CS612 – Lab6

1) The following table summarizes a data set with three attributes *A*, *B*, *C* and two class labels +, −. Build a two-level decision tree.

A	B	C	Number of Instances of +	Number of Instanes of -
T	T	T	5	0
F	T	T	0	40
T	F	T	15	30
F	F	T	20	0
T	T	F	10	15
F	T	F	25	0
T	F	F	15	20
F	F	F	0	5

a) According to the classification error rate, which attribute would be chosen as the first splitting attribute? For each attribute, show the contingency table and the gains in classification error rate.

Ans. First of all, we'll find the error rate for the original data before splitting

$E_{original} = 1 - \max[50/100 , 50/100]$

$= 1 - 50/100$

Original Error rate = $\mathbf{E_{original} = 50/100}$

Now, we will split on **attr 'A'** for finding gain in error rate,

	T	F
+	25	25
-	0	50

$E_{(A = T)} = 1- \max[25/25 , 0/25]$

$= 1 - 25/25$

Error rate = $\mathbf{E_{(A = T)} = 0}$

$E_{(A = F)} = 1 - \max[25/75 , 50/75]$

$= 1 - 50/75$

Error rate = $\mathbf{E_{(A = F)} = 25/75}$

$\Delta_A = E_{original} - 25/100\ E_{(A = T)} - 75/100\ E_{(A = F)}$

Gain in error rate for attr 'A' = $\Delta_A = 25/100$

Now, we will split on **attr 'B'** for finding gain in error rate,

	T	F
+	30	20
-	20	30

$E_{(B = T)}$ = 1- max[30/50 , 20/50]

= 1 – 30/50

$E_{(B = T)}$ = 20/50

$E_{(B = F)}$ = 1 – max[20/50 , 30/50]

= 1 – 30/50

$E_{(B = F)}$ = 20/50

$\Delta_B = E_{original}$ - 50/100 $E_{(B = T)}$ - 50/100 $E_{(B = F)}$

Gain in error rate for attr 'B' $\Delta_B = 10/100$

Now, we will split on attr 'C' for finding gain in error rate,

	T	F
+	25	25
-	25	25

$E_{(C = T)}$ = 1- max[25/50 , 25/50]

= 1 – 25/50

$E_{(C = T)}$ = 25/50

$E_{(C = F)}$ = 1 – max[25/50 , 25/50]

= 1 – 25/50

$E_{(C = F)}$ = 25/50

$\Delta_C = E_{original}$ - 50/100 $E_{(C = T)}$ - 50/100 $E_{(C = F)}$

Δ_C = 0/100

$\Delta_C = 0$

Hence, From all of the above observations, It will select attr 'A' as first splitting because it has the highest gain in error rate.

b) Repeat for the two children of the root node.

Ans. because the A = T node is pure, we need to split it furthermore.

For A=F child node, the training instance are

		Class label	
		+	-
T	T	0	20
F	T	0	5
T	F	25	0
F	F	0	25

The original classification rate of A=F node is:

$\mathbf{E_{original} = 25/75}$

Now, let's find gain in error rate after splitting on **attr 'B',**

	T	F
+	25	0
-	20	30

$E_{(B = T)} = 1 - \max[25/45 , 20/45]$

$= 1 - 25/45$

$\mathbf{E_{(B = T)} = 20/45}$

$E_{(B = F)} = 1 - \max[0/30 , 30/30]$

$\mathbf{E_{(B = F)} = 0}$

$\Delta_B = E_{original} - 45/75 E_{(B = T)} - 20/75 E_{(B = F)}$

$\mathbf{\Delta_B = 5/75}$

Gain in error rate for attr 'B' is 5/75

Now, let's find gain in error rate after splitting on attr 'C',

	T	F

+	0	25
-	25	25

$E_{(C = T)}$ = 1- max[0/25 , 25/25]

= 1 – 25/25

$E_{(C = T)}$ = 0

$E_{(C = F)}$ = 1 – max[25/50 , 25/50]

= 1 – 25/50

$E_{(C = F)}$ = 25/50

$\Delta_C = E_{original}$ - 25/75 $E_{(C = T)}$ - 50/75 $E_{(C = F)}$

$\Delta_C = 0$

Gain in error rate for attr 'C' is 0

Hence, we will split on attr 'B' from the above observations.

c) Repeat parts (a), (b), and (c) using C as the splitting attribute.

Ans. **For child node C = T**, we'll find original error rate before splitting.

$E_{original}$ = 25/50

Now, after splitting on **attr 'A'** to finding gain in error rate,

	T	F
+	25	0
-	0	25

$E_{(A = T)}$ = 0

$E_{(A = F)}$ = 0

Gain in error rate = Δ_A = 25/100

Now, after splitting on **attr 'B'** to finding gain in error rate,

	T	F
+	5	20
-	20	5

$E_{(B = T)}$ **= 5/25**

$E_{(B = F)}$ **= 5/25**

Δ_B **= 15/50**

Hence, we conclude 'A' attr.should be chosed as splitting attribute based on value.

For the child node C = F, error rate before splitting,

$E_{original}$ **= 25/50**

	T	F
+	0	25
-	0	25

$E_{(A = T)}$ **= 0**

$E_{(A = F)}$ **= 25/50**

Δ_A **= 0**

Gain in error rate of attr 'A' is 0

Now, let's check error rate after splitting on **attr 'B',**

	T	F
+	25	0
-	0	25

$E_{(B = T)}$ **= 0**

$E_{(B = F)}$ **= 0**

Δ_B **= 25/50**

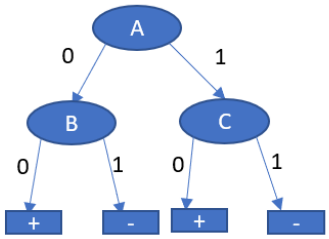
Gain in error rate of attr 'B' is 25/50 = 0.5

As per above obs,For node C=F, **attr 'B' is selected as splitting node. Overall error rate of our tree is zero.**

2) Consider the decision tree shown below:

Training

Instance	A	B	C	Class
1	0	0	0	+
2	0	0	1	+
3	0	1	0	+
4	0	1	1	-
5	1	0	0	-
6	1	0	0	+
7	1	1	0	+
8	1	0	1	+
9	1	1	0	-
10	1	1	0	+

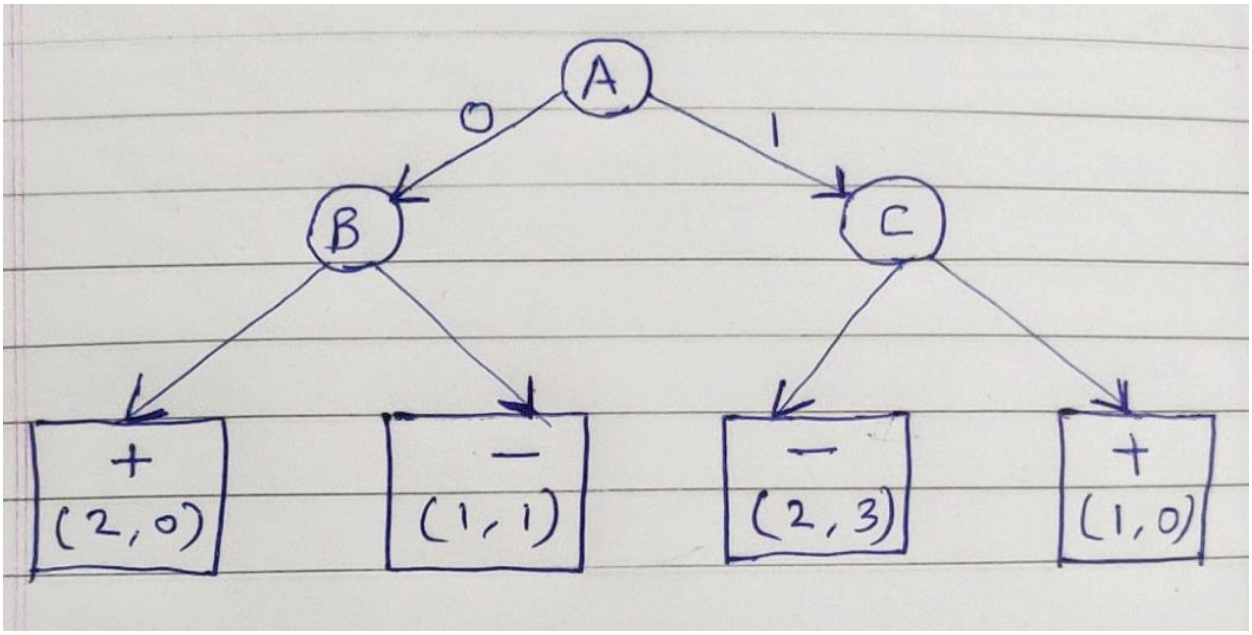


Testing

Instance	A	B	C	Class
11	0	0	0	+
12	0	1	1	-
13	1	1	0	+
14	1	0	1	-
15	1	0	0	+

a) Compute the generalization error rate of the tree using the optimistic approach.

Ans. For all training data, classification is as follows:



Generalization error rate of training data using optimistic approach,

Generalization error rate = 1/10 + 2/10

Generalization error rate = 3/10

Generalization error rate = 0.3

Hence, For training data the generalization error rate using optimistic approach is 0.3

Similarly, For testing data generalization error rate(optimistic approach) is 0/5 = 0

b) Compute the generalization error rate of the tree using the pessimistic approach. (For simplicity, use the strategy of adding a factor of 0.5 to each leaf node.)

Ans 2.b: Total No. of leaf node N = 4

Pessimistic approach to find geralization error rate.

Generalization error rate (Pessimistic approach) = (3 + 0.5*4) / 10

= (3+2) /10

= 5/10

Generalization error rate of tree using (Pessimistic approach) = 0.5

3) Suppose we have the following equation is to find confidence internal.

$$\frac{2 * N * acc + Z_{\alpha/2}^2 \pm Z_{\alpha/2} * \sqrt{(Z_{\alpha/2}^2 + 4*N*acc - 4*N*acc^2}}{2 * (N+Z_{\alpha/2}^2)}$$

1-α	0.99	0.98	0.95	0.9	0.8	0.7	0.5
$Z_{\alpha/2}$	2.58	2.33	1.96	1.65	1.28	1.04	0.67

Where N=1000 is the number of records, acc= 75% is the accuracy, standard normal distribution at confidence level (1-a) is 0.99. Calculate the interval based on the given information.

Ans. Given : N = 1000,

(1-α) = 0.99 .....At Confidence level,

$Z_{\alpha/2}$ = 2.58

Confidence interval

$$= \frac{2 * 1000 * 0.75 \pm (2.58)^2 \pm (2.58) * \sqrt{(2.58)^2 + 4 * 1000 * 0.75 - 4 * 1000 * (0.75)^2}}{2 * (1000 + (0.75)^2)}$$

$$= \frac{1500 \pm 6.65 \pm (2.58) * \sqrt{6.65 + 3000 - 2250}}{2013.3}$$

$$= \frac{1506.65 \pm (2.58)\sqrt{756.65}}{2013.3}$$

$$= \frac{1506.65 \pm (2.58) * 27.50}{2013.3}$$

$$= \frac{1506.6 \pm 70.95}{2013.3}$$

Ans. **Confidence interval = 0.7836 , 0.7131**

4) Suppose model M_A has an error rate of e_1=0.20 when applied to N_1=3000 test records, while model M_B has an error rate of e_2=0.25 when applied to N2=6000 test records. Use the following formula to find the estimated variance of the observed difference in error.

$$\sigma^2_d \cong \hat{\sigma}^2_d = [(e_1 * (1\text{-}e_1)) / n_1] + [(e_2 * (1\text{-}e_2))/n_2]$$

Find If the observed difference in the error "d", the mean dt, its true difference and variance "s2d" with confidence level of 98%. Does the interval span the value zero?

Ans.

Given :- e_1=0.20 , e_2=0.25 , N_1=3000, N_2=6000

s2d = [((0.20) * (1 - 0.20))/3000] + [((0.25) * (1 – 0.25))/6000]

s2d = [0.20 * 0.8 / 3000] + [0.25 * 0.75 / 6000]

s2d = [0.00000533] + [0.00003125]

s2d = 0.00008458

Hence, sd = 0.009196

Now, we'll find mean dt :-

$dt = d \pm Z_{\alpha/2} * sd$

[Confidence interval of dt(98%) = 2.32]

$dt = 0.05 \pm 2.32 * 0.009196$

$dt = 0.05 \pm 0.021$

Ans. Hence, Interval = 0.071 , 0.029