

BIG DATA HW 4A

Nagabharan Nagendran

nxn141730

Q1)

	1	2	3	4	5	6
A		3	1	2	3	
B		5	4		4	
C			2	5		1
D	4	1		4		2

a) $(A, B): I = 2, U = 5$

Similar = $2/5$

Distance = $3/5$

$(A, C): I = 2, U = 5$

Similar = $2/5$

Distance = $3/5$

$(A, D): I = 3, U = 5$

Similar = $3/5$

Distance = $2/5$

$(B, C): I = 1, U = 5$

Similar = $1/5$

Distance = $4/5$

$(B, D): I = 1, U = 6$

Similar = $1/6$

Distance = $5/6$

$(C, D): I = 2, U = 5$

Similar = $2/5$

Distance = $3/5$

b) (A|B):

$$\frac{3 \times 5 + 2 \times 4}{\sqrt{3^2 + 1^2 + 2^2 + 3^2} \times \sqrt{5^2 + 4^2 + 4^2}} = \frac{23}{(4.8 \times 7.55)} = 0.63$$

(A|C):

$$\frac{1 \times 5 + 1 \times 3}{\sqrt{3^2 + 1^2 + 2^2 + 3^2} \times \sqrt{2^2 + 5^2 + 1^2}} = \frac{8}{(4.8 \times 5.48)} = 0.304$$

(A|D):

$$\frac{1 \times 3 + 1 \times 4 + 2 \times 3}{\sqrt{3^2 + 1^2 + 2^2 + 3^2} \times \sqrt{4^2 + 1^2 + 4^2 + 2^2}} = \frac{13}{(4.8 \times 6.08)} = 0.445$$

(B|C):

$$\frac{2 \times 4}{\sqrt{5^2 + 4^2 + 4^2} \times \sqrt{2^2 + 5^2 + 1^2}} = \frac{8}{(7.55 \times 5.48)} = 0.193$$

(B|D):

$$\frac{1 \times 5}{\sqrt{5^2 + 4^2 + 4^2} \times \sqrt{4^2 + 1^2 + 4^2 + 2^2}} = \frac{5}{(7.55 \times 6.08)} = 0.109$$

(C|D):

$$\frac{4 \times 5 + 2 \times 1}{\sqrt{2^2 + 5^2 + 1^2} \times \sqrt{4^2 + 1^2 + 4^2 + 2^2}} = \frac{22}{(5.48 \times 6.08)} = 0.66$$

c> By replacing 3, 4, 5 stars as 1 & 1, 2 stars as 0 we get

	1	2	3	4	5	6	(0,0,0)
A		1				1	1
B		1	1		1		
C				1			(0,0,0)
D	1			1			1

(A) Iteration 1 - Distances

(1,2)	(1,3)	(1,4)	(1,5)	(1,6)
1	1	1/2	1	1
(2,3)	(2,4)	(2,5)	(2,6)	(3,4)
1/2	1	1/2	1/2	1
(3,5)	(3,6)	(4,5)	(4,6)	(5,6)
1	1	1	1	1

Combining 3 & 5 to form a cluster (3,5)

Iteration 2 - Distances

(1,2)	(1,35)	(1,4)	(1,6)
1	1	1/2	1

$(2, 35)$ $(2, 4)$ $(2, 6)$
 $\frac{1}{2}$ $\frac{1}{2}$ $\frac{1}{2}$

$(4, 35)$ $(4, 6)$
 $\frac{1}{2}$ $\frac{1}{2}$

$(6, 35)$
 $\frac{1}{2}$

Combining 1 & 4 to form cluster (1, 4)

Iteration 3 Distances

$(14, 2)$ $(14, 35)$ $(14, 6)$
 $\frac{1}{2}$ $\frac{1}{2}$ $\frac{1}{2}$

$(2, 35)$ $(2, 6)$
 $\frac{1}{2}$ $\frac{1}{2}$

$(6, 35)$
 $\frac{1}{2}$

∴ The final 4 clusters are

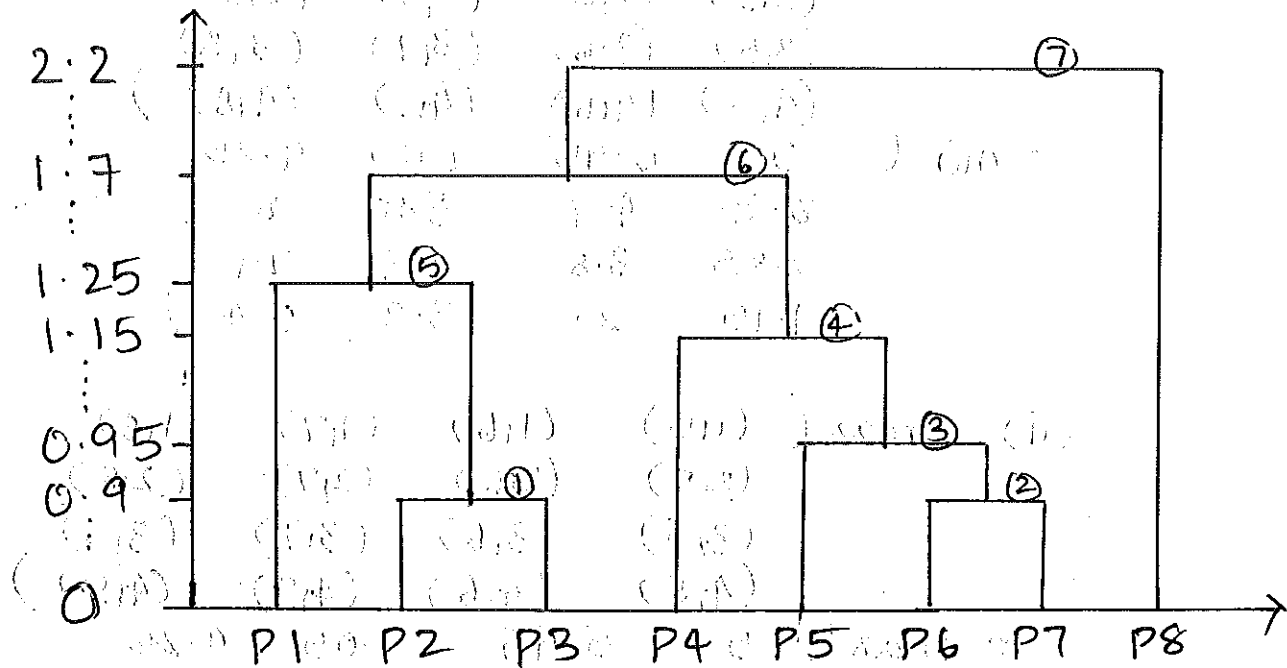
$\langle 1, 4 \rangle$ 2 $\langle 3, 5 \rangle$ 6

$(1, 1)$ $(1, 1)$ $(2, 6, 1)$ $(2, 1)$
 $\frac{1}{2}$ $\frac{1}{2}$ $\frac{1}{2}$ $\frac{1}{2}$

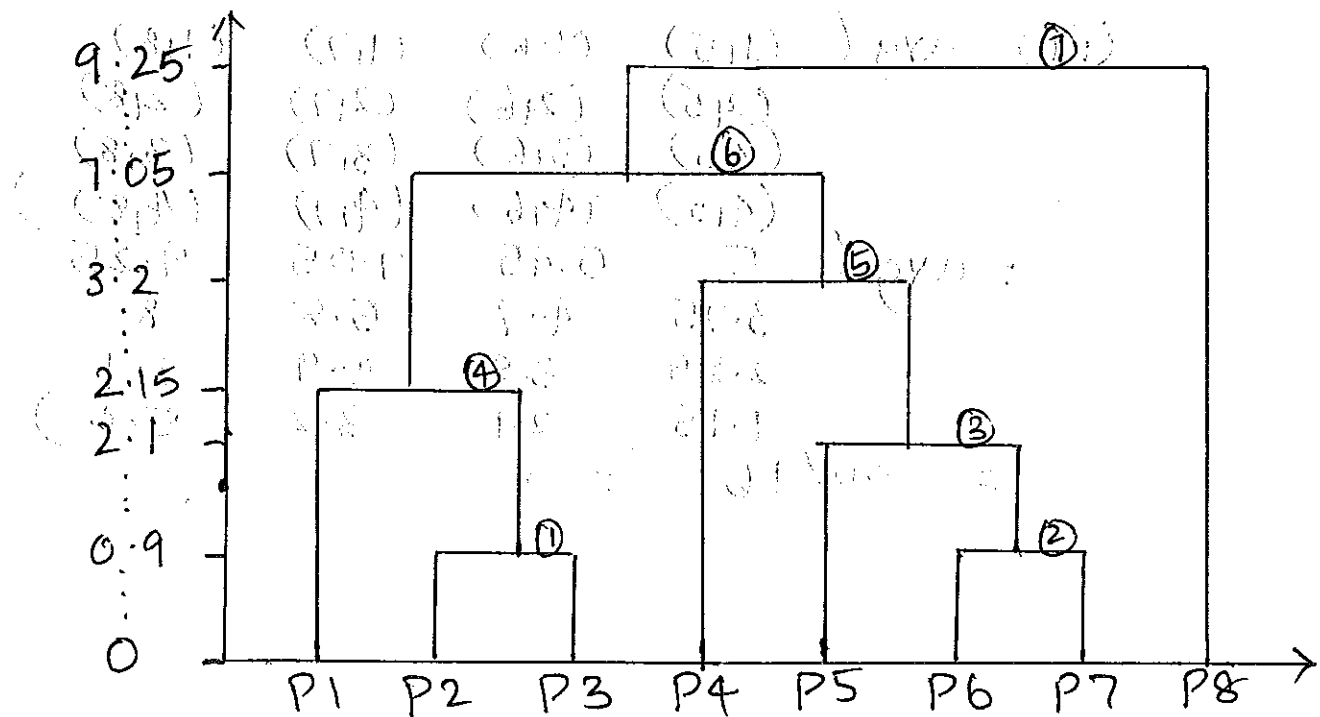
Q2)

a)

(i) Single Link



(ii) Complete Link



b) (i) d { (1,2,3,4), (5,6,7,8) }

$$= \min \begin{pmatrix} (1,5) & (1,6) & (1,7) & (1,8) \\ (2,5) & (2,6) & (2,7) & (2,8) \\ (3,5) & (3,6) & (3,7) & (3,8) \\ (4,5) & (4,6) & (4,7) & (4,8) \end{pmatrix}$$

$$= \min \begin{pmatrix} 5 & 5.95 & 7.05 & 9.25 \\ 3.75 & 4.7 & 5.8 & 8 \\ 2.85 & 3.8 & 4.9 & 7.1 \\ 1.15 & 2.1 & 3.2 & 5.4 \end{pmatrix} = 1.15$$

$$(ii) \max \begin{pmatrix} (1,5) & (1,6) & (1,7) & (1,8) \\ (2,5) & (2,6) & (2,7) & (2,8) \\ (3,5) & (3,6) & (3,7) & (3,8) \\ (4,5) & (4,6) & (4,7) & (4,8) \end{pmatrix}$$

$$= \max \begin{pmatrix} 5 & 5.95 & 7.05 & 9.25 \\ 3.75 & 4.7 & 5.8 & 8 \\ 2.85 & 3.8 & 4.9 & 7.1 \\ 1.15 & 2.1 & 3.2 & 5.4 \end{pmatrix} = 9.25$$

$$(iii) \text{avg} \begin{pmatrix} (1,5) & (1,6) & (1,7) & (1,8) \\ (2,5) & (2,6) & (2,7) & (2,8) \\ (3,5) & (3,6) & (3,7) & (3,8) \\ (4,5) & (4,6) & (4,7) & (4,8) \end{pmatrix}$$

$$= \text{avg} \begin{pmatrix} 5 & 5.95 & 7.05 & 9.25 \\ 3.75 & 4.7 & 5.8 & 8 \\ 2.85 & 3.8 & 4.9 & 7.1 \\ 1.15 & 2.1 & 3.2 & 5.4 \end{pmatrix}$$

$$= 80/16 = 5$$

Iteration 2-

$$\begin{array}{l} P1C1 \\ \sqrt{3^2 + 5^2} \\ 3.04 \end{array}$$

$$\begin{array}{l} P2C1 \\ \sqrt{2^2 + 2.5^2} \\ 3.2 \end{array}$$

$$\begin{array}{l} P3C1 \\ \sqrt{1^2 + 5.5^2} \\ 5.59 \end{array}$$

$$\begin{array}{l} P4C1 \\ \sqrt{1.5^2} \\ 1.5 \end{array}$$

$$\begin{array}{l} P5C1 \\ \sqrt{1^2 + 5.5^2} \\ 5.59 \end{array}$$

$$\begin{array}{l} P6C1 \\ \sqrt{3^2 + 5.5^2} \\ 6.27 \end{array}$$

$$\begin{array}{l} P7C1 \\ \sqrt{3^2 + 5^2} \\ 3.04 \end{array}$$

$$\begin{array}{l} P8C1 \\ \sqrt{4^2 + 1.5^2} \\ 4.27 \end{array}$$

$$\begin{array}{l} P1C2 \\ \sqrt{3.83^2 + 3.17^2} \\ 4.97 \end{array}$$

$$\begin{array}{l} P2C2 \\ \sqrt{2.83^2 + 1.17^2} \\ 3.06 \end{array}$$

$$\begin{array}{l} P3C2 \\ \sqrt{1.83^2 + 1.83^2} \\ 2.59 \end{array}$$

$$\begin{array}{l} P4C2 \\ \sqrt{.83^2 + 2.17^2} \\ 2.32 \end{array}$$

$$\begin{array}{l} P5C2 \\ \sqrt{.17^2 + 1.83^2} \\ 1.84 \end{array}$$

$$\begin{array}{l} P6C2 \\ \sqrt{2.17^2 + 1.83^2} \\ 2.84 \end{array}$$

$$\begin{array}{l} P7C2 \\ \sqrt{2.17^2 + 4.17^2} \\ 4.7 \end{array}$$

$$\begin{array}{l} P8C2 \\ \sqrt{3.17^2 + 2.17^2} \\ 3.84 \end{array}$$

Cluster 1 - {1, 4, 7}

Cluster 2 - {2, 3, 5, 6, 8}

C1 - (5, 2)

C2 - (6, 5, 6)

Iteration 3-

$$\begin{array}{l} P1C1 \\ \sqrt{3^2} \\ 3 \end{array}$$

$$\begin{array}{l} P2C1 \\ \sqrt{2^2 + 2^2} \\ 2.83 \end{array}$$

$$\begin{array}{l} P3C1 \\ \sqrt{1^2 + 5^2} \\ 5.1 \end{array}$$

$$\begin{array}{l} P4C1 \\ \sqrt{1^2} \\ 1 \end{array}$$

$$\begin{array}{l} P5C1 \\ \sqrt{1^2 + 5^2} \\ 5.1 \end{array}$$

$$\begin{array}{l} P6C1 \\ \sqrt{3^2 + 5^2} \\ 5.83 \end{array}$$

$$\begin{array}{l} P7C1 \\ \sqrt{3^2 + 1^2} \\ 3.16 \end{array}$$

$$\begin{array}{l} P8C1 \\ \sqrt{4^2 + 1^2} \\ 4.12 \end{array}$$

Q 3)

P1 P2 P3 P4 P5 P6 P7 P8

1 2 3 4 5 6 7 8

2 1 4 7 3 7 1 3

C1 - (2,2) C2 - (3,5)

Distances:

Iteration 1 -

P1C1 P2C1 P3C1 P4C1 P5C1 P6C1

1.6 1.8 $\sqrt{1^2+2^2}$ $\sqrt{2^2+5^2}$ $\sqrt{3^2+1^2}$ $\sqrt{4^2+5^2}$ $\sqrt{5^2+6^2}$
0 2.24 5.39 3.16 6.41 7.81

$\frac{P7C1}{\sqrt{6^2+1^2}}$
6.08

$\frac{P8C1}{\sqrt{7^2+1^2}}$
7.07

$\frac{P4C2}{\sqrt{1^2+3^2}}$
3.16

$\frac{P2C2}{1}$
1

$\frac{P3C2}{\sqrt{1^2+2^2}}$
2.24

$\frac{P4C2}{\sqrt{2^2+2^2}}$
2.83

$\frac{P5C2}{\sqrt{3^2+2^2}}$
3.61

$\frac{P6C2}{\sqrt{5^2+2^2}}$
5.39

$\frac{P7C2}{\sqrt{5^2+4^2}}$
6.4

$\frac{P8C2}{\sqrt{6^2+2^2}}$
6.32

Cluster 1 = {1, 2, 3, 4, 5}

Cluster 2 = {2, 3, 4, 5, 6, 8}

C1 - (5, 1.5)
C2 - (5.83, 5.17)

P1C2 $\sqrt{4^2+3.6^2}$ 5.83	P2C2 $\sqrt{3^2+1.6^2}$ 3.4	P3C2 $\sqrt{2^2+1.4^2}$ 2.44	P4C2 $\sqrt{1^2+2.6^2}$ 2.78	P5C2 $\sqrt{1.4^2}$ 1.4	P6C2 $\sqrt{2^2+1.4^2}$ 2.44
P7C2 $\sqrt{2^2+4.6^2}$ 5.02	P8C2 $\sqrt{3^2+2.6^2}$ 3.97				

Cluster 1 - {1, 2, 4, 7}, Cluster 2 - {3, 5, 6, 8}

C1 - (4.5, 2.5), C2 - (6.75, 6)

(2.5, 8), (4.5, 12), (6.75, 18), (8.5, 10)

Iteration 4 -

P1C1 $\sqrt{2.5^2+5^2}$ 5.5	P2C1 $\sqrt{1.5^2+1.5^2}$ 2.12	P3C1 $\sqrt{5^2+4.5^2}$ 6.83	P4C1 $\sqrt{.5^2+.8^2}$ 1.14	P5C1 $\sqrt{1.5^2+4.5^2}$ 4.74	P6C1 $\sqrt{3.5^2+4.5^2}$ 5.7
-----------------------------------	--------------------------------------	------------------------------------	------------------------------------	--------------------------------------	-------------------------------------

P7C1 $\sqrt{3.5^2+1.5^2}$ 3.81	P8C1 $\sqrt{4.5^2+5^2}$ 6.75
--------------------------------------	------------------------------------

P1C2 $\sqrt{4.75^2+4^2}$ 6.21	P2C2 $\sqrt{3.75^2+2^2}$ 4.25	P3C2 $\sqrt{2.75^2+1^2}$ 2.93	P4C2 $\sqrt{1.75^2+3^2}$ 3.47	P5C2 $\sqrt{.75^2+1^2}$ 1.25	P6C2 $\sqrt{1.25^2+1^2}$ 1.6
-------------------------------------	-------------------------------------	-------------------------------------	-------------------------------------	------------------------------------	------------------------------------

P7C2 $\sqrt{1.25^2+5^2}$ 5.13	P8C2 $\sqrt{2.25^2+3^2}$ 3.75
-------------------------------------	-------------------------------------

Cluster 1 - {1, 2, 4, 7}, Cluster 2 - {3, 5, 6, 8}
Convergence.

Q4 >

	P1	P2	P3	P4
x_1	5	3	1	4
x_2	2	3	9	1
x_3	7	3	6	2
x_4	4	3	1	7

$$C = (2, 7, 4, 5)$$

$$N = 4$$

$$\text{SUM} = (13, 15, 18, 15)$$

$$\text{SUMSQ} = (51, 95, 98, 75)$$

$$\text{Centroid} = (3.25, 3.75, 4.5, 3.75)$$

Variance -

$$x_1 = 51/4 - (13/4)^2 = 12.75 - 10.56$$

$$= 2.19$$

$$x_2 = 95/4 - (15/4)^2 = 23.75 - 14.06$$

$$= 9.69$$

$$x_3 = 98/4 - (18/4)^2 = 24.5 - 20.25$$

$$= 4.25$$

$$x_4 = 75/4 - (15/4)^2 = 18.75 - 14.06$$

$$= 4.69$$

$$\text{SD} \sigma = \sqrt{V}$$

$$x_1 = 1.48$$

$$x_2 = 3.11$$

$$x_3 = 2.06$$

$$x_4 = 2.17$$

$$d(x, c) = \sqrt{\sum_{i=1}^N \frac{(x_i - c_i)^2}{\sigma_i}}$$

$$d(x_{1,c}) -$$

$$\sqrt{\left(\frac{5-2}{1.48}\right)^2 + \left(\frac{2-7}{3.11}\right)^2 + \left(\frac{7-4}{2.06}\right)^2 + \left(\frac{4-5}{2.17}\right)^2}$$

$$= \sqrt{4.109 + 2.585 + 2.121 + 0.212}$$

$$= \sqrt{9.027} = 3.005$$

$$d(x_{2,c}) -$$

$$\sqrt{\left(\frac{3-2}{1.48}\right)^2 + \left(\frac{3-7}{3.11}\right)^2 + \left(\frac{3-4}{2.06}\right)^2 + \left(\frac{3-5}{2.17}\right)^2}$$

$$= \sqrt{0.457 + 1.654 + 0.236 + 0.849}$$

$$= 1.788$$

$$d(x_{3,c}) -$$

$$\sqrt{\left(\frac{1-2}{1.48}\right)^2 + \left(\frac{9-7}{3.11}\right)^2 + \left(\frac{6-4}{2.06}\right)^2 + \left(\frac{1-5}{2.17}\right)^2}$$

$$= \sqrt{0.457 + 0.414 + 0.943 + 3.398}$$

$$= 2.283$$

$$d(x_{4,c}) -$$

$$\sqrt{\left(\frac{4-2}{1.48}\right)^2 + \left(\frac{1-7}{3.11}\right)^2 + \left(\frac{2-4}{2.06}\right)^2 + \left(\frac{7-5}{2.17}\right)^2}$$

$$= \sqrt{1.826 + 3.722 + 0.943 + 0.849}$$

$$= 2.709$$

$$d(x_{i,c}) : (3.005, 1.788, 2.283, 2.709)$$

Q5)

BFR

- Assumes clusters are normally distributed in each dimension
- Axes are fixed & ellipses at an angle are not allowed.

CURE

- Assumes a Euclidean distance
- Allows clusters to assume any shape

Q6.

$x_1 \quad x_2 \quad x_3 \quad x_4 \quad x$

a e j u +1

b f k v +1

b f j v -1

a f k u +1

a f j u +1

a f k u -1

b e k v +1

a f j v +1

a f k v -1

b e k v +1

$$IE \text{ Entropy} = \frac{-5 \log 5}{10} - \frac{5 \log 5}{10} = 1$$

$x_1: \#a = 6$

$\#b = 4$

a: $\# +1 = 2$

b: $\# +1 = 3$

$\# -1 = 4$

$\# -1 = 1$

$$a: -\frac{2 \log 2}{6} - \frac{4 \log 4}{6}$$

$$b: -\frac{3 \log 3}{4} - \frac{1 \log 1}{4}$$

$$= 0.92$$

$$= 0.81$$

$$IG(x_1) = IE - \frac{6}{10} a - \frac{4}{10} b$$

$$= 1 - 0.876 = 0.124$$

$$X_2: \#e = 3 \quad \#f = 7$$

$$e: \# +1 = 3$$

$$\# -1 = 0$$

$$I_e = 0$$

$$f: \# +1 = 2$$

$$\# -1 = 5$$

$$I_f = -\frac{2}{7} \log \frac{2}{7} - \frac{5}{7} \log \frac{5}{7}$$

$$= 0.87$$

$$IG(X_2) = 1 - \frac{3}{10} I_e - \frac{7}{10} I_f$$

$$= 1 - 0.609$$

$$= 0.391$$

$$X_3: \#j = 4 \quad \#k = 6$$

$$j: \# +1 = 2$$

$$\# -1 = 2$$

$$I_j = 1$$

$$k: \# +1 = 3$$

$$\# -1 = 3$$

$$I_k = 1$$

$$IG(X_3) = 1 - \frac{4}{10} I_j - \frac{6}{10} I_k = 0$$

$$X_4: \#u = 4 \quad \#v = 6$$

$$u: \# +1 = 2$$

$$\# -1 = 2$$

$$I_u = 1$$

$$v: \# +1 = 3$$

$$\# -1 = 3$$

$$I_v = 1$$

$$IG(X_4) = 1 - \frac{4}{10} I_u - \frac{6}{10} I_v = 0$$

X_2 is 1st splitting attribute

$$IG(X_1) = 1 - \frac{5}{10} \log \frac{5}{10} - \frac{5}{10} \log \frac{5}{10} = 1$$

$$X_1: \quad \#a=4 \quad \#b=2$$

$$a: \quad \# +1 = 1 \\ \quad \quad \# -1 = 3$$

$$b: \quad \# +1 = 1 \\ \quad \quad \# -1 = 1$$

$$I_a = -\frac{1}{4} \log \frac{1}{4} - \frac{3}{4} \log \frac{3}{4}$$

$$I_b = -\frac{1}{2} \log \frac{1}{2} - \frac{1}{2} \log \frac{1}{2}$$

$$= 0.81$$

$$I_G(X_1) = 1E - \frac{4}{6} I_a - \frac{2}{6} I_b$$

$$= 1 - 0.873 = 0.13$$

$$X_3: \quad \#j=3 \quad \#k=3$$

$$j: \quad \# +1 = 1 \\ \quad \quad \# -1 = 2$$

$$k: \quad \# +1 = 1 \\ \quad \quad \# -1 = 2$$

$$I_j = -\frac{1}{3} \log \frac{1}{3} - \frac{2}{3} \log \frac{2}{3}$$

$$I_k = -\frac{1}{3} \log \frac{1}{3} - \frac{2}{3} \log \frac{2}{3}$$

$$= 0.92$$

$$= 0.92$$

$$I_G(X_3) = 1E - \frac{3}{6} I_j - \frac{3}{6} I_k$$

$$= 1 - 0.92 = 0.08$$

$$X_4: \quad \#u=2 \quad \#v=4$$

$$u: \quad \# +1 = 1 \\ \quad \quad \# -1 = 1$$

$$v: \quad \# +1 = 1 \\ \quad \quad \# -1 = 3$$

$$I_u = -\frac{1}{2} \log \frac{1}{2} - \frac{1}{2} \log \frac{1}{2}$$

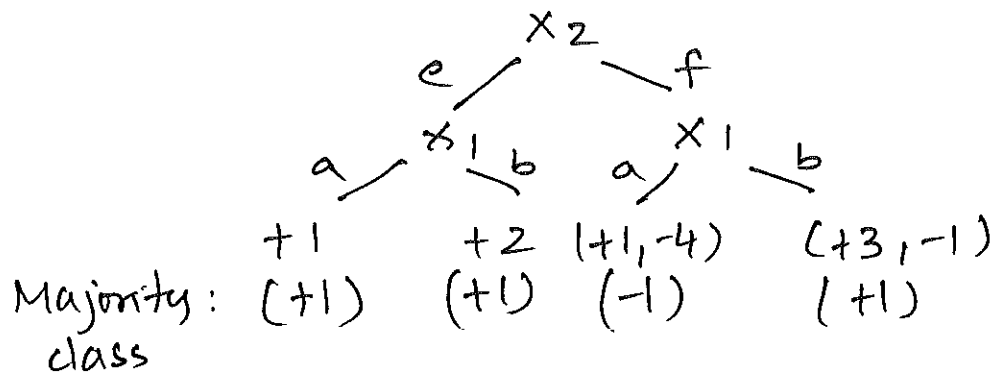
$$I_v = -\frac{1}{4} \log \frac{1}{4} - \frac{3}{4} \log \frac{3}{4}$$

$$= 0.81$$

$$I_G(X_4) = 1E - \frac{2}{6} I_u - \frac{4}{6} I_v = 1 - 0.873 = 0.13$$

x_1 or x_4 can be splitting attribute at level 2 as they have same IG.

Decision Tree:



Classification accuracy: $8/10 = 0.8$ or 80%.