Problem 1 :-

2> a> $1-\left(\frac{10}{20}\right)^2-\left(\frac{10}{20}\right)^2 = 1-(0.5)^2-(0.5)^2 = 0.5$

b> Gini for customer Id

$ID1 = 1-\left(\frac{1}{1}\right)^2 = 0$

$ID2 = 1-\left(\frac{1}{1}\right)^2 = 0$

$\vdots$

$IDn = 1-\left(\frac{1}{1}\right)^2 = 0$

So for customer ID Gini index is zero

c> Gini for Male $= 1-\left(\frac{10}{20}\right)^2-\left(\frac{10}{20}\right)^2 = 0.5$

Gini for female $= 1-\left(\frac{10}{20}\right)^2-\left(\frac{10}{20}\right)^2 = 0.5$

Overall Gini $= \frac{10}{20}\times 0.5 + \frac{10}{20}\times 0.5 = 0.5$

d> Gini for car type

Family type car $= 1-\left(\frac{1}{4}\right)^2-\left(\frac{3}{4}\right)^2 = \frac{3}{8} = 0.375$

Sports car $= 1-\left(\frac{8}{8}\right)^2 = 0$

Luxary car $= 1-\left(\frac{1}{8}\right)^2-\left(\frac{7}{8}\right)^2 = 0.218$

Overall Gini is $\frac{4}{20}(0.375) + 0\times\frac{8}{20} + \frac{8}{20}(0.218) = 0.1625$

e> Ans:

$$= \frac{5}{20}\left(1-\left(\frac{3}{5}\right)^2-\left(\frac{2}{5}\right)^2\right) + \frac{7}{20}\left(1-\left(\frac{3}{7}\right)^2-\left(\frac{4}{7}\right)^2\right) + \frac{4}{20}\left(1-\left(\frac{2}{4}\right)^2-\left(\frac{2}{4}\right)^2\right) + \frac{4}{20}\left(1-\left(\frac{2}{4}\right)^2-\left(\frac{2}{4}\right)^2\right)$$

$$= 0.4914$$

f> If we have a look at c, d, e then Gini for car is lowest so car is a better option to split on.

g> CustomerID is not a kind of data from which we can predict anything. It is like primary key in table in which new ID is assigned to new customers.

3.

a) Entropy for root

Probability(+) = $\frac{4}{9}$ Probability(-) = $\frac{5}{9}$

$$\text{Entropy} = -\left[\frac{4}{9}\log_2\frac{4}{9} + \frac{5}{9}\log_2\frac{5}{9}\right] = 0.9911$$

b)

a_1

a_1	+	−
T	3	1
F	1	4

$$-\frac{4}{9}\left[\frac{3}{4}\log_2\frac{3}{4} + \frac{1}{2}\log_2\frac{1}{4}\right] - \frac{5}{9}\left[\frac{1}{5}\log_2\frac{1}{5} + \frac{4}{5}\log_2\left(\frac{4}{5}\right)\right] = 0.7616$$

Information gain = 0.9911 − 0.7616 = 0.2294

a_2

a_2	+	−
T	2	3
F	2	2

$$-\frac{5}{9}\left[\frac{2}{5}\log_2\frac{2}{5} + \frac{3}{5}\log_2\frac{3}{5}\right] - \frac{4}{9}[0] = 0.9839$$

Information gain = 0.9911 − 0.9839 = 0.0072

c)

a_3	label	Split point	Entropy	Info gain
1.0	+	2.0	0.8484	0.1427
3.0	−	3.5	0.9885	0.0026
4.0	+	4.5	0.9183	0.0728
5.0	−	5.5	0.9839	0.0072
5.0	−			
6.0	+	6.5	0.9728	0.0183
7.0	+	7.5	0.8889	0.1022
7.0	−			

labels	+	−	+	−	−	+	+	−
Sorted Values	1	3	4	5	5	6	7	7

Split points	0.5		2		3.5		4.5		5.5		6.5		7.5	
	≤	>	≤	>	≤	>	≤	>	≤	>	≤	>	≤	>
+	0	4	1	3	1	3	2	2	2	2	3	1	4	0
−	0	5	0	5	1	3	1	3	3	2	3	1	4	0

Entropy in above table ↑

d)

x_1	T	F
+	3	1
−	1	4

$$-\frac{4}{9}\left[\frac{3}{4}\log\frac{3}{4}+\frac{1}{4}\log\frac{1}{4}\right] - \frac{5}{9}\left[\frac{1}{5}\log\frac{1}{5}+\frac{4}{5}\log\frac{4}{5}\right]$$

$$= 0.35 + 0.39$$

$$= 0.74$$

$$I.G = 0.99 - 0.74$$

$$= 0.25$$

x_2	T	F
+	2	2
−	3	2

$$-\frac{4}{9}\left[\frac{2}{4}\log\frac{2}{4}+\frac{2}{4}\log\frac{2}{4}\right] - \frac{5}{9}\left[\frac{3}{5}\log\frac{3}{5}+\frac{2}{5}\log\frac{2}{5}\right]$$

$$= 0.44 + 0.52$$

$$= 0.96$$

$$IG = 0.99 - 0.96$$

$$= 0.03$$

x_3	T	F
+	1	3
−	0	5

$$-\frac{4}{9}\left[\frac{1}{4}\log\frac{1}{4}+\frac{3}{4}\log\frac{3}{4}\right] - \frac{5}{9}\left[\frac{0}{5}\log 0 + \%\log\%\right]$$

$$= 0.83$$

$$IG = 0.99 - 0.83$$

$$= 0.16$$

Information gain on x_1 is greater so its better to split on x_1

e) For $x_1 = 1 - \frac{7}{9} = \frac{2}{9}$

For $x_2 = 1 - \frac{5}{9} = \frac{4}{9}$

So x_1 produces best split.

f) For a_1

$$\text{Gini index} = \frac{4}{9}\left[1-\left(\frac{3}{4}\right)^2-\left(\frac{1}{4}\right)^2\right] + \frac{5}{9}\left[1-\left(\frac{1}{5}\right)^2-\left(\frac{4}{5}\right)^2\right] = 0.344$$

For a_2

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

5.

a) Root Entropy is $-\left(\frac{4}{10}\log_2\frac{4}{10} + \frac{6}{10}\log\frac{6}{10}\right) = 0.9710$

A	+	−
T	4	3
F	0	3

B	+	−
T	3	1
F	1	5

Entropy $= \frac{7}{10}\left(-\frac{4}{7}\log\frac{4}{7} - \frac{3}{7}\log\frac{3}{7}\right)$

$- \frac{3}{10}\left(\frac{3}{3}\log 1\right)$

$= 0.29\ 0.68$

$= \frac{4}{10}\left(\frac{3}{4}\log\frac{4}{3} + \frac{1}{4}\log 4\right) + \frac{6}{10}\left(\frac{1}{6}\log 6 + \frac{5}{6}\log 1.2\right)$

$= 0.70$

Information gain $= 0.9710 - 0.68$ $= 0.9710 - 0.70 = 0.27$

$= 0.29$

looking at information gain attribute A will be chosen.

b) Gini index at node $= 1 - (0.4)^2 - (0.6)^2 = 0.48$

Gini index on splitting A is

$G_{A=T} = 1 - \left(\frac{4}{7}\right)^2 - \left(\frac{3}{7}\right)^2 = 0.4898$

$G_{A=F} = 1 - \left(\frac{3}{3}\right)^2 - \left(\frac{0}{3}\right)^2 = 0$

Entropy $= \frac{7}{10} \times 0.4898 - \frac{3}{10} \times 0 = 0.1371$ —

Gini on splitting on B is

$G_{B=T} = 1 - \left(\frac{1}{4}\right)^2 - \left(\frac{3}{4}\right)^2 = 0.3750$

$G_{B=F} = 1 - \left(\frac{1}{6}\right)^2 - \left(\frac{5}{6}\right)^2 = 0.2778$

Entropy $= \frac{4}{10}(0.3750) - \frac{6}{10}(0.2778)$

Overall Entropy $= 0.48 - \frac{4}{10}(0.3750) - \frac{6}{10}(0.27) = 0.1653$

So Attribute B is chosen to split the node

c) Yes. They have similar range & all are increasing in a same fashion that is monotonous. Their respective entropies and gains are scaled differences or measures, do not necessarily behave in same way as illustrated in above questions that is a & b.

1-2)

a) Precondition: There are equal number of data in each class

so +ve = 50%.

-ve = 50%.

Decision tree predicts every test record to be positive so half of the records are misclassified by tree so error rate = 50%.

= 0.5

b) positive record prediction probability = 0.8

Negative record prediction probability = 0.2

so if we consider 100 records then it is predicting 80 records as positive and 20 as negative but actually there are 50 positive and 50 negative records.

so 20 records which are negative are identified as positive & 30 records which are negative are identified as positive

$$\text{Error rate} = \frac{20}{100} + \frac{30}{100} = 0.2 + 0.3 = 0.5 \text{ or } 50\%.$$

c) positive data = $\frac{2}{3}$ Negative data = $\frac{1}{3}$

Consider size = 60 so Positive = 40 & negative = 20

But it is identifying every instance as positive so it is misidentifying 20 records

$$\text{Error rate} = \frac{20}{60} = \frac{1}{3} = 0.3333 \text{ or } 33.33\%.$$

d) Probability for positive class = $\frac{2}{3}$ & for negative class = $\frac{1}{3}$

so $2N/3$ will be classified wrongly with probability of $1/3$ & $N/3$ will be classified wrongly with probability of $2/3$

$$\text{Error rate} = \left(\frac{2N}{3} \times \frac{1}{3}\right) + \left(\frac{N}{3} \times \frac{2}{3}\right) / N$$

$$= \left(\frac{2N}{9} + \frac{2N}{9}\right) / N$$

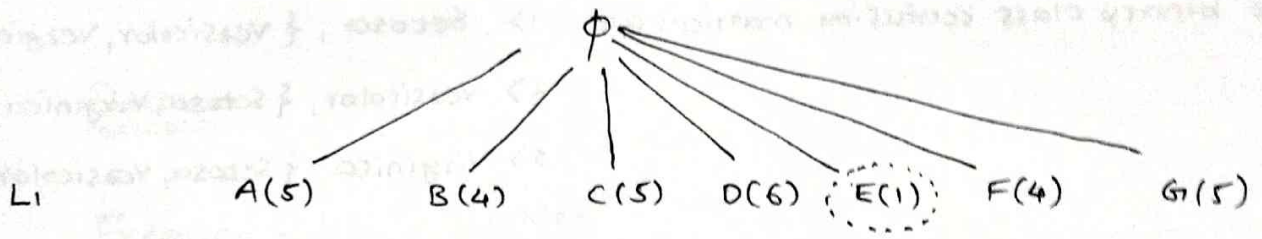
$$= \frac{4}{9} = 0.4444$$

$$= 44.44\%.$$

Q 1.3 Zaki, Chapter 8

tid	itemset
t_1	ABCD
t_2	ACDF
t_3	ACDEG
t_4	ABDF
t_5	BCG
t_6	DFG
t_7	ABG
t_8	CDFG

Value of minsup is given = 3/8



L2

AB	AC	AD	~~AE~~	AF	AG	BC	BD	BF	BG	CD	CF	CG	DF	DG	FG
3	3	4		2	2	2	2	1	2	4	2	3	4	3	2

L3

ABC	ABD	ACD	ACG	ADF	ADG
1	2	3	1	2	1

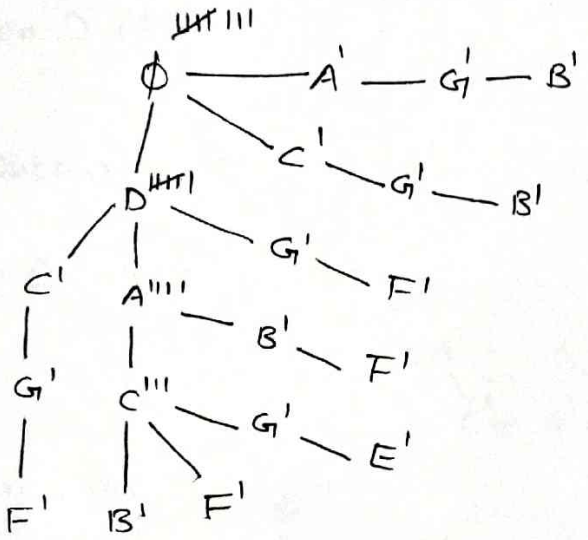
itemset	values
F^1	A, B, C, D, F, G
F^2	AB, AC, AD, CD, CG, DF, DG
F^3	ACD

b) minsup 2/8 FP growth

Frequency D(6) A(5) C(5) G(5) B(4) F(4) E(1)

Original dataset	Converted dataset
ABCD	DACB
ACDF	DACF
ACDEG	DACGE
ABDF	DABF
BCG	CGB
DFG	DGF
ABG	AGB
CDFG	DCGF

Tree →



Projection on A R_A :

A (c=1)

DA (c=4)

Output (R_A) = AD(4)

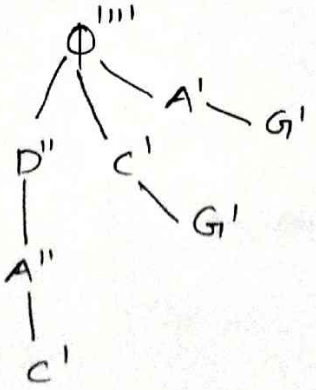
Projection on B R_B

DACB (c=1)

DAB (c=1)

CGB (c=1)

AGB (c=1)



Output → BG(2)

BA(3)

BD(2)

BC(2)

Projection on $(R_B)_A$

DA count = 2

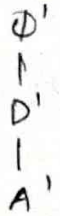
A count = 1

Output = (BAD)(2)

Projection on $(R_B)_C$

DAC (c=1)

C (c=1)

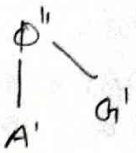


Output = (Φ)

Projection $(R_B)_G$

AG (c=1)

CG (c=1)



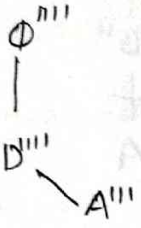
Output = NULL

Projection on C $R_{(C)}$

C (c=1)

DC (c=1)

DAC (c=3)



Output: CD (4)

CA (3)

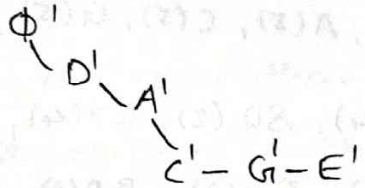
CAD (3)

Projection on D (R_D)

D (c=6)

Output = Φ

Projection on E (R_E):



Output = Φ

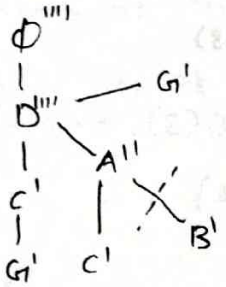
Projection on F (R_F):

DCGF (c=1)

DACF (c=1)

DABF (c=1)

DGF (c=1)



Output: FA(2)

FAD(2)

FD(4)

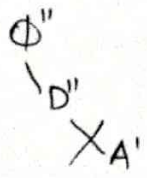
FG(2)

FC(2)

$(R_F)_C$

DC (c=1)

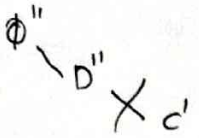
DAC (c=1)



Output: FCD(2)

DCG (c=1)

DG (c=1)



Output = FGD(2)

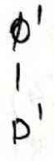
Projection on G R_G

DCG (c=1)
DACG (c=1)
DG (c=1)
CG (c=1)
AG (c=1)

output → GD(3)
GC(3)
GA(2)

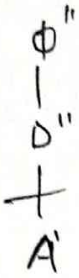
$(R_G)_A$

DA (c=1)
A (c=1)



$(R_G)_C$

CC=1
DAC (c=1)
DC (c=1)



output = GCD(2)

Output = Null

Output

Frequency	Itemset
F^1	D(6), A(5), C(5), G(5), B(4), F(4)
F^2	AD(4), BD(2), CD(4), CA(3), FA(2), GD(3), FD(4), BG(2), BA(3), BC(2), FG(2), FE(2), GA(2), GC(3)
F^3	ABD(2), CAD(3), FAD(2), FCD(2), FGD(2), GCD(2)

4.

tid	itemset
t_1	ACD
t_2	BCE
t_3	ABCE
t_4	BDE
t_5	ABCE
t_6	ABCD

Ans There are three variables present so

possible set for ABE → A, B, E, AB, AE, BE, ABE, ϕ

Rules are

1) $\{A\} \rightarrow \{BE\}$ R.Support = $\frac{S(ABE)}{S(A)} = \frac{2}{4} = 0.5$

2) $\{B\} \rightarrow \{AE\}$ R.support = $\frac{2}{5} = 0.4$

3) $\{E\} \rightarrow \{AB\}$ R.sup = $\frac{2}{4} = 0.5$

4) $\{AB\} \rightarrow \{E\}$ R.sup = $\frac{2}{3} = 0.66$

5) $\{AE\} \rightarrow \{B\}$ R.sup = $\frac{2}{2} = 1$

6) $\{BE\} \rightarrow \{A\}$ R.sup = $\frac{2}{4} = 0.5$

If we consider minsupport value = 0.6 then our possible Rules can be

$\{AB\} \rightarrow \{E\}$

$\{AE\} \rightarrow \{B\}$

1.3

6. a) Item search space in this case is going to be entire item space

so it occupies all the items = size of item taxonomy

= 15

so size of itemspace = 15

b) Consider $X = \{x_1, x_2, \ldots x_n\}$ be a frequent itemset

and we are replacing any element from set with its parent node

→ In the frequent itemset if x_i is replaced by its parent then the frequency of x appearing will increases as reading towards leaf

→ If x_i is such that you need to pass through its parent then

frequency of x is equals to x_{ip}

→ If x_{ip} has more nodes than x_i then will scan x_{ip} while those other childs as well

so frequency of x_{ip} (parent of x) will increase & inturn frequency of itemset x' will increase

so we can say on replacing x with its parent to obtain x',

The support of x' would be more than or equal to support of X

so Answer * (iv) · iv is correct.

Problem 1.4 Multiclass classification :-

Confusion matrix from multiclass.Rmd notebook is

Prediction	Setosa	Versicolor	Virginica
Setosa	8	0	0
Versicolor	0	10	1
Virginica	0	2	9

Data available is Setosa : 8 Versicolor : 12 Virginica : 10

Possible binary class confusion matrices are 1) Setosa , {Versicolor, Verginica}

2) Versicolor, {Setosa, Virginica}

3) Virginica , {Setosa, Versicolor}

1) Setosa , {Versicolor, Verginica} 2) Versicolor, {Setosa, Virginica} 3) Virginica , {Setosa, Versicolor}

Prediction \ Actual	+	−
+	8 TP	0 FP
−	0 FN	22 TN

Prediction \ Actual	+	−
+	10 TP	1 FP
−	2 FN	17 TN

Prediction \ Actual	+	−
+	9 TP	2 FP
−	1 FN	18 TN

$$\text{Accuracy} = \frac{TP+TN}{TP+FP+FN+TN} = \frac{30}{30} = 1 \qquad = \frac{27}{30} = 0.90 \qquad = \frac{27}{30} = 0.90$$

$$\text{Sensitivity} = \frac{TP}{TP+FN} = \frac{8}{8} = 1 \qquad = \frac{10}{12} = 0.833 \qquad = \frac{9}{10} = 0.90$$

$$\text{Specificity} = \frac{TN}{TN+FP} = \frac{22}{22} = 1 \qquad = \frac{17}{18} = 0.944 \qquad = \frac{18}{20} = 0.90$$

$$\text{Precision} = \frac{TP}{TP+FP} = \frac{8}{8} = 1 \qquad = \frac{10}{11} = 0.90 \qquad = \frac{9}{11} = 0.81$$