CS 6350 - BIG DATA ANALYTICS & MGT.
HOMEWORK - 4

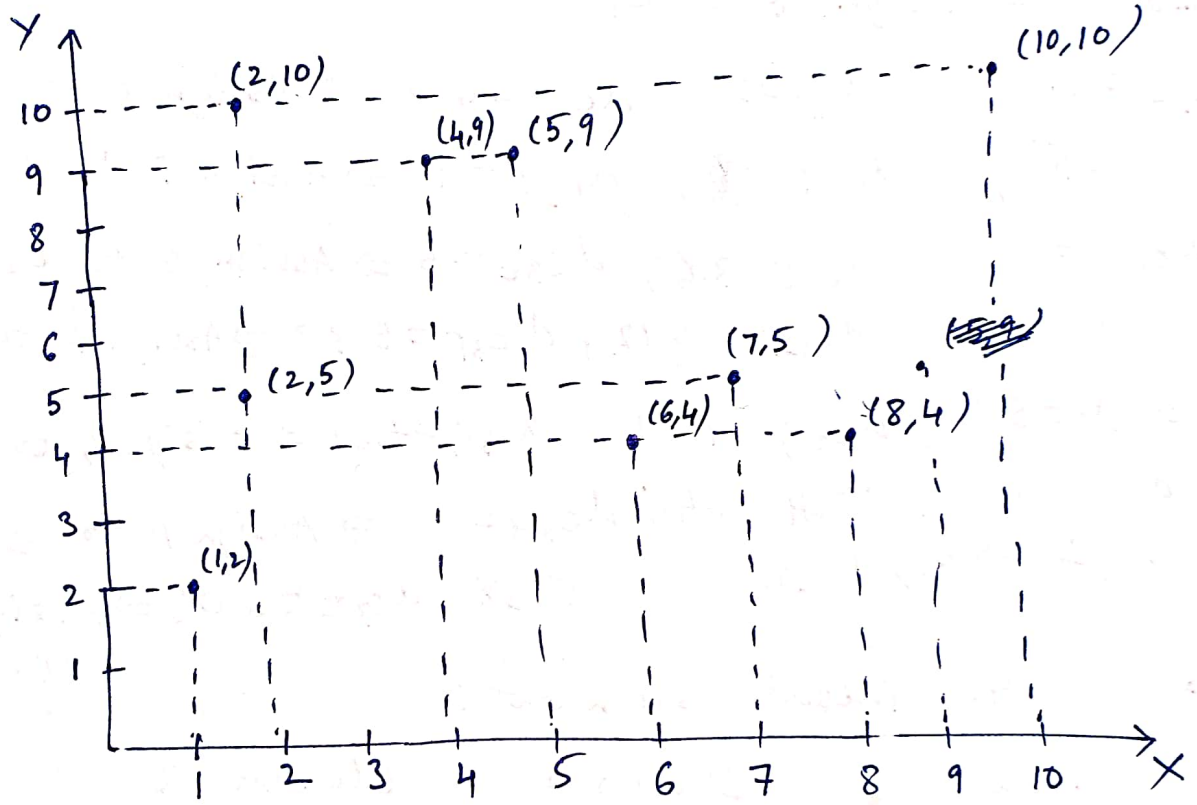
PART-1 : CLUSTERING.

QUESTION A:

Points : (2,10), (2,5), (8,4), (5,9), (7,5), (6,4), (1,2), (4,9), (10,10).

Lets label these points from A to I

(i). Plot the data



Possible clusters:

Cluster-1
(2,5)
(1,2)

Cluster-2 :
(2,10)
(4,9)
(5,9)

Cluster-3
(7,5)
(6,4)
(8,4)
(10,10)

(ii). Initial cluster centres.

$C_1 = (2,5)$

$C_2 = (5,8)$

$C_3 = (4,9)$.

Assign points to nearest clusters.

$d_{C_1A} = 5$, $d_{C_2A} = 3.60$, $d_{C_3A} = 2.24 \Rightarrow$ Assign A to C_3

$d_{C_1B} = 0 \Rightarrow$ Assign B to C_1

$d_{C_1C} = 6.08$, $d_{C_2C} = 5$, $d_{C_3C} = 6.4 \Rightarrow$ Assign C to C_2

$d_{C_1D} = 5$, $d_{C_2D} = 1$, $d_{C_3D} = 1 \Rightarrow$ Assign D to C_2

$d_{C_1E} = 5$ $d_{C_2E} = 3.6$, $d_{C_3E} = 5 \Rightarrow$ Assign E to C_2

$d_{C_1F} = 4.12$, $d_{C_2F} = 4.12$, $d_{C_3F} = 5.38 \Rightarrow$ Assign F to C_2

$d_{C_1G} = 3.16$, $d_{C_2G} = 7.21$ $d_{C_3G} = 7.61 \Rightarrow$ Assign G to C_1

$d_{C_1H} = 4.47$, $d_{C_2H} = 1.41$, $d_{C_3H} = 0 \Rightarrow$ Assign H to C_3

$d_{C_1I} = 9.43$, $d_{C_2I} =$ 5.38, $d_{C_3I} = 6.08 \Rightarrow$ Assign I to C_2

Clusters after initial assignment:

Cluster 1	Cluster 2	Cluster 3
B = (2,5)	D = (5,1)	A = (2,10)
~~C = (8,4)~~	E = (7,5)	H = (4,9)
G = (1,2)	F = (6,4)	
	I = (10,10)	
	C = (8,4)	

Round-1 (contd)

Calculating the centroids of the clusters

$$c_1 = \left(\frac{2+1}{2}, \frac{5+2}{2}\right) = (1.5, 3.5)$$

$$c_2 = \left(\frac{5+7+6+10+8}{5}, \frac{9+5+4+10+4}{5}\right) = (7.2, 6.4)$$

$$c_3 = \left(\frac{2+4}{2}, \frac{10+9}{2}\right) = (3, 9.5)$$

Round-2 :

Assign points to closest centroids

$d_{c_1A} = 6.52$, $d_{c_2A} = 6.32$, $d_{c_3A} = 1.12 \Rightarrow$ Assign A to c_3

$d_{c_1B} = 1.58$, $d_{c_2B} = 5.38$, $d_{c_3B} = 4.6 \Rightarrow$ Assign B to c_1

$d_{c_1C} = 6.52$, $d_{c_2C} = 2.53$, $d_{c_3C} = 7.43 \Rightarrow$ Assign C to C_2

$d_{c_1D} = 6.52$ $d_{c_2D} = 3.4$, $d_{c_3D} = 2.06 \Rightarrow$ Assign D to c_3

$d_{c_1E} = 5.7$, $d_{c_2E} = 1.41$ $d_{c_3E} = 6.02 =$ Assign E to C_2

$d_{c_1F} = 4.53$, $d_{c_2F} = 2.68$ $d_{c_3F} = 6.26 \Rightarrow$ Assign F to C_2

$d_{c_1G} = 1.58$, $d_{c_2G} = 7.60$ $d_{c_3G} = 7.76 \Rightarrow$ Assign G to c_1

$d_{c_1H} = 6.04$, $d_{c_2H} = 4.20$, $d_{c_3H} = 1.12 \Rightarrow$ Assign H to c_3

$d_{c_1I} = 10.7$, $d_{c_2I} = 4.56$ $d_{c_3I} = 7.02 \Rightarrow$ Assign I to C_2

Cluster 1	Cluster 2	Cluster 3
B = (2,5)	C = (8,4)	H = (4,9)
G = (1,2)	E = (7,5)	D = (5,9)
	F = (6,4)	A = (2,10)
	I = (10,10)	

Round-2 (contd)

$C_1 = \left(\frac{2+1}{2}, \frac{5+2}{2}\right) = (1.5, 3.5)$

$C_2 = \left(\frac{8+7+6+10}{4}, \frac{4+5+4+10}{4}\right) = (7.75, 5.75)$

$C_3 = \left(\frac{4+5+2}{3}, \frac{9+9+10}{3}\right) = (3.67, 9.33)$

Round 3:

$d_{c_1 A} = 6.52$, $d_{c_2 A} = 7.15$, $d_{c_3 A} = 1.79 \Rightarrow$ Assign A to C_3

$d_{c_1 B} = 1.58$, $d_{c_2 B} = 5.79$, $d_{c_3 B} = 4.64 \Rightarrow$ Assign B to C_1

$d_{c_1 C} = 6.52$, $d_{c_2 C} = 1.76$, $d_{c_3 C} = 6.86 \Rightarrow$ Assign C to C_2

$d_{c_1 D} = 6.52$, $d_{c_2 D} = 4.25$, $d_{c_3 D} = 1.37 \Rightarrow$ Assign D to C_3

$d_{c_1 E} = 5.7$, $d_{c_2 E} = 1.06$, $d_{c_3 E} = 5.46 \Rightarrow$ Assign E to C_2

$d_{c_1 F} = 4.53$, $d_{c_2 F} = 2.47$, $d_{c_3 F} = 5.81 \Rightarrow$ Assign F to C_2

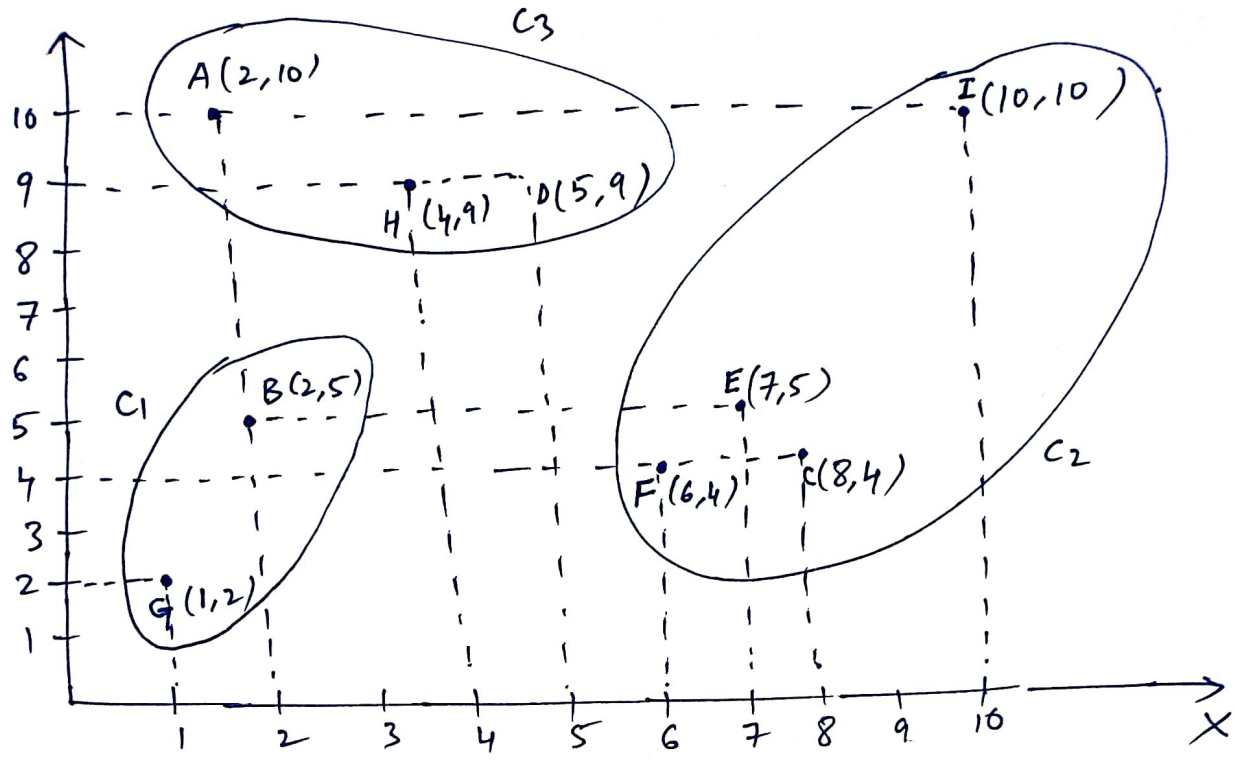
$d_{c_1 G} = 1.58$, $d_{c_2 G} = 7.72$, $d_{c_3 G} = 7.8 \Rightarrow$ Assign G to C_1

$d_{c_1 H} = 6.03$, $d_{c_2 H} = 4.96$, $d_{c_3 H} = 0.46 \Rightarrow$ Assign H to C_3

$d_{c_1 I} = 10.7$, $d_{c_2 I} = 4.8$, $d_{c_3 I} = 6.36 \Rightarrow$ Assign I to C_2

Cluster 1	Cluster 2	Cluster 3
B = (2,5)	C = (8,4)	H = (4,9)
G = (1,2)	E = (7,5)	D = (5,9)
	F = (6,4)	A = (2,10)
	I = (10,10)	

Since the clusters don't change in Round 3, this is the final set of clusters.



Cluster centres :

For C_1 : (1.5, 3.5)

For C_2 : (7.75, 5.75)

for C_3 : (3.67, 9.33)

QUESTION B :

(i). Single link hierarchial clustering.

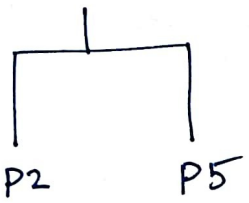
Round-1:

Clusters P2 and P5 are most similar with a value of 0.98.

Initial Clusters :

{P1}, {P2}, {P3}, {P4}, {P5}

Dendrogram :



Updating the proximity matrix.

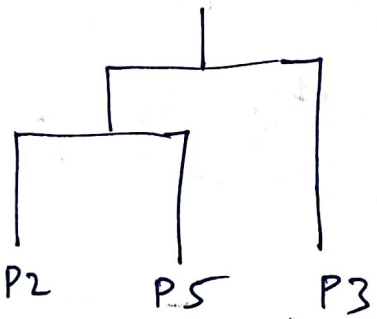
	(P2,P5)	P1	P3	P4
(P2,P5)	1	0.35	0.85	0.76
P1	0.35	1	0.41	0.55
P3	0.85	0.41	1	0.44
P4	0.76	0.55	0.44	1

eg. $sim\{(P2,P5), P1\}$
$= \max\{sim(P2,P1), sim(P5,P1)\}$
$= \max(0.1, 0.35)$
$= 0.35$.

Round-2:

Clusters {P2,P5} and {P3} are most similar with a value of 0.85, and so we merge them.

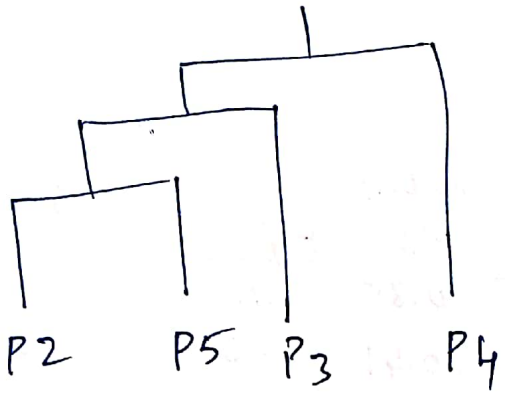
Dendrogram:



Updated proximity matrix.

	(P2, P5, P3)	P1	P4
(P2, P5, P3)	1	0.41	0.76
P1	0.41	1	0.55
P4	0.76	0.55	1

Round-3:
Clusters {P2, P5, P3} and {P4} are most similar with a value of 0.76, hence we merge them.

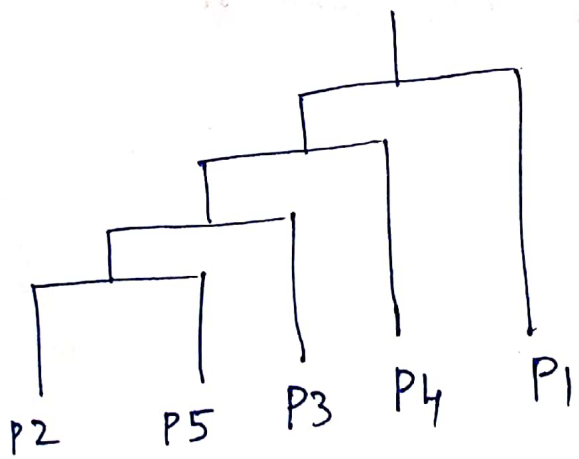


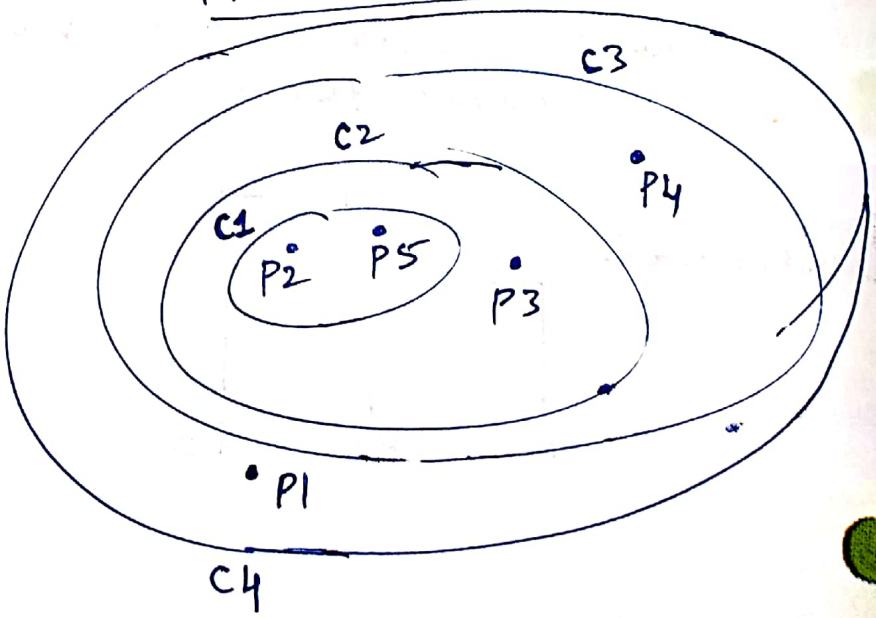
~~Round~~

Updated similarity matrix is not required.

Since only 2 clusters are left, we merge them.

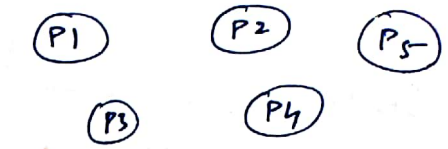
Final hierarchial clustering.





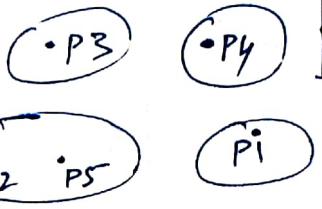
Complete link hierarchial clustering:

Initial clusters:

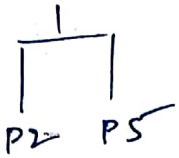


Round-1:

Clusters {P2} and {P5} have the most similarity (value = 0.98), so, we merge them.



Dendrogram:



Updated similarity matrix:

	(P2,P5)	P1	P3	P4
(P2,P5)	1	0.1	0.64	0.47
P1	0.1	1	0.41	0.55
P3	0.64	0.41	1	0.44
P4	0.47	0.55	0.44	1

Example:

sim{(P2,P5),P1}

= min{ sim(P2,P1), sim(P5,P1)}

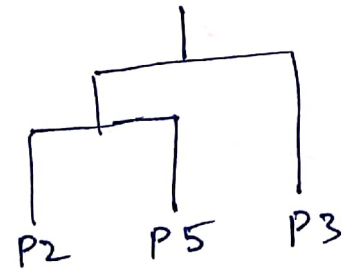
= min(0.1, 0.35)

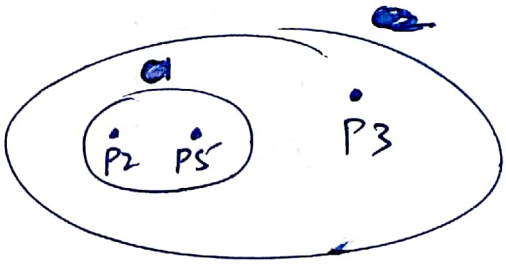
= 0.1

Round-2:

(P2,P5) and P3 have the maximum similarity value = 0.64, so we merge them.

Dendrogram:





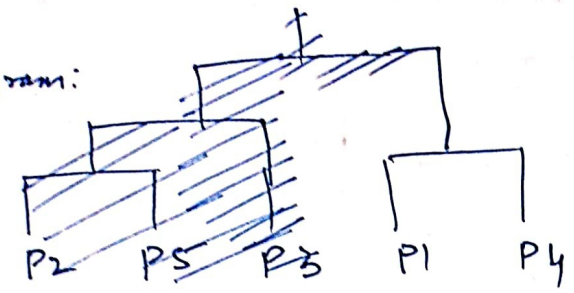
Updated similarity matrix:

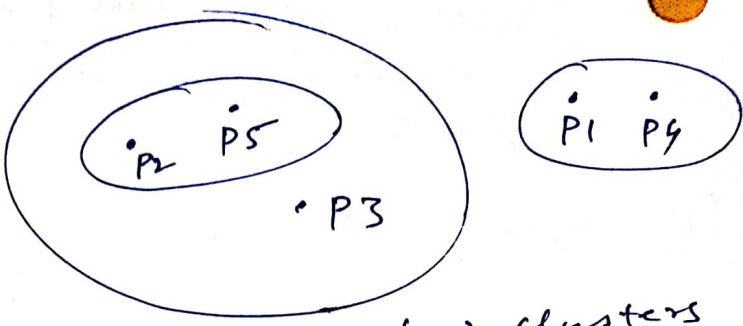
	(P2,P5,P3)	P1	P4
(P2,P5,P3)	1	0.1	0.44
P1	0.1	1	0.55
P4	0.44	0.55	1

Round-3:

Clusters {P1} and {P4} have the highest similarity value of 0.55, so merge them.

Dendrogram:





Updated Similarity matrix: Not required (since only 2 clusters left)

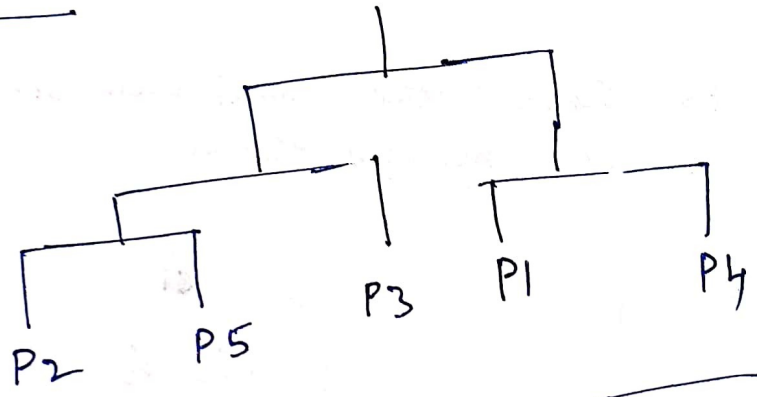
Round-4:

Since only 2 clusters remain:

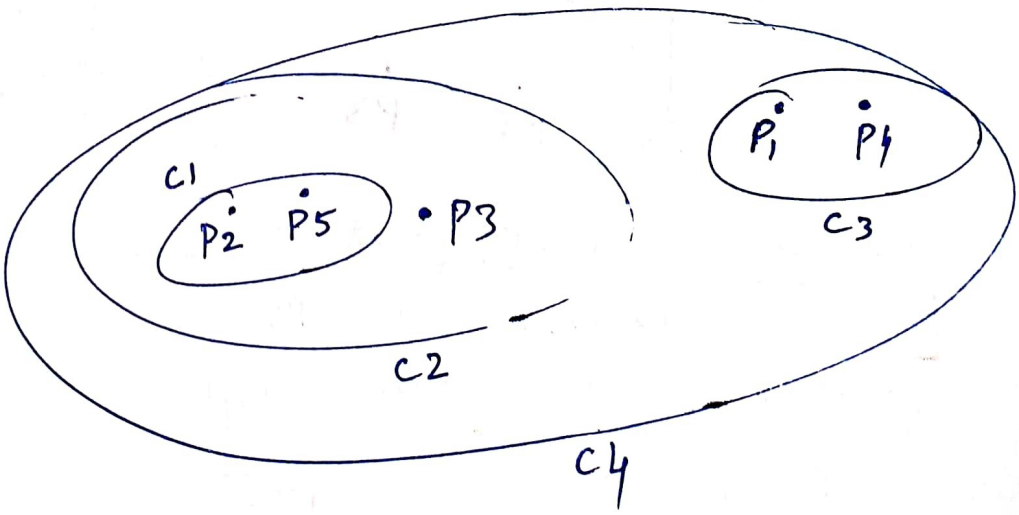
{(P2, P5), P3} and {P1, P4}

we merge them.

Final Dendrogram:



Final hierarchial cluster:



QUESTION C:

(a). Set of points: A = 6, B = 12, C = 18, D = 24
E = 25, F = 28, G = 30, H = 42,
I = 48.

(a). set of centroids:

$c_1 = 5$
$c_2 = 7.5$

$dc_1A = 1$, $dc_2A = 1.5$ ⇒ Assign A to c_1

$dc_1B = 7$, $dc_2B = 4.5$ ⇒ Assign B to c_2

$dc_1C = 13$, $dc_2C = 10.5$ ⇒ assign C to c_2

$dc_1D = 19$, $dc_2D = 16.5$ ⇒ assign D to c_2

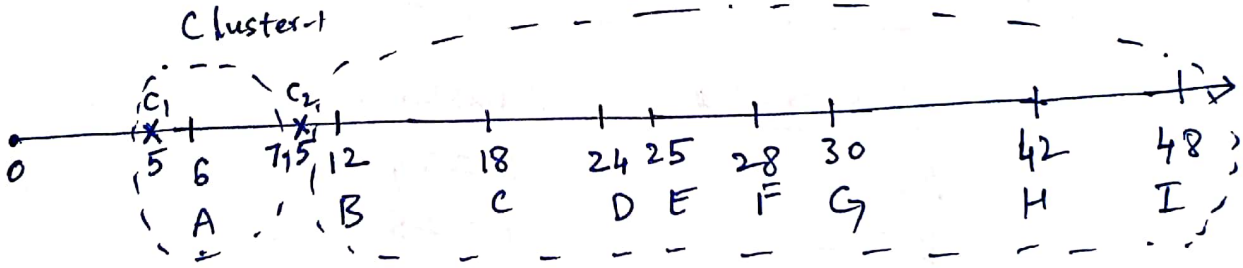
$dc_1E = 20$, $dc_2E = 17.5$ ⇒ assign E to c_2

$dc_1F = 23$, $dc_2F = 20.5$ ⇒ assign F c_2

$dc_1G = 25$, $dc_2G = 22.5$ ⇒ assign G c_2

$dc_1H = 37$, $dc_2H = 34.5$ ⇒ assign H c_2

$dc_1I = 43$, $dc_2I = 40.5$ ⇒ assign I c_2



Cluster 1	Cluster 2
A = 6	B = 12
	C = 18
	D = 24
	E = 25
	F = 28
	G = 30
	H = 42
	I = 48

Total squared error.

$= (1)^2 + (4.5)^2 + (10.5)^2 + (16.5)^2 + (17.5)^2 + (20.5)^2 + (22.5)^2 + (34.5)^2 + (40.5)^2$

$= 4467$

(2). Set of centroids:

$C_1 = 15$

$C_2 = 25$

$d_{C_1 A} = 9$, $d_{C_2 A} = 19$ $\Rightarrow$ Assign A to C_1

$d_{C_1 B} = 3$, $d_{C_2 B} = 13$ $\Rightarrow$ assign B to C_1

$d_{C_1 C} = 3$, $d_{C_2 C} = 7$ $\Rightarrow$ assign C to C_1

$d_{C_1 D} = 9$, $d_{C_2 D} = 1$ $\Rightarrow$ assign D to C_2

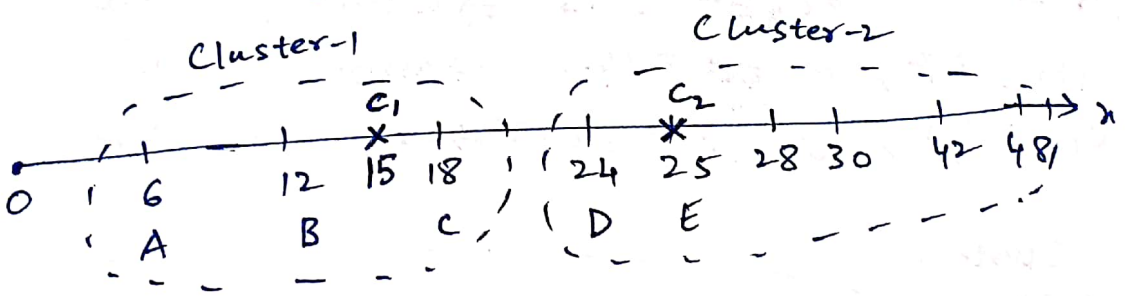
$d_{C_1 E} = 10$, $d_{C_2 E} = 0$ $\Rightarrow$ assign E to C_2

$d_{C_1 F} = 13$, $d_{C_2 F} = 3$ $\Rightarrow$ assign F to C_2

$d_{C_2 G} = 15$, $d_{C_2 G} = 5$ $\Rightarrow$ assign G to C_2

$d_{C_1 H} = 27$, $d_{C_2 H} = 17$ $\Rightarrow$ assign H to C_2

$d_{C_1 I} = 33$, $d_{C_2 I} = 23$ $\Rightarrow$ assign I to to C_2



Cluster-1

A = 6

B = 12

C = 18

Cluster-2

D = 24

E = 25

F = 28

G = 30

H = 42

I = 48.

Total ~~sum of~~ squared errors $= 9^2 + 3^2 + 3^2 + 1^2 + 0^2 + 3^2 + 5^2 + 17^2 + 23^2$

$= 898$.

(b). Centroid set 1

$c_1 = 5$

$c_2 = 7.5$.

Assume the points were allocated as in part a.

Recalculating the new centroids

$c_1 = 6$

$c_2 = \frac{12+18+24+25+28+30+42+48}{8} = 28.37$.

$d_{c_1 A} = 0$, $d_{c_2 A} = 22.37 \Rightarrow$ assign A to c_1

$d_{c_1 B} = 6$, $d_{c_2 B} = 16.37 \Rightarrow$ assign B to c_1

$d_{c_1 C} = 12$, $d_{c_2 C} = 10.37 \Rightarrow$ assign C to c_2

$d_{c_1 D} = 18$, $d_{c_2 D} = 4.37 \Rightarrow$ assign D to c_2

$d_{c_1 E} = 19$, $d_{c_2 E} = 3.37 \Rightarrow$ assign E to c_2

$d_{c_1 F} = 22$, $d_{c_2 F} = 0.37 \Rightarrow$ assign F to c_2

$d_{c_1 G} = 24$, $d_{c_2 G} = 1.63 \Rightarrow$ assign G to c_2

$d_{c_1 H} = 36$, $d_{c_2 H} = 13.63 \Rightarrow$ assign H to c_2

$d_{c_1 I} = 42$, $d_{c_2} = 19.63 \Rightarrow$ assign I to c_2

Cluster-1:	Cluster 2
A = 6	C = 18
B = 12	D = 24
	E = 25
	F = 28
	G = 30
	H = 42
	I = 48

Since there is a change in points in each cluster, centroid set = (5, 7.5) is not a stable solution.

Centroid set - 2 :

Assume that the points were allocated to the 2 clusters as in part (a)

Calculating the centroids (updated)

$$C_1 = \left(\frac{6+12+18}{3}\right), C_2 = \left(\frac{24+25+28+30+42+48}{6}\right)$$

$$C_1 = 12$$

$$C_2 = 32.83$$

Round 2 :

Assign points to the 2 clusters.

$d_{C_1A} = 6$, $d_{C_2A} = 26.83$ ⇒ assign A to C_1

$d_{C_2B} = 0$, $d_{C_2B} = 20.83$ ⇒ assign B to C_1

$d_{C_1C} = 6$, $d_{C_2C} = 14.83$ ⇒ assign C to C_1

$d_{C_1D} = 12$, $d_{C_2D} = 8.83$ ⇒ assign D to C_2

$d_{C_1E} = 13$, $d_{C_2E} = 7.83$ ⇒ assign E to C_2

$d_{C_1F} = 16$, $d_{C_2F} = 4.83$ ⇒ assign F to C_2

$d_{C_1G} = 18$, $d_{C_2G} = 2.83$ ⇒ assign G to C_2

$d_{C_1H} = 30$, $d_{C_2F} = 9.17$ ⇒ assign H to C_2

$d_{C_1I} = 36$, $d_{C_2I} = 15.17$ ⇒ assign I to C_2

Cluster-1	Cluster-2
A = 6	D = 24
B = 12	E = 25
C = 18	F = 28
	G = 30
	H = 42
	I = 48

∴ Centroid set = (15, 25) represents a stable solution.

(c). MIN / single link clustering.

Proximity matrix:

	A	B	C	D	E	F	G	H	I
A	0	6	12	18	19	22	24	36	42
B	6	0	6	12	13	16	18	30	36
C	12	6	0	6	7	10	12	24	30
D	18	12	6	0	1	4	6	18	24
E	19	13	7	1	0	3	5	17	23
F	22	16	10	4	3	0	2	14	20
G	24	18	12	6	5	2	0	12	16
H	36	30	24	18	17	14	12	0	6
I	42	36	30	24	23	20	16	6	0

Initially:

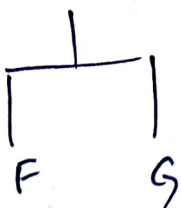
(A) (B) (C) (D) (E) (F) (G) (H) (I)

Round 1:

Clusters containing D and E are closest, so merge them (DE = 1)

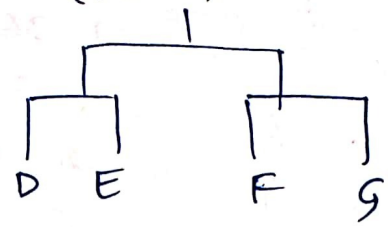
Round-2:

Clusters {F} and {G} are closest so merge them. (FG = 2)



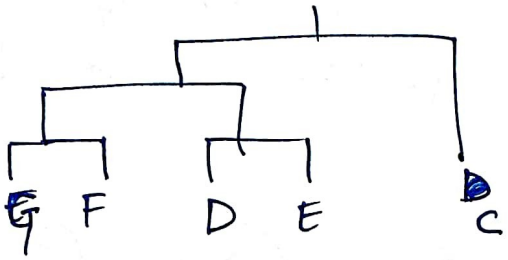
Round-3:

Clusters {D, E} and {F, G} are closest, so merge them (FE=3)



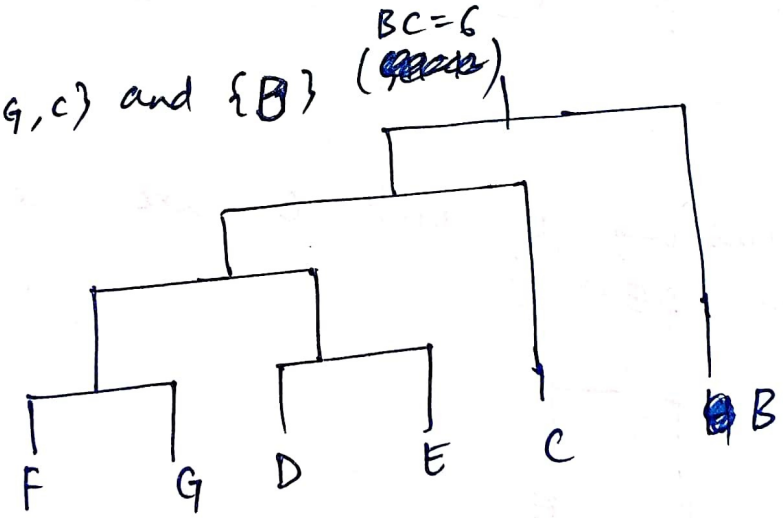
Round-4:

Clusters {C} and {D, E, F, G} are closest, so we merge them (CE=7)



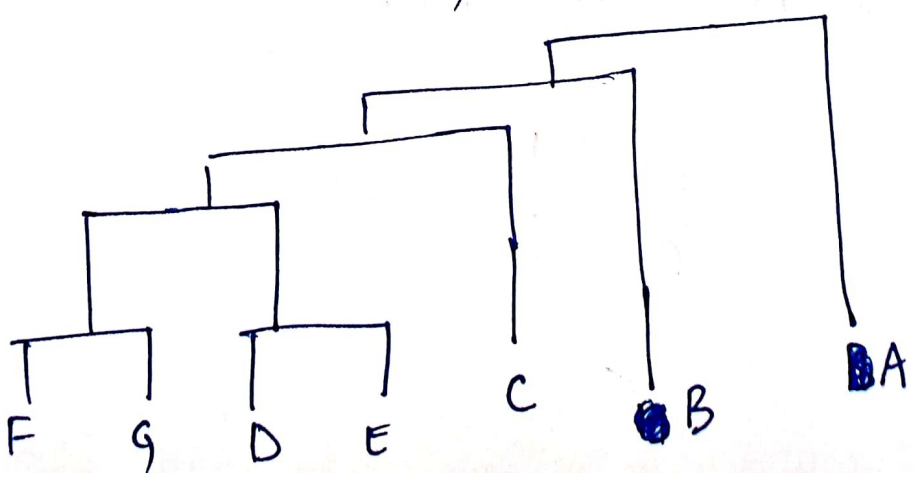
Round-5:

Merge {E, F, D, G, C} and {B} (BC=6)



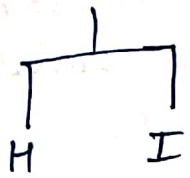
Round-6:

Merge {F, G, D, E, C, H} and {A} (AB=6)



Round-7:

We Merge {H} and {I} since HI = 6 is the lowest distance between any 2 clusters.



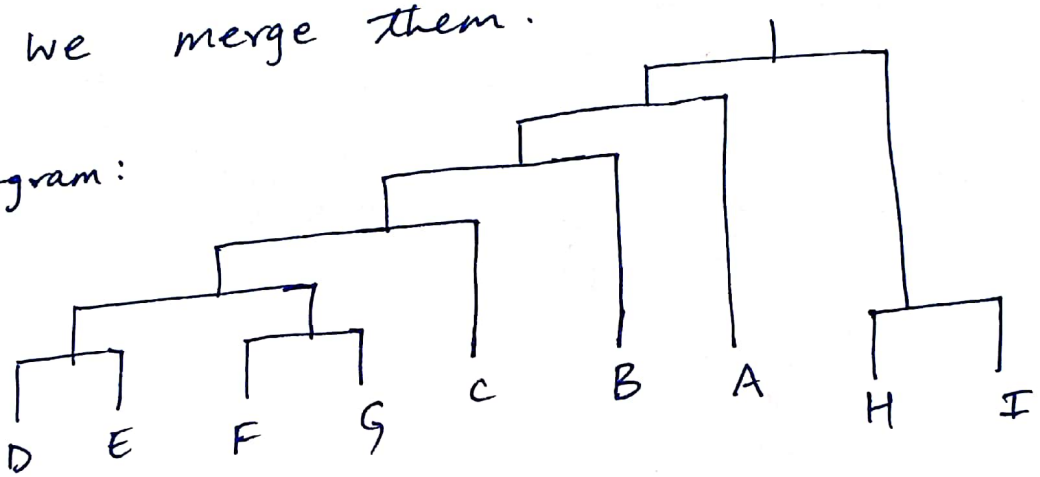
Round-8:

Since only 2 clusters remain:

{A, B, C, D, E, F, G} and {H, I}

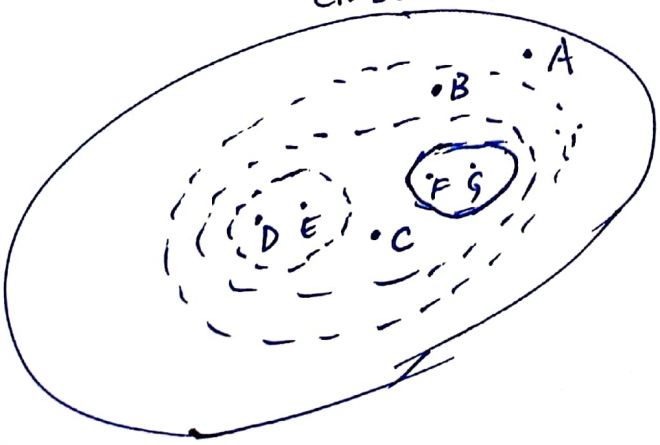
we merge them.

Final dendrogram:

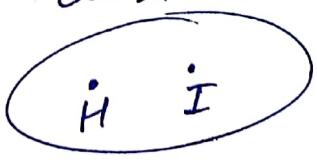


If we want 2 clusters, we must go down one level in the final dendrogram

Cluster-1



Cluster-2



Cluster-1:

A = 6
B = 12
C = 18
D = 24
E = 25
F G = 28
C H = 30

Cluster-2:

H = 42
I = 48

(d). K-means seems to produce the most natural clustering.

(e). The K-means algorithm has the advantage of centroid updation for each cluster in each round, which helps in similar points getting merged into the same cluster.

PART II : CLASSIFICATION

QUESTION D :

Label the examples from 1 to 11.

$S_0 = \{1,2,3,4,5,6,7,8,9,10,11\}$

Target attribute : y

$x_1 \in \{a,b\}$, $x_2 \in \{c,g,u,w\}$, $x_3 \in \{k,s,v\}$

y creates a partition of (5,6)

$\text{Entropy}(S_0) = \text{Entropy}\left(\frac{5}{11}, \frac{6}{11}\right) = 0.517 + 0.476 = 0.9939$

Selecting attribute

$x_1 = a$ creates a partition of (2,3)

$\text{Entropy}(S_0 | x_1 = a) = \text{Entropy}\left(\frac{2}{5}, \frac{3}{5}\right) = 0.823.$

$x_1 = b$ creates a partition of (3,3)

$\text{Entropy}(S_0 | x_1 = b) = \text{Entropy}\left(\frac{3}{6}, \frac{3}{6}\right) = 1$

$\therefore \text{Entropy}(S_0 | x_1) = \frac{5}{11}(0.823) + \frac{6}{11}(1) = 0.9195$ ——— (1).

$x_2 = c$ creates a partition of (3,2)

$\text{Entropy}(S_0 | x_2 = c) = 0.823$

$x_2 = g$ creates a partition of (1,0)

$\text{Entropy}(S_0 | x_2 = g) = \text{Entropy}(1,0) = 0$

$x_2 = u$ creates a partition of (0,1)

$\text{Entropy}(S_0 | x_2 = u) = \text{Entropy}(0,1) = 0$

$x_2 = w$ Creates a partition of (2,2)

$\text{Entropy}(S_0 | x_2 = w) = 1$

$\therefore \text{Entropy}(S_0 | x_2) = \frac{5}{11} \times (0.823) + \frac{1}{11}(0) + \frac{1}{11}(0) + \frac{4}{11}(1)$

$= 0.7377$ ——— (2).

$x_3 = K$ creates a partition of (0,3)

Entropy $(S_0 | x_3 = K) = Ent(0, 1) = 0$

$x_3 = V$ creates a partition of (2,3)

Entropy $(S_0 | x_3 = V)$ = Entropy $\left(\frac{2}{5}, \frac{3}{5}\right) = 0.823$

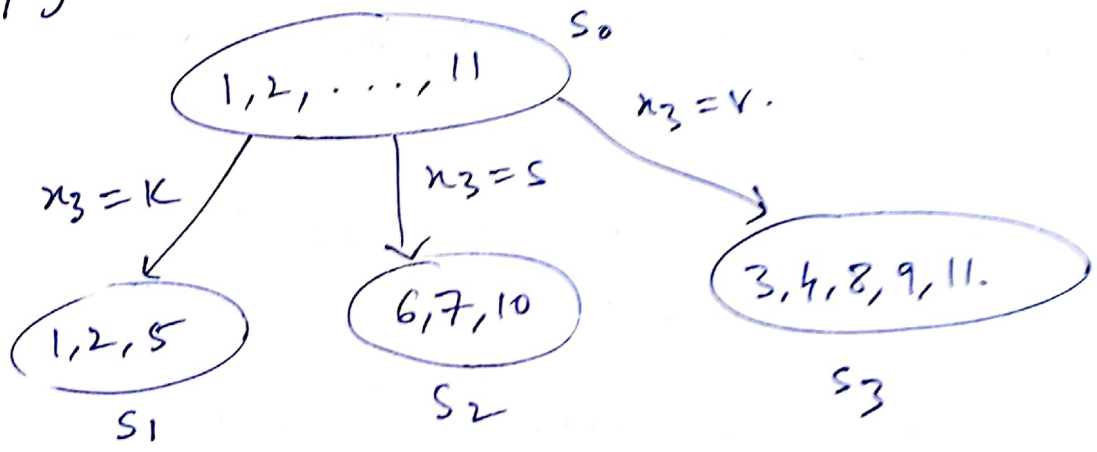
$x_3 = S$ creates a partition (3,0)

Entropy $(S_0 | x_3 = S)$ = Entropy $(1, 0) = 0$

$\therefore$ Entropy $(S_0 | x_3) = \frac{3}{11}(0) \quad \frac{5}{11}(0.823) + \frac{3}{11}(0)$

$= 0.374.$

We split based on the attribute having lowest entropy.



All examples in S_1 have same Y.
All examples in S_2 have same Y

$S_3 = \{3,4,8,9,11\}$.

$x_1 = a$ creates a partition of (1, 1)

Entropy$(S_3 | x_1 = a) = 1$

$x_1 = b$ creates a partition of (1, 2)

Entropy$(S_3 | x_1 = b) = 0.9179$.

Entropy $(S_3 | x_1) = \frac{2}{5}(1) + \frac{3}{5}(0.9179) = 0.9507$ ——— (3)

$x_2 = c$ creates a (1, 1) partition

Entropy$(S_3 | x_3 = c)$ = Entropy $(\frac{1}{2}, \frac{1}{2}) = 1$

$x_2 = g$ creates a partition (0, 1)

Entropy $(S_3 | x_2 = g) = 0$

$x_3 = u$ creates partition (0, 01)

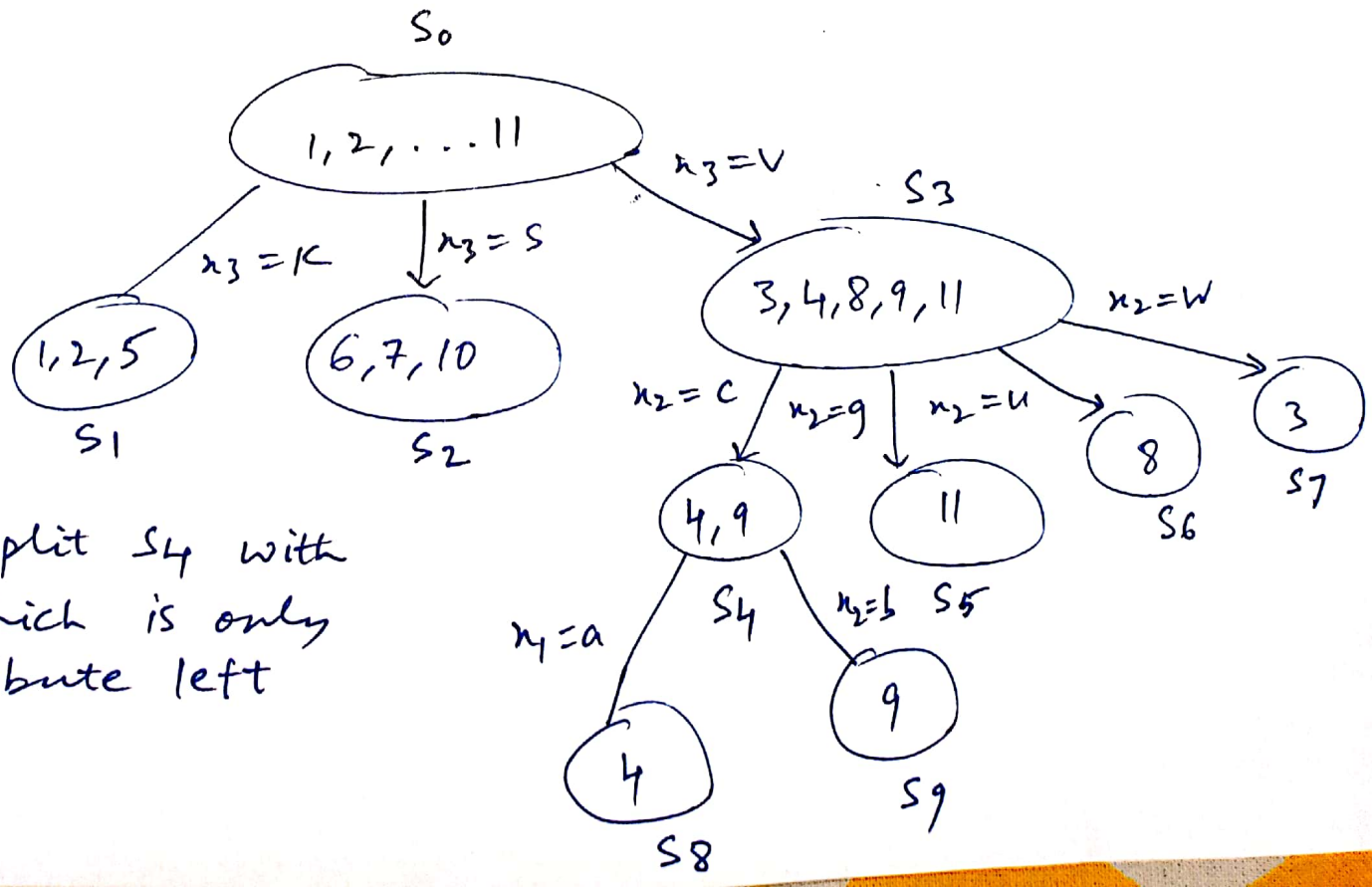
Entropy $(S_3 | x_2 = u) = 0$

$x_3 = w$ creates a (1, 0) partition

Entropy $(S_3 | x_2 = w) = 0$.

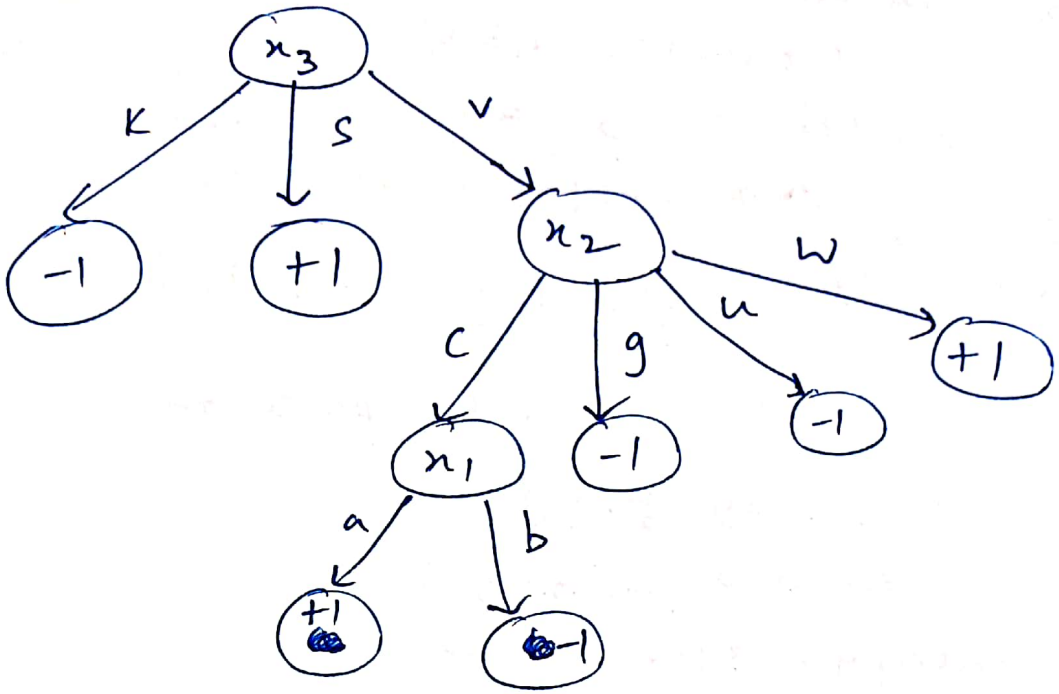
Entropy $(S_3 | x_2) = \frac{2}{5}(1) + \frac{1}{5}(0) + \frac{1}{5}(0) + \frac{1}{5}(0)$

$= 0.4$ ——— (4)



We split S_4 with x_1 which is only attribute left

Final Decision tree :



(b). $x_1 \in \{a, b\}$

$x_2 \in \mathbb{N}$

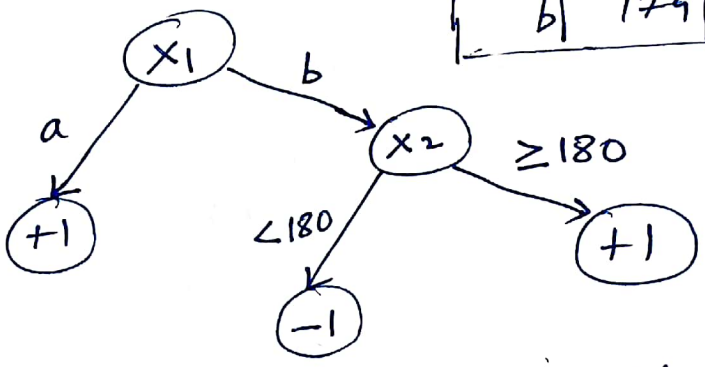
$x_3 \in \{e, f\}$

$x_4 \in \{c, d\}$

$y \in \{+1, -1\}$

x_1	x_2	x_3	x_4	y
b	185	f	d	+1
b	180	f	c	+1
b	170	f	c	-1
b	140	e	d	-1
a	176	f	d	+1
b	179	f	d	-1

i). Decision tree:



The decision tree has a 100% accuracy on the training data.

(ii). $x_1 = b, x_2 = 170, x_3 = f, x_4 = d \Rightarrow$ pred $= -1$ (correct)

$x_1 = a, x_2 = 150, x_3 = f, x_4 = d \Rightarrow$ pred $= +1$ (correct)

$x_1 = b, x_2 = 60, x_3 = f, x_4 = d \Rightarrow$ pred $= -1$ (incorrect)

Test accuracy: 0.667.