Connider datanet

Feature	Sample 1	Sample 2	Sample 3	Sample 4
a	4	8	13	7
b	11	4	5	14

Step: 1

Number of featurer, $n = 2$ (a, b)
Number of nampler, $N = 4$ (nample 1, nample 2, nample 3, nample 4)

Step: 2

Calculating Mean,

$$\bar{a} = \frac{4+8+13+7}{4} = 8$$

$$\bar{b} = \frac{11+4+5+14}{4} = 8.5$$

Step: 3

Calculating covariance matrix, between featurer in the given datanet, ordered featurer are an,

(a, a), (a, b), (b, a), (b, b)

$$cov(a,a) = \frac{1}{N-1} \sum_{k=1}^{N} (a_i - \bar{a})(a_i - \bar{a})$$

$$= \frac{1}{N-1} \sum_{k=1}^{N} (a_i - \bar{a})^2 \longrightarrow \text{for same feature}$$

$$= \frac{1}{4-1} \left[(4-8)^2 + (8-8)^2 + (13-8)^2 + (7-8)^2 \right]$$

$$= 14$$

$$cov(a,b) = \frac{1}{N-1} \sum_{k=1}^{N} (a_i - \bar{a})(b_i - \bar{b})$$

$$= \frac{1}{4-1} \left[(4-8)(11-8.5) + (8-8)(4-8.5) + (13-8)(5-8.5) + (7-8)(14-8.5) \right]$$

$$= -11$$

$$cov(b,a) = \frac{1}{N-1} \sum_{k=1}^{N} (b_i - \bar{b})(a_i - \bar{a})$$

$$= cov(a,b) = -11$$

$$cov(b,b) = \frac{1}{N-1} \sum_{k=1}^{N} (b_i - \bar{b})(b_i - \bar{b})$$

$$= \frac{1}{N-1} \sum_{k=1}^{N} (b_i - \bar{b})^2$$

$$= \frac{1}{4-1} \left[(11-8.5)^2 + (4-8.5)^2 + (5-8.5)^2 + (14-8.5)^2 \right]$$

$$= 23$$

Hence, the covariance matrix can be

$$S = \begin{bmatrix} cov(a,a) & cov(a,b) \\ cov(b,a) & cov(b,b) \end{bmatrix} = \begin{bmatrix} 14 & -11 \\ -11 & 23 \end{bmatrix}$$

Step:4

Calculate Eigen value, Eigen vector, Normalized Eigen vector.

Inorder calculate Eigen value,

$$\det(S - \lambda I) = 0$$

$$I \text{ (Identity matrix)} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

$$\lambda I = \lambda \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} = \begin{bmatrix} \lambda & 0 \\ 0 & \lambda \end{bmatrix}$$

$$\det\left(\begin{bmatrix} 14 & -11 \\ -11 & 23 \end{bmatrix} - \begin{bmatrix} \lambda & 0 \\ 0 & \lambda \end{bmatrix} \right) = 0$$

$$\det \begin{pmatrix} 14-\lambda & -11 \\ -11 & 23-\lambda \end{pmatrix} = 0$$

$$(14-\lambda)(23-\lambda) - (-11 \times -11) = 0$$

$$201 - 37\lambda + \lambda^2 = 0$$

After rearranging,

$$\lambda^2 - 37\lambda + 201 = 0$$

'λ' can be calculated by quadratic equ^n,

$$x = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a} \qquad [a = 1, b = -37, c = 201]$$

$$= \frac{(-37) \pm \sqrt{(37)^2 - 4(1)(201)}}{2(1)}$$

$$= \frac{37 \pm \sqrt{565}}{2}$$

$$= \frac{37 \pm 23.76}{2} \Rightarrow \frac{37 + 23.76}{2}, \frac{37 - 23.76}{2}$$

$$= \frac{60.76}{2}, \frac{13.24}{2}$$

Eigen values $\lambda_1 = 30.38$, $\lambda_2 = 6.62$

So, while arranging in descending order,

$\lambda_1 > \lambda_2 > \ldots\ldots$

Hence,

$\lambda_1 = 30.38$ & $\lambda_2 = 6.62$

We are going to find out Eigen vector for Eigen value,

$\lambda = 30.38$.

$(S - \lambda_1 I)\, U_1 = 0$

[S = covariance matrix
$\lambda_1 = 30.38$, I = Identity matrix
U_1 = Eigen vector of λ_1]

$$\left(\begin{pmatrix} 14 & -11 \\ -11 & 23 \end{pmatrix} - 30.38 \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}\right) U_1 = 0$$

Assume $U_1 = \begin{bmatrix} v_1 \\ v_2 \end{bmatrix}$

Hence,

$$\begin{pmatrix} 14 & -11 \\ -11 & 23 \end{pmatrix} - \begin{pmatrix} 30.38 & 0 \\ 0 & 30.38 \end{pmatrix} \begin{bmatrix} v_1 \\ v_2 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

$$\begin{pmatrix} 14 - 30.38 & -11 \\ -11 & 23 - 30.38 \end{pmatrix} \begin{bmatrix} v_1 \\ v_2 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

$$\begin{pmatrix} -16.38 & -11 \\ -11 & -7.38 \end{pmatrix} \begin{bmatrix} v_1 \\ v_2 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

$$-16.38\, v_1 - 11 v_2 = 0 \quad \text{——— (1)}$$

$$-11 v_1 - 7.38\, v_2 = 0 \quad \text{——— (2)}$$

So, from this if we need to calculate U_1 & U_2,

$$(1) \times 7.38 \Rightarrow 120.88 U_1 - 81.18 U_2 = 0$$

$$(2) \times -11 \Rightarrow +121 U_1 + 81.18 U_2 = 0$$

$$0.12 U_1 = 0$$

$$U_1 = 0$$

then, apply U_1 in (1), then

$$-16.38 \times 0 - 11 U_2 = 0$$

$$U_2 = 0$$

This can't be possible, hence

$$\begin{bmatrix} (14-\lambda_1) & (-11) \\ (-11) & (23-\lambda_1) \end{bmatrix} \begin{bmatrix} U_1 \\ U_2 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

$$(14-\lambda_1) U_1 - 11 U_2 = 0 \longrightarrow (a)$$

$$-11 U_1 + (23-\lambda_1) U_2 = 0 \longrightarrow (b)$$

from (a)

$$(14-\lambda_1) U_1 - 11 U_2 = 0$$

$$(14-\lambda_1) U_1 = 11 U_2$$

$$\frac{U_1}{11} = \frac{U_2}{14-\lambda_1} = A \text{ (Annigning)}$$

Annume $A = 1$

$$\frac{U_1}{11} = \frac{U_2}{14-\lambda_1} = A = 1$$

Hence,

$$\frac{U_1}{11} = 1 \Rightarrow U_1 = 11.$$

$$\frac{U_2}{14-\lambda_1} = 1 \Rightarrow U_2 = 14-\lambda_1 = 14-30.38 = -16.38$$

Hence Eigen vector for $\lambda_1 \Rightarrow \begin{bmatrix} U_1 \\ U_2 \end{bmatrix} = \begin{bmatrix} 11 \\ -16.38 \end{bmatrix}$

Then if we want to normalize the eigen vector,

$$n_1 = \begin{bmatrix} 11 / \sqrt{11^2+16.38^2} \\ -16.38 / \sqrt{11^2+16.38^2} \end{bmatrix}$$

$$= \begin{bmatrix} \frac{11}{19.73} \\ -16.38/19.73 \end{bmatrix} = \begin{bmatrix} 0.5575 \\ -0.8302 \end{bmatrix}$$

Now, calculate eigen vector for $\lambda_2 = 6.62$

$$(S-\lambda_2 I) U_2 = 0$$

$$\begin{bmatrix} (14-\lambda_2) & -11 \\ -11 & (23-\lambda_2) \end{bmatrix} \begin{bmatrix} U_1 \\ U_2 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

$$(14-\lambda_2) U_1 - 11 U_2 = 0 \longrightarrow (c)$$

$$-11 U_1 - (23-\lambda_2) U_2 = 0 \longrightarrow (d)$$

from (c),

$$(14-\lambda_2) U_1 - 11 U_2 = 0$$

$$\frac{U_1}{11} = \frac{U_2}{14-\lambda_2} = B \text{ (Assume)}$$

Annume $B = 7$,

$$\frac{U_1}{11} = \frac{U_2}{14-\lambda_2} = B = 1$$

Hence, $\frac{U_1}{11} = 1 \Rightarrow U_1 = 11$

$$\frac{U_2}{14-\lambda_2} = 1 \Rightarrow U_2 = 14 - \lambda_2 = 7.38$$

Hence, Eigen vector for $\lambda_2 = \begin{bmatrix} 11 \\ 7.38 \end{bmatrix}$

If we want to normalize eigen vector,

$$n_2 = \begin{bmatrix} 11/\sqrt{11^2+7.38^2} \\ 7.38/\sqrt{11^2+7.38^2} \end{bmatrix} = \begin{bmatrix} 0.8308 \\ 0.5574 \end{bmatrix}$$

Step 5

New datanet,

Feature	Sample 1	Sample 2	Sample 3	Sample 4
a	4	8	13	7
b	11	4	5	14

Int PC	P_{11}	P_{12}	P_{13}	P_{14}
	Sample 1	Sample 2	Sample 3	Sample 4

$$P_{11} = n_1^T \begin{bmatrix} 4-8 \\ 11-8.5 \end{bmatrix}$$

$$= \begin{bmatrix} 0.5575 & -0.8302 \end{bmatrix} \begin{bmatrix} -4 \\ 2.5 \end{bmatrix}$$

$$= (-2.23 - 2.0755) = -4.305$$

$$P_{12} = n_1^T \begin{bmatrix} 8-8 \\ 4-8.5 \end{bmatrix} \Rightarrow (0.5575 \quad -0.8302) \begin{pmatrix} 0 \\ -4.5 \end{pmatrix}$$

$$= 0 + 3.7359 = 3.7359$$

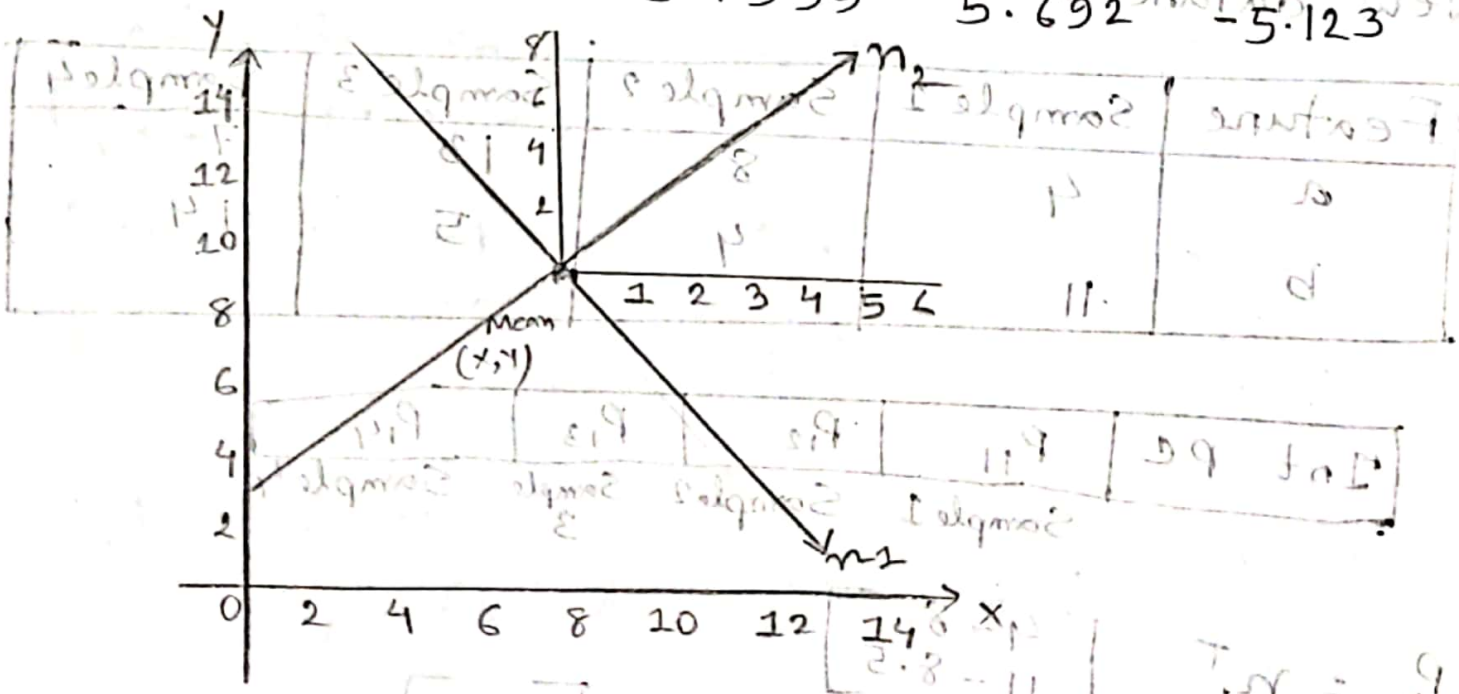
$$P_{13} = n_1^T \begin{bmatrix} 13-8 \\ 5-8.5 \end{bmatrix} \Rightarrow (0.5575 \quad -0.8302) \begin{pmatrix} 5 \\ -3.5 \end{pmatrix}$$

$$= 2.787 + 2.905 = 5.692$$

$$P_{14} = n_1^T \begin{bmatrix} 7-8 \\ 4-8.5 \end{bmatrix} \Rightarrow (0.5575 \quad -0.8302) \begin{pmatrix} -1 \\ 5.5 \end{pmatrix}$$

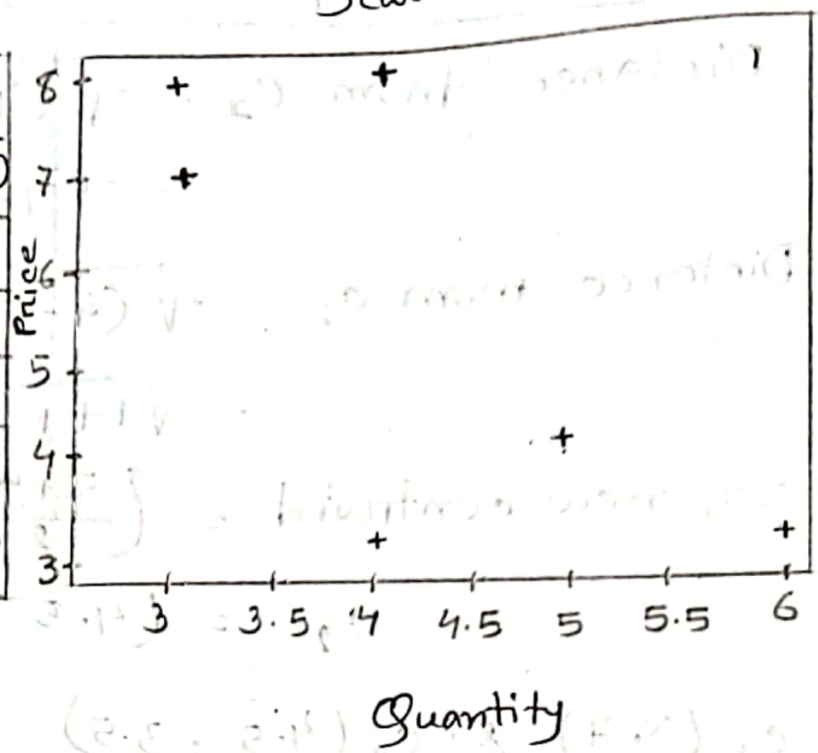
$$= -0.5575 - 4.5661 = -5.123$$

	Sample 1	Sample 2	Sample 3	Sample 4
PC 1 =	-4.305	3.7359	5.692	-5.123



	A	B	C
	Productn	Quantity	Price(k)
1	Facewanh	3	7
2	Cream	5	4
3	Shoen	4	3
4	Bagr	4	8
5	Jacket	6	3
6	Shirt	3	8

Scatter Plot



$C_1 = (3,7)$ & $C_2 = (5,4)$

For first data point (3,7) • Facewanh :-

Dintance from C_1 = 0 (C_1)

Dintance from C_2 = $\sqrt{(5-3)^2 + (4-7)^2}$

= $\sqrt{4+9}$ = 3.60

For second data point (5,4) • Cream :-

Dintance from C_1 = $\sqrt{(5-3)^2 + (4-7)^2}$

= 3.60

Dintance from C_2 = 0 (C_2)

For third data point (4,3) • Shoen :-

Distance from $c_1 = \sqrt{(4-3)^2 + (3-7)^2}$

$= \sqrt{1+16} = 4.12$

Distance from $c_2 = \sqrt{(4-5)^2 + (3-4)^2}$

$= \sqrt{1+1} = 1.41$ (c_2)

So, new centroid $= \left(\frac{5+4}{2}, \frac{4+3}{2}\right)$

$c_2 = (4.5, 3.5)$

$c_1 (3,7)$ & $c_2 (4.5, 3.5)$

For 4th data point (4,8) • Bagn :-

Distance from $c_1 = \sqrt{(4-3)^2 + (8-7)^2}$

$= \sqrt{2} = 1.41$ (c_1)

Distance from $c_2 = \sqrt{(4-4.5)^2 + (8-3.5)^2}$

$= 4.53$

So, new centroid $= \left(\frac{3+4}{2}, \frac{7+8}{2}\right)$

$c_1 = (3.5, 7.5)$

$c_1 (3.5, 7.5)$ & $c_2 (4.5, 3.5)$

For 5th data point (6,3) • Jacket :-

Distance from $c_1 = \sqrt{(6-3.5)^2 + (3-7.5)^2}$

$= 5.15$

Dintance from $c_2 = \sqrt{(6-4.5)^2 + (3-3.5)^2}$

$= 1.58$ (c_2)

For new centroid $= \frac{5+4+6}{3}, \frac{4+3+3}{3}$

$c_2 = (5, 3.33)$

$c_1 = (3.5, 7.5)$ & $c_2 = (5, 3.3)$

For 6th data point (3, 8) • Shirt :-

Dintance from $c_1 = \sqrt{(3-3.5)^2 + (8-7.5)^2}$

$= 0.70$ (c_1)

Dintance from $c_2 = \sqrt{(3-5)^2 + (8-3.33)^2}$

$= 2.48$

So, new centroid $= \left(\frac{3+4+3}{3}, \frac{7+8+8}{3}\right)$

$c_1 = (3.33.., 7.67)$

Sample No.	X	Y
P1	0.40	0.53
P2	0.22	0.38
P3	0.35	0.32
P4	0.26	0.19
P5	0.08	0.41
P6	0.45	0.30

Problem Definition :-

For the given dataset find the clusters using a single link technique. Use Euclidean distance & draw the Dendrogram.

Step 1 :- Compute the distance matrix

- So we have to find the Euclidean distance bet^n each & every points.
- Let $A(X_1, Y_1)$ & $B(X_2, Y_2)$ are two points.
- Then Euclidean distance bet^n

$$d(A,B) = \sqrt{(X_2 - X_1)^2 + (Y_2 - Y_1)^2}$$

$$d(P_1, P_2) = \sqrt{(0.22 - 0.40)^2 + (0.38 - 0.53)^2}$$
$$= 0.23$$

$$d(P_1, P_3) = \sqrt{(0.35 - 0.40)^2 + (0.32 - 0.53)^2}$$
$$= 0.22$$

$$d(P_2, P_3) = \sqrt{(0.35 - 0.22)^2 + (0.32 - 0.38)^2}$$
$$= 0.14$$

Similarly $d(P_1, P_4) = 0.37$, $d(P_2, P_4) = 0.19$, $d(P_3, P_4) = 0.13$

$d(P_1, P_5) = 0.34$, $d(P_2, P_5) = 0.14$, $d(P_3, P_5) = 0.28$

$d(P_4, P_5) = 0.23$, $(P_1, P_6) = 0.24$, $(P_2, P_6) = 0.24$, $(P_3, P_6) = 0.10$, $(P_4, P_6) = 0.22$, $(P_5, P_6) = 0.39$.

	P_1	P_2	P_3	P_4	P_5	P_6
P_1	0					
P_2	0.23	0				
P_3	0.22	0.14	0			
P_4	0.37	0.19	0.13	0		
P_5	0.34	0.14	0.28	0.23	0	
P_6	0.24	0.24	0.10	0.22	0.39	0

Step 2:- Merging the two closest members.

- Here the minimum value is 0.10 & hence we combine P_3 & P_6 (as 0.10 came in the P_6 row & P_3 column).
- Now, form clusters of elements corresponding to the minimum value & update the distance matrix.

Now we will update the Dintance Matrix:

	P_1	P_2	P_3	P_4	P_5	P_6
P_1	0					
P_2	0.23	0				
P_3	0.22	0.14	0			
P_4	0.37	0.19	0.13	0		
P_5	0.34	0.14	0.28	0.23	0	
P_6	0.24	0.24	0.10	0.22	0.39	0

	P_1	P_2	P_3, P_6	P_4	P_5
P_1	0				
P_2	0.23	0			
P_3, P_6	0.22	0.14	0		
P_4	0.37	0.19	0.13	0	
P_5	0.34	0.14	0.28	0.23	0

(P_3, P_6)

Now we will repeat the same process.

Merge two closest members of the two clusters.

The minimum value is 0.13 & hence we combine P_3, P_6 & P_4

Now we will update the Dintance Matrix:

	P_1	P_2	P_3, P_6, P_4	P_5
P_1	0			
P_2	0.23	0		
P_3, P_6, P_4	0.22	0.14	0	
P_5	0.34	0.14	0.28	0

$\{(P_3, P_6), P_4\}$

Again we will update the Distance matrix :-

	P_1	P_2, P_5	P_3, P_6, P_4
P_1	0		
P_2, P_5	0.23	0	
P_3, P_6, P_4	0.22	0.14	0

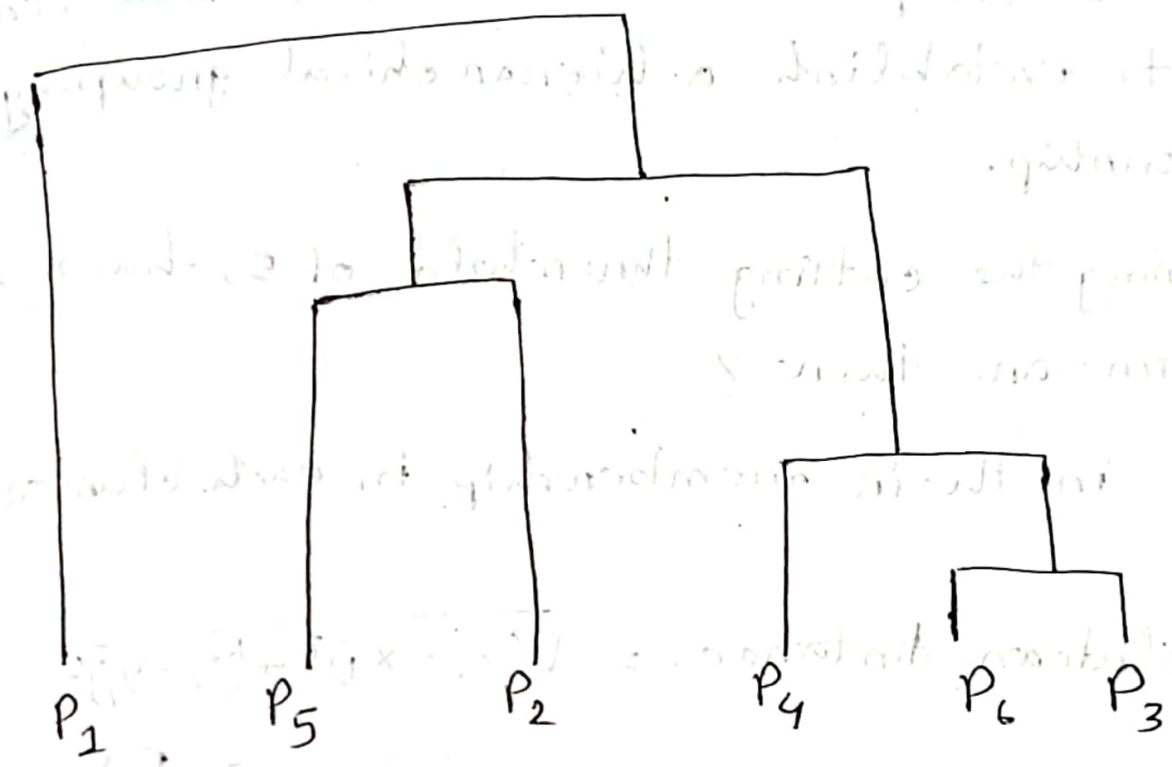
$\{(P_3, P_6), P_4\}$ & (P_2, P_5)

	P_1	P_2, P_5, P_3, P_6, P_4
P_1	0	
P_2, P_5, P_3, P_6, P_4	0.22	0

$[\{(P_3, P_6), P_4\} \rightarrow (P_2, P_5)], P_1$

So now we have reached to the solution, the dendrogram for thone question will be an follown :-

$[\{(P_3, P_6), P_4\}, (P_2, P_5)], P_1$



Dendogram of the cluster formed.

Complete Linkage - Agglomerative Clustering

Given a one-dimensional data net {1, 5, 8, 10, 2}, Use the agglomerative clustering algorithm with the complete link with Euclidean distance to establish a hierarchical grouping relationship.

- By using the cutting threshold of 5, how many clusters are there?
- What is their membership in each cluster?

$$\text{Euclidean distance} = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$$

$$\text{Euclidean distance} = \sqrt{(x_2 - x_1)^2} \quad \left[\text{For One dimension datanet}\right]$$

⇒ In order to use the agglomerative algo, we need to calculate the distance matrix

One-dimensional data net {1, 5, 8, 10, 2}

	1	5	8	10	2
1	0	4	7	9	1
5	4	0	3	5	3
8	7	3	0	2	6
10	9	5	2	0	8
2	1	3	6	8	0

Replace the actual data to column & row number,

	1	2	3	4	5
1	0	4	7	9	1
2	4	0	3	5	3
3	7	3	0	2	6
4	9	5	2	0	8
5	1	3	6	8	0

- From the distance matrix, we can find the distance betn point 1 & 5 is smallest.
- Therefore, we merge them together with their distance as the threshold.
- Then, we update the distance matrix by using

the cluster $\{2, 5\}$

- Using the complete link, we can re-calculate the distance betⁿ this cluster & other points.

$d(2, \{1, 5\}) = \max\{d(2,1), d(2,5)\} = \max\{4, 3\} = 4$

$d(3, \{1, 5\}) = \max\{d(3,1), d(3,5)\} = \max\{7, 6\} = 7$

$d(4, \{1, 5\}) = \max\{d(4,1), d(4,5)\} = \max\{9, 8\} = 9$

Let the 1st column (row) denote the distances betⁿ this cluster & other points, we have the following distance matrix:

	1,5	2	3	4
1,5	0	4	7	9
2	4	0	3	5
3	7	3	0	2
4	9	5	2	0

- From the above distance matrix, we can see the distance betⁿ points 3 & 4 is smallest
- Hence, they merge together to form a cluster $\{3, 4\}$.

- Using the complete link, we have the distance betⁿ different points / clusters as follows :-

$$d(\{1,5\},\{3,4\}) = \max\{d(\{1,5\},3), d(\{1,5\},4)\} = \max\{7,9\} = 9$$

$$d(2,\{3,4\}) = \max\{d(2,3), d(2,4)\} = \max\{3,5\} = 5$$

Thus we can update the distance matrix where now 2 corresponds to point 2, rows 1 & 3 correspond to clusters $\{1,5\}$ & $\{3,4\}$ as follows :

	1,5	2	3,4
1,5	0	4	9
2	4	0	5
3,4	9	5	0

- Following the same procedure we merge point 2 with the cluster $\{1,5\}$ to form $\{1,2,5\}$ and update the distance matrix as follows :-

[1,5], 2 [3,4]

[1,5], 2	0	9
[3,4]	9	0

After increasing the distance threshold to 9, all clusters would merge.

Based on all above distance matrixces, we draw the dendrogram tree as follows