac{2}{3}\log\frac{2}{3}) \\ &= 0.8113 - 0.6887 = 0.1226 \text{ bits} \end{aligned}$$

If we rather split on C, the information gain is

$$\begin{aligned} I(Y;C|A=1) &= H(Y|A=1) - \frac{1}{2}H(Y|C=0,A=1) - \frac{1}{2}H(Y|C=2,A=1) \\ &= H(Y|A=1) - \frac{1}{2} * (-\frac{1}{2}\log\frac{1}{2} - \frac{1}{2}\log\frac{1}{2}) - \frac{1}{2} * 0 \\ &= 0.8113 - 0.5 = 0.3113 \text{ bits} \end{aligned}$$

Thus, attribute C provides a larger information gain and hence the second splitting criterion is attribute C.

h.) Draw your completed Decision Tree.

Solution : The complete tree is shown in Fig. 3

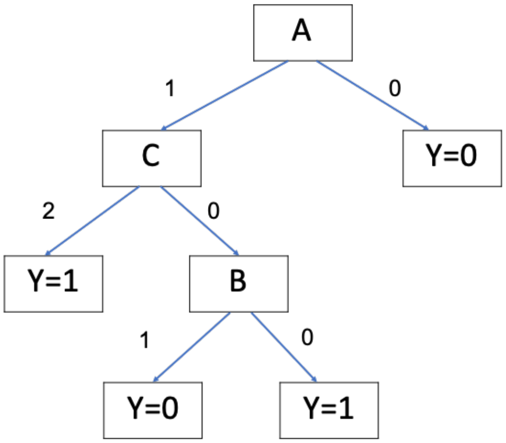


Figure 3: Decision tree for prob 3 .

4. 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 dataID (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.

i) Compute the information gain if we use ID as the splitting attribute.

Solution : Using ID as the splitting attribute will give us 20 leaf nodes which are all pure with zero entropy. Hence the information gain is same as the original entropy of the dataset with 10 positive and 10 negative data. Thus the **information gain is 1 bits**.

ii) Repeat part *(i)* using Handedness as the splitting attribute.

Solution : From part *(i)*, we have $H(Y) = 1$ bits. Let us represent left handedness as $A = 0$ and right handedness as $A = 1$. Thus information gain for splitting attribute

A is

$$\begin{aligned} I(Y;A) &= H(Y) - \frac{n(A=0)}{n} H(Y|A=0) - \frac{n(A=1)}{n} H(Y|A=1) \\ &= H(Y) - \frac{1}{2}(-\frac{1}{10}\log\frac{1}{10} - \frac{9}{10}\log\frac{9}{10}) - \frac{1}{2}(-\frac{1}{10}\log\frac{1}{10} - \frac{9}{10}\log\frac{9}{10}) \\ &= 1 - 0.2345 - 0.2345 = 0.5310 \text{ bits} \end{aligned}$$

iii) Based on your answers in parts *(i)* and *(ii)*, which attribute will be chosen according to information gain?

Solution : Since splitting on ID gives a larger information gain than splitting on handedness, if information gain is used as the criterion, the ID attribute will be chosen for splitting.

iv) Repeat part *(i)* using gain ratio (instead of information gain).

Solution : Since we have already calculated the information gain, we next compute the $splitINFO$ to find the gain ratio for ID as attribute

$$\begin{aligned} splitINFO(ID) &= -\sum_{i=1}^{20} \frac{1}{20}\log\frac{1}{20} \\ &= 4.3219 \end{aligned}$$

Thus the gain ratio for ID is $IG/splitINFO = 1/4.3219 = 0.2314$.

v) Repeat part *(ii)* using gain ratio (instead of information gain).

Solution : To compute the gain ratio, we first compute the $splitINFO$ for handedness as

$$\begin{aligned} splitINFO(Handedness) &= -\sum_{i=1}^{2} \frac{1}{2}\log\frac{1}{2} \\ &= 1 \end{aligned}$$

Thus the gain ratio is $IG/splitINFO = 0.5310/1 = 0.5310$.

vi) Based on your answers in parts *(iv)* and *(v)*, which attribute will be chosen according to gain ratio?

Solution : Since the gain ratio for handedness is better than the ID, handedness will be chosen as a splitting attribute.

5. Model evaluation and statistics

5-1 Consider the given dataset for disease detection.

a.) Draw the corresponding ROC curves for both classifiers on the same plot.

Solution : The ROC curves for the two classifiers are shown in Fig. 4

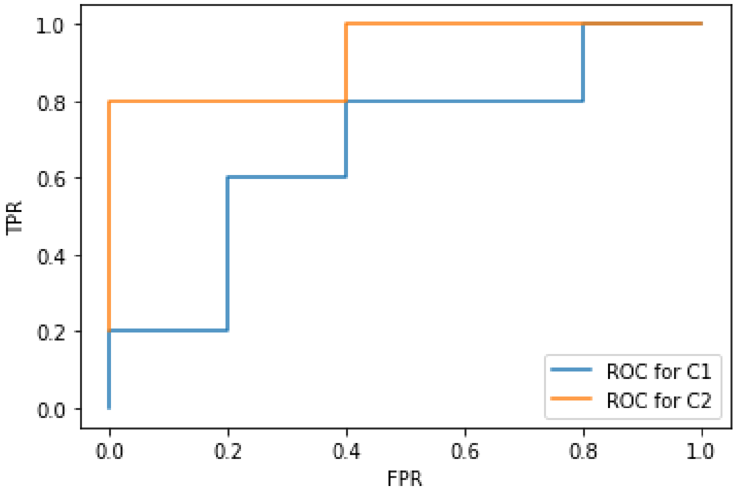


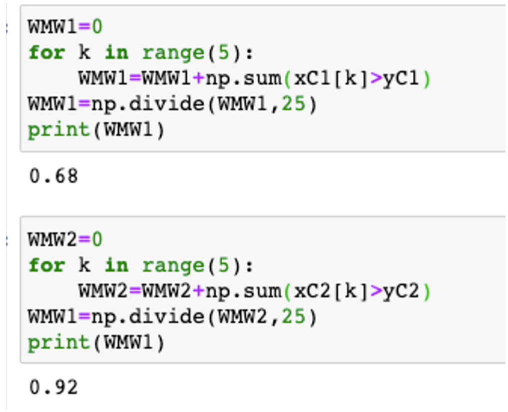
Figure 4: Decision tree for prob 3 .

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

Solution : The area under the ROC curve for classifier $C1$ is 0.68, while that for $C2$ is 0.92 units. Thus $C2$ has a larger area under ROC curve.

c.) Compute the Wilcoxon Mann Whitney statistic for both classifiers. Which classifier has a larger WMW value? Based on your answers, state the relationship between WMW and the ROC curve.

Solution : The provided formula for WMW was implemented via the following code



```
WMW1=0
for k in range(5):
    WMW1=WMW1+np.sum(xC1[k]>yC1)
WMW1=np.divide(WMW1,25)
print(WMW1)
```

```
0.68
```

```
WMW2=0
for k in range(5):
    WMW2=WMW2+np.sum(xC2[k]>yC2)
WMW1=np.divide(WMW2,25)
print(WMW1)
```

```
0.92
```

Again, Classifier $C2$ has a larger WMW value. In fact, here we observe that the WMW values are same as the area under the ROC curve.

5-2 Consider the given confusion matrix for a 1000 points dataset. Calculate the following metrics.

Solution : From the given dataset we have

$$\begin{aligned}
\text{True Positive, } TP &= 345\\
\text{False Positive, } FP &= 195\\
\text{True Negative, } TN &= 235\\
\text{False Negative, } FN &= 225
\end{aligned}$$

We next compute the required metrics:

$$\text{a. Precision} = \frac{TP}{TP+FP} = \frac{345}{345+195} = 0.6389$$

$$\text{b. Recall} = \frac{TP}{TP+FN} = \frac{345}{345+225} = 0.6053$$

$$\text{c. False Pos. Rate} = \frac{FP}{TN+FP} = \frac{195}{195+235} = 0.4535$$

$$\text{d. F-score} = \frac{2TP}{2TP+FP+FN} = \frac{2*345}{2*345+195+225} = 0.6216$$

$$\text{e. Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} = \frac{345+235}{345+195+235+225} = 0.58$$