Machine Learning

(21CSC305P)

Clustering

Clustering is an unsupervised Machine Learning-based Algorithm that comprises a group of data points into clusters so that the objects belong to the same group. The process of making a group of abstract objects into classes of similar objects is known as clustering.

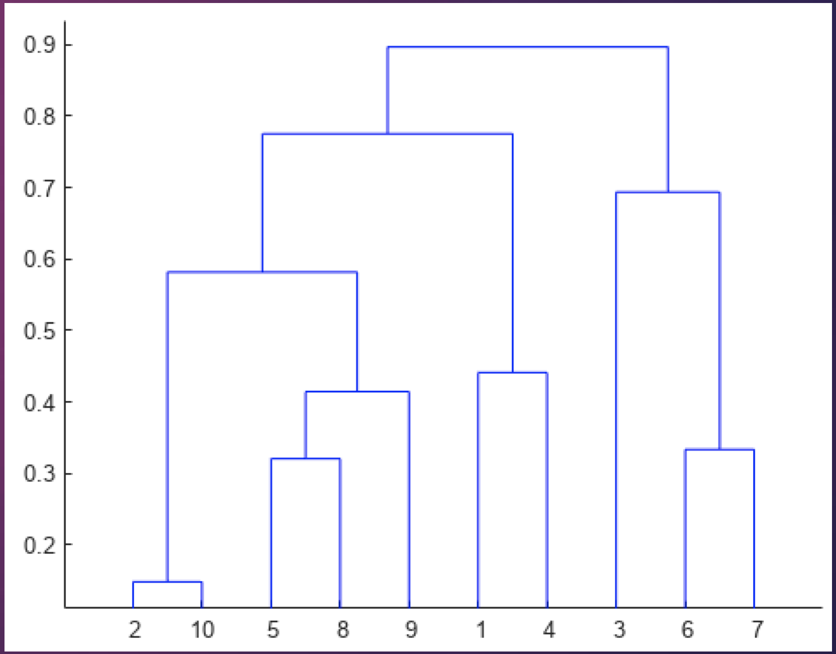
- **Clustering Methods:**

1. Partitioning Method
2. Hierarchical Method

Hierarchical Clustering

Hierarchical clustering is a popular method for grouping objects. It create groups so that objects within a group are similar to each other and different from objects in other groups.

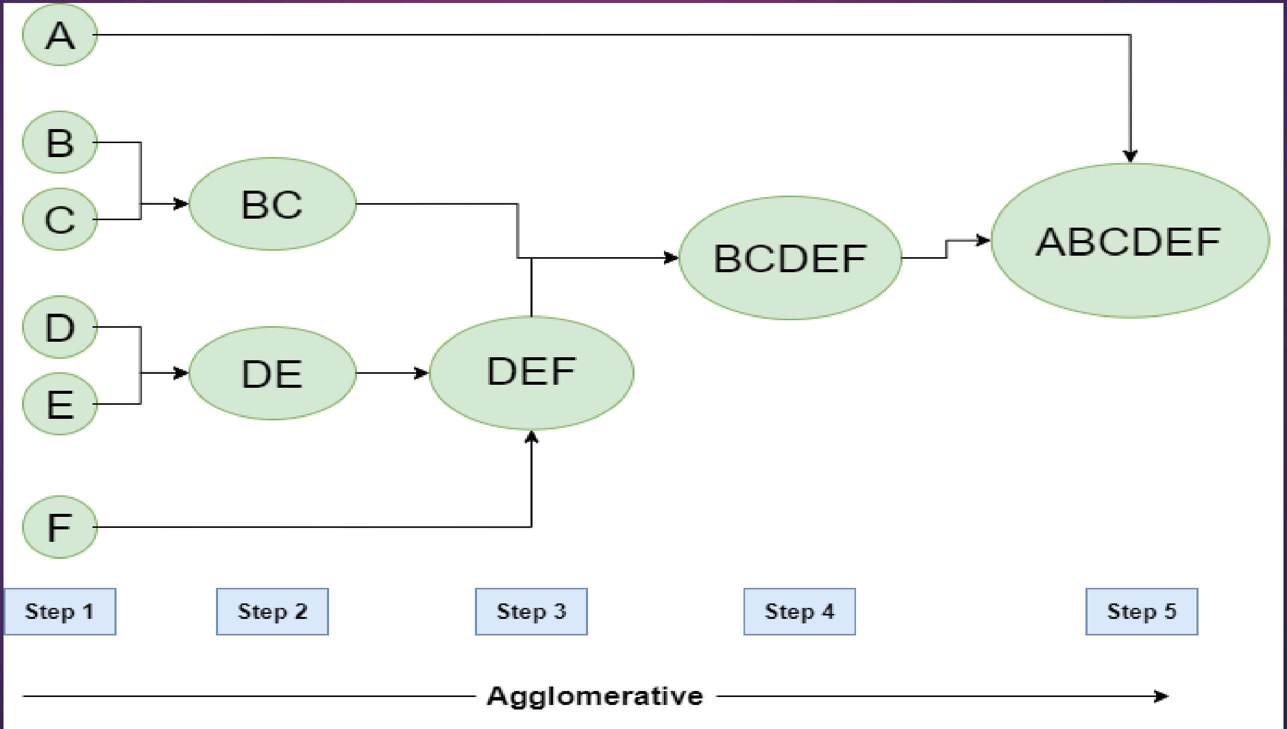
- Why Hierarchical Clustering?
- What is Dendrogram?



Hierarchical Clustering Types

Agglomerative: (Bottom-up method)

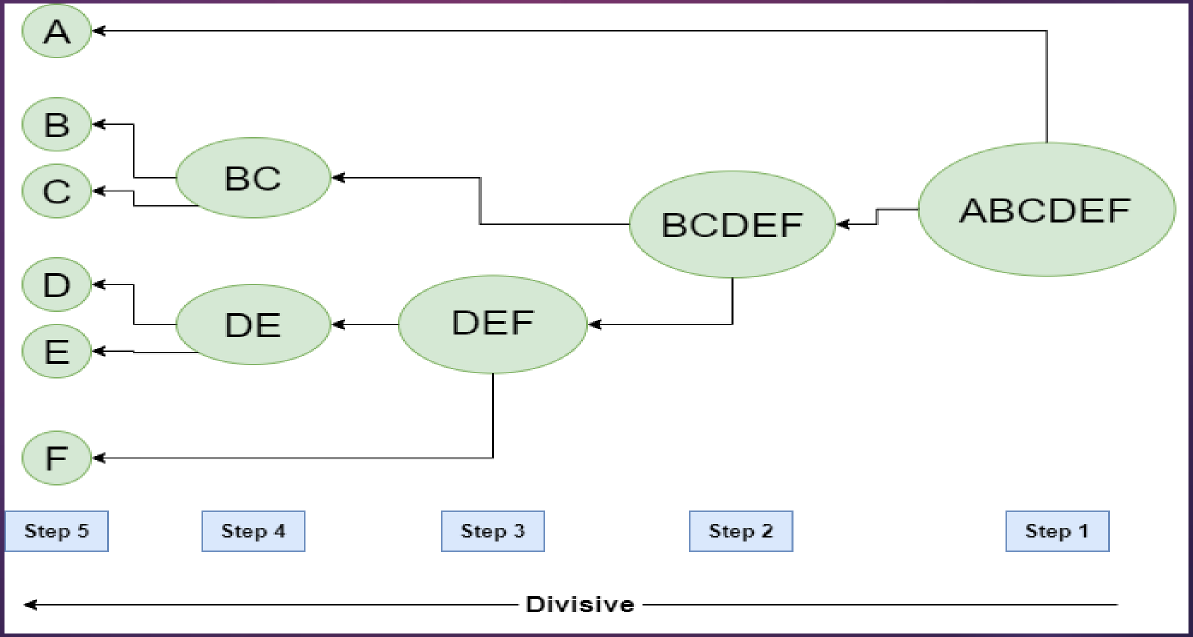
Initially consider every data point as an individual cluster and at every step, merge the nearest pairs of the cluster.



Cont'd

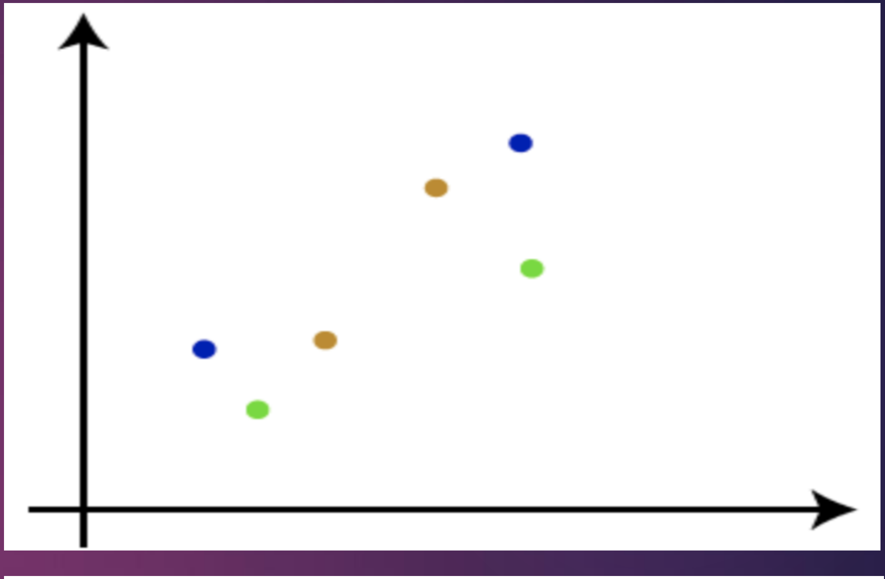
Divisive: (Top-down method)

In Divisive Hierarchical clustering, we take into account all of the data points as a single cluster and in every iteration, we separate the data points from the clusters which aren’t comparable.

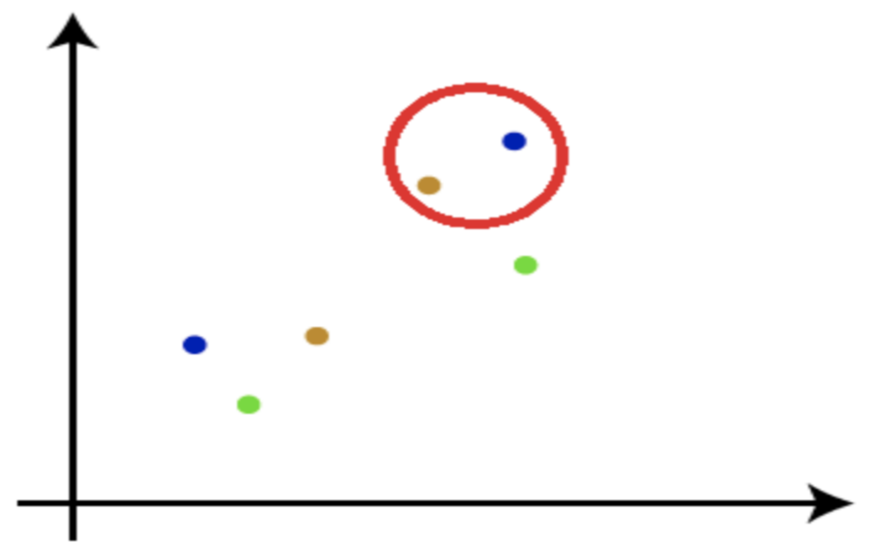


Agglomerative Hierarchical Clustering

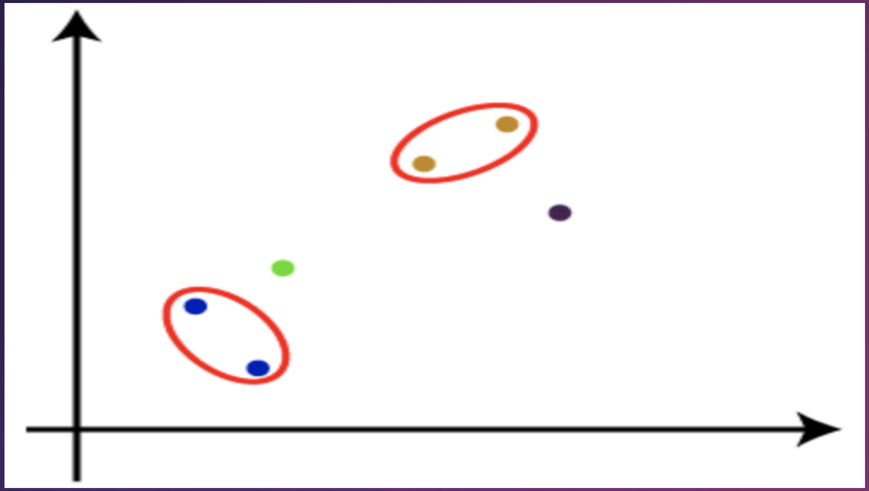
Step-1: Create each data point as a single cluster. Let's say there are N data points, so the number of clusters will also be N.



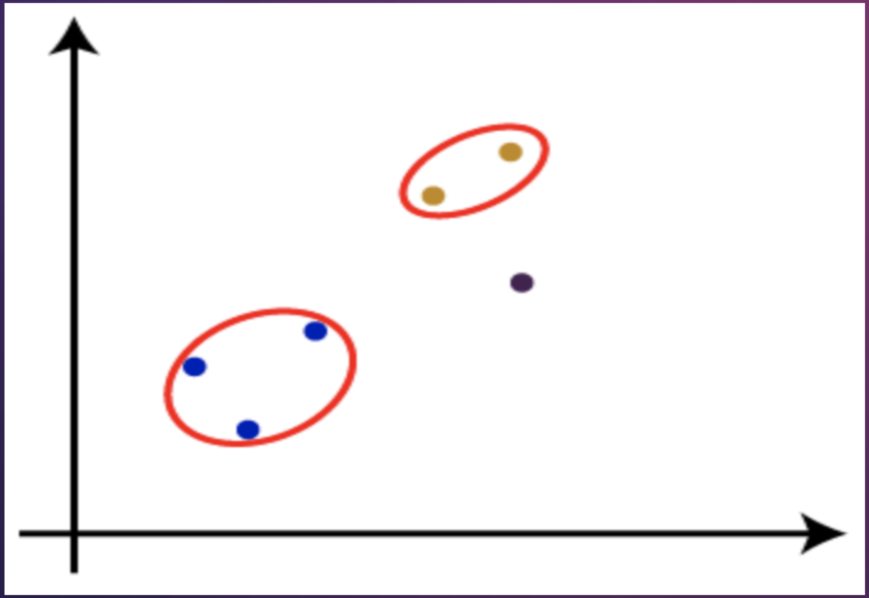
Step-2: Take two closest data points or clusters and merge them to form one cluster. So, there will now be N-1 clusters.



Cont'd

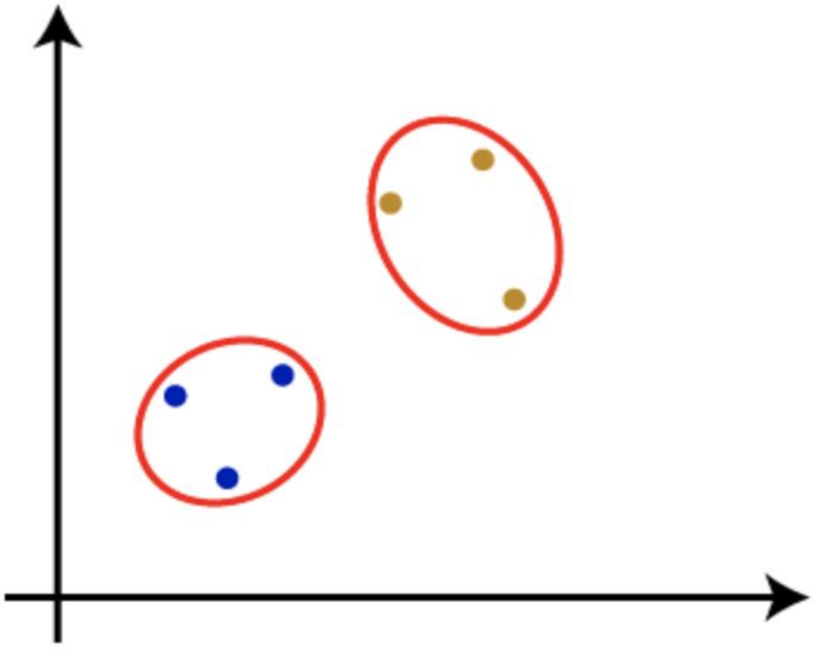


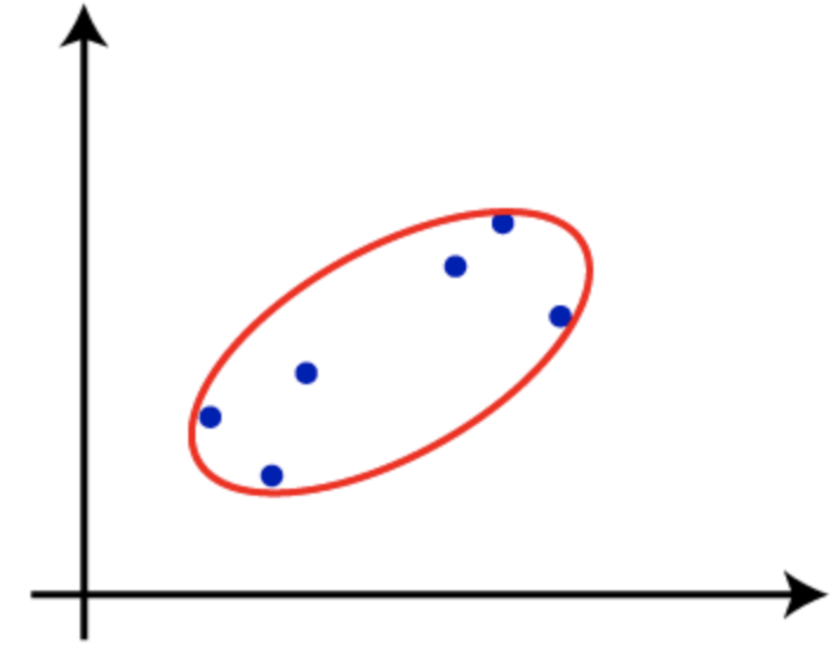
Step-3: Again, take the two closest clusters and merge them together to form one cluster. There will be N-2 clusters.



Step-4: Repeat Step 3 until only one cluster left. So, we will get the following clusters.

Cont'd

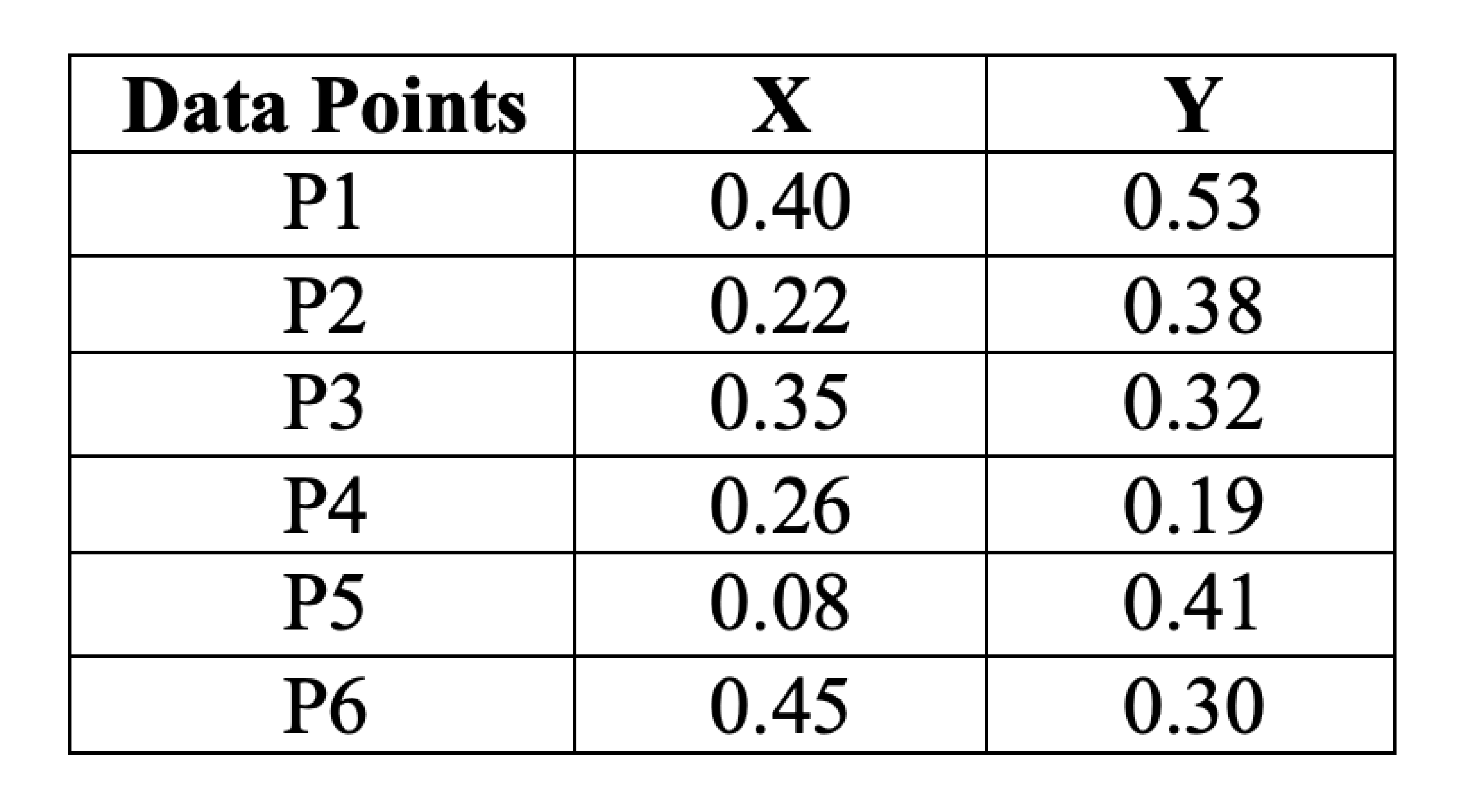




Step-5: Once all the clusters are combined into one big cluster, develop the dendrogram to divide the clusters as per the problem.

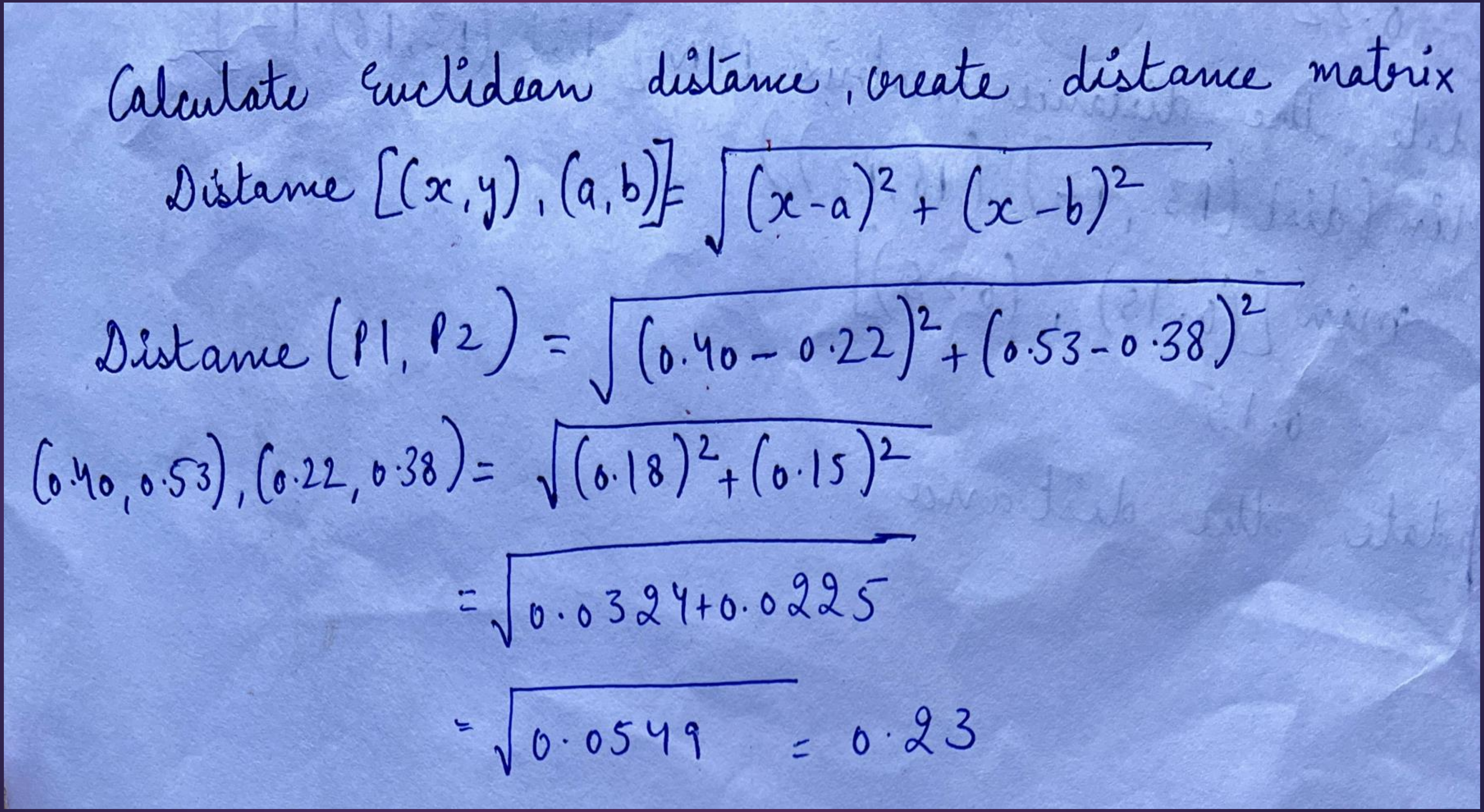
Implementation

- Data Set:



Data Points	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

Cont'd



Calculate Euclidean distance, create distance matrix

Distance $[(x,y),(a,b)] = \sqrt{(x-a)^2 + (x-b)^2}$

Distance $(P1, P2) = \sqrt{(0.40-0.22)^2 + (0.53-0.38)^2}$

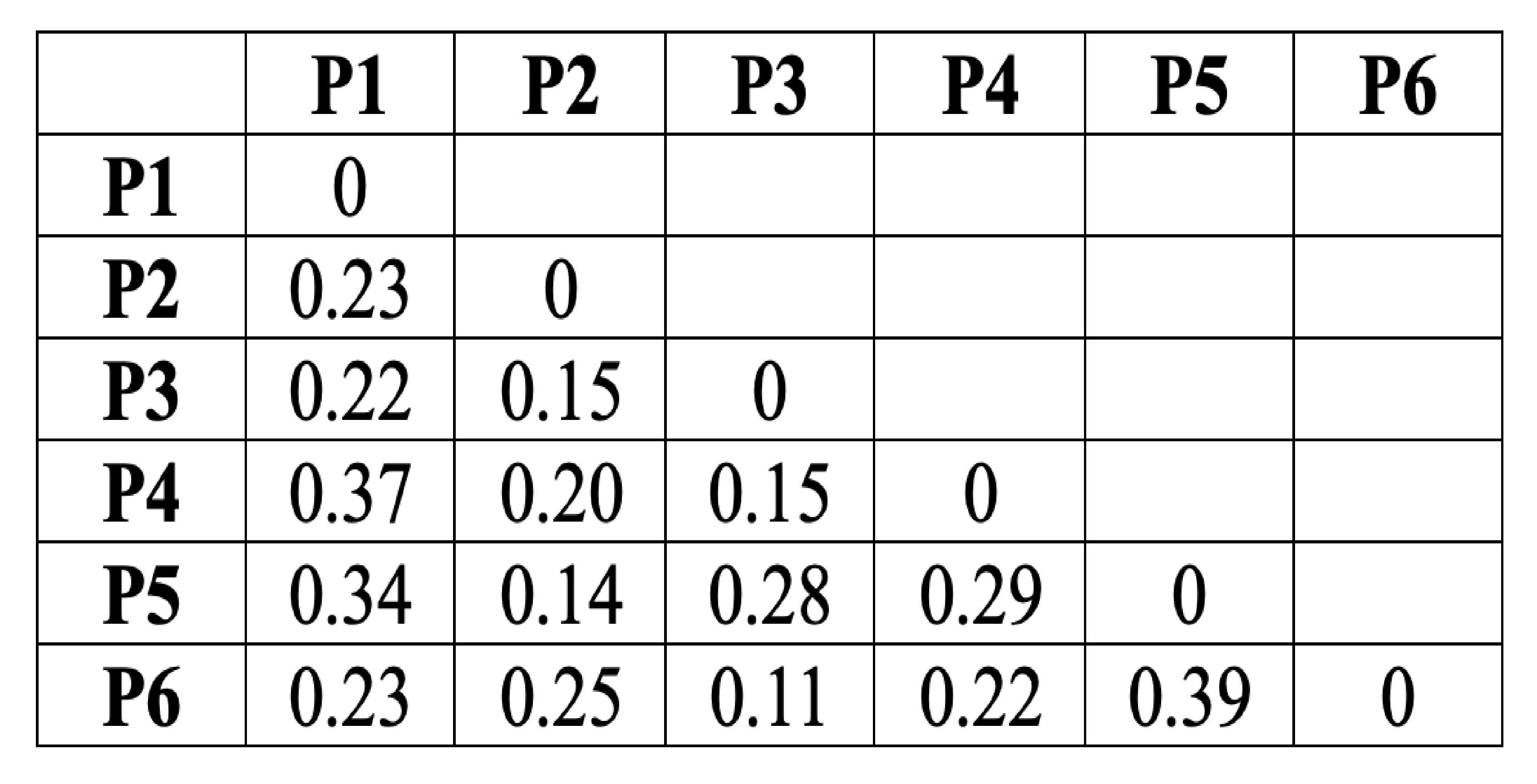
$(0.40, 0.53), (0.22, 0.38) = \sqrt{(0.18)^2 + (0.15)^2}$

$= \sqrt{0.0324+0.0225}$

$= \sqrt{0.0549} = 0.23$

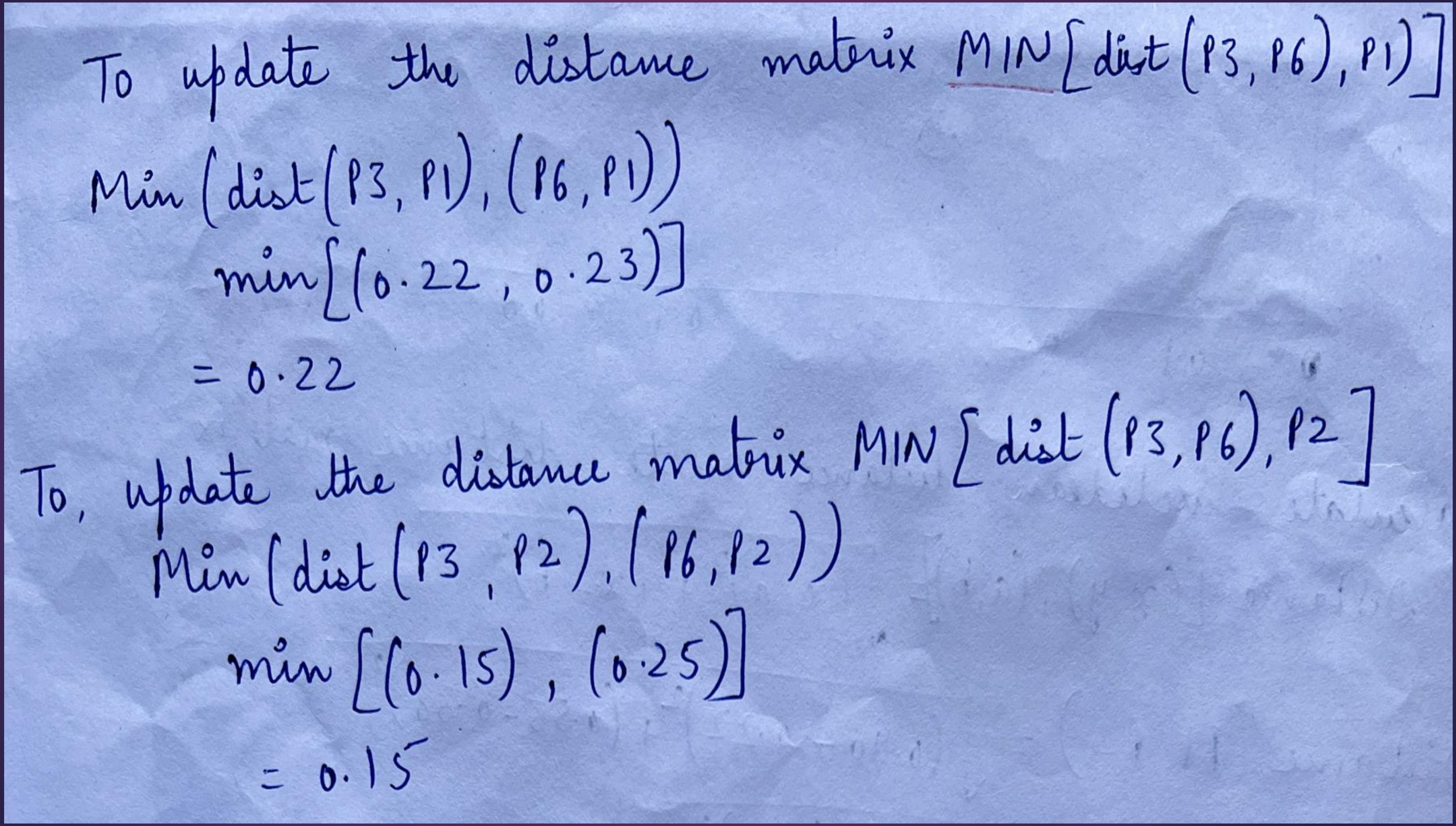
Cont'd

The distance matrix is:



	P1	P2	P3	P4	P5	P6
P1	0					
P2	0.23	0				
P3	0.22	0.15	0			
P4	0.37	0.20	0.15	0		
P5	0.34	0.14	0.28	0.29	0	
P6	0.23	0.25	0.11	0.22	0.39	0

Cont'd



To update the distame matrix MIN[dist(P3, P6), P1)]

Min (dist(P3, P1), (P6, P1))

min[(0.22 , 0.23)]

= 0.22

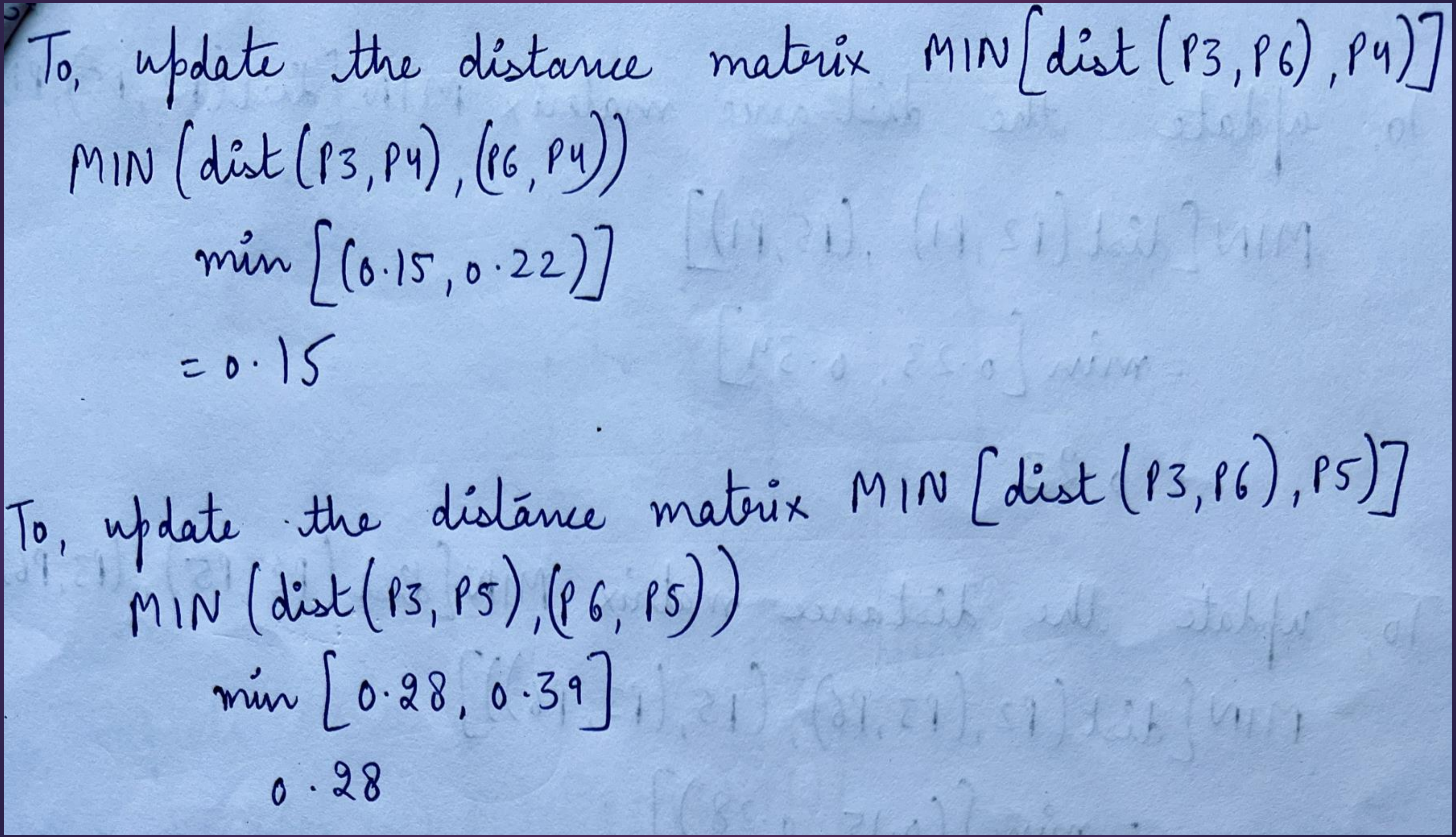
To, update the distance matrix MIN [dist (P3, P6), P2]

Min (dist (P3 , P2), (P6, P2))

min [(0.15) , (0.25)]

= 0.15

Cont'd



To, update the distance matrix MIN[dist(P3,P6),P4)]

MIN(dist(P3,P4),(P6,P4))

min[(0.15,0.22)]

=0.15

To, update the distance matrix MIN[dist(P3,P6),P5)]

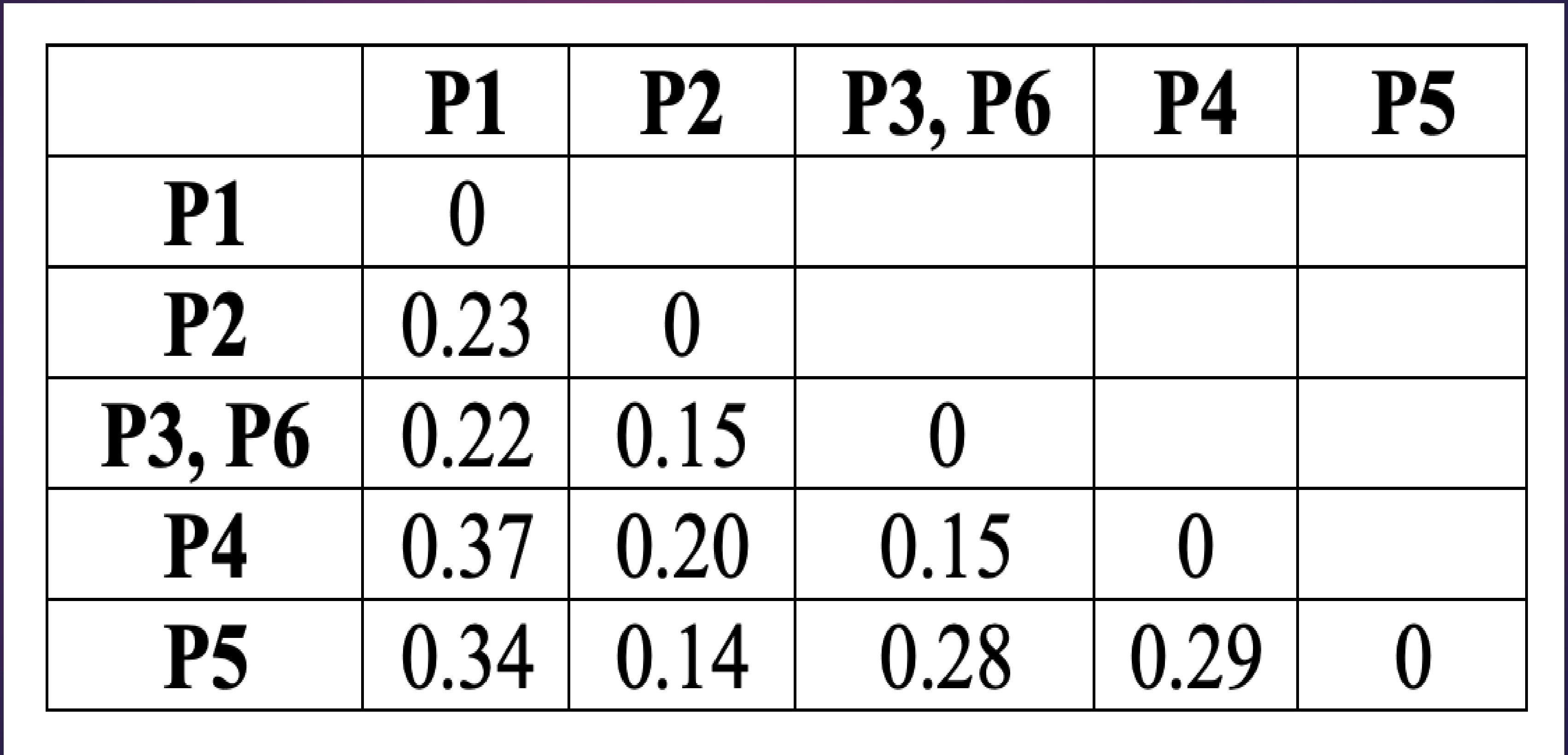
MIN(dist(P3,P5),(P6,P5))

min[0.28, 0.39]

0.28

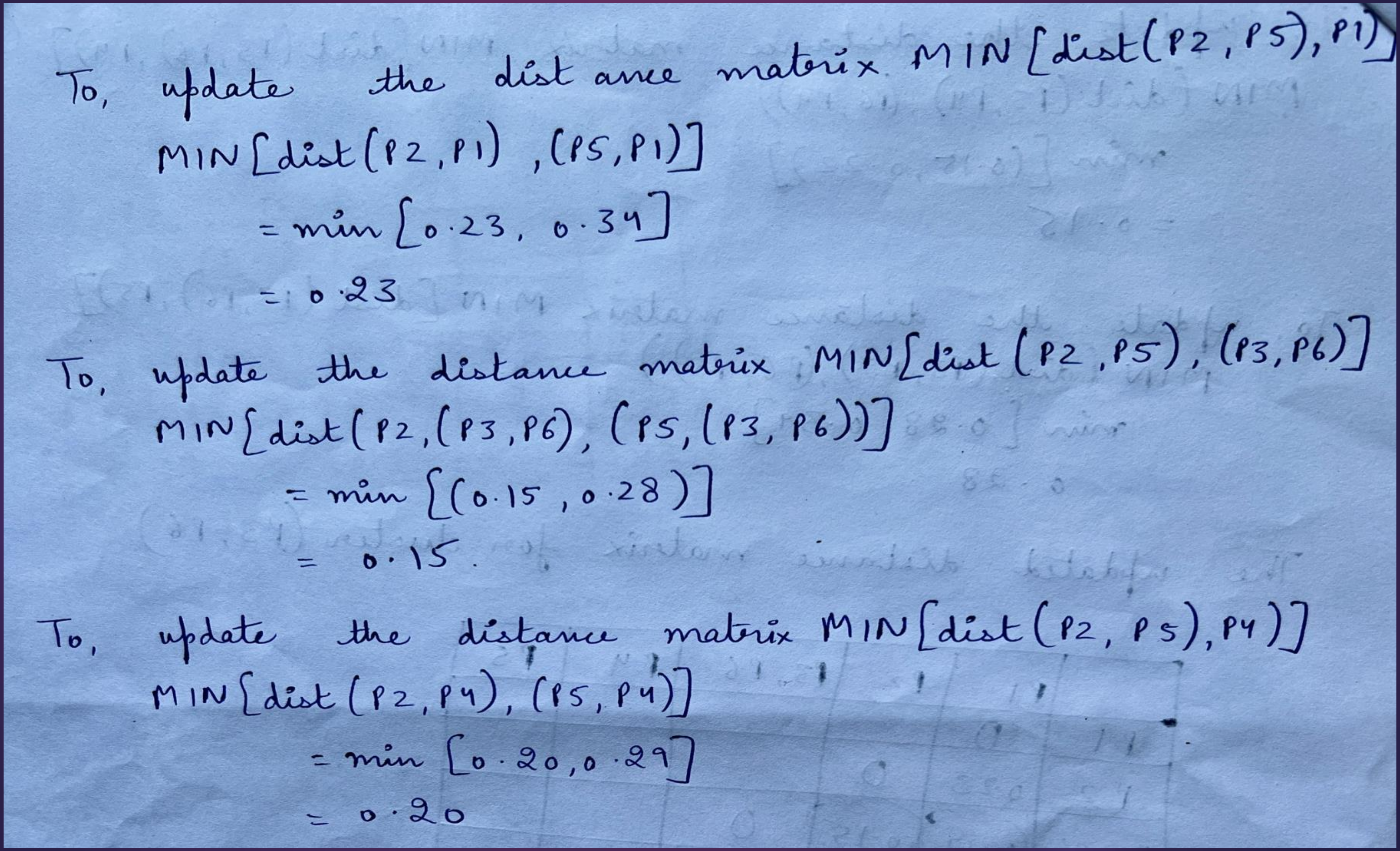
Cont'd

The updated distance matrix for clusters (P3, P6):



	P1	P2	P3, P6	P4	P5
P1	0				
P2	0.23	0			
P3, P6	0.22	0.15	0		
P4	0.37	0.20	0.15	0	
P5	0.34	0.14	0.28	0.29	0

Cont'd



To, update the dist ance matrix MIN [dist(P2, P5), P1)]

MIN [dist(P2, P1) , (P5, P1)]

= min [0.23, 0.34]

= 0.23

To, update the distance matrix MIN[dist (P2, P5), (P3, P6)]

MIN[dist(P2, (P3, P6), (P5, (P3, P6))]

= min [(0.15, 0.28)]

= 0.15.

To, update the distance matrix MIN[dist(P2, P5), P4)]

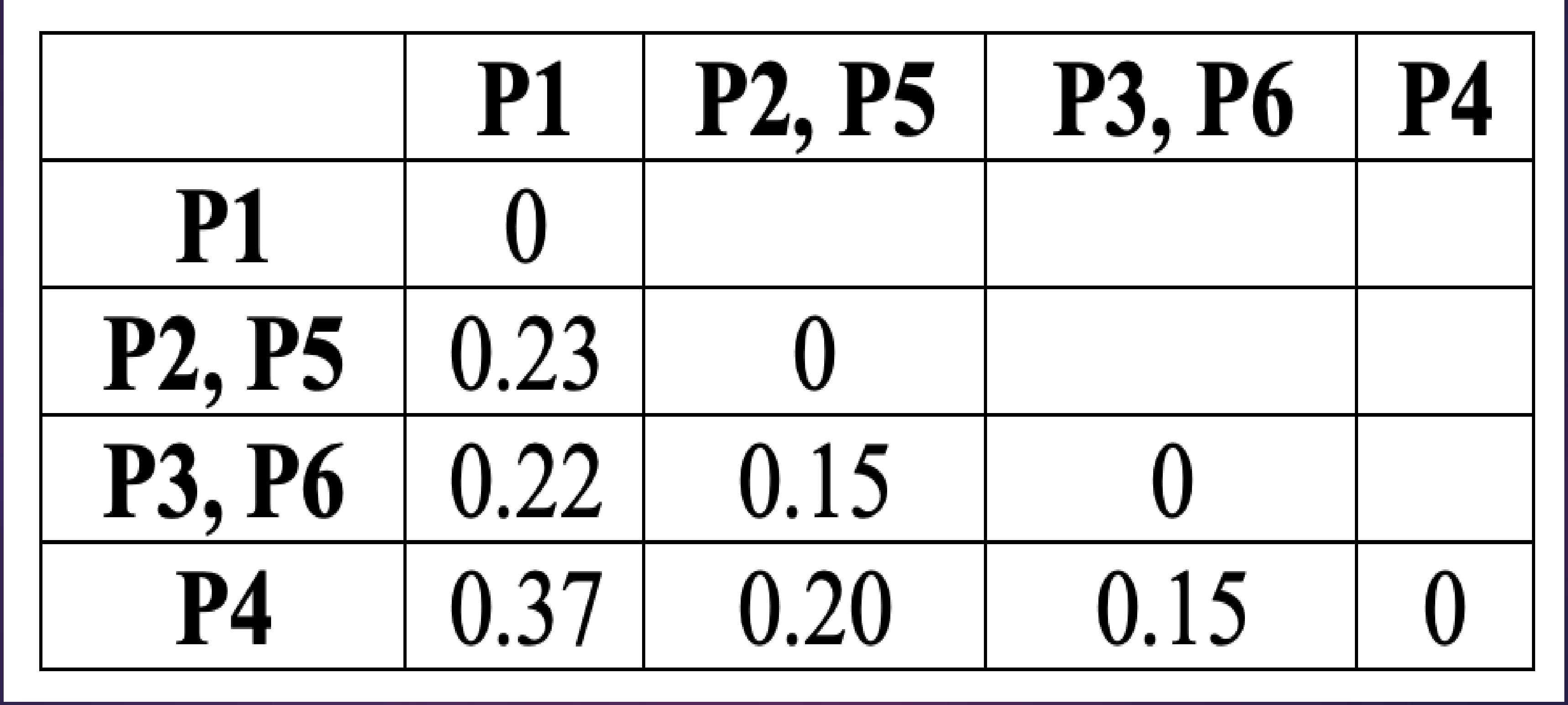
MIN[dist(P2, P4), (P5, P4)]

= min [0.20, 0.29]

= 0.20

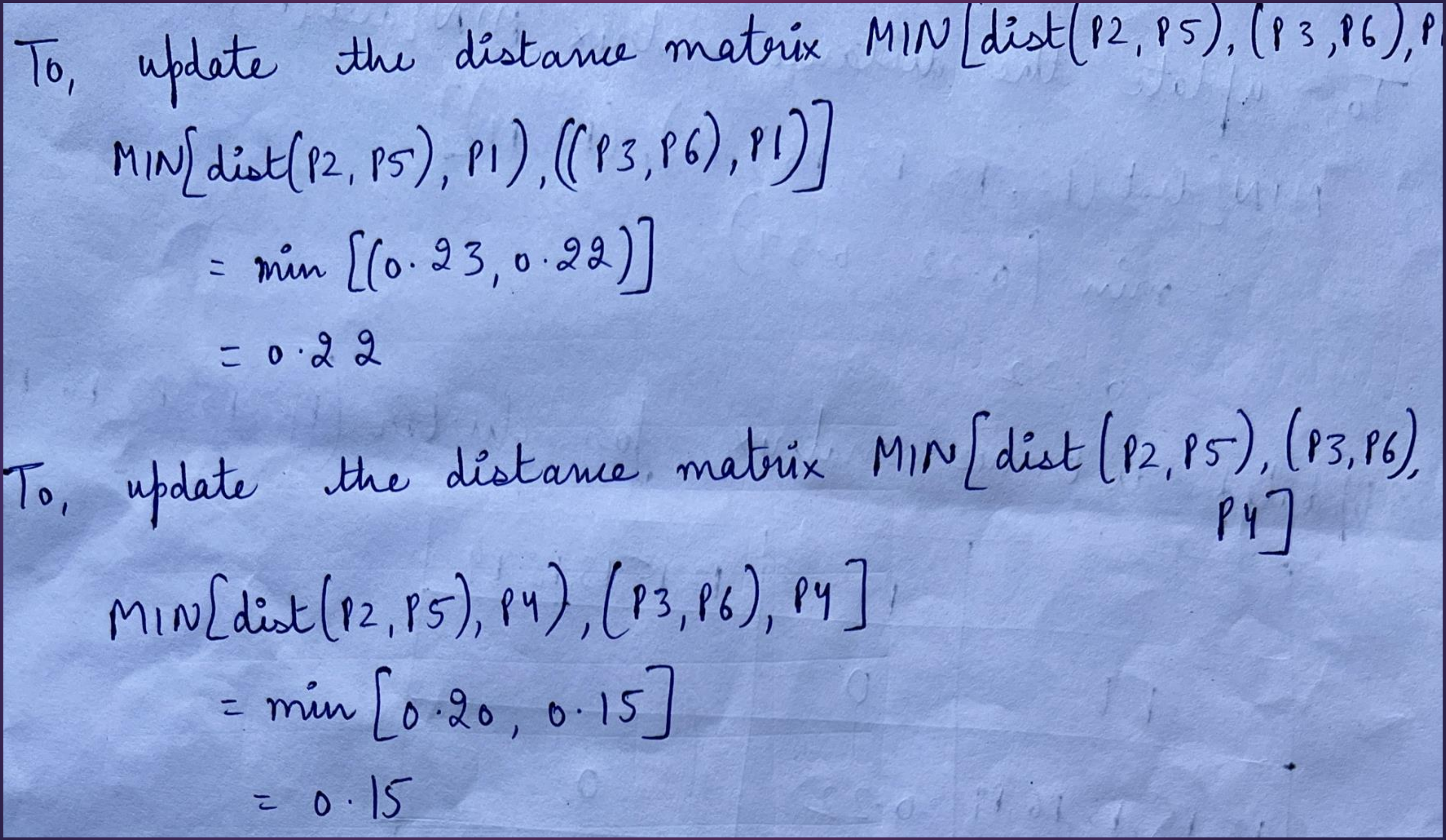
Cont'd

Updated distance matrix for cluster (P2, P5):



	P1	P2, P5	P3, P6	P4
P1	0			
P2, P5	0.23	0		
P3, P6	0.22	0.15	0	
P4	0.37	0.20	0.15	0

Cont'd



To, update the distance matrix MIN[dist(P2, P5), (P3,P6), P

MIN[dist(P2, P5), P1), ((P3, P6), P1)]

= min [(0.23, 0.22)]

= 0.22

To, update the distance matrix MIN[dist (P2, P5), (P3, P6), P4]

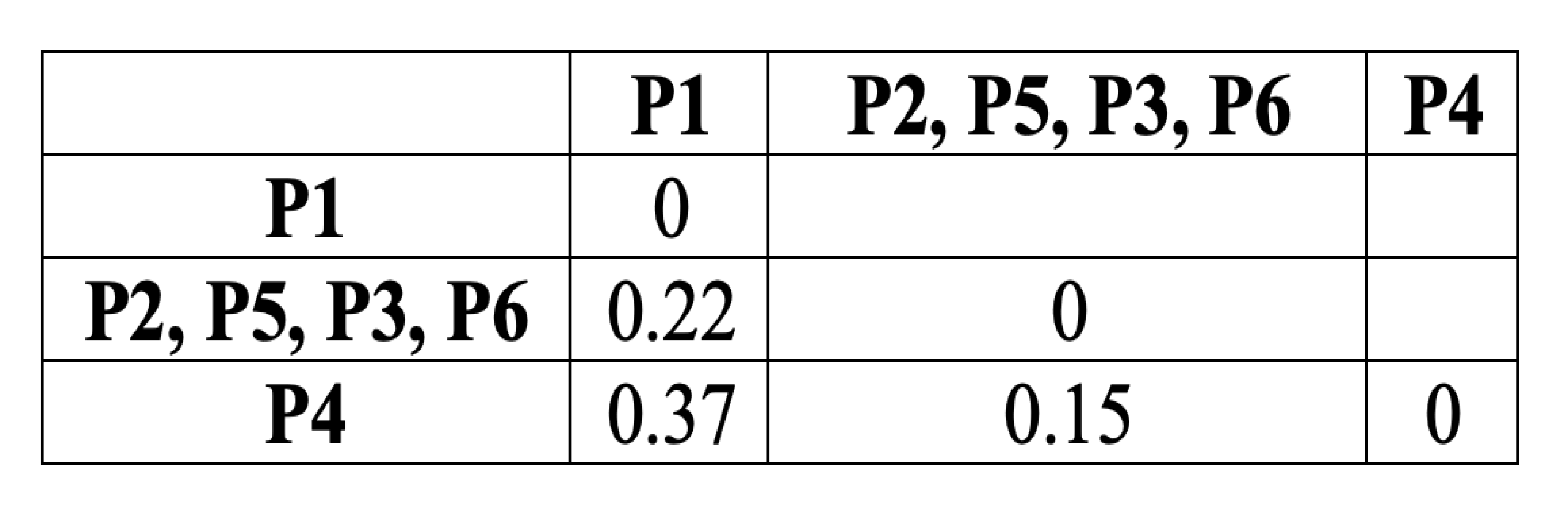
MIN[dist(P2, P5), P4), (P3, P6), P4]

= min [0.20, 0.15]

= 0.15

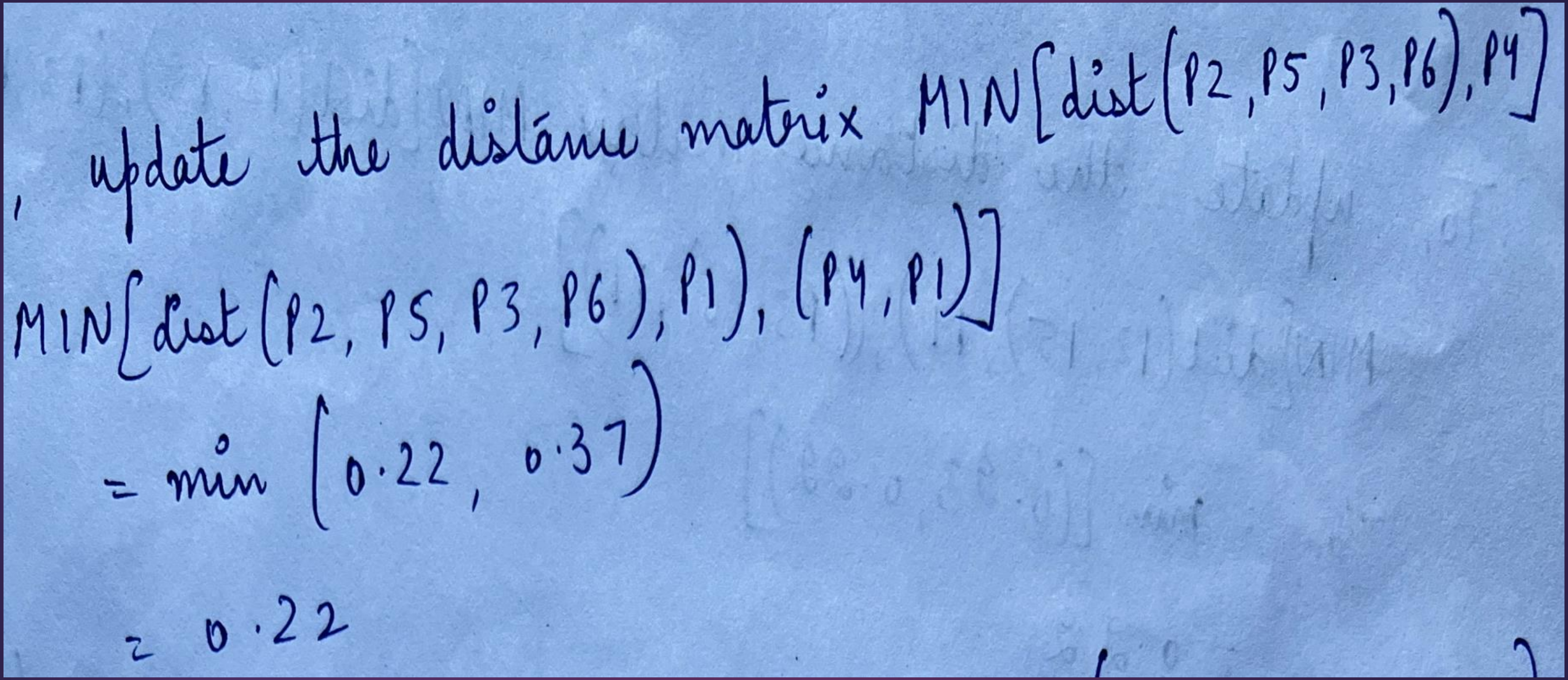
Cont'd

Updated Matrix for cluster (P2, P5, P3, P6):



	P1	P2, P5, P3, P6	P4
P1	0		
P2, P5, P3, P6	0.22	0	
P4	0.37	0.15	0

Cont'd



update the distance matrix MIN[dist(P2, P5, P3, P6), P4]

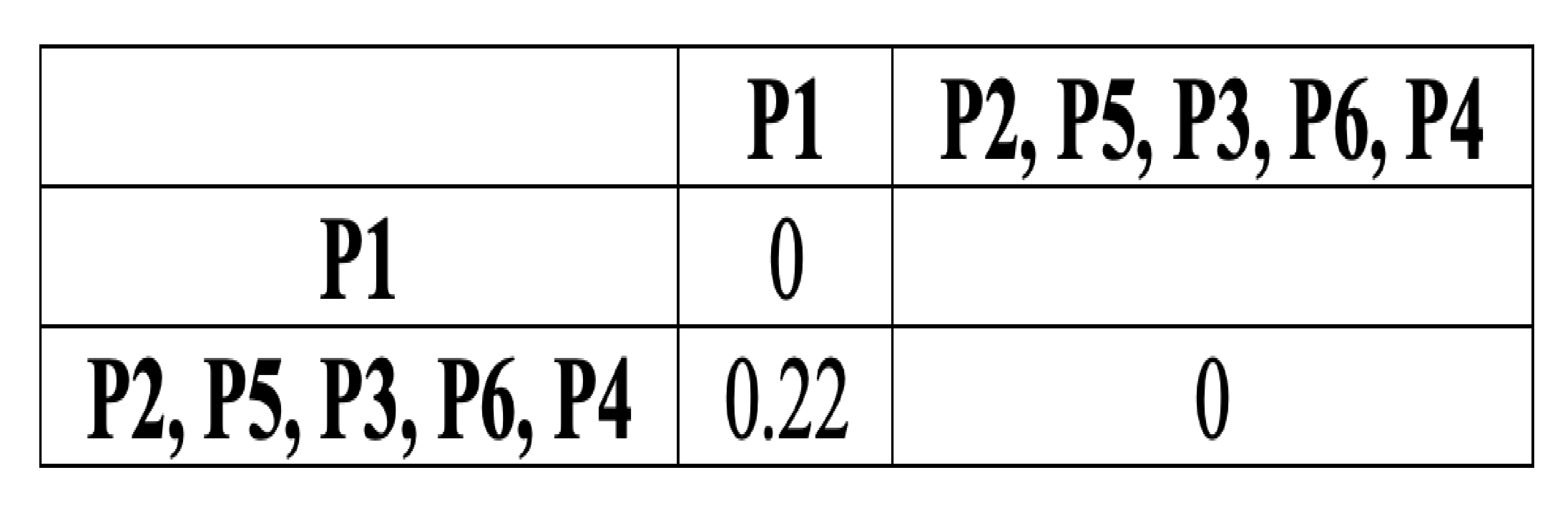
MIN[dist(P2, P5, P3, P6), P1), (P4, P1)]

= min (0.22, 0.37)

= 0.22

Cont'd

Updated Matrix for cluster (P2, P5, P3, P6, P4):



	P1	P2, P5, P3, P6, P4
P1	0	
P2, P5, P3, P6, P4	0.22	0