Name: Soumyadip Roy
Roll. no: 2361019

Subject: Data Mining
Paper code: AIML2101

Assignment (SET-1)

1. i) Mean = $\frac{13+15+16+16+19+20+20+21+22+22+25+25+25+25+30+33+33+35+35+35+36+40+45+46+52+70}{27}$

$= \frac{809}{27} = 29.96 \approx 30$ (Ans).

Median = $\frac{27+1}{2}$ = 14th value = 25 (Ans)

ii) Mode = 25 (unimodal since it has a single mode at 25).

iii) Mid-Range = $\frac{13+70}{2}$ = 41.5 (Ans).

iv) Q1 (first-quartile) = $\frac{13+1}{2}$ = 7th value of lower half (as seen from median).

= 20.5 (Ans).

Q3 (Third-quartile) = $\frac{13+1}{2}$ = 7th value of upper half

= 35.0 (Ans).

v) Five-number summary:

Minimum = 13
Q1 = 20.5
Median = 25.0
Q3 = 35.0
Maximum = 70.

2. i) Create Bins of depth 3 :	Smoothing by mean:
1st bin : 13, 15, 16	14.67
2nd bin : 16, 19, 20	18.33
3rd bin : 20, 21, 22	21.00
4th bin : 22, 25, 25	24.00
5th bin : 25, 25, 30	26.67
6th bin : 33, 33, 35	33.67
7th bin : 35, 35, 35	35.00
8th bin : 35, 36, 40	37.00
9th bin : 45, 46, 52	47.67
10th bin : 70.	70

smoothed data :

14.67, 14.67, 14.67, 18.33, 18.33, 18.33, 21.00, 21.00, 21.00, 24.00, 24.00, 24.00, 26.67, 26.67, 26.67, 33.67, 33.67, 33.67, 35.00, 35.00, 35.00, 37.00, 37.00, 37.00, 47.67, 47.67, 47.67, 70

ii) IQR = Q3 − Q1 .

From 1 (iv) we have :

Q3 = 35 and Q1 = 20.5

∴ IQR = 14.5

Lower bound ~~Q1 − 1.5 X IQR~~ = Q1 − 1.5 X IQR = 20.5 − 1.5 X 14.5
= 20.5 − 21.75 = −1.25

Upper bound = Q3 + 1.5 X IQR = 35 + 1.5 X 14.5 = 35 + 21.75 = 56.75

Outlier from the given data is 70 (∵ 70 > 56.75).

3.

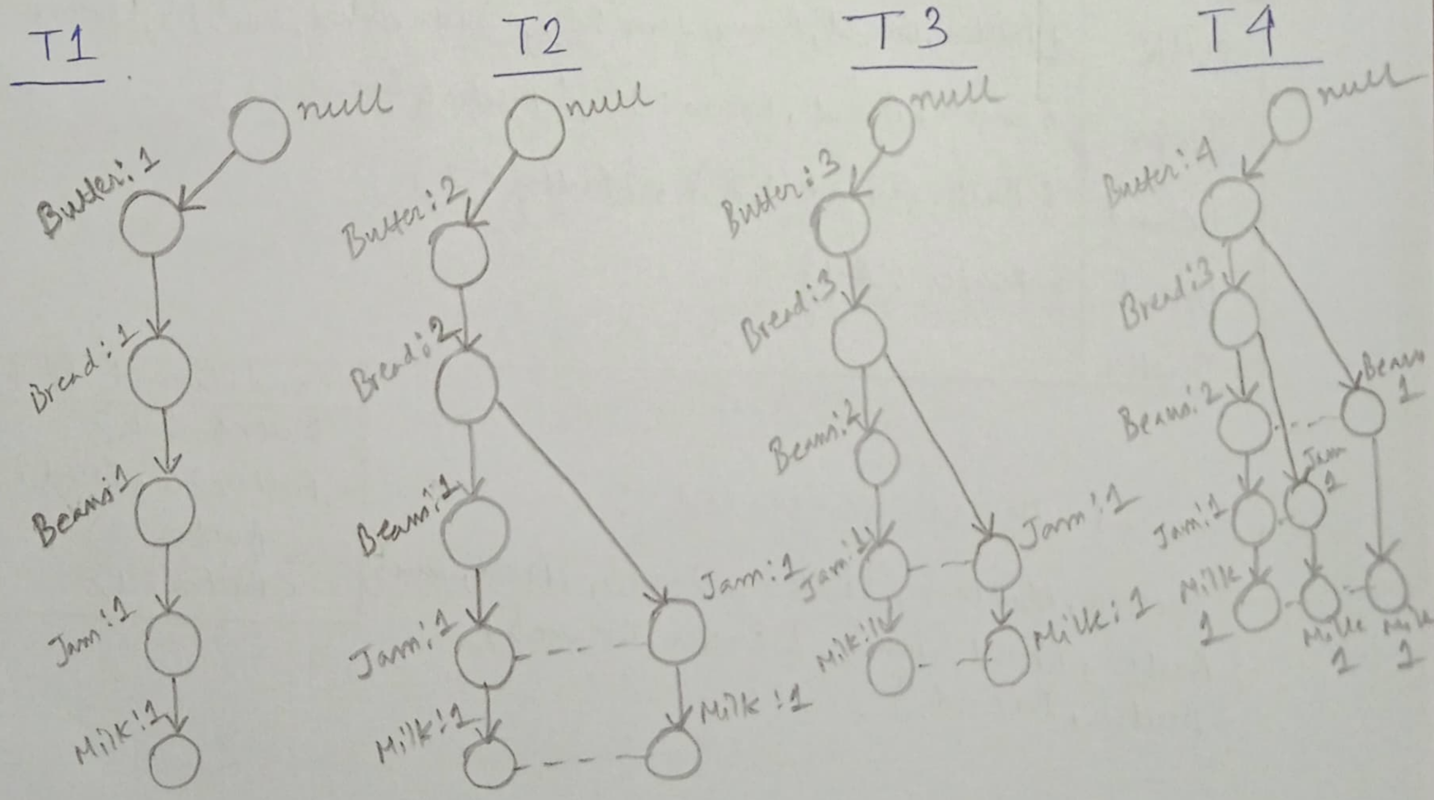
Items	Apple	Beans	Banana	Bread	Butter	Jam	Milk	Onion	Potato	Shamp
support count	2	3	1	4	5	3	3	1	2	1

~~L = {{Beans: 3}, {Bread:~~

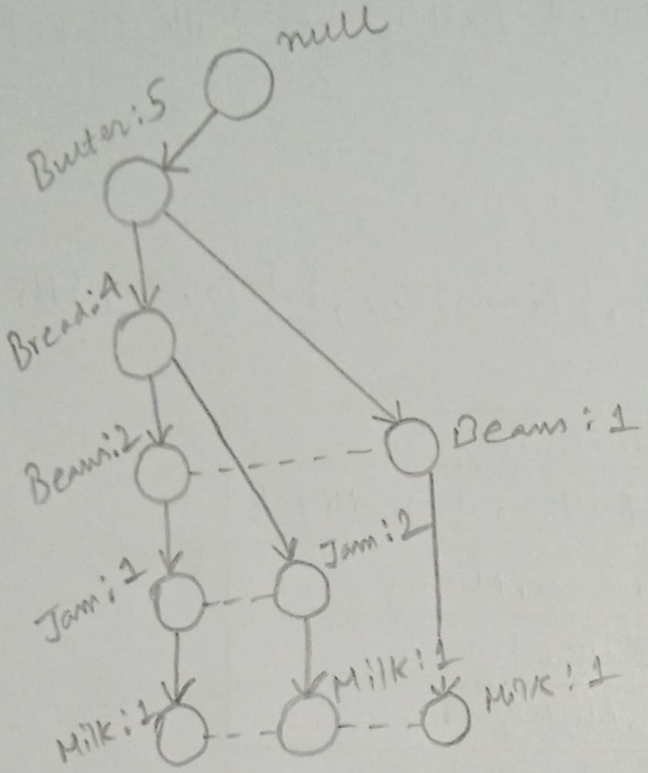
L = {{Butter: 5}, {Bread: 4}, {Beans: 3}, {Jam: 3}, {Milk: 3}}.

TID	Items Purchased
T1	{Beans, Bread, Butter, Jam, Milk}
T2	{Bread, Butter, Jam, Milk}
T3	{Beans, Bread, Butter}
T4	{Beans, Butter, Milk}
T5	{Bread, Butter, Jam}

Transactions:



T5



Items	Conditional patterns Base	Conditional FP-tree
Milk	{Butter, Bread, Beans, Jam : 1}, {Butter, Bread, Jam : 1}, {Butter, Beans:1}	< Butter : 3 >
Jam	{Butter, Bread, Beans : 1}, {Butter, Bread : 2}	<Butter : 3, Bread : 3>
Beans	{Butter, Bread : 2}, {Butter : 1}	< Butter : 3>
Bread	{Butter : ~~3~~4}	< Butter : 4>
Butter		

∴ Frequent items generated :

{Butter, Milk}, {Butter, Jam}, {Bread, Jam}, {Butter, Bread, Jam}, {Butter, Beans}, {Butter, Bread}.

4. $\text{Info}(D) = -\frac{3}{6}\log_2\frac{3}{6} - \frac{3}{6}\log_2\frac{3}{6}$

$= 0.5 + 0.5 = 1$

$\text{Info}_{sex}(D) = \frac{5}{6}\times\left(-\frac{3}{5}\log_2\frac{3}{5} - \frac{2}{5}\log_2\frac{2}{5}\right)$
$+ \frac{1}{6}\times\left(-\frac{1}{1}\log_1\frac{1}{1} - 0\right)$

$= \frac{5}{6}\times\left(0.442 + 0.529\right) = 0.809$

$\text{Gain}(sex) = 1 - 0.809 = 0.191$

$\text{Info}_{mask}(D) = \frac{3}{6}\times\left(-\frac{2}{3}\log_2\frac{2}{3} - \frac{1}{3}\log_2\frac{1}{3}\right)$
$+ \frac{3}{6}\times\left(-\frac{1}{3}\log_2\frac{1}{3} - \frac{2}{3}\log_2\frac{2}{3}\right)$

$= 2\times\frac{3}{6}\times\left(0.390 + 0.528\right)$

$= 0.918$

$\text{Gain}(mask) = 1 - 0.918 = 0.082$

$\text{Info}_{cape}(D) = \frac{4}{6}\times\left(-\frac{1}{4}\log_2\frac{1}{4} - \frac{3}{4}\log_2\frac{3}{4}\right) + 0$

$= \frac{4}{6}\times\left(0.5 + 0.311\right) = 0.540$

$\text{Gain}(cape) = 1 - 0.540 = 0.460$

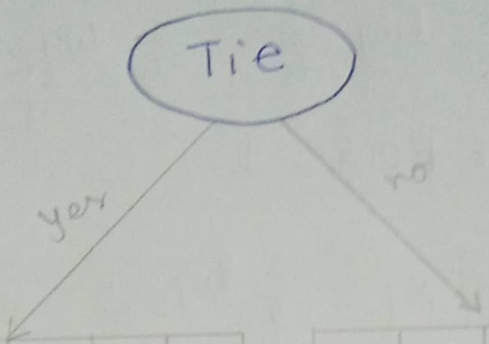
$\text{Info}_{tie}(D) = \text{Info}_{ears}(D) = \frac{2}{6}\times\left(-\frac{1}{2}\log_2\frac{1}{2} - \frac{1}{2}\log_2\frac{1}{2}\right)$
$+ \frac{4}{6}\times\left(-\frac{2}{2}\log_2\frac{2}{2} - \frac{2}{2}\log_2\frac{2}{2}\right)$

$= 0.33$

$\text{Gain}(ties = \text{Gain}(ears) = 1 - 0.33 = 0.67$

$\text{Info}_{smokes}(D) = \cancel{\frac{5}{6}\times\left(\frac{3}{7}\right.}\ 0.809 \quad \therefore \text{Gain}(smokes) = 0.191$

Hence, either <u>tie</u> or <u>ears</u> can be selected as the root node.



sex	mask	cape	ears	smokes	class
male	no	no	no	no	Good
male	no	no	no	yes	Bad

sex	mask	cape	ears	smokes	class
male	Yes	yes	yes	no	Good
male	yes	yes	no	no	Good
female	yes	no	yes	no	Bad
male	no	no	no	no	Bad

$Info(D) = 1$

$Info_{sex}(D) = Info_{mask}(D)$

$= Info_{cape}(D) = Info_{ears}(D) = 1$

$\therefore$ Their gain $= 0$

$\therefore$ sub node is smokes.

$Info(D) = 1$.

$Info_{smokes}(D) = Info_{cape}(D) = 1$

$\therefore Gain(smokes) = Gain(cape) = 0$

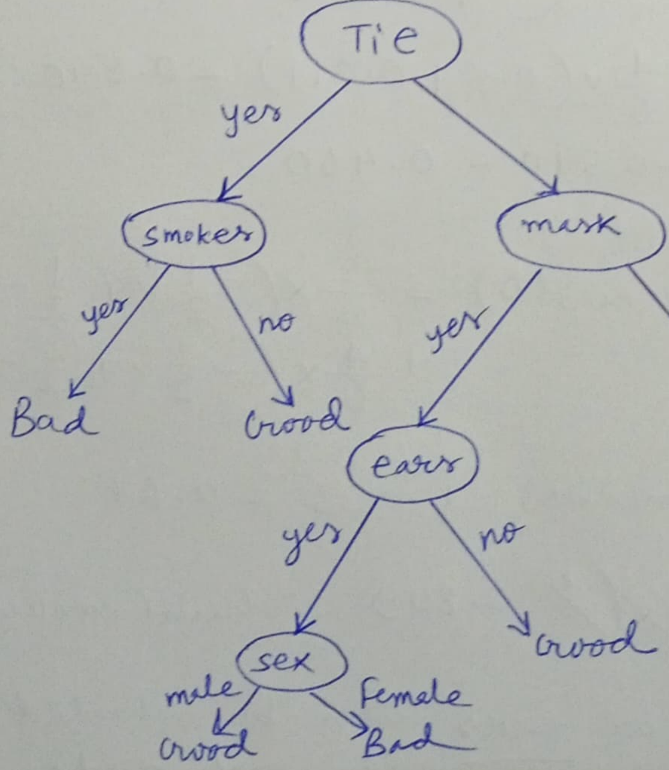
$$Info_{mask}(D) = \frac{3}{4} \times \left(-\frac{2}{3}\log_2\frac{2}{3} - \frac{1}{3}\log_2 1/3\right) + \frac{1}{4} \times 0$$

$$= 3/4 \times (0.390 + 0.528)$$

$$= 0.689$$

$\therefore Gain(mask) = 0.312$

$Info_{ears}(D) = 1 \therefore Gain = 0$



sex	cape	ears	smoke	class
male	Y	Y	N	G
male	Y	N	N	G
female	N	Y	N	B

$Info(D) = 0.918$.

$Info_{ears}(D) = 0.667$

$Gain(ears) = 0.251$

M Y N G
F N N B

5. i. Entropy $= -\frac{5}{9}\log_2\frac{5}{9} - \frac{4}{9}\log_2\frac{4}{9}$

$= 0.470 + 0.521 = 0.991$

ii. $a_1 = T$

Entropy $(T) = -\frac{2}{4}\log_2\frac{2}{4} - \frac{2}{4}\log_2\frac{2}{4} = 1$

$a_1 = f$

Entropy $(f) = -\frac{3}{5}\log_2\frac{3}{5} - \frac{2}{5}\log_2\frac{2}{5} \approx 0.971$

Entropy after split $= \frac{4}{9} \times 1 + \frac{5}{9} \times 0.971 = 0.985$

$IG(a_1) = 0.991 - 0.985 = 0.006$

$a_2 = T$

Entropy $(T) = -\frac{3}{5}\log_2\frac{3}{5} - \frac{2}{5}\log_2\frac{2}{5} = 0.971$

$a_2 = f$

Entropy $(f) = -\frac{2}{4}\log_2\frac{2}{4} - \frac{2}{4}\log_2\frac{2}{4} = 1$

Entropy after split $= \frac{5}{9} \times 0.971 + \frac{4}{9} \times 1 = 0.985$

$IG(a_2) = 0.006$

iii. Possible splits : 1.0, 3.0, 4.0, 5.0, 6.0, 7.0, 8.0

iv. As the entropy split of a_1 and a_2 is same i.e. 0.006 which is quite small. Hence, a_3 is the best split according to IG.