Clustering: It is similar to classification in that data is grouped. However unlike classification. The groups are not predefined The groups are Called clusters.
each cluster is a set of like elements of raw different clusters are not alike

Clustering Can be Considered the most important unsupervised learning problem. The definition of clustering Could be "The Process of Organizing objects into groups whose members are similar in some way". A cluster is therefore a collection of objects which are "similar" between them and are "dissimilar" to the objects belonging to other clusters.

العنقده هي عمليه تصنيف ولكن الفرق بينها وبين
classification هو ان المجموعه تكون معرفه مسبقا فيه
وتكون predefined لهذا تكون supervised
(تعلم موجه) بينما في Cluster (العنقده) لا يوجد بيانا
معرفه مسبقاً (not predefined) ولهذا يسمى تعلم غير
موجه unsupervised لأن العناقيد تكتب بعد فحص
البيانات
وكل عنقود هو مجموعه من العناصر المتشابهه التي تختلف عن عناصر
العناقيد الأخرى

Def: Given a database $D = \{t_1, \ldots, t_n\}$ of tuples and an integer k.

The clustering problem is to define a mapping $f : D \to \{1, 2, \ldots, k\}$, where each t_i is assigned to one cluster C_j, $j = 1, 2, \ldots, k$.
A cluste C_j contains precisely those tuples Mapped to it.

i.e. $C_j = \{t_i : f(t_i) = C_j,\ i = 1, 2, \ldots, n,\ t_i \in D\}$
This means that given a cluster C_j $\forall\, t_{il}, t_{im} \in C_j$ and $t_i \notin C_j$
$dist(t_{il}, t_{jm}) \le dist(t_{il}, t_i)$
K هو عدد العناقيد

العنقدة: هي عملية وضع العناصر المتشابهة ضمن عنقود واحد و هذا العنقود تختلف عن باقي العناقيد.
∴ كل عنقود يضم عناصر وهذه العناصر تكون متشابهة مع بعضها البعض و في نفس الوقت مختلفة عن بقية العناصر من حيث التشابه.

هكذا مسائل تظهر عندما يكون لدينا عدد كبير من المتغيرات مثل الفضاء حيث يتم قياس عدة امور ومن ثم تحديد التشابه مثل تحديد اي كوكب اقرب الى كوكب الارض.

او مثلا يتم اخذ دول العالم و ادخال متغيرات مثل (مساحة كل دولة و عدد السكان و معدل الدخل الشهري و عدد الاطباء في كل دولة و تحديد عددهم في كل دولة) وبعد استخدام العنقدة بتبين ان دول اوروبا الغربية وامريكا و كندا واليابان واستراليا تقع ضمن عنقود و دول اوربا الشرقية وامريكا اللاتينية في عنقود واحد. والهند لوحدها ضمن عنقود.

Example:

objects	X_1	X_2
O_1	3	65
O_2	5	70
O_3	14	97
O_4	2	60
O_5	11	90

Find the cluster?

Solution: Denotes the distance between objects O_i and O_j by d_{ij}

$$d_{ij} = \| O_i - O_j \| = \sqrt{\underbrace{[o_i - o_j]^2}_{x_1} + \underbrace{[o_i - o_j]^2}_{x_2}}$$

عدد المسافات التي يجب حسابها $= \dfrac{n(n-1)}{2} = \dfrac{5(4)}{2} = 10$

$d_{12} = \sqrt{(3-5)^2 + (65-70)^2} = 5.4$

$d_{13} = \sqrt{(3-14)^2 + (65-97)^2} = 33.8$

$d_{14} = \sqrt{(3-2)^2 + (65-60)^2} = 5.1$

d_{15}

d_{23}

d_{24}

d_{25}

d_{34}

$d_{35} = \sqrt{(14-11)^2 + (97-90)^2} = 7.9$

$d_{45} = \sqrt{(2-11)^2 + (60-90)^2} = 31$

The distance Matrix

$$D = \begin{array}{c} \\ O_1 \\ O_2 \\ O_3 \\ O_4 \\ O_5 \end{array} \begin{array}{c} \begin{array}{ccccc} O_1 & O_2 & O_3 & O_4 & O_5 \end{array} \\ \begin{bmatrix} 0 & 5.4 & 33.8 & 5.1 & 26.3 \\ 5.4 & 0 & 28.5 & 10.4 & 20.9 \\ 33.8 & 28.5 & 0 & 38.9 & 7.9 \\ 5.1 & 10.4 & 38.9 & 0 & 31 \\ 26.3 & 20.9 & 7.9 & 31 & 0 \end{bmatrix} \end{array}$$

A similarity measure between objects O_i and O_j is given by

$$S_{ij} = \left(1 - \frac{d_{ij}}{d_{max}}\right) * 100\%$$

$$d_{max} = 38.9$$

$$S_{12} = \left(1 - \frac{5.4}{38.9}\right) * 100 = 86.1$$

$$S_{13} = \left(1 - \frac{33.8}{38.9}\right) * 100 = 13.1$$

$$S_{14} = \left(1 - \frac{5.1}{38.9}\right) * 100 = 86.9$$

$$S_{15} = \left(1 - \frac{26.3}{38.9}\right) * 100 = 32.4$$

$$S_{23} = \left(1 - \frac{28.5}{38.9}\right) * 100 = 26.7$$

$$S_{24} = \left(1 - \frac{10.9}{38.9}\right) * 100 = 73.3$$

$$S_{25} = \left(1 - \frac{20.9}{38.9}\right) * 100 = 0$$

$$S_{34} = \left(1 - \frac{7.9}{38.9}\right) * 100 = 79.7$$

$$S_{45} = \left(1 - \frac{31}{38.9}\right) * 100 = 20.3$$

$$S_{11} = S_{22} = S_{33} = S_{44} = S_{55} = \left(1 - \frac{0}{38.9}\right) * 100 = 100$$

Hence, the similarity matrix is given by

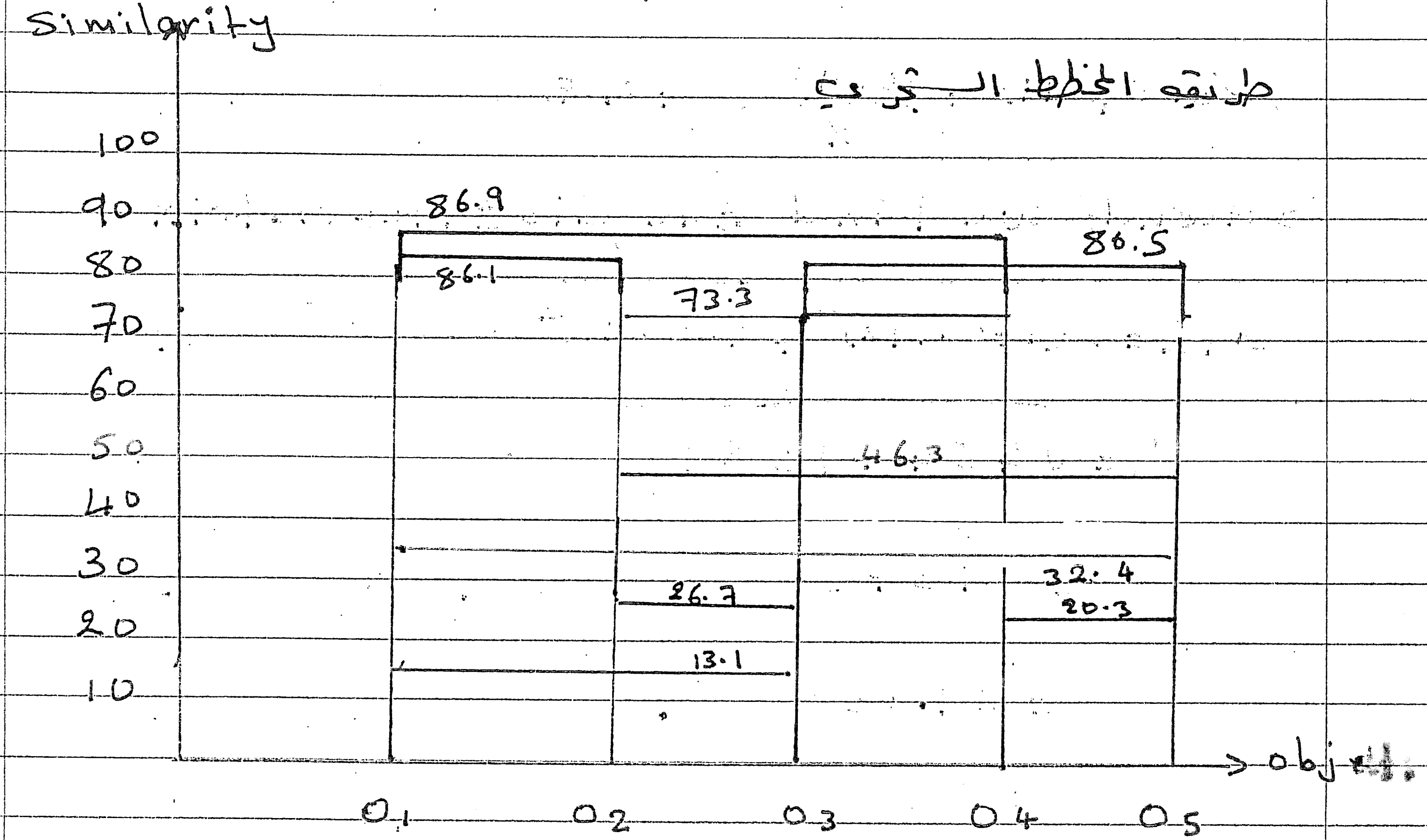
$$S_m = \begin{array}{c|ccccc} & O_1 & O_2 & O_3 & O_4 & O_5 \\ \hline O_1 & 100 & 86.1 & 13.1 & 86.9 & 32.4 \\ O_2 & & 100 & 26.7 & 73.3 & 46.3 \\ O_3 & & & 100 & 0 & 79.7 \\ O_4 & & & & 100 & 20.3 \\ O_5 & & & & & 100 \end{array}$$

نبحث عن اكبر تشابه بين O_1 وباقي العناصر

$$C_1 = \{O_1, O_2, O_4\}$$

$$C_2 = \{O_3, O_5\}$$

طريقة المخطط الشجري



H.w

classify the following players according to their Highets and weights (using clusters

	X_1 (Hight cm)	X_2 weight (kg)
O1	170	60
O2	192	93
O3	163	65
O4	165	62
O5	163	63
O6	187	94
O7	193	102
O8	192	92
O9	193	97
O10	162	52

solution: we first find the number of distance

$$\text{No. of dis} = \frac{10(9)}{2} = 45$$

second: we find the distance matrix

$$d_{12} = \sqrt{(170-192)^2 + (60-93)^2}$$

$$= \sqrt{484 + 1089}$$

$$= \sqrt{1573}$$

$$= 39.66$$

$$d_{13} = \sqrt{(170-163)^2 + (60-65)^2}$$

$$= 8.6$$

$d_{14} = 5.385$

$d_{15} = 7.6$

$d_{16} = 38$

$$D = \begin{array}{c|cccccccccc} & O_1 & O_2 & O_3 & O_4 & O_5 & O_6 & O_7 & O_8 & O_9 & O_{10} \\ \hline O_1 & 0 & 39.7 & 8.6 & 5.4 & 7.6 & 38 & 47.9 & 38.8 & 43.6 & 11.6 \\ O_2 & & 0 & 40.3 & 41.1 & 41.7 & 5 & 12.04 & 1 & 4.1 & 50.8 \\ O_3 & & & 0 & 3.6 & 2 & 37.6 & 47.6 & 39.6 & 43.9 & 13 \\ O_4 & & & & 0 & 2.2 & 38.8 & 48.8 & 40.4 & 41.8 & 10.4 \\ O_5 & & & & & 0 & 39.2 & 49.2 & 41 & 45.3 & 11 \\ O_6 & & & & & & 0 & 11 & 5.4 & 6.7 & 48.8 \\ O_7 & & & & & & & 0 & 10.05 & 5 & 38.8 \\ O_8 & & & & & & & & 0 & 5 & 50 \\ O_9 & & & & & & & & & 0 & 45.3 \\ O_{10} & & & & & & & & & & 0 \end{array}$$

$$S = 100 \times \left(1 - \frac{d_{ij}}{d_{Max}}\right)$$

$d_{Max} = 50.8$

$S_{11} = S_{22} = S_{33} = S_{44} = S_{55} = S_{66} = S_{77} = S_{88} = S_{99} = S_{1010} = 100$

$$
S = \begin{array}{c} \\ O_1 \\ O_2 \\ O_3 \\ O_4 \\ O_5 \\ O_6 \\ O_7 \\ O_8 \\ O_9 \\ O_{10} \end{array}
\begin{array}{c}
\begin{array}{cccccccccc} O_1 & O_2 & O_3 & O_4 & O_5 & O_6 & O_7 & O_8 & O_9 & O_{10} \end{array} \\
\left[\begin{array}{cccccccccc}
100 & 21.85 & 83.07 & 89.37 & 85.04 & 25.19 & 5.71 & 23.62 & 85.83 & 77.76 \\
 & 100 & 20.28 & 19.09 & 17.91 & 90.16 & 76.29 & 98.03 & 91.92 & 0 \\
 & & 100 & 29.91 & 96.06 & 25.98 & 6.29 & 22.05 & 13.58 & 74.41 \\
 & & & 100 & 95.67 & 23.62 & 3.94 & 20.47 & 11.8 & 79.1 \\
 & & & & 100 & 22.83 & 3.15 & 17.29 & 10.83 & 78.35 \\
 & & & & & 100 & 78.35 & 89.37 & 86.81 & 3.94 \\
 & & & & & & 100 & 80.22 & 90.15 & 23.62 \\
 & & & & & & & 100 & 90.15 & 1.57 \\
 & & & & & & & & 100 & 10.83 \\
 & & & & & & & & & 100
\end{array}\right]
\end{array}
$$

$C_1 = \{O_1, O_3, O_4, O_5, O_9, O_{10}\}$

$C_2 = \{O_2, O_7, O_8, O_6\}$

Hierachical Clustering: is a widely used Method for detecting clusters and seek to bulid a hierachy of clusters or is a techique to group objects based on distance or similarity.

ⓐ single link Method: single link algorithum is an example of agglomerative hierachical clustering method

we recall that is a bottom-up strategy Compare each point with each point. Each object is placed in a separate cluster and at each step we merge the closet pair of cluster

Example: Let D be a distance matrix find the clusters using
① single link method

$$D = \begin{array}{c|ccccc} & O_1 & O_2 & O_3 & O_4 & O_5 \\ \hline O_1 & 0 & 5.4 & 33.8 & 5.1 & 26.3 \\ O_2 & & 0 & 28.5 & 10.4 & 20.9 \\ O_3 & & & 0 & 38.9 & 7.9 \\ O_4 & & & & 0 & 31 \\ O_5 & & & & & 0 \end{array}$$

Solution: First step
we find the smallest distance in D

The smallest distance = 5.1 (O_1, O_4)

$C_1 = \{O_1, O_4\}$

$dis(C_1, O_2) = \min(d_{O_1,O_2}, d_{O_4,O_2})$

$= \min(5.4, 10.4)$

$= 5.4$

$dis(C_1, O_3) = \min(d_{O_1,O_3}, d_{O_4,O_3})$

$= \min(33.8, 38.9)$

$= 33.8$

$dis(C_1, O_5) = \min(d_{O_1,O_5}, d_{O_4,O_5})$

$= \min(26.3, 31)$

$= 26.3$

$$D_1 = \begin{array}{c|cccc} & C_1 & O_2 & O_3 & O_5 \\ \hline C_1 & 0 & 5.4 & 33.8 & 26.3 \\ O_2 & & 0 & 28.5 & 20.9 \\ O_3 & & & 0 & 7.9 \\ O_5 & & & & 0 \end{array}$$

The smallest distance in $D_1 = 5.4$

$C_2 = \{C_1, O_2\}$

$d_{C_2,O_3} = \min\{d_{C_1,O_3}, d_{O_2,O_3}\}$

$= \min(33.8, 28.5)$

$= 28.5$

$$d_{C_2,O_5} = \min(d_{C_1,O_5}, d_{O_2,O_5})$$

$$= \min(26.3, 20.9)$$

$$= 20.9$$

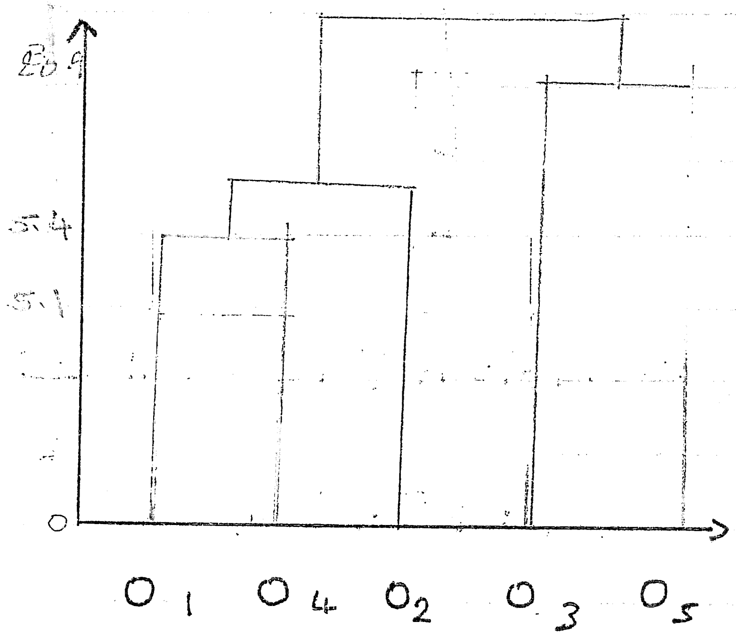
$$D_2 = \begin{array}{c|ccc} & C_2 & O_3 & O_5 \\ \hline C_2 & 0 & 28.5 & 20.9 \\ O_3 & & 0 & 7.9 \\ O_5 & & & 0 \end{array}$$

The smallest distance in $D_2 = 7.9$

$C_3 = \{O_3, O_5\}$

$d_{C_2,C_3} = \min(d_{C_2,O_3}, d_{C_2,O_5}) = \min(28.5, 20.9) = 20.9$

$$D_3 = \begin{array}{c|cc} & C_2 & C_3 \\ \hline C_2 & 0 & 20.9 \\ C_3 & & 0 \end{array}$$



(b) Complete-link Method

This method also called farthest neighbor clustering method is the oppisite of single linkage, distance between groups is now defined as the distance between the most distance pair of object, one from each group.

$D(r,s) = Max\{d(i,j)$: where object i in the cluster r and object j in the cluster$\}$

Example: Find the cluster of the distance Matrix by using Complete-linkage

$$D = \begin{array}{ccccc} O_1 & O_2 & O_3 & O_4 & O_5 \end{array}$$

$$\begin{bmatrix} 0 & 5.4 & 33.8 & 5.1 & 26.3 \\ & 0 & 28.5 & 10.4 & 20.9 \\ & & 0 & 38.9 & 7.9 \\ & & & 0 & 31 \\ & & & & 0 \end{bmatrix}$$

Solution: The smallest distance in $D = 5.1$

$C_1 = \{O_1, O_4\}$

Now, we find the Maximum distance between C_1 and the other elements in D

$d_{c_1,O_2} = Max(d_{O_1,O_2}, d_{O_4,O_2})$

$= Max(5.4, 10.4)$

$= 10.4$

$d_{c_1,O_3} = Max(d_{O_1,O_3}, d_{O_4,O_3})$

$= Max(33.8, 38.9)$

$= 38.9$

$d_{c_1,O_5} = Max(d_{O_1,O_5}, d_{O_4,O_5})$

$= Max(26.3, 31)$

$= 31$

$$D_1 = \begin{array}{c} \\ C_1 \\ O_2 \\ O_3 \\ O_5 \end{array} \begin{array}{c} \begin{array}{cccc} C_1 & O_2 & O_3 & O_5 \end{array} \\ \begin{bmatrix} 0 & 10.4 & 38.9 & 31 \\ & 0 & 28.5 & 20.9 \\ & & 0 & 7.9 \\ & & & 0 \end{bmatrix} \end{array}$$

The smallest distance in $D_1 = 7.9$

$C_2 = \{O_3, O_5\}$

$d_{c_1,c_2} = Max(d_{c_1,O_3}, d_{c_1,O_5})$

$= Max(38.9, 31)$

$= 38.9$

$d_{c_2,O_2} = Max(d_{O_3,O_2}, d_{O_5,O_2})$

$= Max(28.5, 20.9)$

$= 28.5$

$$D_2 = \begin{array}{c|ccc} & C_1 & O_2 & C_2 \\ \hline C_1 & 0 & 10.4 & 38.9 \\ O_2 & & 0 & 28.5 \\ C_2 & & & 0 \end{array}$$

The smallest distance in $D_2 = 10.4$

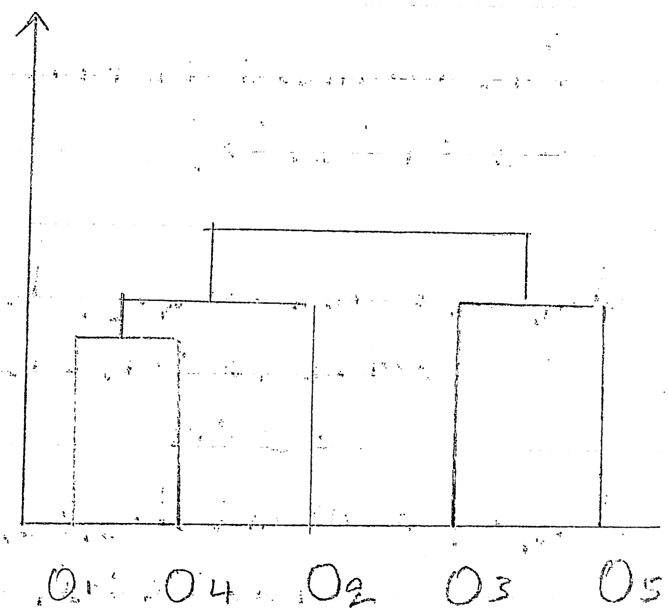
$C_3 = \{C_1, O_2\}$

$$d_{C_3, C_2} = Max(d_{C_1, C_2}, d_{O_2, C_2})$$

$$= Max(38.9, 28.5)$$

$$= 28.5$$

$$D_3 = \begin{array}{c|cc} & C_2 & C_3 \\ \hline C_2 & 0 & 28.5 \\ C_3 & & 0 \end{array}$$



4/12/2022 Sunday

Example: Let D be a distance matrix. Find the cluster by using

① single-linkage cluster
② Complete-linkage cluster

Solution:

$$D = \begin{array}{c|ccccc} & x_1 & x_2 & x_3 & x_4 & x_5 \\ \hline x_1 & 0 & 0.5 & 2.24 & 3.35 & 3 \\ x_2 & & 0 & 2.5 & 3.61 & 3.04 \\ x_3 & & & 0 & 1.12 & 1.41 \\ x_4 & & & & 0 & 1.5 \\ x_5 & & & & & 0 \end{array}$$

① Single-linkage

The smallest distance in $D = 0.5$

$C_1 = \{x_1, x_2\}$

$$d_{C_1, x_3} = \min(d_{x_1, x_3}, d_{x_2, x_3}) = \min(2.24, 2.5) = 2.24$$

$$d_{C_1, x_4} = \min(d_{x_1, x_4}, d_{x_2, x_4}) = \min(3.35, 3.61) = 3.35$$

$$d_{c_1, x_5} = \min(d_{x_1, x_5}, d_{x_2, x_5})$$

$$= \min(3, 3.04)$$

$$= 3$$

$$D_1 = \begin{array}{c|cccc} & C_1 & X_3 & X_4 & X_5 \\ \hline C_1 & 0 & 2.24 & 3.35 & 3 \\ X_3 & & 0 & 1.12 & 1.41 \\ X_4 & & & 0 & 1.5 \\ X_5 & & & & 0 \end{array}$$

The smallest distance in $D_1 = 1.12$

$$C_2 = \{X_3, X_4\}$$

$$d_{c_2, c_1} = \min(d_{c_1, x_3}, d_{c_1, x_4})$$

$$= \min(2.24, 3.35)$$

$$= 2.24$$

$$d_{c_2, x_5} = \min(d_{x_3, x_5}, d_{x_4, x_5})$$

$$= \min(1.41, 1.5)$$

$$= 1.41$$

$$D_2 = \begin{array}{c|ccc} & C_1 & C_2 & X_5 \\ \hline C_1 & 0 & 2.24 & 3 \\ C_2 & & 0 & 1.41 \\ X_5 & & & 0 \end{array}$$

The smallest distance in $D_2 = 1.41$

$C_3 = \{C_2, X_5\}$

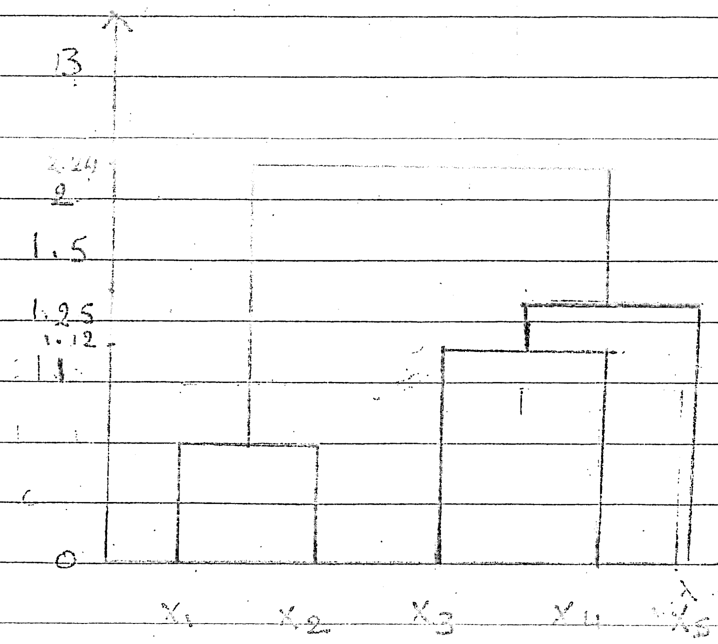
$d_{C_3,C_1} = \min(d_{C_2,C_1}, d_{C_1,X_5})$

$= \min(2.24, 3)$

$= 2.24$

$$D_3 = \begin{array}{c|cc} & C_1 & C_3 \\ \hline C_1 & 0 & 2.24 \\ C_3 & & 0 \end{array}$$

$C_4 = \{C_1, C_3\}$



② Complete-linkage cluster

① The smallest distance in D = 0.5

$C_1 = \{x_1, x_2\}$

$d_{c_1, x_3} = Max(d_{x_1, x_3}, d_{x_2, x_3})$

$= Max(2.24, 2.5)$

$= 2.5$

$d_{c_1, x_4} = Max(d_{x_1, x_4}, d_{x_2, x_4})$

$= Max(3.35, 3.6)$

$= 3.6$

$d_{c_1, x_5} = Max(d_{x_1, x_5}, d_{x_2, x_5})$

$= Max(3, 3.04)$

$= 3$

$$D_1 = \begin{array}{c} \\ C_1 \\ x_3 \\ x_4 \\ x_5 \end{array} \begin{array}{c} \begin{array}{cccc} C_1 & x_3 & x_4 & x_5 \end{array} \\ \begin{bmatrix} 0 & 2.5 & 3.6 & 3 \\ & 0 & 1.12 & 1.41 \\ & & 0 & 1.5 \\ & & & 0 \end{bmatrix} \end{array}$$

The smallest distance in $D_1 = 1.12$

$C_2 = \{x_3, x_4\}$

$$d_{c_2, c_1} = Max(d_{c_1, x_3}, d_{c_1, x_4})$$

$$= Max(2.24, 3.35)$$

$$= 3.35$$

$$d_{c_2, x_5} = Max(d_{x_3, x_5}, d_{x_4, x_5})$$

$$= Max(\cancel{1.41, 1.5})$$

$$= 1.5$$

$$D_2 = \begin{array}{c|ccc} & C_1 & C_2 & X_5 \\ \hline C_1 & 0 & 3.35 & 3 \\ C_2 & & 0 & 1.5 \\ X_5 & & & 0 \end{array}$$

The smallest distance in $D_2 = 1.5$

$$C_3 = \{C_2, X_5\}$$

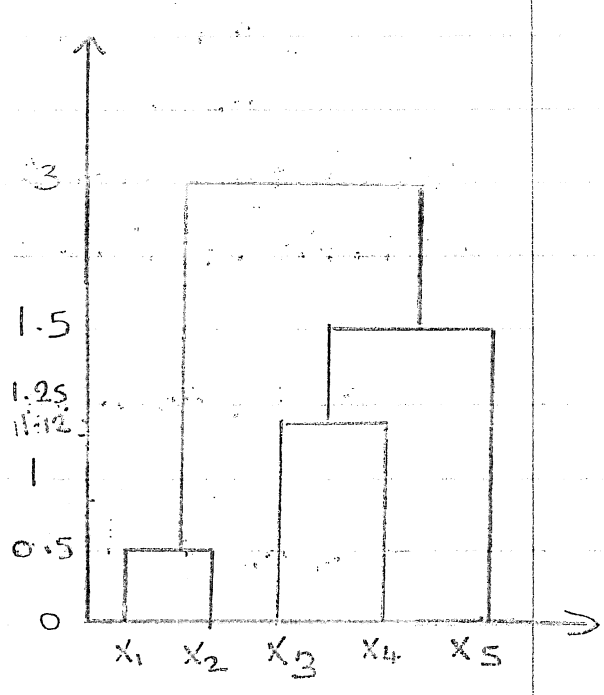
$$d_{C_3, C_1} = Max(d_{C_2, C_1}, d_{C_1, X_5})$$

$$= Max(2.24, 3)$$

$$= 3$$

$$D_3 = \begin{array}{c|cc} & C_1 & C_3 \\ \hline C_1 & 0 & 3 \\ C_3 & & 0 \end{array}$$

$$C_4 = \{C_1, C_3\}$$



The Average -Linkage.

In the average linkage approach, the distance between two clusters A and B is defined as the average of the $n_A n_B$ distances between the n_A points in A and the n_B points in B:

$$D(A,B) = \frac{1}{n_A n_B} \sum_{i=1}^{n_A} \sum_{j=1}^{n_B} d(y_i, y_j)$$

where the sum is over all y_i in A and all y_j in B. At each step, we join the two clusters with the smallest distance

Example: Let

$$D = \begin{array}{c} \\ x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \end{array} \begin{array}{c} \begin{array}{ccccc} x_1 & x_2 & x_3 & x_4 & x_5 \end{array} \\ \begin{bmatrix} 0 & 0.5 & 2.24 & 3.35 & 3 \\ & 0 & 2.5 & 3.61 & 3.04 \\ & & 0 & 1.12 & 1.41 \\ & & & 0 & 1.5 \\ & & & & 0 \end{bmatrix} \end{array}$$

Find the cluster by using average linkage.

solution: The smallest distance in D = 0.5

$C_1 = \{x_1, x_2\}$

$$d_{C_1, x_3} = d(C_1, x_3) = \frac{1}{2}\left(d(x_1, x_3) + d(x_2, x_3)\right)$$

$= 0.5(2.24 + 2.5)$
$= 2.37$

$$d_{c_1, x_4} = 0.5(d_{x_1, x_4} + d_{x_2, x_4})$$
$$= 0.5(3.35 + 3.61)$$
$$= 3.48$$

$$d_{c_1, x_5} = 0.5(d_{x_1, x_5} + d_{x_2, x_5})$$
$$= 0.5(3 + 3.04)$$
$$= 3.02$$

$$D_1 = \begin{array}{c|cccc} & C_1 & X_3 & X_4 & X_5 \\ \hline C_1 & 0 & 2.37 & 3.48 & 3.02 \\ X_3 & & 0 & 1.12 & 1.41 \\ X_4 & & & 0 & 1.5 \\ X_5 & & & & 0 \end{array}$$

The smallest distance in $D_1 = 1.12$

$C_2 = \{X_3, X_4\}$

$$d_{c_1, c_2} = 0.5(d_{c_1, x_3} + d_{c_1, x_4})$$
$$= 0.5(2.37 + 3.48)$$
$$= 2.925$$

OR

$$d_{c_1, c_2} = \frac{1}{4}(d_{x_1, x_4} + d_{x_2, x_4} + d_{x_1, x_3} + d_{x_2, x_3})$$
$$= \frac{1}{4}(3.35 + 2.24 + 3.61 + 2.5)$$

$$d_{C_2,X_5} = 0.5(d_{X_3,X_5} + d_{X_4,X_5})$$

$$= 0.5(1.41 + 1.5)$$
$$= 1.46$$

$$D_2 = \begin{array}{c|ccc} & C_1 & C_2 & X_5 \\ \hline C_1 & 0 & 2.925 & 3.02 \\ C_2 & & 0 & 1.46 \\ X_5 & & & 0 \end{array}$$

The smallest distance in $D_2 = 1.46$

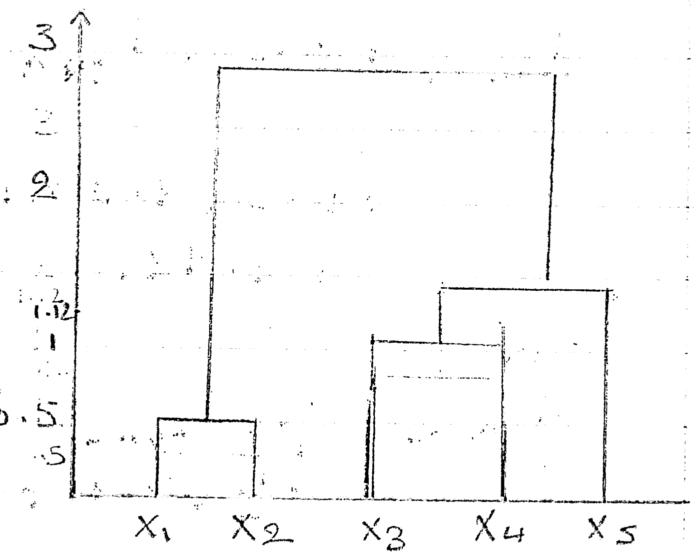
$C_3 = \{C_2, X_5\}$

$$d_{C_1,C_3} = 0.5(d_{C_1,C_2 + C_1,X_5})$$

$$= 0.5(2.925 + 3.02)$$
$$= 2.9725$$

$$D_3 = \begin{array}{c|cc} & C_1 & C_3 \\ \hline C_1 & 0 & 2.9725 \\ C_3 & & 0 \end{array}$$

$C_4 = \{C_1, C_3\}$



Example: Let X and Y are two vectors as Follows:

$$X = [1 \quad 1 \quad 3 \quad 4 \quad 4]$$
$$y = [2 \quad 2.5 \quad 1 \quad 0.5 \quad 2]$$

Find the distance matrix and find the cluster by using average linkage.

solution:

we used ecludean distance :

$d_{11} = d_{22} = d_{33} = d_{44} = d_{55} = 0$

$d_{12} = \sqrt{(1-1)^2 + (2-2.5)^2} = 0.5$

$d_{13} = \sqrt{(1-3)^2 + (2-1)^2} = \sqrt{5} = 2.2361$

$d_{14} = \sqrt{(1-4)^2 + (2-0.5)^2} = 3.354$

$d_{15} = \sqrt{(1-4)^2 + (2-2)^2} = 3$

$d_{23} = \sqrt{(1-3)^2 + (2.5-1)^2} = 2.5$

$d_{24} = \sqrt{(1-4)^2 + (2.5-0.5)^2} = 3.6$

$d_{25} = \sqrt{(1-4)^2 + (2.5-2)^2} = 3.04$

$d_{34} = \sqrt{(3-4)^2 + (1-0.5)^2} = 1.118$

$d_{35} = \sqrt{(3-4)^2 + (1-2)^2} = \sqrt{2} = 1.414$

$d_{45} = \sqrt{(4-4)^2 + (0.5-2)^2} = 1.5$

$$D = \begin{array}{c|ccccc} & X_1 & X_2 & X_3 & X_4 & X_5 \\ \hline X_1 & 0 & 0.5 & 2.2361 & 3.35 & 3 \\ X_2 & & 0 & 2.5 & 3.6 & 3.04 \\ X_3 & & & 0 & 1.118 & 1.414 \\ X_4 & & & & 0 & 1.5 \\ X_5 & & & & & 0 \end{array}$$

The smallest distance in $D = 0.5$

$C_1 = \{X_1, X_2\}$

$$d_{C_1, X_3} = 0.5(2.2361 + 2.5) = 2.36805$$

$$d_{C_1, X_4} = 0.5(3.35 + 3.6) = 3.475$$

$$d_{C_1, X_5} = 0.5(3 + 3.04) = 3.02$$

$$D_1 = \begin{array}{c|cccc} & C_1 & X_3 & X_4 & X_5 \\ \hline C_1 & 0 & 2.36805 & 3.475 & 3.02 \\ X_3 & & 0 & 1.118 & 1.414 \\ X_4 & & & 0 & 1.5 \\ X_5 & & & & 0 \end{array}$$

The smallest distance in $D_1 = 1.118$

$C_2 = \{X_3, X_4\}$

$$d_{c_2,c_1} = 0.5\,(d_{c_1,x_3} + d_{c_1,x_4})$$

$$= 0.5\,(2.36805 + 3.475)$$

$$= 2.922$$

$$d_{c_2,x_5} = 0.5\,(1.414 + 1.5)$$

$$= 1.457$$

$$D_2 = \begin{array}{c} \\ c_1 \\ c_2 \\ x_5 \end{array} \begin{array}{c} \begin{array}{ccc} c_1 & c_2 & x_5 \end{array} \\ \begin{bmatrix} 0 & 2.929 & 3.02 \\ & 0 & 1.457 \\ & & 0 \end{bmatrix} \end{array}$$

The smallest distance in $D_2 = 1.457$

$$c_3 = \{c_2, x_5\}$$

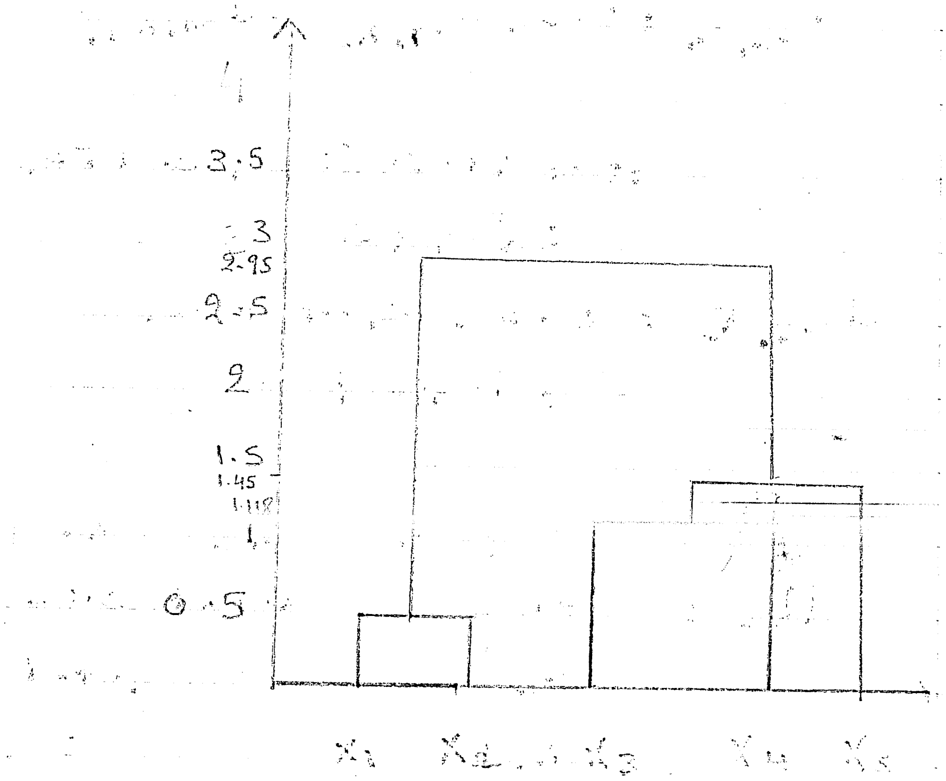
$$d_{c_3,c_1} = 0.5\,(d_{c_2,c_1} + d_{c_1,x_5})$$

$$= 0.5\,(2.929 + 3.02)$$

$$= 2.9745$$

$$D_3 = \begin{array}{c} \\ c_1 \\ c_3 \end{array} \begin{array}{c} \begin{array}{cc} c_1 & c_3 \end{array} \\ \begin{bmatrix} 0 & 2.9745 \\ & 0 \end{bmatrix} \end{array}$$

$$\therefore c_4 = \{c_1, c_3\}$$



By using Matlab program

```
X=[1 2; 1 2.5; 3 1; 4 0.5; 4 2];
y = pdist(x)
k = squareform(y)
Z = linkage(k)
dendrogram(Z)
```

ملاحظات

١- ركز على المصفوفه لاشيء غير المصفوفه (صديقنا سبونج بوب)

٢- الايعاز pdist لحساب المسافه الاقليديه واذا اردنا حساب اي نوع اخر من المسافات فيكون الايعاز بالشكل

```
y = pdist(X, 'cityblock')
```

قانون cityblock

$$d_M(x_i, x_j) = \sum_{k=1}^{n} |x_{ik} - x_{jk}|$$

حيث تصبح قيم D بالشكل

$$D = \begin{array}{ccccc} X_1 & X_2 & X_3 & X_4 & X_5 \end{array}$$
$$\begin{bmatrix} 0 & 0.5 & 3 & 4.5 & 3 \\ & 0 & 3.5 & 5 & 3.5 \\ & & 0 & 1.5 & 2 \\ & & & 0 & 1.5 \\ & & & & 0 \end{bmatrix}$$

٢- اذا اتشابهت قيمتان او اكثر في مصفوفة المسافة فهنا نلجأ لحساب مصفوفة التشابه ك ثم نجد العناقيد

الايعاز dendrogram لرسم المخطط الشجري للعناقيد

٣- اضافة لاستخدام برنامج [illegible] يكون بالشكل

بعد ادخال المتغيرات كأعمدة نذهب الى

Analysis → Correlate → Distance →
between cases → ok

هنا تظهر المصفوفة وايضا الرسم.

H.w

$$D = \begin{array}{c|cccc} & A & B & C & D \\ \hline A & 0 & 5 & 3 & 2 \\ B & & 0 & 4 & 2 \\ C & & & 0 & 1 \\ D & & & & 0 \end{array}$$

Find the cluster by using

1- single-link
2- complete link
3- Average-link.

The Centroid Algorithm

The centroid algorithm is also referred as the "unweighted pair group method using centroids" .Two clusters are merged based on the distance of their centroids (mean, median, mods),defined as

$$m_i = \frac{1}{n}\sum_{x \in i} x$$

Where n_i is the number of data point belonging to the **cluster**

$$d(C_k,\{C_i,C_j\}) = \frac{n_i}{n_i+n_j} d(C_k, C_i) + \frac{n_j}{n_i+n_j} d(C_k, C_j) - \frac{n_i n_j}{(n_i+n_j)^2} d(C_i, C_j)$$

this definition is equivalent to the calculation of the squared Euclidean distance between the centroid of the two clusters

$$d(C_k,\{C_i,C_j\}) = \|m_k - m_{ij}\|^2$$

Applying the centroid algorithm to the dataset, the first stage is still the same as in other algorithm, x_1, and x_2 are merged.

Example : find the cluster of the distance matrix A

$$\begin{array}{cc} & \begin{array}{ccccc} x_1 & x_2 & x_3 & x_4 & x_5 \end{array} \\ A = \begin{array}{c} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \end{array} & \begin{bmatrix} 0 & 0.5 & 2.24 & 3.35 & 3 \\ & 0 & 2.5 & 3.61 & 3.04 \\ & & 0 & 1.12 & 1.41 \\ & & & 0 & 1.5 \\ & & & & 0 \end{bmatrix} \end{array}$$

Solution: After x_1 and x_2 are merged, the distance are:

$d(\{x_1,x_2\},x_3)=0.5[(d(x_1,x_3)+ (d(x_2,x_3))] -0.25\, d(x_1,x_2)$

$=2.245$

$d(\{x_1,x_2\},x_4)=0.5[(d(x_1,x_4)+ (d(x_2,x_4))]-0.25\, d(x_1,x_2)$

$=3.355$

$d(\{x_1,x_2\},x_5)=0.5[(d(x_1,x_5)+ (d(x_2,x_5))]-0.25\, d(x_1,x_2)$

$=2.895$

And the similarity matrix becomes

$$A_1 = \begin{array}{c} \\ C_1 \\ x_3 \\ x_4 \\ x_5 \end{array} \begin{array}{c} \begin{array}{cccc} C_1 & x_3 & x_4 & x_5 \end{array} \\ \begin{bmatrix} 0 & 2.245 & 3.355 & 2.895 \\ & 0 & 1.12 & 1.41 \\ & & 0 & 1.5 \\ & & & 0 \end{bmatrix} \end{array}$$

At the second stage , x_3 and x_4 will be merged , and the distances are :

$d(\{x_1,x_2\},\{x_3,x_4\})=0.5(d\{x_1,x_2\},x_3)+d(\{x_1,x_2\},x_4)-0.25(d\{x_3,x_4\})$

$=0.5(2.245+3.355)-0.25(1.12)$

$= 2.52$

$d(\{x_3,x_4\},x_5)=0.5(1.41+1.5)-0.25(1.12)$

$=1.175$

And the dissimilarity matrix becomes

$$A3= \begin{array}{c} \\ C_1 \\ C_2 \\ x_5 \end{array} \begin{array}{c} \begin{array}{ccc} C_1 & C_2 & x_5 \end{array} \\ \begin{bmatrix} 0 & 2.52 & 2.895 \\ & 0 & 1.175 \\ & & 0 \end{bmatrix} \end{array}$$

At the third stage, x_5 will be merged with $\{x_3,x_4\}$ and the distance is :

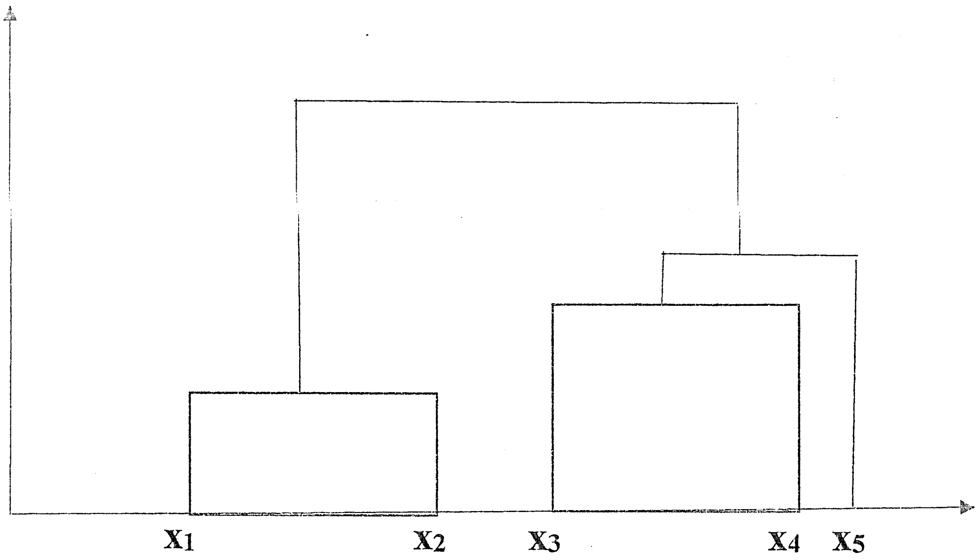
$$d(\{x_1,x_2\},\{x_3,x_4,x_5\})=(2/3)(2.52)+(1/3)(2.895)-(2/9)(1.175)$$

$$=2.384$$

And the similarity matrix becomes

$$A_4=\begin{array}{c c} & \begin{array}{cc} x_1x_2 & x_3x_4x_5 \end{array} \\ \begin{array}{c} x_1x_2 \\ x_3x_4x_5 \end{array} & \begin{bmatrix} 0 & 2.384 \\ & 0 \end{bmatrix} \end{array}$$

The whole process of this clustering can be represented by the dendrogram



K-means Clustering Method

The aim of k-means algorithum is to partition the data into k-clusters so that the within group sum of squence

Input data
$D = \{t_1, t_2, \ldots, t_n\}$ // set of elements
// numbers of designed cluster

Output
k : C // set of cluster

1. Assign each item t_i to the cluster which has the closest mean
2. Calculate new Means for each cluster until convergence criteria is Met.

Example: suppose that we are given the following items to cluster

$S = \{2, 4, 10, 3, 12, 20, 30, 11, 25\}$
and suppose that the number of clusters is $k = 2$.

1- Assign the Means to the first two values i.e, set $\mu_1 = 2$ and $\mu_2 = 4$

2. Using the Euclidean distance we find the elements of each cluster

$C_1 = \{2, 3\}$ $C_2 = \{4, 10, 12, 20, 30, 11, 25\}$

Recalculate the means of each cluster

$$\mu_1 = \frac{2+3}{2} = 2.5$$

$$\mu_2 = \frac{4+10+12+20+30+4+25}{7} = 16$$

4. Again we use the Euclidean distance to find the elements of each cluster

$C_1 = \{2, 3, 4\}$

$C_2 = \{10, 12, 20, 30, 11, 25\}$

Note: stop we you get threshold

$|\mu_1 - \mu_2| = 18$

ملاحظات: خطوات الحل كالآتي:

1. تحديد عدد العناقيد المطلوبة
2. نأخذ اقرب المسافات بين العناصر ونضعها في عنقود
3. تحديد المعدل لكل عنقود
4. نأخذ العناصر بالتسلسل وليس حسب الاكبر والاصغر
5. عند اخذ العنصر نحدد هل هو اقرب الى M_1 ام M_2 فاذا كان اقرب الى M_1 ينقل الى C_1 واذا كان اقرب الى M_2 يبقى في C_2 ولا نغير شيء
6. نتوقف عندما لا ينتقل عنصر من عنقود الى آخر

Iteration	C_1	C_2	M_1	M_2	$r=\lvert M$
0	{2}	{4,10,12,3,20,30,11,25}	2	4	2
1	{2,3}	{4,10,12,20,30,11,25}	2.5	16	13.5
2	{2,3,4}	{10,12,20,30,11,25}	3	18	15
3	{2,3,4,10}	{12,20,30,11,25}	4.75	19.6	14.8
4	{2,3,4,10,12}	{20,30,11,25}	6.2	21.5	15.3
5	{2,3,4,10,12}	{20,30,11,25}	6.2	21.5	15.3
6	{2,3,4,10,12}	{20,30,11,25}	6.2	21.5	15.3
7	{2,3,4,10,12,11}	{20,30,25}	7	25	18
8	{2,3,4,10,12,11}	{20,30,25}	7	25	18

$\therefore C_1 = \{2, 3, 4, 10, 11\}$

$C_2 = \{20, 30, 25\}$

Example 2

Let $x = \{2, 5, 10, 14, 17\}$

$y = \{3, 8, 12, 11, 13\}$

$d_1 = \sqrt{(2-5)^2 + (3-8)^2} = 5.83$

$d_2 = \sqrt{(2-10)^2 + (3-12)^2} = 12.04$

$d_3 = \sqrt{(2-14)^2 + (3-11)^2} = \sqrt{144 + 64} = 14.42$

$d_4 = \sqrt{(2-17)^2 + (3-13)^2} = 18.02$

$D = \{5.83, 12.04, 14.42, 18.02\}$

Iteration: $|\mu_1 - \mu_2| = 7.25$

0	$\{5.83\}$	$\{12.04, 14.42, 18.02\}$	5.83	14.82	6.2
1	$\{5.83, 12.04\}$	$\{14.42, 18.02\}$	8.93	16.22	7.29
2	$\{5.83, 12.04, 14.42\}$	$\{18.02\}$	10.76	18.02	7.25

$C_1 = \{5.83, 14.42, 12.04\}$

$C_2 = \{18.02\}$