CS 5990 (Advanced Data Mining) - Assignment #2
Maximum Points: 100 pts.

Bronco ID: 017314921
Last Name: Panda
First Name: Subham

Answers:

1) Binary Classification Problem:

a) The entropy of the collection of training examples with respect to the class attribute

$$Entropy = -\sum_{i=0}^{c-1} p_i(t) log_2 p_i(t)$$

$$p(+) = \frac{4}{9} = 0.44$$

$$p(-) = \frac{5}{9} = 0.55$$

$$\therefore\ E = -0.44\, log_2\, 0.44 - 0.55\, log_2\, 0.55$$

$$= -0.44(-1.1844) - 0.55(-0.8625)$$

$$= 0.52 + 0.47$$

$$= 0.9911$$

b) The information gains of $a1$ and $a2$ relative to the training examples is
Information Gain = P-M

Information Gain:

$$Gain_{split} = Entropy(p) - \sum_{i=1}^{k} \frac{n_i}{n} Entropy(i)$$

Parent node, p is split into k partitions (children)
n_i is number of records in child node i

$$Entropy(p) = 0.9911$$

$$E(a_1) = \frac{4}{9}\left[-\frac{3}{4}log_2\left(\frac{3}{4}\right) - \frac{1}{4}log_2\left(\frac{1}{4}\right)\right] + \frac{5}{9}\left[-\frac{1}{5}log_2\left(\frac{1}{5}\right) - \frac{4}{5}log_2\left(\frac{4}{5}\right)\right]$$
$$= 0.7616$$

$$\therefore\ Information\ Gain(a_1) = Entropy(p) - E(a_1)$$
$$= 0.9911 - 0.7616 = 0.2294$$

$$E(a_2) = \frac{4}{9}\left[-\frac{2}{4}log_2\left(\frac{2}{4}\right) - \frac{2}{4}log_2\left(\frac{2}{4}\right)\right] + \frac{5}{9}\left[-\frac{2}{5}log_2\left(\frac{2}{5}\right) - \frac{3}{5}log_2\left(\frac{3}{5}\right)\right]$$
$$= 0.9839$$

$$\therefore\ Information\ Gain(a_2) = Entropy(p) - E(a_2)$$
$$= 0.9911 - 0.9839 = 0.0072$$

c) For a_3, the information gain for every possible split using the efficient computation technique which includes sorting, splitting, filling class distribution

$For\ a_3$	$+$	$-$	$+$	$-\ -$	$+$	$-\ +$	$-$
$Sorted\ Values$	1.0	3.0	4.0	5.0 5.0	6.0	7.0 7.0	8.0
$Split\ Values$	0.5	2.0	3.5	4.5	5.5	6.5	7.5
$+$		1 3	1 3	2 2	2 2	3 1	4 0
$-$		0 5	1 4	1 4	3 2	3 2	4 1
$Entropy(a_3)$		0.8484	0.9885	0.9183	0.9838	0.9728	0.8889
$Information\ Gain(a_3)$		0.1427	0.0026	0.0728	0.0072	0.0183	0.1022

$\therefore$ The best split for a_3 occurs at split value 2.0 with an information gain of 0.1427

d) According to the information gain a_1 produces the best split.

e)

$$Gain\ Ratio = \frac{Gain_{split}}{Split\ Info} \qquad Split\ Info = -\sum_{i=1}^{k} \frac{n_i}{n} log_2 \frac{n_i}{n}$$

$SI(a_1) = -\frac{4}{9} log_2 \frac{4}{9} - \frac{5}{9} log_2 \frac{5}{9}$

$= -0.4444(-1.17) - 0.5555(-0.848)$

$= 0.991$

$SI(a_2) = 0.991$

$SI(a_3) = -\frac{1}{9} log_2 \frac{1}{9} - \frac{8}{9} log_2 \frac{8}{9}$

$= -0.1111(-3.17) - 0.8888(-0.16993)$

0.503

$GR(a_1) = \frac{0.2294}{0.991} = 0.231$

$GR(a_2) = \frac{0.0072}{0.991} = 0.0072$

$GR(a_3) = \frac{0.1427}{0.991} = 0.2836$

$\therefore a_3\ has\ best\ split\ according\ to\ gain\ ratio\ as\ per\ information\ gain.$

f)

$$Gini\ Index = 1 - \sum_{i=0}^{c-1} p_i(t)^2$$

$$GI(a_1) = \frac{4}{9}\left[1 - \left(\frac{3}{4}\right)^2 - \left(\frac{1}{4}\right)^2\right] + \frac{5}{9}\ [1 - \left(\frac{1}{5}\right)^2 - \left(\frac{4}{5}\right)^2 = 0.3444$$

$$GI(a_2) = \frac{5}{9}\left[1 - \left(\frac{2}{5}\right)^2 - \left(\frac{3}{5}\right)^2\right] + \frac{4}{9}\ [1 - \left(\frac{2}{4}\right)^2 - \left(\frac{2}{4}\right)^2 = 0.4889$$

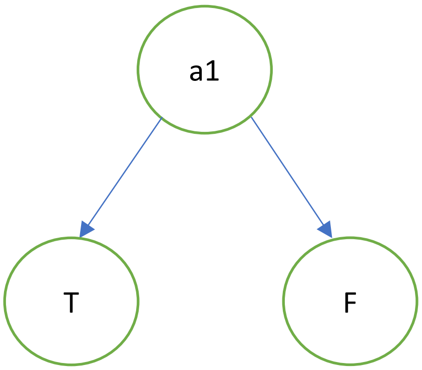
$$\therefore\ Gini\ Index\ of\ a_1\ is\ smaller\ \&\ it\ produces\ better\ split.$$

g)

Classification Error Rate

$$Classification\ error = 1 - \max[p_i(t)]$$

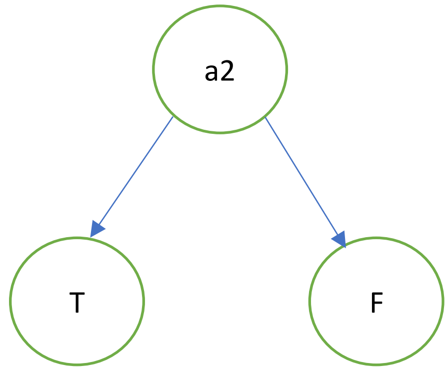
But using shortcut method:



+	-		+	-
3	1		1	4

Minority = 1+1 =2

$E(t)a_1 = \frac{2}{9} = 0.222$



+	-		+	-
2	3		2	2

Minority = 1+1 =2

$E(t)a_2 = \frac{4}{9} = 0.444$

$$Since\ E(t)\ of\ a_1\ is\ less\ than\ a_2$$

$\therefore a_1 will\ provide\ best\ split\ according\ to\ the\ classification\ error\ rate.$

2) Two-level decision tree:

a)

X	C_1	C_2
0	60	60
1	40	40

$$Class_{err}(X) = \frac{40 + 40}{200} = 0.4$$

Y	C_1	C_2
0	40	60
1	60	40

$$Class_{err}(Y) = \frac{40 + 40}{200} = 0.4$$

Z	C_1	C_2
0	30	70
1	70	30

$$Class_{err}(Z) = \frac{30 + 30}{200} = 0.3$$

Z is the lowest in terms of classification error and is chosen as the splitting attribute at level 1.

Now, **Z** = 0

X	C_1	C_2
0	15	45
1	25	25

Y	C_1	C_2
0	15	45
1	15	25

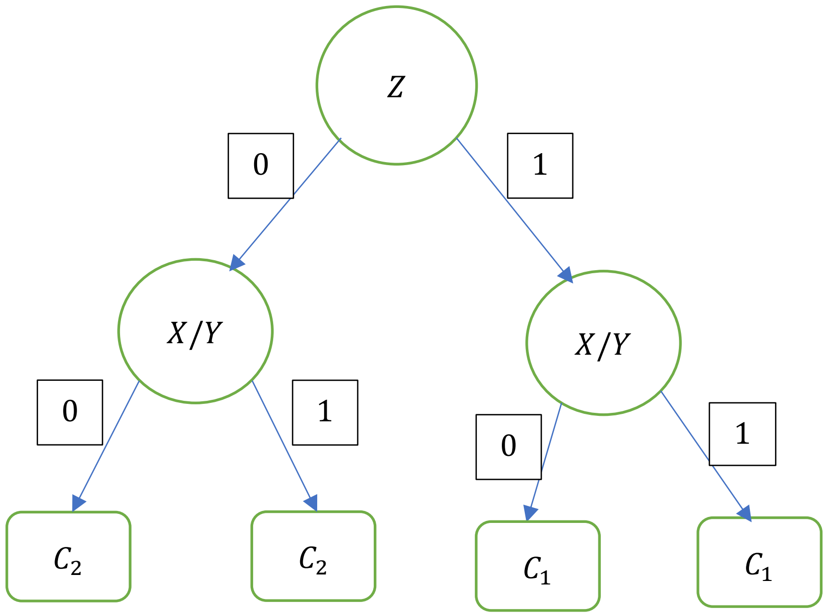
$$Class_{err}Z_0(X, Y) = \frac{15 + 15}{100} = 0.3$$

Now, **Z** = 1

X	C_1	C_2
0	45	15
1	25	15

Y	C_1	C_2
0	25	15
1	45	15

$$Class_{err}Z_1(X, Y) = \frac{15 + 15}{100} = 0.3$$



Z
0
1
X/Y
X/Y
0
1
0
1
C2
C2
C1
C1

b) Overall error rate of the induced tree is $\frac{(15+15+15+15)}{200} = 0.3$

3) Model decision tree:

Ground Truth	$\boldsymbol{B+}$		$\boldsymbol{B-}$		$\boldsymbol{C+}$		$\boldsymbol{C-}$	
Predicted	$+$	$-$	$+$	$-$	$+$	$-$	$+$	$-$
Training - ID	$\mathbf{1, 2}$	$\boldsymbol{NA}$	$\mathbf{3}$	$\mathbf{4}$	$\mathbf{5, 6}$	$\mathbf{7, 9, 10}$	$\mathbf{8}$	$\boldsymbol{NA}$
Validation - ID	$\mathbf{11}$	$\boldsymbol{NA}$	$\mathbf{12}$	$\boldsymbol{NA}$	$\mathbf{13, 15}$	$\boldsymbol{NA}$	$\boldsymbol{NA}$	$\mathbf{14}$
Test - ID	$\mathbf{23}$	$\mathbf{21}$	$\mathbf{16}$	$\boldsymbol{NA}$	$\mathbf{17, 22}$	$\boldsymbol{NA}$	$\mathbf{19}$	$\mathbf{18, 20}$

a) The generalization error rate of the tree using the optimistic approach

$$err_{tr} = \frac{5}{10} = 0.5$$

b)

Ground Truth	$B+$		$B-$		$C+$		$C-$	
Predicted	$+$	$-$	$+$	$-$	$+$	$-$	$+$	$-$
Training - ID	$\mathbf{1,2}$	$\boldsymbol{NA}$	$\mathbf{3}$	$\mathbf{4}$	$\mathbf{5,6}$	$\mathbf{7,9,10}$	$\mathbf{8}$	$\boldsymbol{NA}$
Validation - ID	$\mathbf{11}$	$\boldsymbol{NA}$	$\mathbf{12}$	$\boldsymbol{NA}$	$\mathbf{13,15}$	$\boldsymbol{NA}$	$\boldsymbol{NA}$	$\mathbf{14}$
Test - ID	$\mathbf{23}$	$\mathbf{21}$	$\mathbf{16}$	$\boldsymbol{NA}$	$\mathbf{17,22}$	$\boldsymbol{NA}$	$\mathbf{19}$	$\mathbf{18,20}$

The generalization error rate of the tree using the pessimistic approach with factor of 0.5

$$err_{gen}(T) = err(T) + \Omega \times \frac{k}{N_{train}}$$

$$err_{gen}(T) = \frac{5}{10} + 0.5\left(\frac{4}{10}\right) = 0.7$$

c)

Ground Truth	$B+$		$B-$		$C+$		$C-$	
Predicted	$+$	$-$	$+$	$-$	$+$	$-$	$+$	$-$
Training - ID	$\mathbf{1,2}$	NA	$\mathbf{3}$	$\mathbf{4}$	$\mathbf{5,6}$	$\mathbf{7,9,10}$	$\mathbf{8}$	NA
Validation - ID	$\mathbf{11}$	NA	$\mathbf{12}$	NA	$\mathbf{13,15}$	NA	NA	$\mathbf{14}$
Test - ID	$\mathbf{23}$	$\mathbf{21}$	$\mathbf{16}$	NA	$\mathbf{17,22}$	NA	$\mathbf{19}$	$\mathbf{18,20}$

The generalization error rate of the tree using the validation set

$$err_{gen}(V) = \frac{1}{5} = 0.2$$

d) Accuracy of the model on the test set

	ID	$Total$
$True\ Positive\ (TP)$	17,22,23	3
$True\ Negative\ (TN)$	18,20	2
$False\ Positive\ (FP)$	16,19	2
$False\ Negative\ (FN)$	21	1

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} = \frac{3 + 2}{3 + 2 + 2 + 1} = 0.625$$

e)

	ID	$Total$
$True\ Positive\ (TP)$	17,22,23	3
$True\ Negative\ (TN)$	18,20	2
$False\ Positive\ (FP)$	16,19	2
$False\ Negative\ (FN)$	21	1

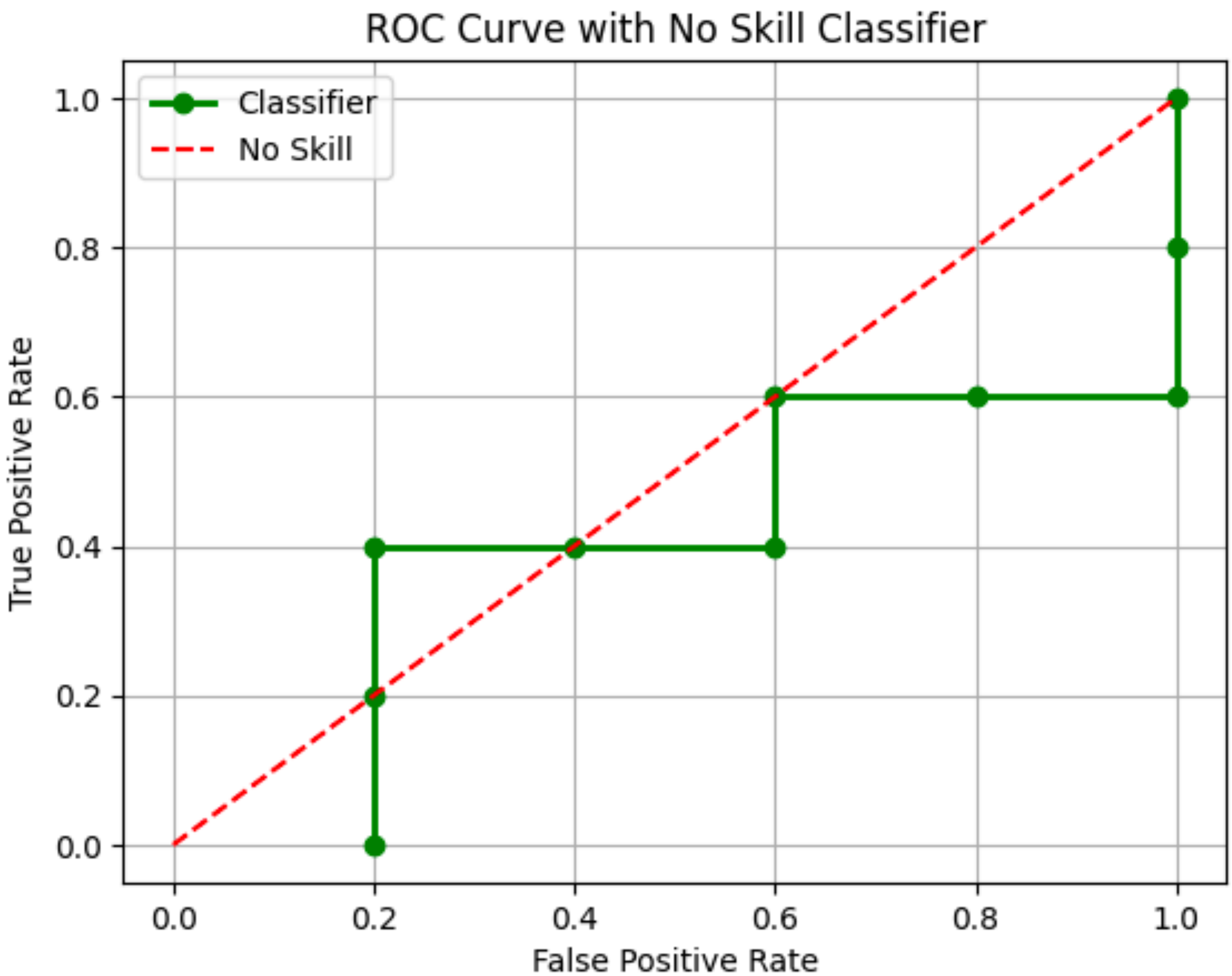
$$Precision\ (P) = \frac{TP}{TP + FN} = \frac{3}{3+2} = 0.6$$

$$Recall\ (R) = \frac{TP}{TP + FN} = \frac{3}{3+1} = \frac{3}{4} = 0.75$$

$$F1\ Score = 2 * \frac{PR}{P+R} = \frac{2(0.6)(0.75)}{0.6 + 0.75} = \frac{0.9}{1.35} = 0.667$$

4) ROC Curve:

Class	+	+	−	−	+	−	−	+	+	−
Threshold >=	**0.01**	**0.03**	**0.04**	**0.05**	**0.09**	**0.31**	**0.38**	**0.45**	**0.61**	**0.68**
TP	5	4	3	3	3	2	2	2	1	0
FP	5	5	5	4	3	3	2	1	1	1
TN	0	0	0	1	2	2	3	4	4	4
FN	0	1	2	2	2	3	3	3	4	5
TPR	1.0	0.8	0.6	0.6	0.6	0.4	0.4	0.4	0.2	0.0
FPR	1.0	1.0	1.0	0.8	0.6	0.6	0.4	0.2	0.2	0.2



5) Python program (decision_tree.py)

6) **Python program (roc_curve.py)**