11/2/18

Data Mining
Homework - 2

A20406657
Avinash Vellingiri

Problem 1.2:

2)a) Compute the gini index for the overall collection of training examples.

20 observation → 10 are class 0
→ 10 are class 1

$$\text{gini index} = 1 - \left[\left(\frac{10}{20}\right)^2 + \left(\frac{10}{20}\right)^2\right] = 1 - [0.25 + 0.25]$$

$$= 1 - 0.5 = 0.5$$

gini index = 0.5

b) gini index for the customer ID attribute.

Since the customer ID's are unique (i.e) no relation between the customer ID values.

∴ The gini index for Customer ID is Zero

c) Gini index for Gender attribute.

We have 10 observations as Male and 10 observations as Female

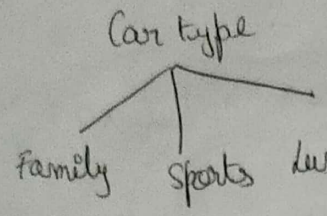
	M	F
C_0	6	4
C_1	4	6

$$\text{Gini\_Index (Male)} = 1 - \left[\left(\frac{6}{10}\right)^2 + \left(\frac{4}{10}\right)^2\right] = 0.48$$

$$\text{Gini\_Index (Female)} = 1 - \left[\left(\frac{4}{10}\right)^2 + \left(\frac{6}{10}\right)^2\right] = 0.48$$

$$\text{Overall gini Index} = \frac{10}{20} \times 0.48 + \frac{10}{20} \times 0.48 = 0.48$$

d) Gini index for car type using multiway split.



	Family	Sports	Luxury
C_0	1	8	1
C_1	3	0	7

$$\text{Gini(Family)} = 1 - \left[\left(\frac{1}{4}\right)^2 + \left(\frac{3}{4}\right)^2\right] = 0.375$$

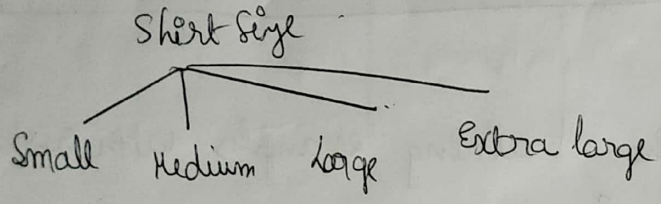
$$\text{Gini(Sports)} = 1 - \left[\left(\frac{8}{8}\right)^2 + \left(\frac{0}{8}\right)^2\right] = 0$$

$$\text{Gini(Luxury)} = 1 - \left[\left(\frac{1}{8}\right)^2 + \left(\frac{7}{8}\right)^2\right] = 0.21875$$

$$\text{Overall Gini} = \frac{4}{20} \times 0.375 + \frac{8}{20} \times 0 + \frac{8}{20} \times 0.21875$$

$$= 0.075 + 0.0875 = \underline{0.1625}$$

e) Gini index for Shirt size using multiway split.



	Small	Medium	Large	Extra large
C_0	3	3	2	2
C_1	2	4	2	2

$$\text{Gini(Small)} = 1 - \left[\left(\frac{3}{5}\right)^2 + \left(\frac{2}{5}\right)^2\right] = 0.48$$

$$\text{Gini(Medium)} = 1 - \left[\left(\frac{3}{7}\right)^2 + \left(\frac{4}{7}\right)^2\right] = 0.4898$$

$$\text{Gini(Large)} = 1 - \left[\left(\frac{2}{4}\right)^2 + \left(\frac{2}{4}\right)^2\right] = 0.5$$

$$\text{Gini(Extra large)} = 1 - \left[\left(\frac{2}{4}\right)^2 + \left(\frac{2}{4}\right)^2\right] = 0.5$$

Overall Gini Index of shirt size

$$= \frac{5}{20} \times 0.48 + \frac{15}{20} \times 0.4898 + \frac{4}{20} \times 0.5 + \frac{4}{20} \times 0.5$$

$$= 0.49143$$

f) which among Gender, Car Type, Shirt Size is better?

Car type is better since it has the lowest gini Index.

g) why Customer ID should not be used as the attribute test condition?

Customer ID is unique. Each customer is assigned with a unique ID called the Customer ID. We can't predict the customer ID attribute because new customers are given unique ID.

---

3)

a) Entropy of the training examples with respect to the positive class?

$$\text{Entropy(positive class)} = -\left[\frac{4}{9}\log_2\frac{4}{9} + \frac{5}{9}\log_2\frac{5}{9}\right]$$

$$= 0.9911$$

a_1

	+	−
T	3	1
F	1	4

a_2

	+	−
T	2	3
F	2	2

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

$= 0.7616$

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

$= 0.9839$

Information Gain(a_1) = $0.9911 - 0.7616 = 0.2294$

Information gain(a_2) = $0.9911 - 0.9839 = 0.0072$

c) For a_3 find information gain for every possible split

	1		3		4		5		6		7	
Split →	2		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

Split 2

$\frac{1}{9}\left[-\left(\frac{1}{1}\log_2 1 + \frac{0}{1}\log_2 0\right)\right] + \frac{8}{9}\left[-\left(\frac{3}{8}\log_2\frac{3}{8} + \frac{5}{8}\log_2\frac{5}{8}\right)\right]$

Entropy = 0.8484

Information gain = $0.9911 - 0.8484 = 0.1427$

Split 3.5

$\frac{2}{9}\left[-\left(\frac{1}{2}\log_2\frac{1}{2} + \frac{1}{2}\log_2\frac{1}{2}\right)\right] + \frac{7}{9}\left[-\left(\frac{3}{7}\log_2\frac{3}{7} + \frac{4}{7}\log_2\frac{4}{7}\right)\right]$

Entropy = 0.9885

Information gain = $0.9911 - 0.9885 = 0.0026$

Split 4.5 :-

$$\frac{3}{9}\left[-\left(\frac{2}{3}\log_2 2/3 + \frac{1}{3}\log_2 1/3\right)\right] + \frac{6}{9}\left[-\left(\frac{2}{6}\log_2 \frac{2}{6} + \frac{4}{6}\log_2 4/6\right)\right]$$

Entropy = 0.9183

Info gain = 0.9911 − 0.9183 = 0.0728

Split 5.5 :-

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

Entropy = 0.9839

Info gain = 0.9911 − 0.9839 = 0.0072.

Split 6.5 :-

$$\frac{6}{9}\left[-\left(\frac{3}{6}\log_2 3/6 + 3/6\log_2 3/6\right)\right] + \frac{3}{9}\left[-\left(\frac{1}{3}\log_2 1/3 + \frac{2}{3}\log_2 2/3\right)\right]$$

Entropy = 0.9728

Info gain = 0.9911 − 0.9728 = 0.0183.

Split 7.5 :-

$$\frac{8}{9}\left[-\left(\frac{4}{8}\log_2 4/8 + \frac{4}{8}\log_2 4/8\right)\right] + \frac{1}{8}\left[-(0\log_2 0 + 1\log_2 1)\right]$$

Entropy = 0.8889

Info gain = 0.9911 − 0.8889 = 0.1022

Best split for a_3 is at position split 2

Here the information gain is maximum.

d) From the information gains of attributes a_1, a_2, a_3.

$a_1 = 0.2294$; $a_2 = 0.0072$; $a_{3\,max} = 0.1427$

a_1 is the best attribute to split as it has high Information gain.

e) a_1

Error rate (T) $= 1 - \left[\frac{3}{4}, \frac{1}{4}\right]_{max} = 1 - \frac{3}{4} = 1/4$

Error rate (F) $= 1 - Max\left[\frac{1}{5}, \frac{4}{5}\right] = 1 - 4/5 = 1/5$.

Error $(a_1) = \frac{4}{9}(1/4) + \frac{5}{9}(1/5) = \underline{\underline{2/9}}$

a_2

Error rate (T) $= 1 - [Max\left[\frac{2}{4}, \frac{2}{4}\right] = 1 - 1/2 = 1/2$

Error rate (F) $= 1 - Max\left[\frac{3}{5}, \frac{2}{5}\right] = 1 - 3/5 = 2/5$

Error $(a_2) = \frac{4}{9}(1/2) + \frac{5}{9}(2/5) = 4/9$

a_1 is best than a_2 as a_2 has higher error rate.

f) what is the best split according to the Gini index?

a_1

Gini (+ve) $= 1 - \left[\left(\frac{3}{4}\right)^2 + \left(\frac{1}{4}\right)^2\right] = 0.375$

Gini (-ve) $= 1 - \left[\left(\frac{1}{5}\right)^2 + \left(\frac{4}{5}\right)^2\right] = 0.32$

Gini $(a_1) = \frac{4}{9} \times 0.375 + \frac{5}{9} \times 0.32 = 0.344$

a2

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

$$= 0.4889$$

$\therefore$ Gini(a_1) < Gini(a_2). a_1 is a better split.

---

5) a) Information gain when splitting on A and B.

+ve 4 out of 10, −ve 6 out of 10.

A	+	−
T	4	3
F	0	3

$$E_A(+ve) = -\left[\frac{4}{4}\log_2 4/4 + \frac{0}{4}\log_2 0/4\right]$$

A	T	F
+	4	0
−	3	3

B	T	F
+	3	1
−	1	5

$$E_{tot} = -\left[\frac{4}{10}\log_2\frac{4}{10} + \frac{6}{10}\log_2 6/10\right]$$

$$E_{tot} = 0.9710$$

Split on A:-

$$E_T = -\left[\frac{4}{7}\log_2 4/7 + \frac{3}{7}\log_2 3/7\right] = 0.9852$$

$$E_F = -\left[\frac{0}{3}\log_2\frac{0}{3} + \frac{3}{3}\log_2 3/3\right] = 0$$

$$\Delta = E_{tot} - \left[\frac{7}{10} * 0.9852\right] = 0.9710 - 0.68964$$

$$\Delta_{info} = 0.2813$$

Split on B:-

$$E_T = -\left[\frac{3}{4}\log_2 3/4 + 1/4 \log_2 1/4\right] = 0.8113$$

$$E_F = -\left[\frac{1}{6}\log_2 1/6 + \frac{5}{6}\log_2 5/6\right] = 0.65$$

$$\Delta = 0.9710 - \left[\frac{4}{10} * 0.8113 + \frac{6}{10} * 0.65\right]$$

$$\Delta = 0.2569$$

Information gain for "A" is higher. So it is better to split at A.

b) $G_{tot} = 1 - \left[\left(\frac{4}{10}\right)^2 + \left(\frac{6}{10}\right)^2\right] = 0.48$

Splitting at A:-

$$G_T = 1 - \left[\left(\frac{4}{7}\right)^2 + \left(3/7\right)^2\right] = 0.4898$$

$$G_F = 1 - \left[\left(\frac{0}{3}\right)^2 + \left(3/3\right)^2\right] = 0$$

$$\Delta = 0.48 - \left[\frac{7}{10} * 0.4898\right] = 0.1371$$

Splitting at B:-

$$G_T = 1 - \left[\left(\frac{3}{4}\right)^2 + \left(\frac{1}{4}\right)^2\right] = 0.3750$$

$$G_F = 1 - \left[\left(\frac{1}{6}\right)^2 + \left(5/6\right)^2\right] = 0.2778$$

$$\Delta = 0.48 - \left[\frac{4}{10} * 0.3750 + \frac{6}{10} * 0.2778\right]$$

$$= 0.1633$$

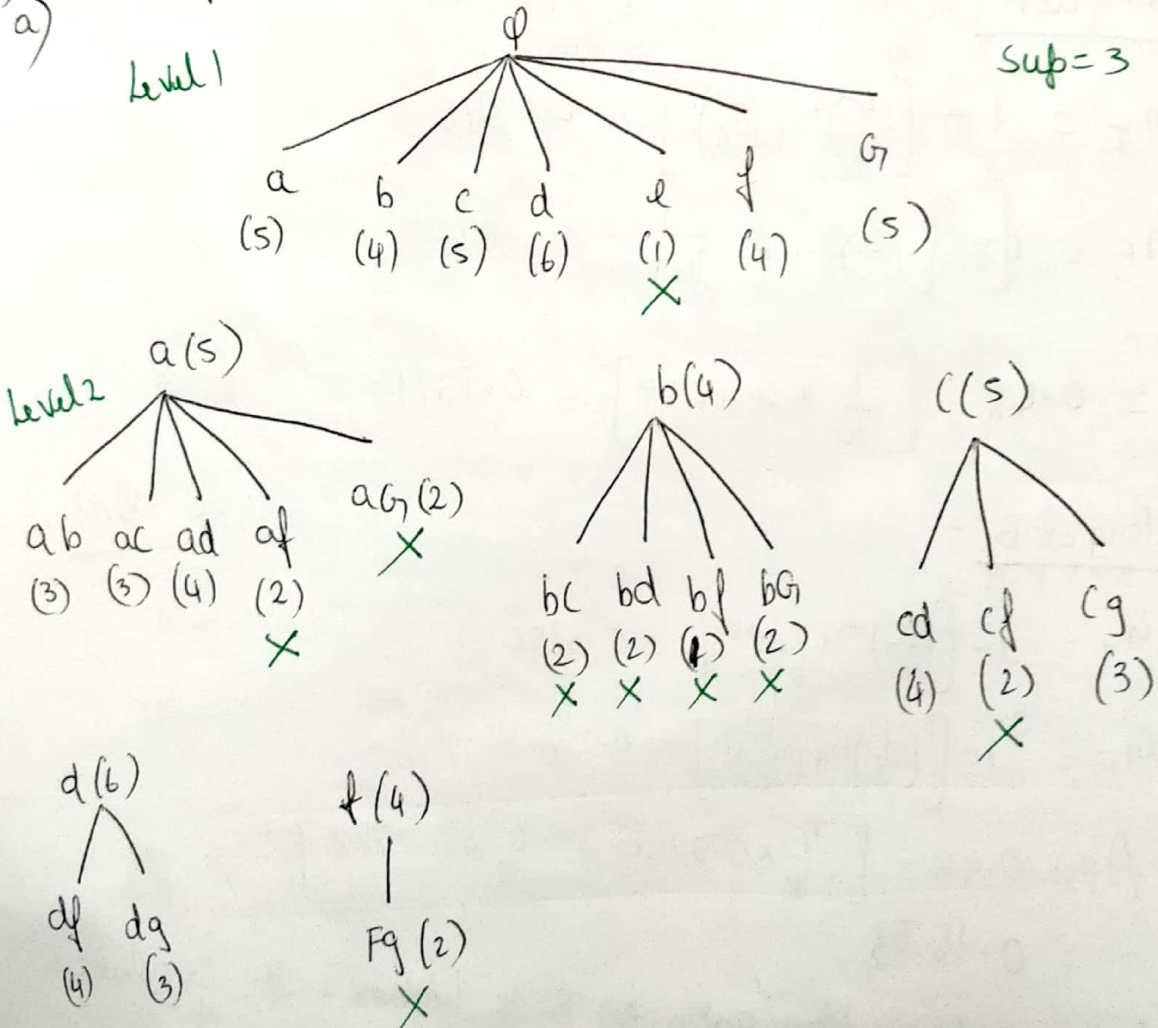
Since information gain for B is higher. It is better to split at attribute B. Algorithm picks "B"

c) Yes, An attribute selection can vary based upon gini index or entropy "Δ" calculation we can see from the results of (a) and (b) of questions that both the method picked different attributes.

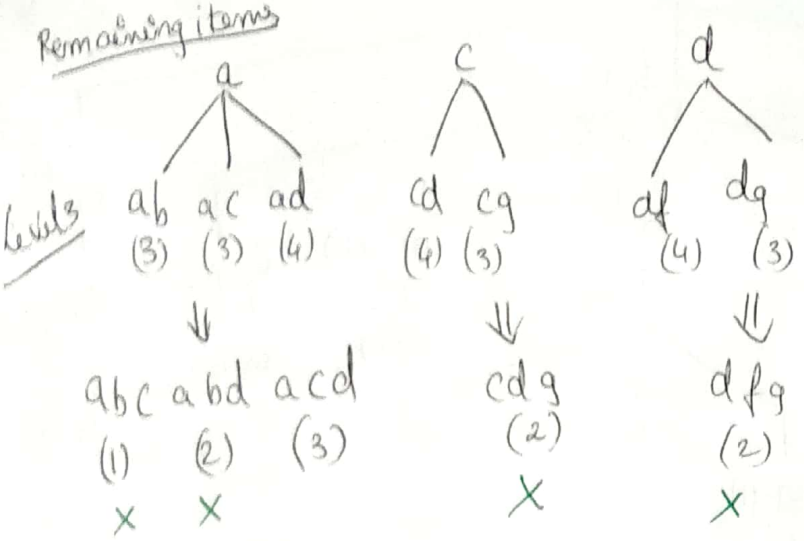
Because the gain calculated "Δ" are scaled differently Entropy follows a long scale while gini index follows a polynomial scale.

Problem 1.4 (Zaki)

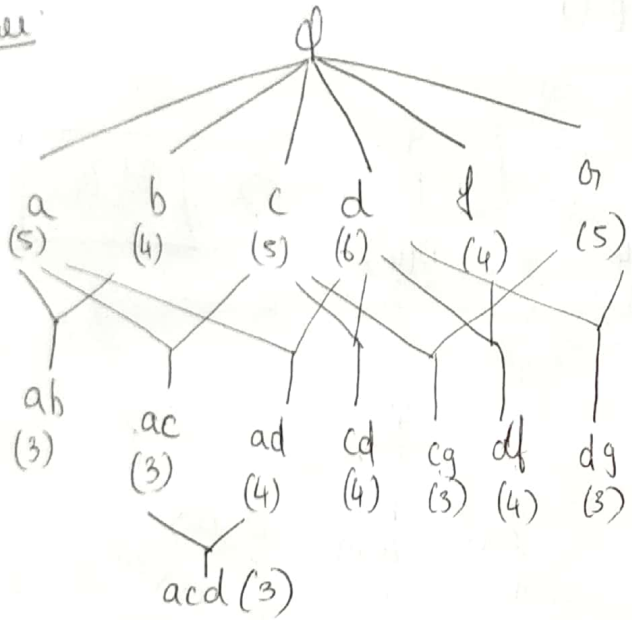
1) a) Minsup = 3/8, frequent patterns using apriori



Remaining items



Final tree



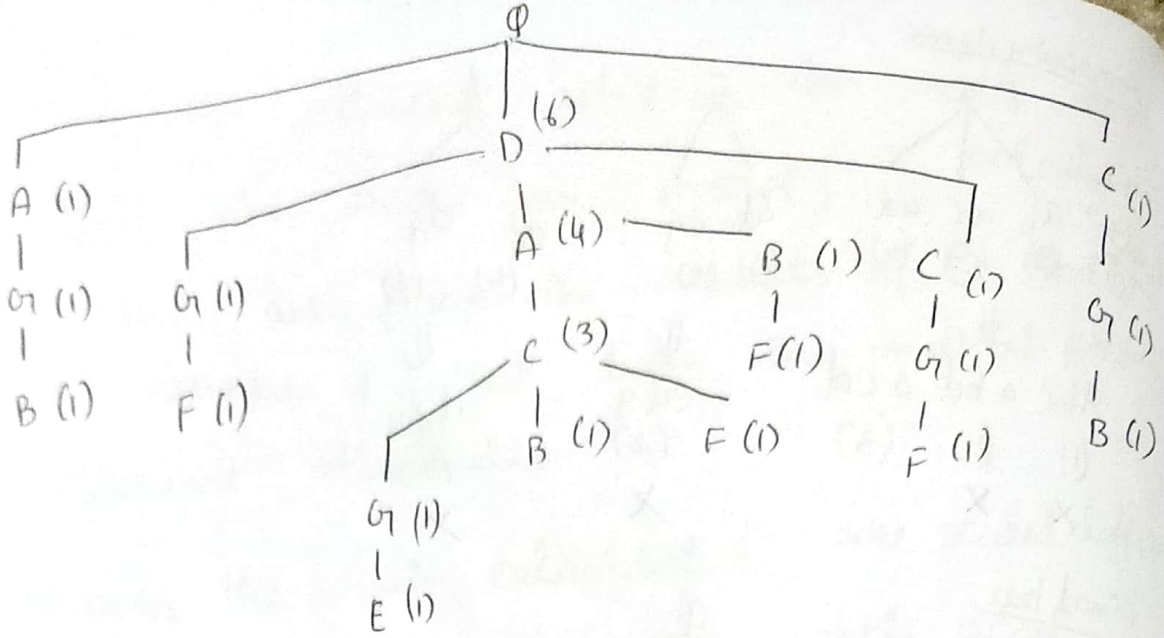
Frequent Item Sets "14 items" = { d(6), g(5), a(5), c(5), b(4), f(4), df(4), ad(4), cd(4), ab(3), cg(3), dg(3), ac(3), acd(3) }

1b)

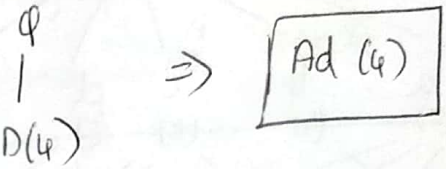
D	A	C	G	B	F	E
6	(5)	(5)	(5)	(4)	(4)	(1)
						x

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

D(6), A(5), C(5), G(5)
B(4), F(4)



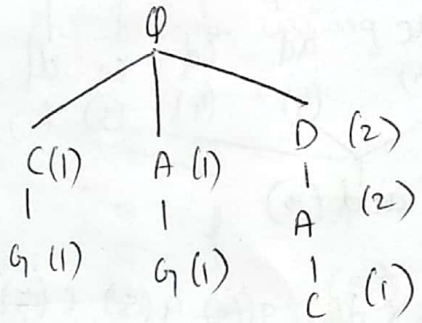
Projection trees:-



R_A D, count(A) = 4

R_B

AG, count(B) = 1
DAC, count(B) = 1
DA, count(B) = 1
CG, count(B) = 1



{D, A, C} ⇒ { D(2), A(2), C(1), DA(2), DC(1), AC(1) }
× ×

{A, G} ⇒ { A(1), G(1), AG(1) }
×

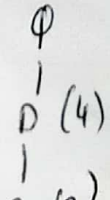
{C, G} ⇒ { C(1), G(1), CG(1) }
×

⇒ { D(2), A(3), C(2), G(2), DA(2) }

BD(2), BA(3), BC(2), BG(2), BDA(2)

R_C

DA, count(C) = 3
D, count(C) = 1



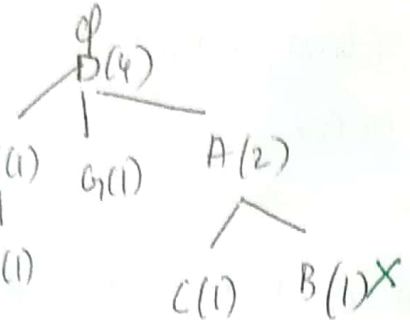
{D(4), A(3), DA(3)}

CD(4), CA(3), CDA(3)

$R_D \Rightarrow$ None

R_E
DACG, count(E) = 1 $\Rightarrow$ None all have count 1

R_F
DG, count(F) = 1
DAC, count(F) = 1
DAB, count(F) = 1
DCG, count(F) = 1



{D, A} $\Rightarrow$ { D(4), A(2), DA(2) }

FD(4), FA(2), FDA(2)

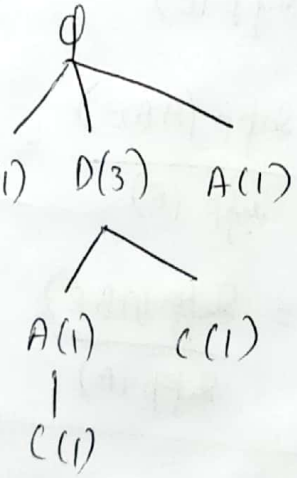
{D, G} $\Rightarrow$ { D(4), G(1), DG(1) }

{D, C, G} = { D(4), C(1), G(1), DC(1), DG(1), CG(1), DCG(1) }

{DAC} = { D(4), A(2), C(1), DA(2), DC(1), AC(1), DAC(1) }

FG(2), FDG(2), FDC(2), FC(2)

R_G
A, count(G) = 1
D, count(G) = 1
DA, count(G) = 1
DC, count(G) = 1



{A(1)}, {C(1)}, {D(3), C(1), DC(1)}, {D(3), A(1), C(1), DA(1), DC(1), AC(1), DAC(1)}

$\Rightarrow$ GA(2), GC(3), GD(3), GDC(2)

Total no of frequent item sets ⇒ 26 sets

D(6), A(5), C(5), G(5), B(4), F(4), AD(4), BD(2), BA(3), B(2), BG(2), BDA(2), CD(4), CA(3), CDA(3), FD(4), FA(2), FDA(2), FG(2), FDG(2), FDC(2), FC(2), GA(2), G(3), GD(3), GDC(2).

---

4) ABE ⇒ {A(4), B(5), E(4), AB(3), AE(2), BE(4)}

$$C(BE \rightarrow A) = \frac{Supp(ABE)}{Supp(BE)} = \frac{2}{4} = 0.5$$

$$C(AE \rightarrow B) = \frac{Supp(ABE)}{Supp(AE)} = \frac{2}{2} = 1$$

$$C(AB \rightarrow E) = \frac{Supp(ABE)}{Supp(AB)} = 2/3 = 0.66$$

$$C(E \rightarrow AB) = \frac{Supp(ABE)}{Supp(E)} = \frac{2}{4} = 0.5$$

$$C(B \rightarrow AE) = \frac{Supp(ABE)}{Supp(B)} = 2/5 = 0.4$$

$$C(A \rightarrow BE) = \frac{Supp(ABE)}{Supp(A)} = 2/4 = 0.5$$

---

6) here k=11

a) No of itemsets $= 2^k - 1 = 2^{11} - 1 = 2047$

b) (iv) More than or equal to support of X