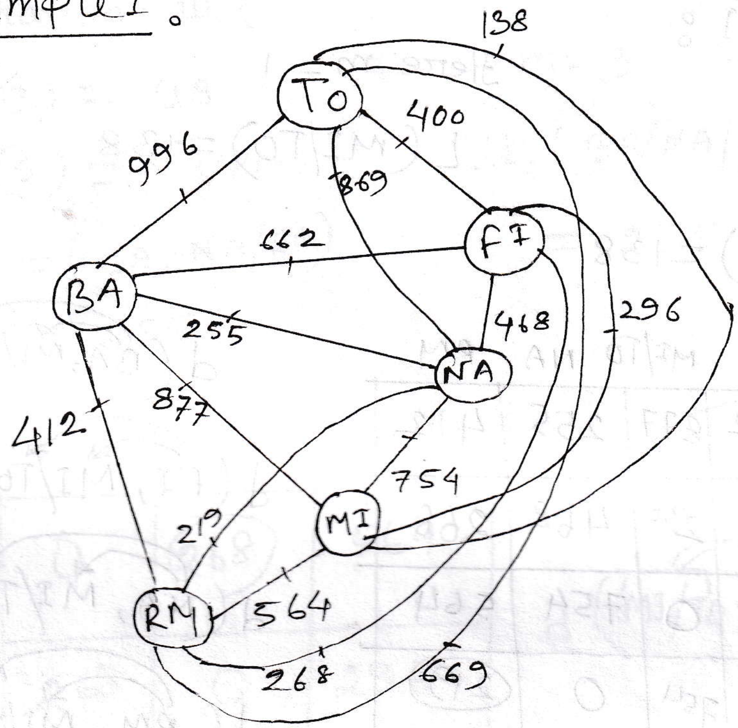
Hierarchical clustering

Example I:



Show a hierarchical Clustering of distances in km between some Italian cities.

Ans: Input distance matrix

	BA	FI	MI	NA	RM	TO
BA	0	662	877	255	412	996
FI	662	0	296	468	268	400
MI	877	296	0	754	564	138
NA	255	468	754	0	219	869
RM	412	268	564	219	0	669
TO	996	400	138	869	669	0

Here, Sequence Number, $m = 0$

$\therefore L(BA) = 0$ | $L(MI) = 0$ | $L(RM) = 0$

$L(FI) = 0$ | $L(NA) = 0$ | $L(TO) = 0$

Iteration-I :

(r) = MI

(s) = TO

Here m = 1

∴ L(MI/TO) = 138

∴ d(MI, TO) = 138

	BA	FI	MI/TO	NA	RM
BA	0	662	877	255	412
FI	662	0	296	468	268
MI/TO	877	296	0	754	564
NA	255	468	754	0	219
RM	412	268	564	219	0

d(BA, MI/TO) = min(877, 996)
= 877

d(FI, MI/TO) = min(296, 400)
= 296

d(NA, MI/TO) = min(754, 869)
= 754

d(RM, MI/TO) = min(564, 669)
= 564

Iteration-II :

Here, (r) = NA

(s) = RM

d(NA, RM) = 219

m = 2

L(NA/RM) = 219

	BA	FI	MI/TO	NA/RM
BA	0	662	877	255
FI	662	0	296	268
MI/TO	877	296	0	564
NA/RM	255	268	564	0

d(BA, NA/RM) = min(255, 412)
= 255

d(FI, NA/RM) = min(468, 268)
= 268

d(MI/TO, NA/RM) = min(754, 564)
= 564

Iteration - III :

$(r) = BA$ $\quad m = 3$

$(s) = NA/RM$ $\quad L(BA/NA/RM)$

$d = (BA, NA/RM)$ $\quad = 255$

$= 255$

	BA/NA/RM	FI	MI/TO
BA/NA/RM	0	268	564
FI	268	0	296
MI/TO	564	296	0

$d(FI, BA/NA/RM) = \min(662, 268)$

$= 268$

$d(MI/TO, BA/NA/RM) = \min(877, 564)$

$= 564$

Iteration-IV :

Here, $(r) = BA/NA/RM$ $\quad m = 4$

$(s) = FI$ $\quad L(BA/NA/RM/FI)$

$d(BA/NA/RM, FI)$ $\quad = 268$

$= 268$

	BA/NA/RM/FI	MI/TO
BA/NA/RM/FI	0	296
MI/TO	296	0

$d(MI/TO, BA/NA/RM/FI)$

$= \min(564, 296)$

$= 296$

Iteration – V:

Here $(r) = BA/NA/RM/FI$

$(s) = MI/TO$

$m = 5$

$L(BA/NA/RM/FI/MI/TO) = 296$

$d(BA/NA/RM/FI, MI/TO)$

$= 296$

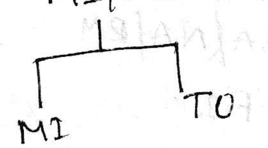
	BA/NA/RM/FI/MI/TO
BA/NA/RM/FI/MI/TO	0

Summary:

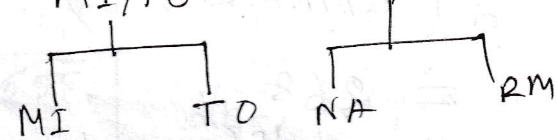
Hierarchical Tree:

Initial: BA FI MI NA RM TO

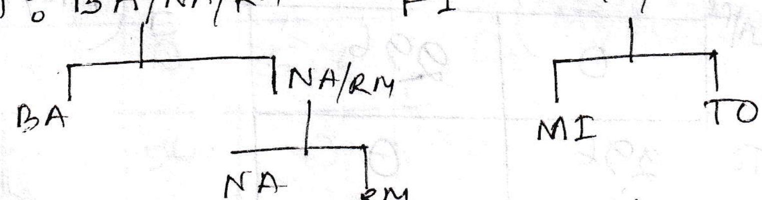
Iteration-I: BA FI MI/TO NARM



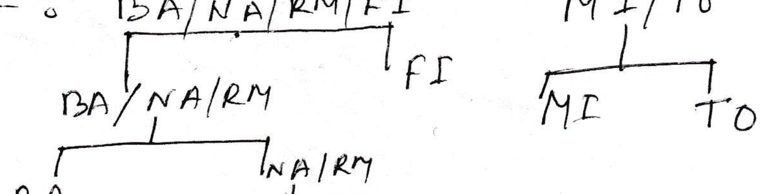
Iteration-II: BA FI MI/TO NA/RM



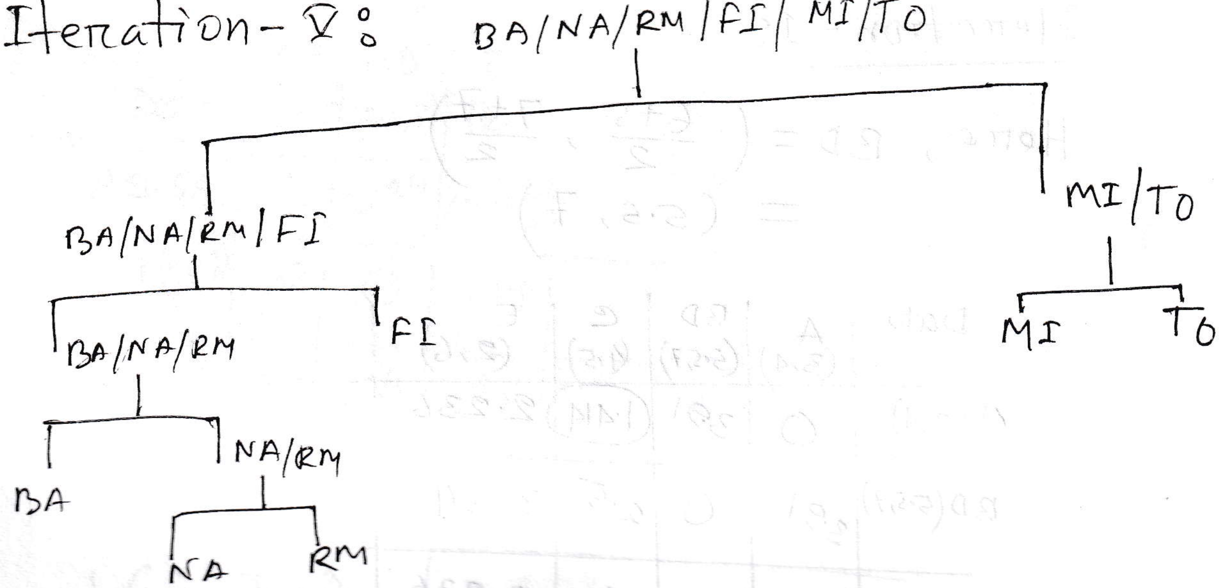
Iteration-III: BA/NA/RM FI MI/TO



Iteration-IV: BA/NA/RM/FI MI/To



Iteration-V:



Example 2: Show the Mechanism of hierarchical clustering using the following data.

(3,4) (6,7) (4,5) (5,7) (2,6)

Soln: Let,

A ≡ (3,4)
B ≡ (6,7)
C ≡ (4,5)
D ≡ (5,7)
E ≡ (2,6)

Input Distance matrix

Data	A (3,4)	B (6,7)	C (4,5)	D (5,7)	E (2,6)
A (3,4)	0	4.242	1.414	3.605	2.236
B (6,7)	4.242	0	2.828	1.0	4.123
C (4,5)	1.414	2.828	0	2.236	2.236
D (5,7)	3.605	1.0	2.236	0	3.162
E (2,6)	2.236	4.123	2.236	3.162	0

Iteration - I:

Here, $BD = \left(\frac{6+5}{2}, \frac{7+7}{2}\right)$
$= (5.5, 7)$

Data	A (3,4)	BD (5.5,7)	C (4,5)	E (2,6)
A(3,4)	0	3.91	1.414	2.236
BD(5.5,7)	3.91	0	2.5	3.64
C(4,5)	1.414	2.5	0	2.236
E(2,6)	2.236	3.64	2.236	0

$$d(A, BD) = \sqrt{(3-5.5)^2 + (4-7)^2}$$
$$= \sqrt{(-2.5)^2 + 3^2}$$
$$= \sqrt{6.25 + 9}$$
$$= \sqrt{15.25}$$
$$= 3.91$$

$$d(BD, C) = \sqrt{(5.5-4)^2 + (7-5)^2}$$
$$= \sqrt{(1.5)^2 + 2^2}$$
$$= \sqrt{2.5 + 4}$$
$$= \sqrt{6.25}$$
$$= 2.5$$

$$d(BD, E) = \sqrt{(5.5-2)^2 + (7-6)^2}$$
$$= \sqrt{3.5^2 + 1^2}$$
$$= \sqrt{12.25 + 1}$$
$$= \sqrt{13.25} = 3.64$$

Iteration - II:

Here, $AC = \left(\frac{3+4}{2}, \frac{4+5}{2}\right)$
$= (3.5, 4.5)$

Data	AC (3.5,4.5)	BD (5.5,7)	E (2,6)
AC(3.5,4.5)	0	3.20	2.12
BD(5.5,7)	3.20	0	3.64
E(2,6)	2.12	3.64	0

$$d(AC, BD) = \sqrt{(3.5-5.5)^2 + (4.5-7)^2}$$
$$= \sqrt{(-2)^2 + 2.5^2}$$
$$= \sqrt{4 + 6.25}$$
$$= \sqrt{10.25}$$
$$= 3.20$$

$$d(AC, E) = \sqrt{(3.5-2)^2 + (4.5-6)^2}$$
$$= \sqrt{1.5^2 + (-1.5)^2}$$
$$= \sqrt{2.25 + 2.25}$$
$$= \sqrt{4.5}$$
$$= 2.12$$

Iteration-III

Here, $(AC, E) = \left(\frac{3.5+2}{2}, \frac{4.5+6}{2}\right)$
$= (2.75, 5.25)$

Data	ACE (2.75,5.25)	BD (5.5,7)
ACE (2.75,5.25)	0	3.26
BD (5.5,7)	3.26	0

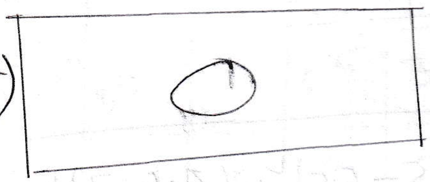
$$d(ACE, BD) = \sqrt{(2.75-5.5)^2 + (5.25-7)^2}$$
$$= \sqrt{(-2.75)^2 + (-1.75)^2}$$
$$= \sqrt{7.5625 + 3.0625}$$
$$= \sqrt{10.625}$$
$$= 3.26$$

Iteration- IV :

Here, $(ACE, BD) = \left(\frac{2.75+5.5}{2}, \frac{5.25+7}{2}\right)$
$= (4.125, 6.125)$

ACEBD(4.125, 6.125)

ACEBD (4.125,6.125)

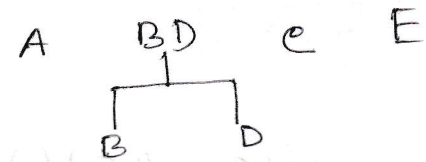


Summary:

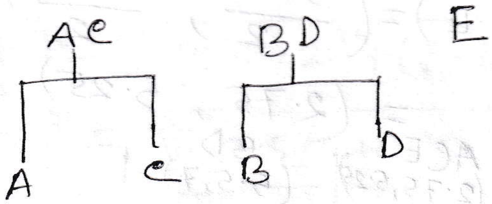
Hierarchical Tree:

Initial : A B C D E

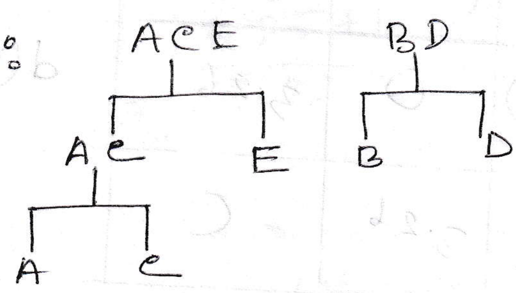
Iteration-I :



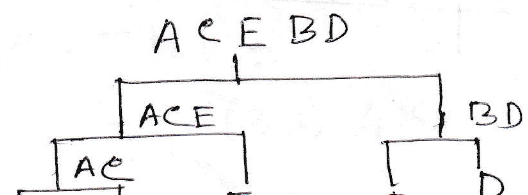
Iteration- II :



Iteration – III :



Iteration- IV :



Hierarchical Algorithm :

Step 1: Begin with the disjoint clustering having level $L(0) = 0$ and sequence number $m = 0$.

Step 2: Find the least dissimilar pair of clusters in the current clustering, say pair (r), (s), according to

$d[(r),(s)] = \min d[(i),(j)]$.

Where the minimum is over all pairs of clusters in the current clustering.

Step 3: Increment the sequence number: $m = m+1$.

Merge clusters (r) and (s) into a single cluster to form the clustering m.

Set the level of this clustering to

$L(m) = d[(r),(s)]$.

Step 4: Update the proximity matrix, D, by deleting the rows and columns corresponding to clusters (r) and (s) and adding a row and column corresponding to the newly formed cluster. The proximity between the new cluster, denoted (r,s) and old cluster (k) is defined in this way

$d[(k),(r,s)] = \min d[(k),(r)], d[(k),(s)]$

Step 5: If all objects are in one cluster, stop. Else, go to step 2.

Example: Cluster the data (3,4) (6,7) and (4,6) into two cluster using the Fuzzy c-means clustering algorithm.

Number of cluster 2 $(c=2)$

Ans:

Step 1:

$$U^{(0)} = \begin{array}{c} i=1 \\ i=2 \\ i=3 \\ \\ \end{array} \begin{bmatrix} 0.8 & 0.2 \\ 0.3 & 0.7 \\ 0.9 & 0.1 \\ j=1 & j=2 \end{bmatrix}$$

Number of data set 3 $(N=3)$

[Total of each row must not greater than 1]

Here, $U_{11} = 0.8$ $\quad U_{12} = 0.2$

$U_{21} = 0.3$ $\quad U_{22} = 0.7$

$U_{31} = 0.9$ $\quad U_{32} = 0.1$

Iteration-1:

Step 2: c_1 and c_2 must be calculated because $c=2$.

c_1 Calculation:

(1st dim)
$$c_1 = \frac{U_{11} * 3 + U_{21} * 6 + U_{31} * 4}{U_{11} + U_{21} + U_{31}}$$

$$= \frac{0.8 * 3 + 0.3 * 6 + 0.9 * 4}{0.8 + 0.3 + 0.9}$$

$$= \frac{2.4 + 1.8 + 3.6}{2.0}$$

$$= \frac{7.8}{2.0}$$

$$= 3.9$$

(2nd dim) $$c_1 = \frac{U_{11} * 4 + U_{21} * 7 + U_{31} * 6}{U_{11} + U_{21} + U_{31}}$$

$$= \frac{0.8 * 4 + 0.3 * 7 + 0.9 * 6}{0.8 + 0.3 + 0.9}$$

$$= \frac{3.2 + 2.1 + 5.4}{2.0}$$

$$= \frac{10.7}{2.0} = 5.35$$

$$\therefore \quad c_1 \equiv (3.9, 5.35)$$

c_2 Calculation:

(1st dim) $$c_2 = \frac{U_{12} * 3 + U_{22} * 6 + U_{32} * 4}{U_{12} + U_{22} + U_{32}}$$

$$= \frac{0.2 * 3 + 0.7 * 6 + 0.1 * 4}{0.2 + 0.7 + 0.1}$$

$$= \frac{0.6 + 4.2 + 0.4}{1.0}$$

$$= \frac{5.20}{1.0} = 5.2$$

(2nd dim) $$c_2 = \frac{U_{12} * 4 + U_{22} * 7 + U_{32} * 6}{U_{12} + U_{22} + U_{32}}$$

$$= \frac{0.2 * 4 + 0.7 * 7 + 0.1 * 6}{0.2 + 0.7 + 0.1}$$

$$= \frac{0.8 + 4.9 + 0.6}{1.0}$$

$$= \frac{6.3}{1.0} = 6.3$$

Step 3:

When $i=1, j=1$; $u_{ij} = \dfrac{1}{\sum_{k=1}^{c} \left(\dfrac{\|x_i - c_j\|}{\|x_i - c_k\|} \right)^{\frac{2}{m-1}}}$

$$\Rightarrow u_{11} = \frac{1}{\sum_{k=1}^{2} \left(\frac{\|x_1 - c_1\|}{\|x_1 - c_k\|} \right)^2} \quad \left[\begin{array}{l} \text{if } m=2 \\ c=2 \end{array} \right]$$

$$= \frac{1}{\left(\frac{\|x_1 - c_1\|}{\|x_1 - c_1\|} \right)^2 + \left(\frac{\|x_1 - c_1\|}{\|x_1 - c_2\|} \right)^2}$$

$$= \frac{1}{1 + \left(\frac{\|x_1 - c_1\|}{\|x_1 - c_2\|} \right)^2}$$

$$= \frac{1}{1 + \left(\frac{\|(3,4) - (3.9, 5.35)\|}{\|(3,4) - (5.2, 6.3)\|} \right)^2}$$

$$= \frac{1}{1 + \frac{(3-3.9)^2 + (4-5.35)^2}{(3-5.2)^2 + (4-6.3)^2}}$$

$$= \frac{1}{1 + \frac{(-0.9)^2 + (-1.35)^2}{(-2.2)^2 + (2.3)^2}}$$

1

$$= \frac{1}{1+0.26}$$

$$\Rightarrow u_{ij} = u_{11} = \frac{1}{1+0.26}$$

$$= \frac{1}{1.26}$$

$$\therefore u_{11} = 0.794$$

When $i=1, j=2$; $u_{ij} = \frac{1}{\sum_{k=1}^{c} \left(\frac{\|x_i - c_j\|}{\|x_i - c_k\|} \right)^{\frac{2}{m-1}}}$

$$\Rightarrow u_{12} = \frac{1}{\sum_{k=1}^{2} \left(\frac{\|x_1 - c_2\|}{\|x_1 - c_k\|} \right)^{2}} \text{; [if } m=2, c=2]$$

$$= \frac{1}{\left(\frac{\|x_1 - c_2\|}{\|x_1 - c_1\|} \right)^2 + \left(\frac{\|x_1 - c_2\|}{\|x_1 - c_2\|} \right)^2}$$

$$= \frac{1}{\left(\frac{\|x_1 - c_2\|}{\|x_1 - c_1\|} \right)^2 + 1}$$

$$= \frac{1}{\left\{ \frac{\|(3,4) - (5.2, 6.3)\|}{\|(3,4) - (3.9, 5.35)\|} \right\}^2 + 1}$$

$$U_{12} = \frac{1}{\left\{\frac{\|(3,4)-(5.2,6.3)\|}{\|(3,4)-(3.9,5.35)\|}\right\}^2 + 1}$$

$$= \frac{1}{\frac{(3-5.2)^2+(4-6.3)^2}{(3-3.9)^2+(4-5.35)^2}+1}$$

$$= \frac{1}{\frac{(-2.2)^2+(-2.3)^2}{(-0.9)^2+(-1.35)^2}+1}$$

$$= \frac{1}{\frac{10.13}{2.6325}+1} = \frac{1}{3.85+1} = \frac{1}{4.85} = 0.206$$

When $i=2,\ j=1$; $U_{ij} = \frac{1}{\sum_{k=1}^{c}\left(\frac{\|x_i - c_j\|}{\|x_i - c_k\|}\right)^{\frac{2}{m-1}}}$

$$\Rightarrow U_{21} = \frac{1}{\sum_{k=1}^{2}\left(\frac{\|x_2 - c_1\|}{\|x_2 - c_k}\right)^2}$$ [if $m=2$, $c=2$]

$$= \frac{1}{\left(\frac{\|x_2-c_1\|}{\|x_2-c_1\|}\right)^2 + \left(\frac{\|x_2-c_1\|}{\|x_2-c_2\|}\right)^2}$$

$$\Rightarrow U_{21} = \frac{1}{\left(\frac{\|x_2 - c_1\|}{\|x_2 - c_1\|}\right)^2 + \left(\frac{\|x_2 - c_1\|}{\|x_2 - c_2\|}\right)^2}$$

$$= \frac{1}{1 + \left\{\frac{\|(6,7) - (3.9, 5.35)\|}{\|(6,7) - (5.2, 6.3)\|}\right\}^2}$$

$$= \frac{1}{1 + \frac{(6-3.9)^2 + (7-5.35)^2}{(6-5.2)^2 + (7-6.3)^2}}$$

$$= \frac{1}{1 + \frac{2.1^2 + 1.65^2}{0.8^2 + 0.7^2}}$$

$$= \frac{1}{1 + \frac{7.1325}{1.13}}$$

$$= \frac{1}{1 + 6.312} = \frac{1}{7.312} = 0.137$$

When $i=2, j=2; \; U_{ij} = \dfrac{1}{\sum_{k=1}^{c} \left(\dfrac{\|x_i - c_j\|}{\|x_i - c_k\|} \right)^{\frac{2}{m-1}}}$

$$\Rightarrow U_{22} = \frac{1}{\sum_{k=1}^{2} \left(\frac{\|x_2 - c_2\|}{\|x_2 - c_k\|} \right)^2} \quad \left[\text{if } m=2, \; c=2\right]$$

$$= \frac{1}{\left(\frac{\|x_2 - c_2\|}{\|x_2 - c_1\|} \right)^2 + \left(\frac{\|x_2 - c_2\|}{\|x_2 - c_2\|} \right)^2}$$

$$= \frac{1}{\left\{ \frac{\|(6,7) - (5.2, 6.3)\|}{\|(6,7) - (3.9, 5.35\|} \right\}^2 + 1}$$

$$= \frac{1}{\frac{(6-5.2)^2 + (7-6.3)^2}{(6-3.9)^2 + (7-5.35)^2} + 1}$$

$$= \frac{1}{\frac{0.8^2 + 0.7^2}{2.1^2 + 1.65^2} + 1} = \frac{1}{\frac{1.13}{7.1325} + 1}$$

$$= \frac{1}{0.158 + 1} = \frac{1}{1.158}$$

When $i=3, j=1$; $U_{ij} = \dfrac{1}{\sum_{K=1}^{c} \left(\dfrac{\|x_i - c_j\|}{\|x_i - c_K\|} \right)^{\frac{2}{m-1}}}$

$$\Rightarrow U_{31} = \frac{1}{\sum_{K=1}^{2} \left(\dfrac{\|x_3 - c_1\|}{\|x_3 - c_K\|} \right)^2} \quad \left[\begin{array}{l} \text{if } m=2 \\ c=2 \end{array} \right]$$

$$= \frac{1}{\left(\dfrac{\|x_3 - c_1\|}{\|x_3 - c_1\|} \right)^2 + \left(\dfrac{\|x_3 - c_1\|}{\|x_3 - c_2\|} \right)^2}$$

$$= \frac{1}{1 + \left\{ \dfrac{\|(4,6) - (3.9, 5.35)\|}{\|(4,6) - (5.2, 6.3)\|} \right\}^2}$$

$$= \frac{1}{1 + \dfrac{(4-3.9)^2 + (6-5.35)^2}{(4-5.2)^2 + (6-6.3)^2}}$$

$$= \frac{1}{1 + \dfrac{0.1^2 + 0.65^2}{(-1.2)^2 + (-0.3)^2}}$$

$$= \frac{1}{\quad} = \frac{1}{\quad} = \frac{1}{\quad} = 0.780$$

When $i=3, j=2; U_{ij} = \frac{1}{\sum_{k=1}^{c} \left(\frac{\|x_i - c_j\|}{\|x_i - c_k\|} \right)^{\frac{2}{m-1}}}$

$$\Rightarrow U_{32} = \frac{1}{\sum_{k=1}^{2} \left(\frac{\|x_3 - c_2\|}{\|x_3 - c_k\|} \right)^{2}} \quad \left[\text{if } m=2,\ c=2\right]$$

$$= \frac{1}{\left(\frac{\|x_3 - c_2\|}{\|x_3 - c_1\|} \right)^{2} + \left(\frac{\|x_3 - c_2\|}{\|x_3 - c_2\|} \right)^{2}}$$

$$= \frac{1}{\left\{ \frac{\|(4,6) - (5.2, 6.3)\|}{\|(4,6) - (3.9, 5.35)\|} \right\}^{2} + 1}$$

$$= \frac{1}{\frac{(4-5.2)^2 + (6-6.3)^2}{(4-3.9)^2 + (6-5.35)^2} + 1}$$

$$= \frac{1}{\frac{1.53}{\ldots} + 1} = \frac{1}{3.537 + 1}$$

$$\therefore U^{(1)} = \begin{bmatrix} 0.794 & 0.206 \\ 0.137 & 0.863 \\ 0.780 & 0.220 \end{bmatrix}$$

Step 4:

$$|U_{11}^{(1)} - U_{11}^{(0)}| = |0.794 - 0.8| = 0.006$$

$$|U_{12}^{(1)} - U_{12}^{(0)}| = |0.206 - 0.2| = 0.006$$

$$|U_{21}^{(1)} - U_{21}^{(0)}| = |0.137 - 0.3| = 0.163$$

$$|U_{22}^{(1)} - U_{22}^{(0)}| = |0.863 - 0.7| = 0.163$$

$$|U_{31}^{(1)} - U_{31}^{(0)}| = |0.780 - 0.9| = 0.12$$

$$|U_{32}^{(1)} - U_{32}^{(0)}| = |0.22 - 0.1| = 0.12$$

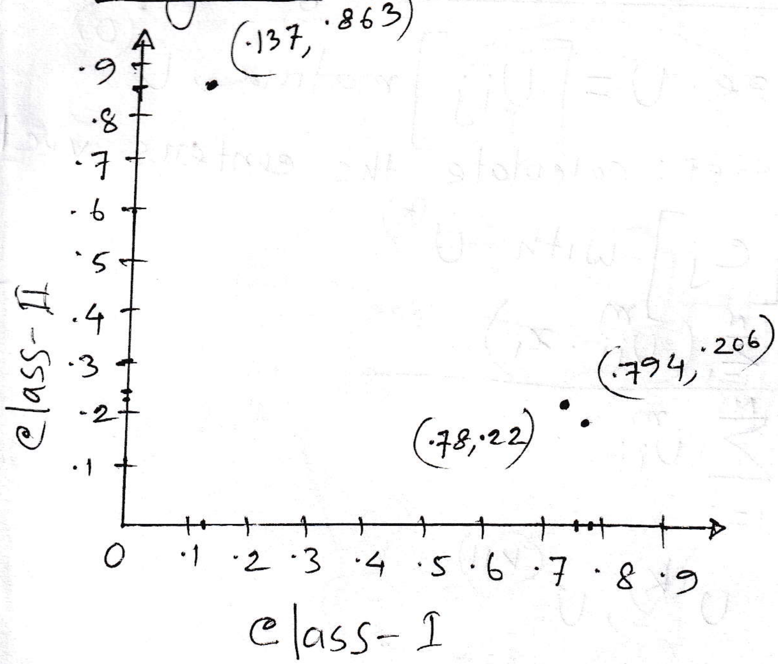
$$\max(0.006, 0.163, 0.12) = 0.163 < \epsilon$$ [if $\epsilon = 3$] [if $\epsilon = 0.3$]

Algorithm stop.

Output:

Data	Class-I	Class-II
(3,4)	79.4 %	20.6 %
(6,7)	13.7 %	86.3 %
(4,6)	78 %	22 %

Diagram:



Fuzzy c-means clustering algorithm

step1: Initialize $U = [U_{ij}]$ matrix, $U^{(0)}$

step2: At K-step: calculate the centers vectors $c^{(K)} = [c_j]$ with $U^{(K)}$

$$c_j = \frac{\sum_{i=1}^{N} \left(U_{ij}^{m} \cdot x_i\right)}{\sum_{i=1}^{N} U_{ij}^{m}}$$

step3: Update $U^{(K)}, U^{(K+1)}$

$$U_{ij} = \frac{1}{\sum_{K=1}^{c} \left(\frac{\|x_i - c_j\|}{\|x_i - c_K\|}\right)^{\frac{2}{m-1}}}$$

step4: If $\|U^{(K+1)} - U^{(K)}\| < \epsilon$ then stop; otherwise return to step 2.

Parameters:
Fuzzyness co-efficient $m=2$ (m any real number greater than 1.)
Termination condition, $\max_{ij}\left\{\left|U_{ij}^{(K+1)} - U_{ij}^{(K)}\right|\right\} < \epsilon$

where ϵ is the termination criterion between 0 and 1, where as K are the iterations steps.
U_{ij} = degree of membership of x_i in the cluster j.
x_i = i th of d-dimensional measured data.
C_j = d-dimension center of the cluster.
... ... similarity between