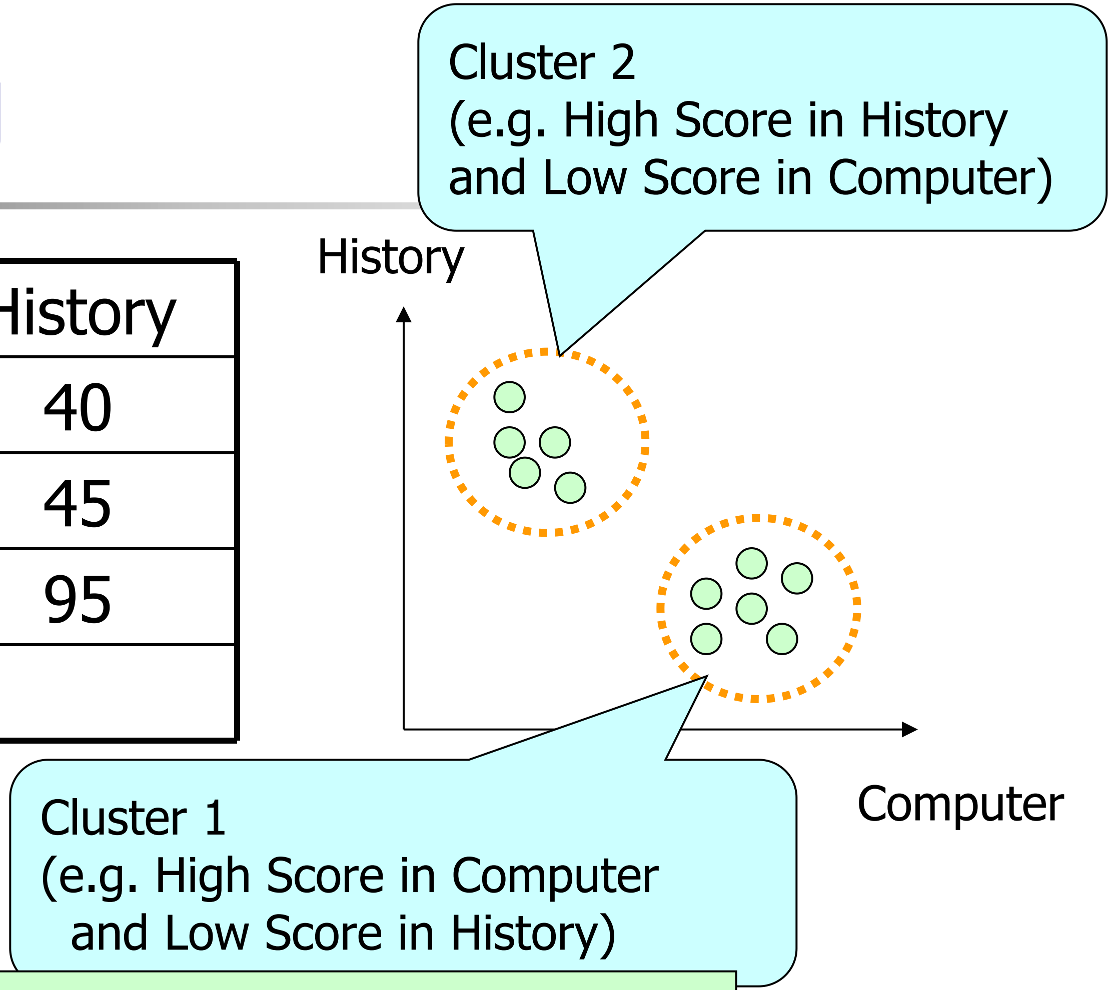
CSIT5210

Clustering

Prepared by Raymond Wong
Some parts of this notes are borrowed from LW Chan's notes
Presented by Raymond Wong
raywong@cse

Clustering

	Computer	History
Raymond	100	40
Louis	90	45
Wyman	20	95
…	…	…



Problem: to find all clusters

Why Clustering?

- Clustering for Utility
  - Summarization
  - Compression

Why Clustering?

- Clustering for Understanding
  - Applications
    - Biology
      - Group different species
    - Psychology and Medicine
      - Group medicine
    - Business
      - Group different customers for marketing
    - Network
      - Group different types of traffic patterns
    - Software
      - Group different programs for data analysis

Clustering Methods

- K-means Clustering
  - Original k-means Clustering
  - Sequential K-means Clustering
  - Forgetful Sequential K-means Clustering
- Hierarchical Clustering Methods
  - Agglomerative methods
  - Divisive methods – polythetic approach and monothetic approach

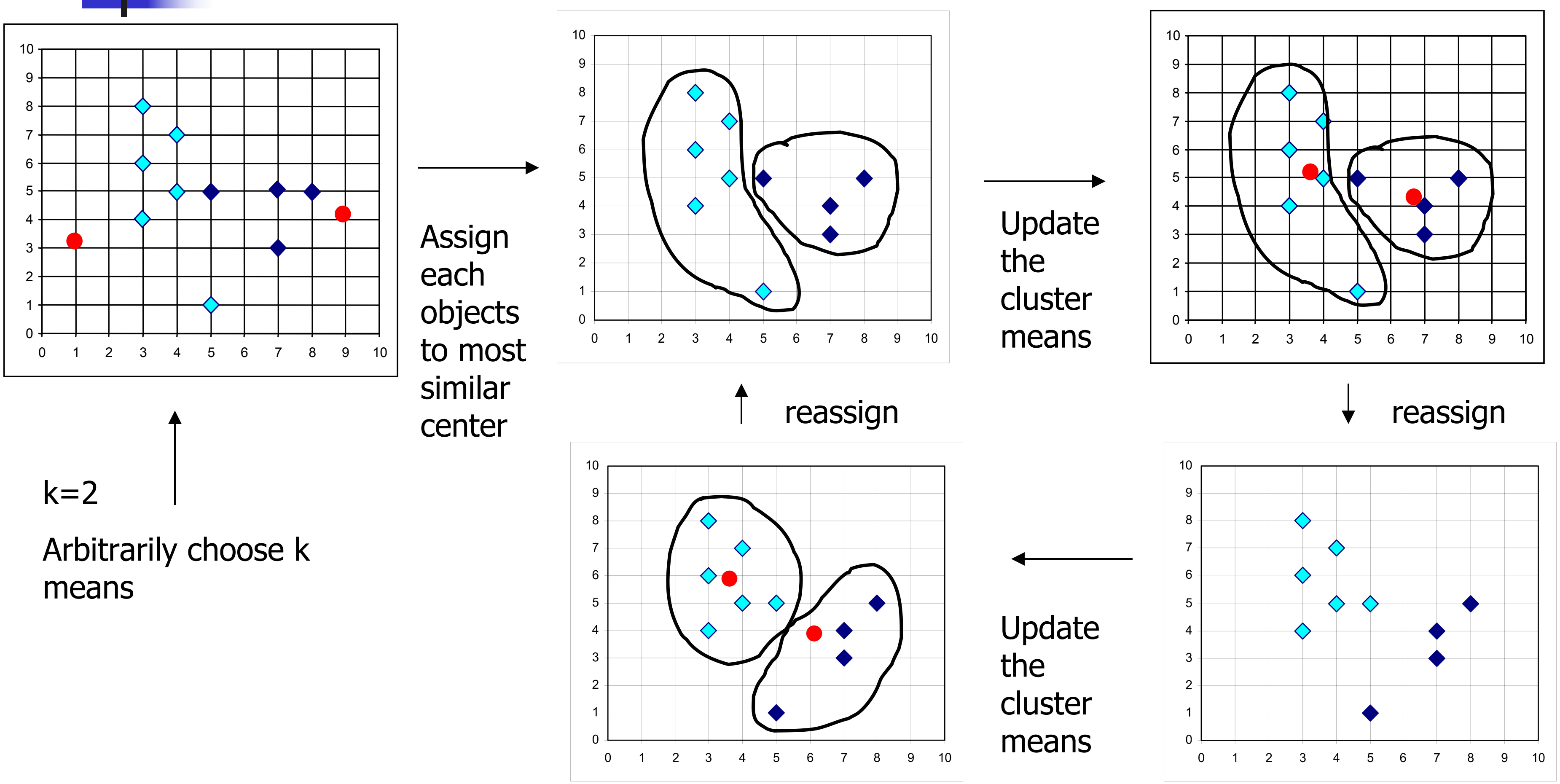
K-mean Clustering

- Suppose that we have n example feature vectors $x_1, x_2, \ldots, x_n$, and we know that they fall into k compact clusters, $k < n$
- Let m_i be the mean of the vectors in cluster i.
- we can say that x is in cluster i if distance from x to m_i is the minimum of all the k distances.

Procedure for finding k-means

- Make initial guesses for the means m_1, m_2, .., m_k
- Until there is no change in any mean
  - Assign each data point to the cluster whose mean is the nearest
  - Calculate the mean of each cluster
  - For i from 1 to k
    - Replace m_i with the mean of all examples for cluster i

Procedure for finding k-means



Initialization of k-means

- The way to initialize the means was not specified. One popular way to start is to randomly choose k of the examples
- The results produced depend on the initial values for the means, and it frequently happens that suboptimal partitions are found. The standard solution is to try a number of different starting points

Disadvantages of k-means

- Disadvantages
  - In a "bad" initial guess, there are no points assigned to the cluster with the initial mean m_i.
  - The value of k is not user-friendly. This is because we do not know the number of clusters before we want to find clusters.

Clustering Methods

- K-means Clustering
  - Original k-means Clustering
  - Sequential K-means Clustering
  - Forgetful Sequential K-means Clustering
- Hierarchical Clustering Methods
  - Agglomerative methods
  - Divisive methods – polythetic approach and monothetic approach

Sequential k-Means Clustering

- Another way to modify the k-means procedure is to update the means one example at a time, rather than all at once.
- This is particularly attractive when we acquire the examples over a period of time, and we want to start clustering before we have seen all of the examples
- Here is a modification of the k-means procedure that operates sequentially

Sequential k-Means Clustering

- Make initial guesses for the means m_1, m_2, …, m_k
- Set the counts n_1, n_2, .., n_k to zero
- Until interrupted
  - Acquire the next example, x
  - If m_i is closest to x
    - Increment n_i
    - Replace m_i by $m_i + (1/n_i) \cdot (x - m_i)$

Clustering Methods

- K-means Clustering
  - Original k-means Clustering
  - Sequential K-means Clustering
  - Forgetful Sequential K-means Clustering
- Hierarchical Clustering Methods
  - Agglomerative methods
  - Divisive methods – polythetic approach and monothetic approach

Forgetful Sequential k-means

- This also suggests another alternative in which we replace the counts by constants. In particular, suppose that a is a constant between 0 and 1, and consider the following variation:

- Make initial guesses for the means $m_1, m_2, \ldots, m_k$
- Until interrupted
  - Acquire the next example x
  - If m_i is closest to x, replace m_i by $m_i+a(x-m_i)$

Forgetful Sequential k-means

- The result is called the “forgetful” sequential k-means procedure.
- It is not hard to show that m_i is a weighted average of the examples that were closest to m_i, where the weight decreases exponentially with the “age” to the example.
- That is, if m_0 is the initial value of the mean vector and if x_j is the j-th example out of n examples that were used to form m_i, then it is not hard to show that

$$m_n = (1-a)^n m_0 + a \sum_{k=1}^{n} (1-a)^{n-k} x_k$$

Forgetful Sequential k-means

- Thus, the initial value m_0 is eventually forgotten, and recent examples receive more weight than ancient examples.
- This variation of k-means is particularly simple to implement, and it is attractive when the nature of the problem changes over time and the cluster centres "drift".

Clustering Methods

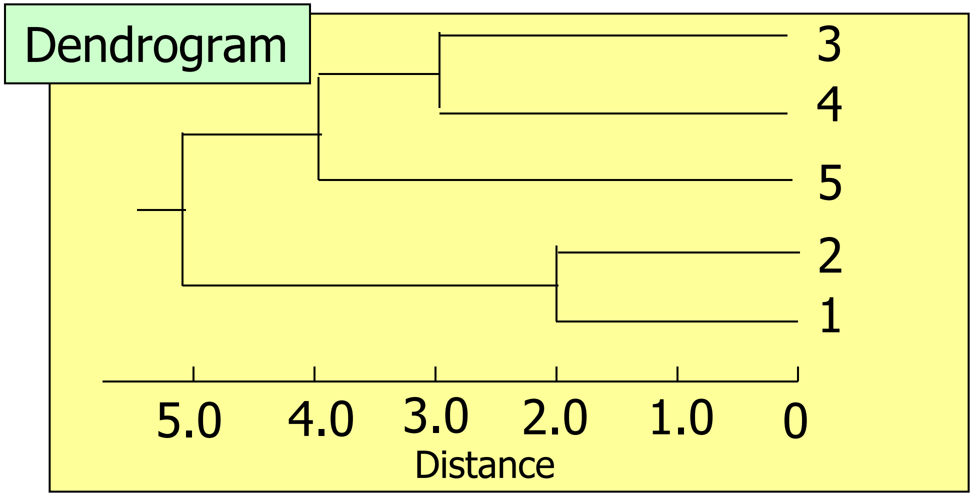
- K-means Clustering
  - Original k-means Clustering
  - Sequential K-means Clustering
  - Forgetful Sequential K-means Clustering
- Hierarchical Clustering Methods
  - Agglomerative methods
  - Divisive methods – polythetic approach and monothetic approach

Hierarchical Clustering Methods

- The partition of data is not done at a single step.
- There are two varieties of hierarchical clustering algorithms
  - Agglomerative – successively fusions of the data into groups
  - Divisive – separate the data successively into finer groups

Dendrogram

- Hierarchic grouping can be represented by two-dimensional diagram known as a **dendrogram**.

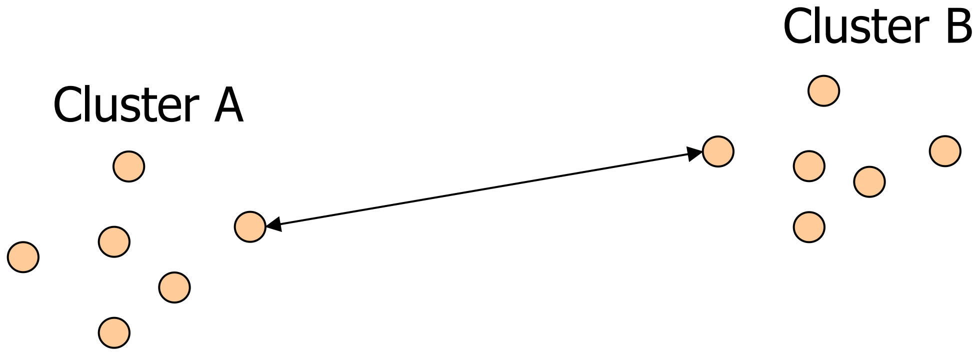


Distance

- Single Linkage
- Complete Linkage
- Group Average Linkage
- Centroid Linkage
- Median Linkage

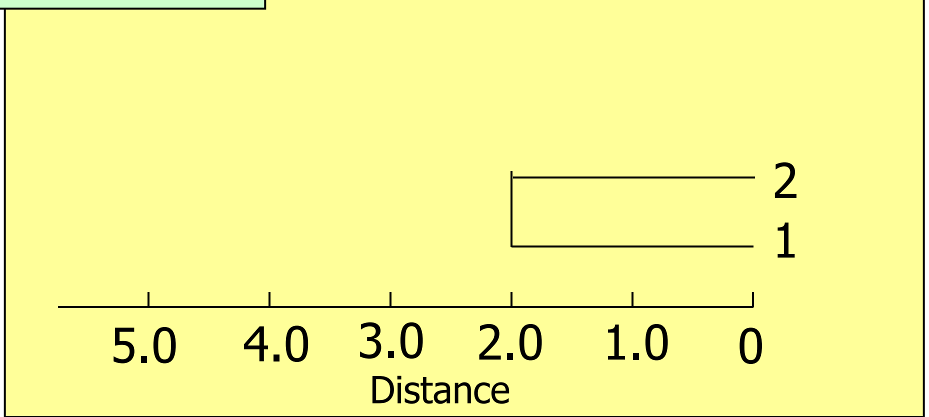
Single Linkage

- Also, known as the **nearest neighbor** technique
- Distance between groups is defined as that of the closest pair of data, where only pairs consisting of one record from each group are considered



$$
\begin{array}{c}
\\ 1 \\ 2 \\ 3 \\ 4 \\ 5
\end{array}
\begin{array}{c}
\begin{array}{ccccc} 1 & 2 & 3 & 4 & 5 \end{array} \\
\begin{pmatrix}
0.0 & & & & \\
2.0 & 0.0 & & & \\
6.0 & 5.0 & 0.0 & & \\
10.0 & 9.0 & 4.0 & 0.0 & \\
9.0 & 8.0 & 5.0 & 3.0 & 0.0
\end{pmatrix}
\end{array}
\qquad
\begin{array}{c}
\\ (12) \\ 3 \\ 4 \\ 5
\end{array}
\begin{array}{c}
\begin{array}{cccc} (12) & 3 & 4 & 5 \end{array} \\
\begin{pmatrix}
0.0 & & & \\
5.0 & 0.0 & & \\
9.0 & 4.0 & 0.0 & \\
8.0 & 5.0 & 3.0 & 0.0
\end{pmatrix}
\end{array}
$$

Dendrogram

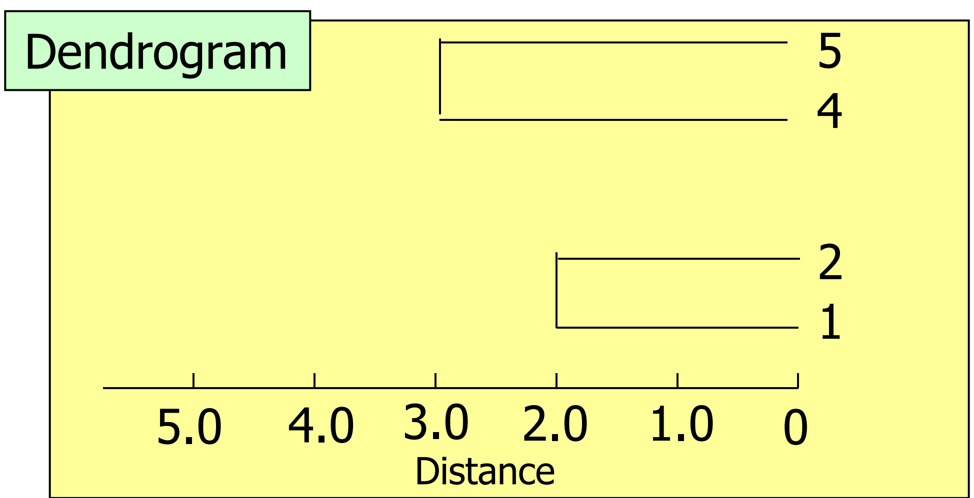


$$
\begin{array}{cc}
 & \begin{array}{cccc} (12) & 3 & 4 & 5 \end{array} \\
\begin{array}{c} (12) \\ 3 \\ 4 \\ 5 \end{array} &
\begin{pmatrix}
0.0 & & & \\
5.0 & 0.0 & & \\
9.0 & 4.0 & 0.0 & \\
8.0 & 5.0 & 3.0 & 0.0
\end{pmatrix}
\end{array}
$$

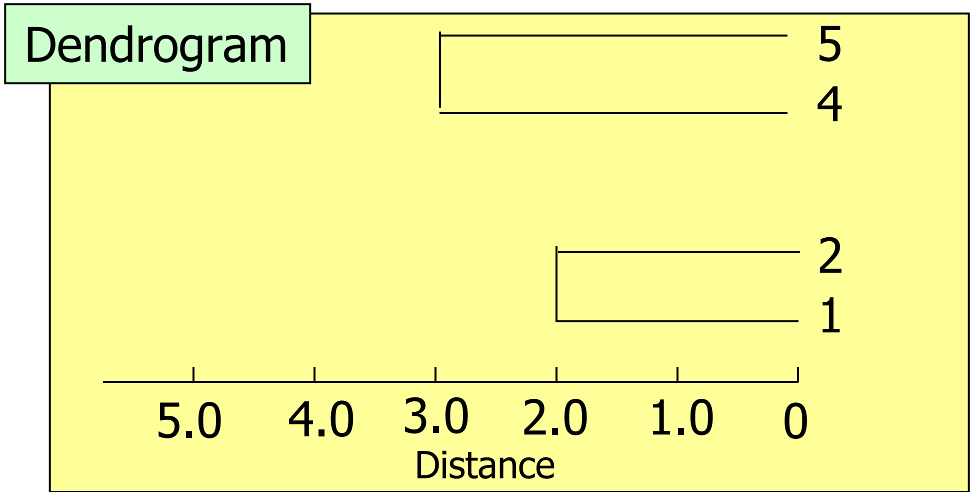
Dendrogram

$$
\begin{array}{c|cccc}
 & (12) & 3 & 4 & 5 \\
\hline
(12) & 0.0 & & & \\
3 & 5.0 & 0.0 & & \\
4 & 9.0 & 4.0 & 0.0 & \\
5 & 8.0 & 5.0 & 3.0 & 0.0
\end{array}
$$

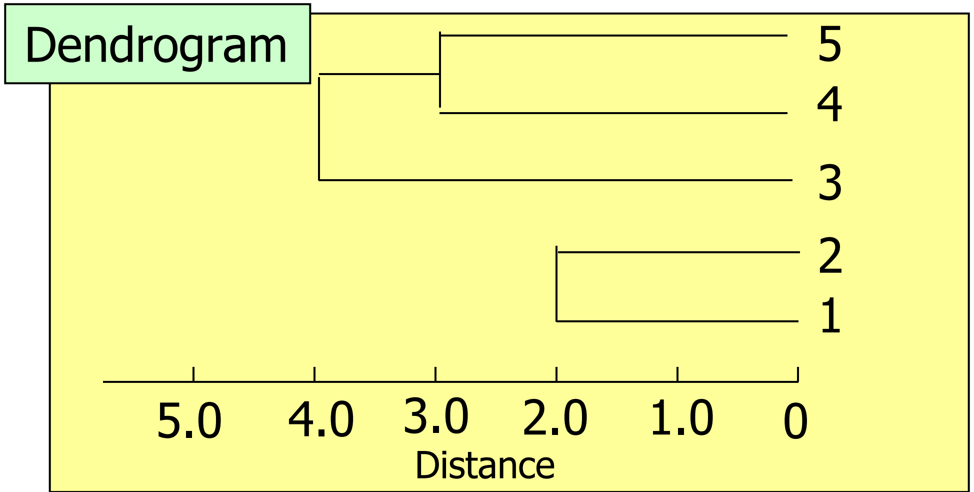
$$
\begin{array}{c|ccc}
 & (12) & 3 & (4\ 5) \\
\hline
(12) & 0.0 & & \\
3 & 5.0 & 0.0 & \\
(4\ 5) & 8.0 & 4.0 & 0.0
\end{array}
$$



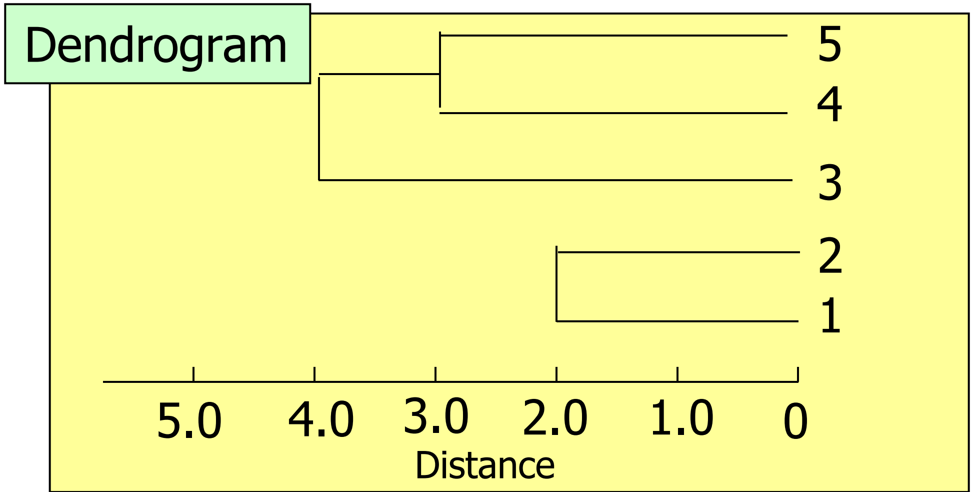
$$\begin{array}{cc} & \begin{array}{ccc} (12) & 3 & (4\ 5) \end{array} \\ \begin{array}{c} (12) \\ 3 \\ (4\ 5) \end{array} & \begin{pmatrix} 0.0 & & \\ 5.0 & 0.0 & \\ 8.0 & 4.0 & 0.0 \end{pmatrix} \end{array}$$



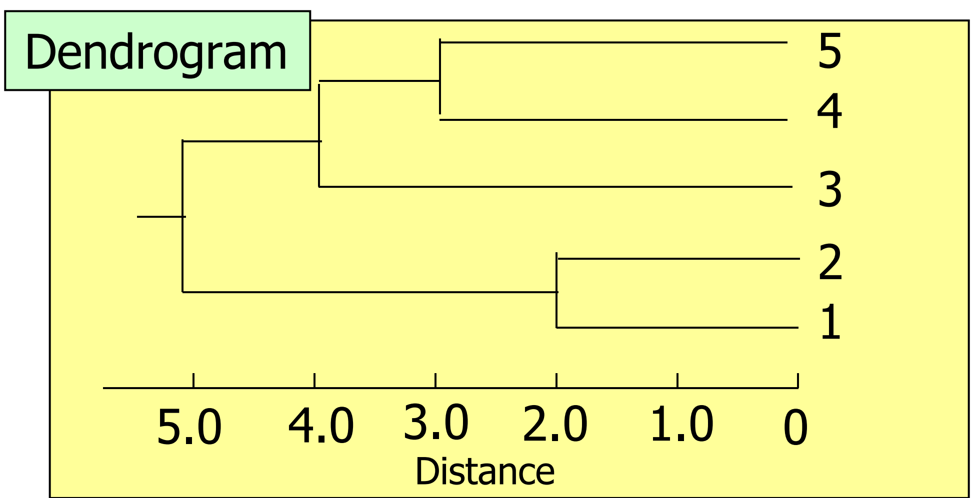
$$
\begin{array}{c|c}
 & \begin{array}{ccc} (12) & 3 & (4\ 5) \end{array} \\
\begin{array}{c} (12) \\ 3 \\ (4\ 5) \end{array} & \begin{pmatrix} 0.0 & & \\ 5.0 & 0.0 & \\ 8.0 & 4.0 & 0.0 \end{pmatrix}
\end{array}
\qquad
\begin{array}{c|c}
 & \begin{array}{cc} (12) & (3\ 4\ 5) \end{array} \\
\begin{array}{c} (12) \\ (3\ 4\ 5) \end{array} & \begin{pmatrix} 0.0 & \\ 5.0 & 0.0 \end{pmatrix}
\end{array}
$$



$$\begin{array}{cc} & \begin{array}{cc}(12) & (3\ 4\ 5)\end{array} \\ \begin{array}{c}(12) \\ (3\ 4\ 5)\end{array} & \begin{pmatrix} 0.0 & \\ 5.0 & 0.0 \end{pmatrix} \end{array}$$



$$
\begin{array}{ccc}
 & (12) & (3\ 4\ 5) \\
\begin{array}{c}(12)\\(3\ 4\ 5)\end{array} & \multicolumn{2}{c}{\begin{pmatrix} 0.0 & \\ 5.0 & 0.0 \end{pmatrix}}
\end{array}
$$

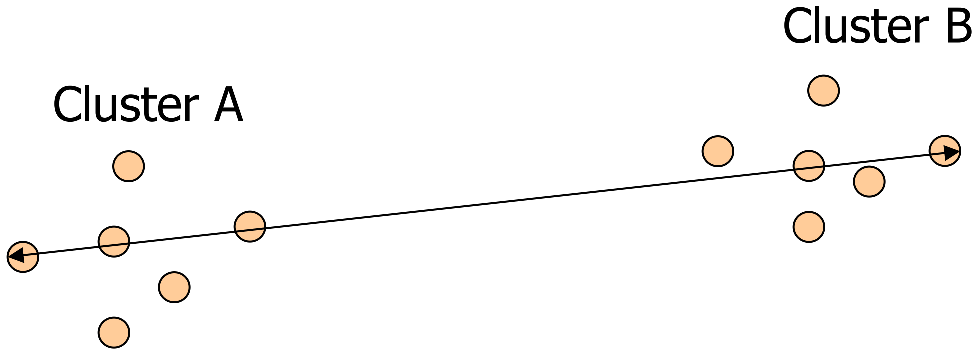


Distance

- Single Linkage
- Complete Linkage
- Group Average Linkage
- Centroid Linkage
- Median Linkage

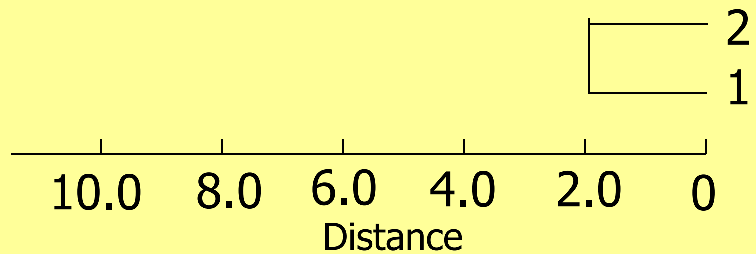
Complete Linkage

- The distance between two clusters is given by the distance between their most distant members

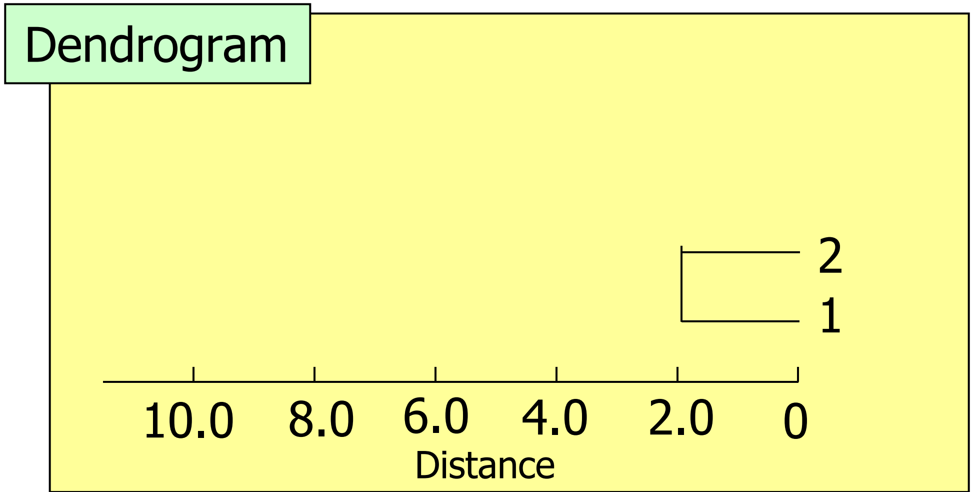


$$
\begin{array}{c|ccccc}
 & 1 & 2 & 3 & 4 & 5 \\
\hline
1 & 0.0 & & & & \\
2 & 2.0 & 0.0 & & & \\
3 & 6.0 & 5.0 & 0.0 & & \\
4 & 10.0 & 9.0 & 4.0 & 0.0 & \\
5 & 9.0 & 8.0 & 5.0 & 3.0 & 0.0
\end{array}
\qquad
\begin{array}{c|cccc}
 & (12) & 3 & 4 & 5 \\
\hline
(12) & 0.0 & & & \\
3 & 6.0 & 0.0 & & \\
4 & 10.0 & 4.0 & 0.0 & \\
5 & 9.0 & 5.0 & 3.0 & 0.0
\end{array}
$$

Dendrogram

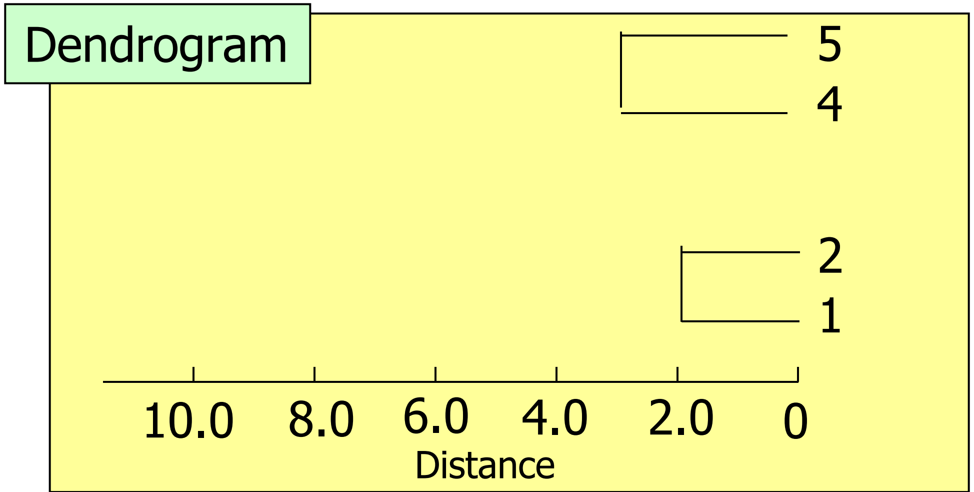


$$\begin{array}{c} \\ (12) \\ 3 \\ 4 \\ 5 \end{array} \begin{array}{c} \begin{array}{cccc} (12) & 3 & 4 & 5 \end{array} \\ \begin{pmatrix} 0.0 & & & \\ 6.0 & 0.0 & & \\ 10.0 & 4.0 & 0.0 & \\ 9.0 & 5.0 & 3.0 & 0.0 \end{pmatrix} \end{array}$$

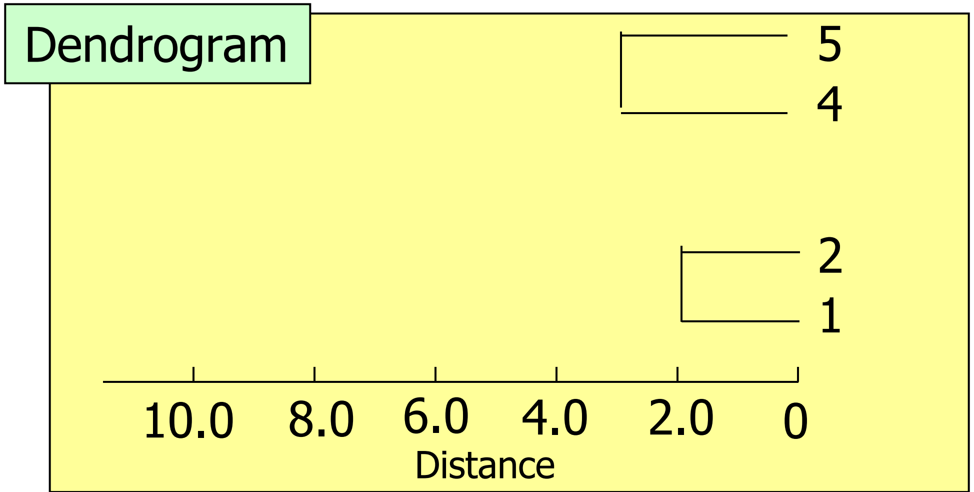


$$
\begin{array}{c} \\ (12) \\ 3 \\ 4 \\ 5 \end{array}
\begin{array}{c} \begin{array}{cccc} (12) & 3 & 4 & 5 \end{array} \\ \begin{pmatrix} 0.0 & & & \\ 6.0 & 0.0 & & \\ 10.0 & 4.0 & 0.0 & \\ 9.0 & 5.0 & 3.0 & 0.0 \end{pmatrix} \end{array}
$$

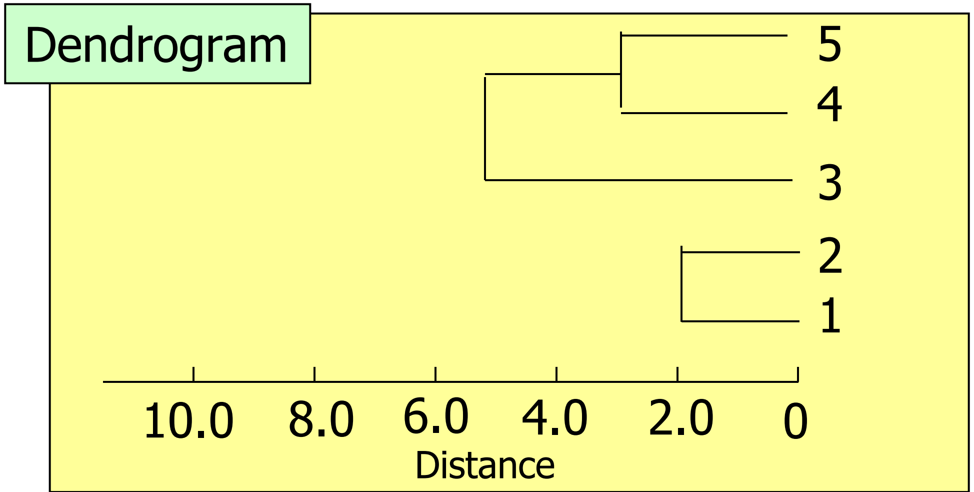
$$
\begin{array}{c} \\ (12) \\ 3 \\ (4\ 5) \end{array}
\begin{array}{c} \begin{array}{ccc} (12) & 3 & (4\ 5) \end{array} \\ \begin{pmatrix} 0.0 & & \\ 6.0 & 0.0 & \\ 10.0 & 5.0 & 0.0 \end{pmatrix} \end{array}
$$



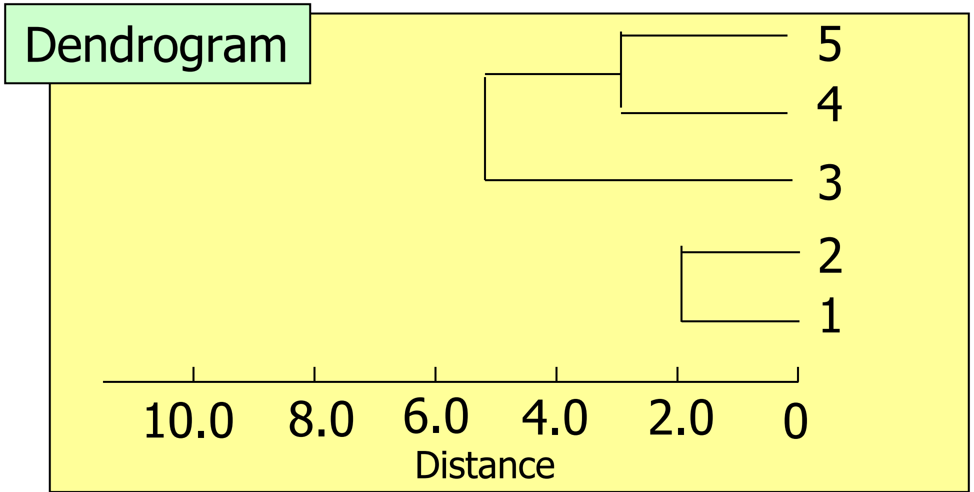
$$
\begin{array}{cc}
 & \begin{array}{ccc} (12) & 3 & (4\ 5) \end{array} \\
\begin{array}{c} (12) \\ 3 \\ (4\ 5) \end{array} &
\begin{pmatrix} 0.0 & & \\ 6.0 & 0.0 & \\ 10.0 & 5.0 & 0.0 \end{pmatrix}
\end{array}
$$



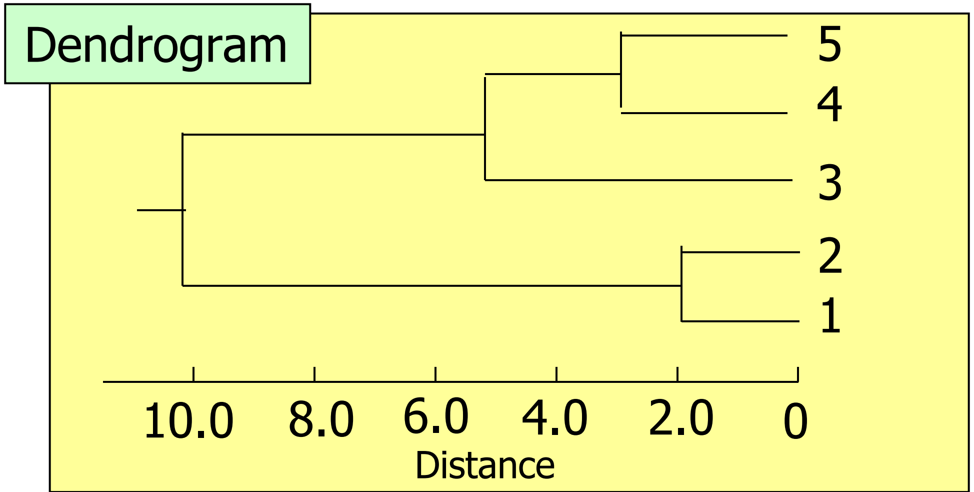
$$
\begin{array}{c c}
 & \begin{array}{ccc} (12) & 3 & (4\ 5) \end{array} \\
\begin{array}{c} (12) \\ 3 \\ (4\ 5) \end{array} & \begin{pmatrix} 0.0 & & \\ 6.0 & 0.0 & \\ 10.0 & 5.0 & 0.0 \end{pmatrix}
\end{array}
\qquad
\begin{array}{c c}
 & \begin{array}{cc} (12) & (3\ 4\ 5) \end{array} \\
\begin{array}{c} (12) \\ (3\ 4\ 5) \end{array} & \begin{pmatrix} 0.0 & \\ 10.0 & 0.0 \end{pmatrix}
\end{array}
$$



$$
\begin{array}{ccc}
 & (12) & (3\ 4\ 5) \\
\begin{array}{c} (12) \\ (3\ 4\ 5) \end{array} & \multicolumn{2}{c}{\begin{pmatrix} 0.0 & \\ 10.0 & 0.0 \end{pmatrix}}
\end{array}
$$



$$\begin{array}{c c} & \begin{array}{cc} (12) & (3\ 4\ 5) \end{array} \\ \begin{array}{c} (12) \\ (3\ 4\ 5) \end{array} & \begin{pmatrix} 0.0 & \\ 10.0 & 0.0 \end{pmatrix} \end{array}$$

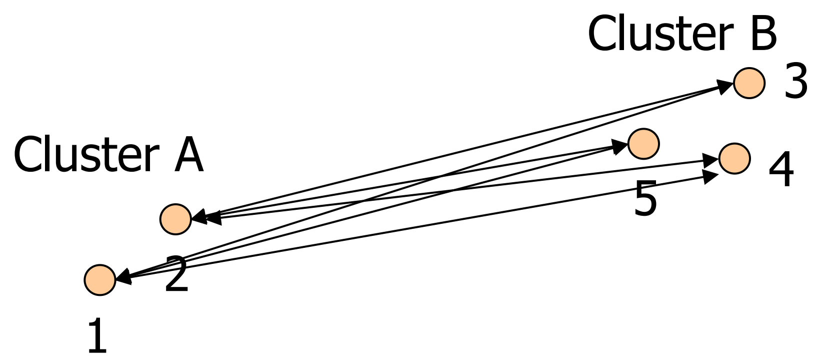


Distance

- Single Linkage
- Complete Linkage
- Group Average Linkage
- Centroid Linkage
- Median Linkage

Group Average Clustering

- The distance between two clusters is defined as the average of the distances between all pairs of records (one from each cluster).
- $d_{AB} = 1/6\ (d_{13} + d_{14} + d_{15} + d_{23} + d_{24} + d_{25})$

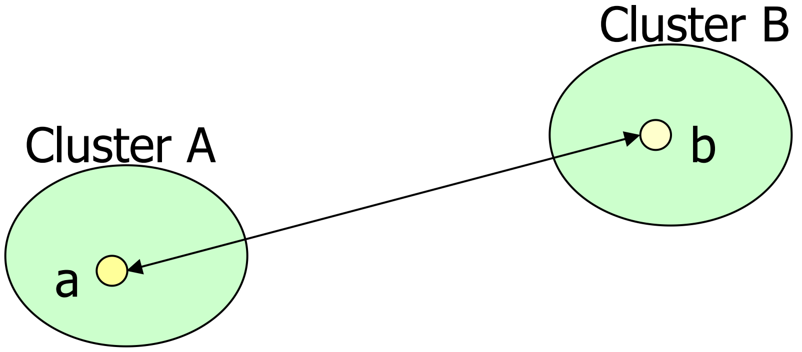


Distance

- Single Linkage
- Complete Linkage
- Group Average Linkage
- Centroid Linkage
- Median Linkage

Centroid Clustering

- The distance between two clusters is defined as the distance between the mean vectors of the two clusters.
- $d_{AB} = d_{ab}$
- where a is the mean vector of the cluster A and b is the mean vector of the cluster B.

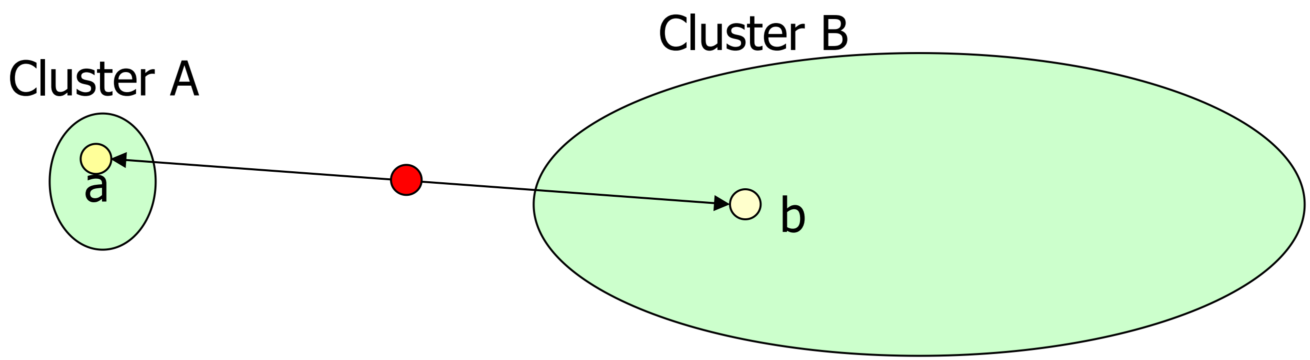


Distance

- Single Linkage
- Complete Linkage
- Group Average Linkage
- Centroid Linkage
- Median Linkage

Median Clustering

- Disadvantage of the Centroid Clustering: When a large cluster is merged with a small one, the centroid of the combined cluster would be closed to the large one, ie. The characteristic properties of the small one are lost
- After we have combined two groups, the mid-point of the original two cluster centres is used as the centre of the newly combined group



Clustering Methods

- K-means Clustering
  - Original k-means Clustering
  - Sequential K-means Clustering
  - Forgetful Sequential K-means Clustering
- Hierarchical Clustering Methods
  - Agglomerative methods
  - Divisive methods – polythetic approach and monothetic approach

Divisive Methods

- In a divisive algorithm, we start with the assumption that all the data is part of one cluster.
- We then use a distance criterion to divide the cluster in two, and then subdivide the clusters until a stopping criterion is achieved.
  - Polythetic – divide the data based on the values by all attributes
  - Monothetic – divide the data on the basis of the possession of a single specified attribute

Polythetic Approach

- Distance
  - Single Linkage
  - Complete Linkage
  - Group Average Linkage
  - Centroid Linkage
  - Median Linkage

Polythetic Approach

$$
\begin{array}{c|ccccccc}
 & 1 & 2 & 3 & 4 & 5 & 6 & 7 \\
\hline
1 & \multicolumn{7}{l}{} \\
\end{array}
$$

$$
\begin{array}{c}
 \\ 1 \\ 2 \\ 3 \\ 4 \\ 5 \\ 6 \\ 7
\end{array}
\begin{array}{c}
\begin{array}{ccccccc} 1 & 2 & 3 & 4 & 5 & 6 & 7 \end{array} \\
\begin{pmatrix}
0 & & & & & & \\
10 & 0 & & & & & \\
7 & 7 & 0 & & & & \\
30 & 23 & 21 & 0 & & & \\
29 & 25 & 22 & 7 & 0 & & \\
38 & 34 & 31 & 10 & 11 & 0 & \\
42 & 36 & 36 & 13 & 17 & 9 & 0
\end{pmatrix}
\end{array}
$$

D(1, *) = 26.0

D(2, *) = 22.5

D(3, *) = 20.7

D(4, *) = 17.3

D(5, *) = 18.5

D(6, *) = 22.2

D(7, *) = 25.5

A = {1 }

B = {2, 3, 4, 5, 6, 7}

Polythetic Approach

$$
\begin{array}{c|ccccccc}
 & 1 & 2 & 3 & 4 & 5 & 6 & 7 \\
\hline
1 & \multicolumn{7}{l}{\left(\begin{array}{ccccccc}
0 & & & & & & \\
10 & 0 & & & & & \\
7 & 7 & 0 & & & & \\
30 & 23 & 21 & 0 & & & \\
29 & 25 & 22 & 7 & 0 & & \\
38 & 34 & 31 & 10 & 11 & 0 & \\
42 & 36 & 36 & 13 & 17 & 9 & 0
\end{array}\right)}
\end{array}
$$

	1	2	3	4	5	6	7
1	0						
2	10	0					
3	7	7	0				
4	30	23	21	0			
5	29	25	22	7	0		
6	38	34	31	10	11	0	
7	42	36	36	13	17	9	0

D(2, A) = 10

D(3, A) = 7

D(4, A) = 30

D(5, A) = 29

D(6, A) = 38

D(7, A) = 42

A = {1 }

B = {2, 3, 4, 5, 6, 7}

Polythetic Approach

$$
\begin{array}{c|c}
 & \begin{array}{ccccccc} 1 & 2 & 3 & 4 & 5 & 6 & 7 \end{array} \\
\begin{array}{c} 1 \\ 2 \\ 3 \\ 4 \\ 5 \\ 6 \\ 7 \end{array} &
\begin{pmatrix}
0 & & & & & & \\
10 & 0 & & & & & \\
7 & 7 & 0 & & & & \\
30 & 23 & 21 & 0 & & & \\
29 & 25 & 22 & 7 & 0 & & \\
38 & 34 & 31 & 10 & 11 & 0 & \\
42 & 36 & 36 & 13 & 17 & 9 & 0
\end{pmatrix}
\end{array}
$$

A = {1 }

B = {2, 3, 4, 5, 6, 7}

D(2, A) = 10	D(2, B) = 25.0
D(3, A) = 7	D(3, B) = 23.4
D(4, A) = 30	D(4, B) = 14.8
D(5, A) = 29	D(5, B) = 16.4
D(6, A) = 38	D(6, B) = 19.0
D(7, A) = 42	D(7, B) = 22.2

Polythetic Approach

$$
\begin{array}{cc}
 & \begin{array}{ccccccc} 1 & 2 & 3 & 4 & 5 & 6 & 7 \end{array} \\
\begin{array}{c} 1 \\ 2 \\ 3 \\ 4 \\ 5 \\ 6 \\ 7 \end{array} &
\begin{pmatrix}
0 & & & & & & \\
10 & 0 & & & & & \\
7 & 7 & 0 & & & & \\
30 & 23 & 21 & 0 & & & \\
29 & 25 & 22 & 7 & 0 & & \\
38 & 34 & 31 & 10 & 11 & 0 & \\
42 & 36 & 36 & 13 & 17 & 9 & 0
\end{pmatrix}
\end{array}
$$

D(2, A) = 10	D(2, B) = 25.0	Δ_2 = 15.0
D(3, A) = 7	D(3, B) = 23.4	Δ_3 = 16.4
D(4, A) = 30	D(4, B) = 14.8	Δ_4 = -15.2
D(5, A) = 29	D(5, B) = 16.4	Δ_5 = -12.6
D(6, A) = 38	D(6, B) = 19.0	Δ_6 = -19.0
D(7, A) = 42	D(7, B) = 22.2	Δ_7 = -19.8

A = {1, 3 }

B = {2, ~~3,~~ 4, 5, 6, 7}

Polythetic Approach

$$\begin{array}{cc} & \begin{array}{ccccccc} 1 & 2 & 3 & 4 & 5 & 6 & 7 \end{array} \\ \begin{array}{c} 1 \\ 2 \\ 3 \\ 4 \\ 5 \\ 6 \\ 7 \end{array} & \begin{pmatrix} 0 & & & & & & \\ 10 & 0 & & & & & \\ 7 & 7 & 0 & & & & \\ 30 & 23 & 21 & 0 & & & \\ 29 & 25 & 22 & 7 & 0 & & \\ 38 & 34 & 31 & 10 & 11 & 0 & \\ 42 & 36 & 36 & 13 & 17 & 9 & 0 \end{pmatrix} \end{array}$$

D(2, A) = 10	D(2, B) = 25.0	Δ_2 = 15.0
D(3, A) = 7	D(3, B) = 23.4	Δ_3 = 16.4
D(4, A) = 30	D(4, B) = 14.8	Δ_4 = -15.2
D(5, A) = 29	D(5, B) = 16.4	Δ_5 = -12.6
D(6, A) = 38	D(6, B) = 19.0	Δ_6 = -19.0
D(7, A) = 42	D(7, B) = 22.2	Δ_7 = -19.8

A = {1, 3 }

B = {2, 4, 5, 6, 7}

Polythetic Approach

$$
\begin{array}{c|ccccccc}
 & 1 & 2 & 3 & 4 & 5 & 6 & 7 \\
\hline
1 & 0 & & & & & & \\
2 & 10 & 0 & & & & & \\
3 & 7 & 7 & 0 & & & & \\
4 & 30 & 23 & 21 & 0 & & & \\
5 & 29 & 25 & 22 & 7 & 0 & & \\
6 & 38 & 34 & 31 & 10 & 11 & 0 & \\
7 & 42 & 36 & 36 & 13 & 17 & 9 & 0
\end{array}
$$

D(2, A) = 8.5

D(4, A) = 25.5

D(5, A) = 25.5

D(6, A) = 34.5

D(7, A) = 39.0

A = {1, 3 }

B = {2, 4, 5, 6, 7}

Polythetic Approach

$$
\begin{array}{c|ccccccc}
 & 1 & 2 & 3 & 4 & 5 & 6 & 7 \\
\hline
1 & 0 & & & & & & \\
2 & 10 & 0 & & & & & \\
3 & 7 & 7 & 0 & & & & \\
4 & 30 & 23 & 21 & 0 & & & \\
5 & 29 & 25 & 22 & 7 & 0 & & \\
6 & 38 & 34 & 31 & 10 & 11 & 0 & \\
7 & 42 & 36 & 36 & 13 & 17 & 9 & 0
\end{array}
$$

D(2, A) = 8.5

D(2, B) = 29.5

D(4, A) = 25.5

D(4, B) = 13.2

D(5, A) = 25.5

D(5, B) = 15.0

D(6, A) = 34.5

D(6, B) = 16.0

D(7, A) = 39.0

D(7, B) = 18.75

A = {1, 3 }

B = {2, 4, 5, 6, 7}

Polythetic Approach

$$
\begin{array}{c c}
 & \begin{array}{ccccccc} 1 & 2 & 3 & 4 & 5 & 6 & 7 \end{array} \\
\begin{array}{c} 1 \\ 2 \\ 3 \\ 4 \\ 5 \\ 6 \\ 7 \end{array} &
\begin{pmatrix}
0 & & & & & & \\
10 & 0 & & & & & \\
7 & 7 & 0 & & & & \\
30 & 23 & 21 & 0 & & & \\
29 & 25 & 22 & 7 & 0 & & \\
38 & 34 & 31 & 10 & 11 & 0 & \\
42 & 36 & 36 & 13 & 17 & 9 & 0
\end{pmatrix}
\end{array}
$$

D(2, A) = 8.5	D(2, B) = 29.5	Δ_2 = 21.0
D(4, A) = 25.5	D(4, B) = 13.2	Δ_4 = -12.3
D(5, A) = 25.5	D(5, B) = 15.0	Δ_5 = -10.5
D(6, A) = 34.5	D(6, B) = 16.0	Δ_6 = -18.5
D(7, A) = 39.0	D(7, B) = 18.75	Δ_7 = -20.25

A = {1, 3, 2 }

B = {~~2,~~ 4, 5, 6, 7}

Polythetic Approach

$$
\begin{array}{c|ccccccc}
 & 1 & 2 & 3 & 4 & 5 & 6 & 7 \\
\hline
1 & 0 & & & & & & \\
2 & 10 & 0 & & & & & \\
3 & 7 & 7 & 0 & & & & \\
4 & 30 & 23 & 21 & 0 & & & \\
5 & 29 & 25 & 22 & 7 & 0 & & \\
6 & 38 & 34 & 31 & 10 & 11 & 0 & \\
7 & 42 & 36 & 36 & 13 & 17 & 9 & 0
\end{array}
$$

D(2, A) = 8.5	D(2, B) = 29.5	Δ_2 = 21.0
D(4, A) = 25.5	D(4, B) = 13.2	Δ_4 = -12.3
D(5, A) = 25.5	D(5, B) = 15.0	Δ_5 = -10.5
D(6, A) = 34.5	D(6, B) = 16.0	Δ_6 = -18.5
D(7, A) = 39.0	D(7, B) = 18.75	Δ_7 = -20.25

A = {1, 3 , 2 }

B = { 4, 5, 6, 7}

Polythetic Approach

$$
\begin{array}{c|ccccccc}
 & 1 & 2 & 3 & 4 & 5 & 6 & 7 \\
\hline
1 & 0 & & & & & & \\
2 & 10 & 0 & & & & & \\
3 & 7 & 7 & 0 & & & & \\
4 & 30 & 23 & 21 & 0 & & & \\
5 & 29 & 25 & 22 & 7 & 0 & & \\
6 & 38 & 34 & 31 & 10 & 11 & 0 & \\
7 & 42 & 36 & 36 & 13 & 17 & 9 & 0
\end{array}
$$

D(4, A) = 24.7

D(5, A) = 25.3

D(6, A) = 34.3

D(7, A) = 38.0

A = {1, 3, 2 }

B = {4, 5, 6, 7}

Polythetic Approach

$$\begin{array}{c|ccccccc}
 & 1 & 2 & 3 & 4 & 5 & 6 & 7 \\
\hline
1 & 0 & & & & & & \\
2 & 10 & 0 & & & & & \\
3 & 7 & 7 & 0 & & & & \\
4 & 30 & 23 & 21 & 0 & & & \\
5 & 29 & 25 & 22 & 7 & 0 & & \\
6 & 38 & 34 & 31 & 10 & 11 & 0 & \\
7 & 42 & 36 & 36 & 13 & 17 & 9 & 0
\end{array}$$

D(4, A) = 24.7

D(4, B) = 10.0

D(5, A) = 25.3

D(5, B) = 11.7

D(6, A) = 34.3

D(6, B) = 10.0

D(7, A) = 38.0

D(7, B) = 13.0

A = {1, 3, 2 }

B = {4, 5, 6, 7}

Polythetic Approach

$$
\begin{array}{c|ccccccc}
 & 1 & 2 & 3 & 4 & 5 & 6 & 7 \\
\hline
1 & \multicolumn{7}{l}{\left(\begin{array}{ccccccc} 0 & & & & & & \\ 10 & 0 & & & & & \\ 7 & 7 & 0 & & & & \\ 30 & 23 & 21 & 0 & & & \\ 29 & 25 & 22 & 7 & 0 & & \\ 38 & 34 & 31 & 10 & 11 & 0 & \\ 42 & 36 & 36 & 13 & 17 & 9 & 0 \end{array}\right)}
\end{array}
$$

	1	2	3	4	5	6	7
1	0						
2	10	0					
3	7	7	0				
4	30	23	21	0			
5	29	25	22	7	0		
6	38	34	31	10	11	0	
7	42	36	36	13	17	9	0

D(4, A) = 24.7 D(4, B) = 10.0 Δ_4 = -14.7

D(5, A) = 25.3 D(5, B) = 11.7 Δ_5 = -13.6

D(6, A) = 34.3 D(6, B) = 10.0 Δ_6 = -24.3

D(7, A) = 38.0 D(7, B) = 13.0 Δ_7 = -25.0

All differences are negative. The process would continue on each subgroup separately.

A = {1, 3, 2 }

B = {4, 5, 6, 7}

Clustering Methods

- K-means Clustering
  - Original k-means Clustering
  - Sequential K-means Clustering
  - Forgetful Sequential K-means Clustering
- Hierarchical Clustering Methods
  - Agglomerative methods
  - Divisive methods – polythetic approach and monothetic approach

Monothetic

It is usually used when the data consists of **binary** variables.

	A	B	C
1	0	1	1
2	1	1	0
3	1	1	1
4	1	1	0
5	0	0	1

Monothetic

It is usually used when the data consists of **binary** variables.

	A	B	C
1	0	1	1
2	1	1	0
3	1	1	1
4	1	1	0
5	0	0	1

B \ A	1	0
1	a=3	b=1
0	c=0	d=1

Chi-Square Measure

$$\chi_{AB}^{2} = \frac{(ad-bc)^2 N}{(a+b)(a+c)(b+d)(c+d)}$$

$$= \frac{(3-0)^2 \cdot 5}{4 \cdot 3 \cdot 2 \cdot 1}$$

= 1.875

B \ A	1	0
1	a=3	b=1
0	c=0	d=1

...etic

Chi-Square Measure

It is usually used when the data consists of **binary** variables.

	A	B	C
1	0	1	1
2	1	1	0
3	1	1	1
4	1	1	0
5	0	0	1

$$\chi_{AB}^{2} = \frac{(ad-bc)^2 N}{(a+b)(a+c)(b+d)(c+d)}$$

$$= \frac{(3-0)^2 \cdot 5}{4 \cdot 3 \cdot 2 \cdot 1}$$

$= 1.875$

B \ A	1	0
1	a=3	b=1
0	c=0	d=1

etic

Chi-Square Measure

It is usually used when the data consists of **binary** variables.

Attr.	AB		
a	3		
b	1		
c	0		
d	1		
N	5		
χ^2	1.87		

	A	B	C
1	0	1	1
2	1	1	0
3	1	1	1
4	1	1	0
5	0	0	1

$$\chi_{AB}^{2} = \frac{(ad-bc)^2 N}{(a+b)(a+c)(b+d)(c+d)}$$

$$= \frac{(3-0)^2 \cdot 5}{4 \cdot 3 \cdot 2 \cdot 1}$$

= 1.875

B \ A	1	0
1	a=3	b=1
0	c=0	d=1

etic

Chi-Square Measure

It is usually used when the data consists of **binary** variables.

	A	B	C
1	0	1	1
2	1	1	0
3	1	1	1
4	1	1	0
5	0	0	1

Attr.	AB	AC	BC
a	3	1	2
b	1	2	1
c	0	2	2
d	1	0	0
N	5	5	5
χ^2	1.87	2.22	0.83

For attribute A,

$$\chi_{AB}^2 + \chi_{AC}^2 = 4.09$$

For attribute B,

$$\chi_{AB}^2 + \chi_{BC}^2 = 2.70$$

For attribute C,

$$\chi_{AC}^2 + \chi_{BC}^2 = 3.05$$

B \ A	1	0
1	a=3	b=1
0	c=0	d=1

etic

Chi-Square Measure

It is usually used when the data consists of **binary** variables.

Attr.	AB	AC	BC
a	3	1	2
b	1	2	1
[illegible]	[illegible]	[illegible]	[illegible]
[illegible]	[illegible]	[illegible]	[illegible]
χ^2	[illegible]7	2.22	0.83

We choose attribute A for dividing the data into two groups. {2, 3, 4}, and {1, 5}

	A	B	C
1	0	1	1
2	1	1	0
3	1	1	1
4	1	1	0
5	0	0	1

For attribute A,

$$\chi_{AB}^2 + \chi_{AC}^2 = 4.09$$

For attribute B,

$$\chi_{AB}^2 + \chi_{BC}^2 = 2.70$$

For attribute C,

$$\chi_{AC}^2 + \chi_{BC}^2 = 3.05$$