Clustering

Data Mining

Prof. Sujee Lee

Department of Systems Management Engineering

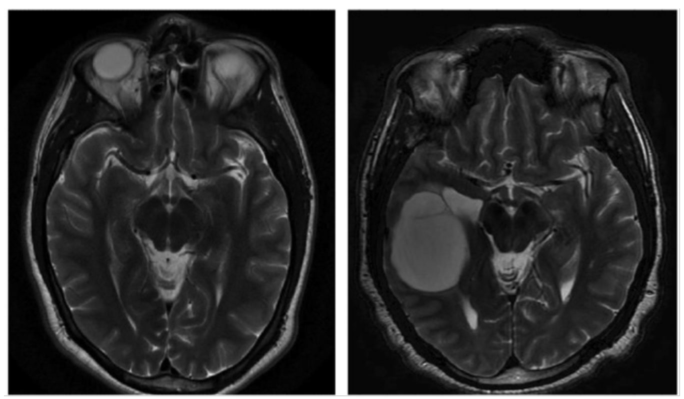
Sungkyunkwan University

Clustering

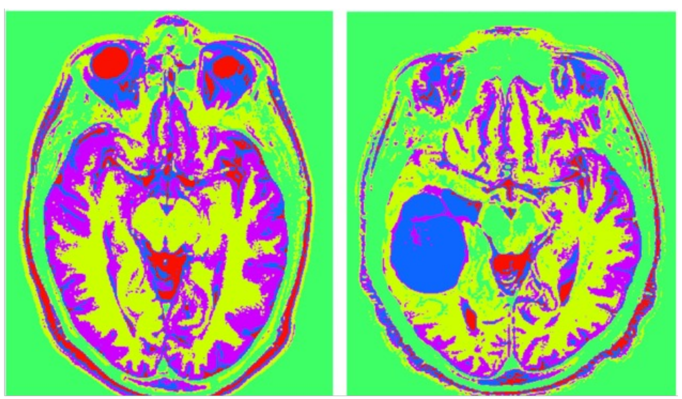
- **What Is Clustering?**
  - **Clustering** is a **process of partitioning** a set of data (or objects) into a set of meaningful sub-classes, called **clusters**
    - Help users understand the natural grouping or structure in a data set
  - **Cluster**: a collection of data objects that are "similar" to one another and thus can be treated collectively as one group
  - **Clustering**: **unsupervised** classification, no predefined classes
  - Used either as a stand-alone tool to get insight into data distribution or as a preprocessing step for other algorithms

Applications of Clustering

- **Applications**
  - **Market Segmentation**
    - By utilizing information such as customers’ purchasing patterns, residence, occupation, and income, customers can be divided into various groups.
    - Advertising and marketing strategies can then be tailored to the characteristics of each group.
  - **Image Segmentation**
    - Image data at the pixel level is divided into multiple segments, simplifying the original image to extract more meaningful information or facilitate analysis.



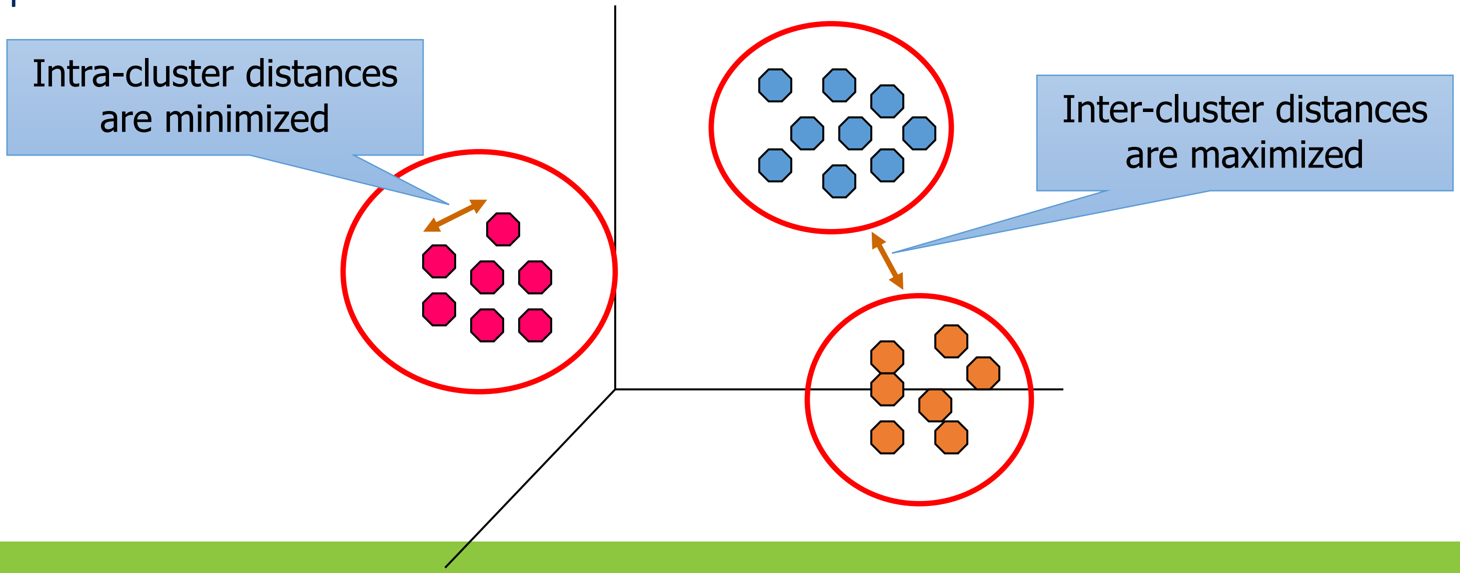
Original MRI Images



Segmented Images

Quality of Clustering

- **What Is Good Clustering?**
  - A good clustering method will produce high quality clusters in which:
    - the intra-class (that is, intra-cluster) similarity is high.
    - the inter-class similarity is low.
  - The quality of a clustering result also depends on both the similarity measure used by the method and its implementation.
  - The quality of a clustering method is also measured by its ability to discover some or all of the hidden patterns.



Mathematical Notation

- **Mathematical Notation**
  - A set of *n* objects: $S = \{O_1, O_2, \ldots, O_n\}$
  - Each object has *p* attributes (or variables)

$$O_i : \mathbf{x}_i = (x_{i1}, x_{i2}, \ldots, x_{ip})$$

  - A clustering result (or solution) is a partition of S:

$$P = \{C_1, C_2, \ldots, C_k\} \quad \text{where} \quad \bigcup_{i=1}^{k} C_i = S$$

$$C_i \cap C_j = \Phi, \text{ for } 1 \leq i \neq j \leq k$$

C_i : *i*-th cluster

k : number of clusters

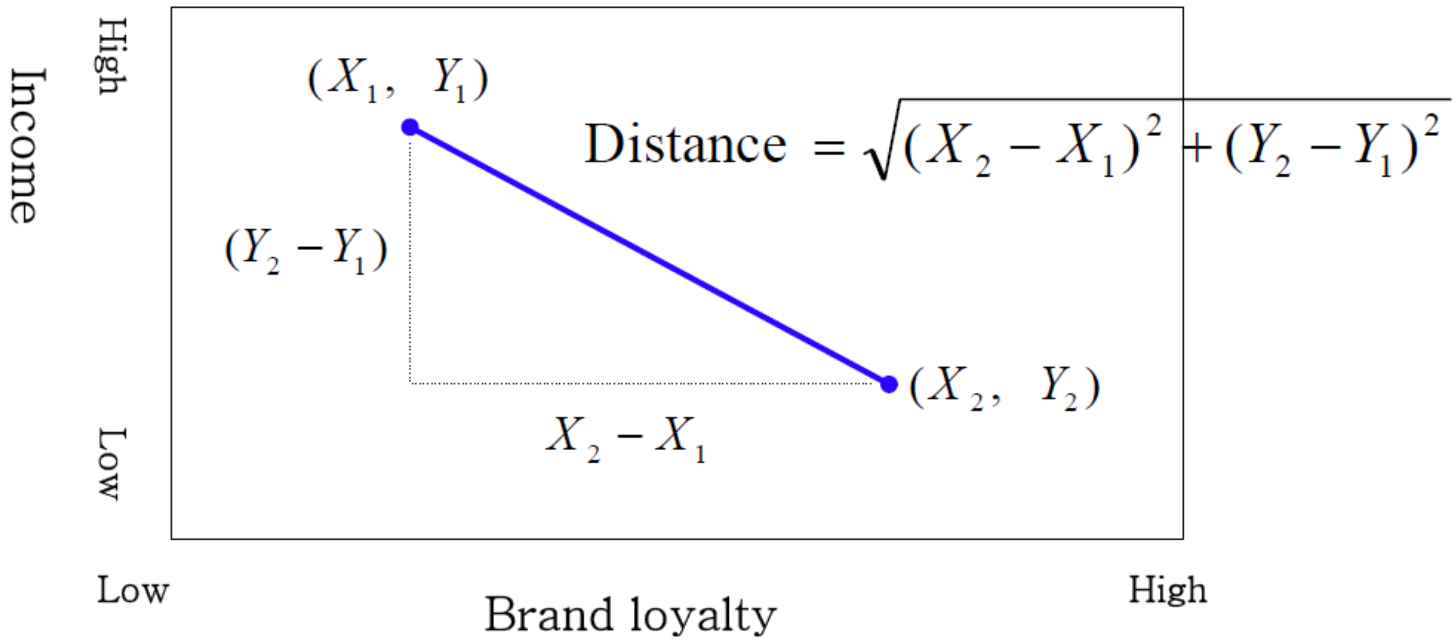
  - The objective of cluster analysis is to group objects into clusters such that each cluster is as **homogeneous** as possible with respect to the clustering variables

Steps in Cluster Analysis

- **Steps in Cluster Analysis**
  - Select a measure of dissimilarity
  - Select a clustering method: hierarchical / non-hierarchical
  - Decide the number of clusters
  - Interpret the result

Distance (or Dissimilarity) Measures

- **Euclidean distance (L2 norm)**



$$d_{ij} = d(\mathbf{x}_i, \mathbf{x}_j) = \left(\sum_{k=1}^{p} |x_{ik} - x_{jk}|^2 \right)^{\frac{1}{2}}$$

Distance (or Dissimilarity) Measures

- **Manhattan distance (L1 Norm)**

$$d_{ij} = d(\mathbf{x}_i, \mathbf{x}_j) = \sum_{k=1}^{p} |x_{ik} - x_{jk}|$$

- **Minkowski distance**

$$d_{ij} = d(\mathbf{x}_i, \mathbf{x}_j) = \left(\sum_{k=1}^{p} |x_{ik} - x_{jk}|^m \right)^{\frac{1}{m}}$$

- **Standardized Minkowski distance**

$$d_{ij} = d(\mathbf{x}_i, \mathbf{x}_j) = \left(\sum_{k=1}^{p} \left| \frac{x_{ik} - x_{jk}}{s_k} \right|^m \right)^{\frac{1}{m}}, \; s_k = \sqrt{\frac{\sum_{a=1}^{n} (x_{ak} - \bar{x}_k)^2}{n-1}}$$

Distance (or Dissimilarity) Measures

- **Cosine Distance**

$$Cosine\ Similarity(x_i, x_j) = \frac{\sum_{k=1}^{p} x_{ik} x_{jk}}{\sqrt{\sum_{k=1}^{p} x_{ik}^2}\sqrt{\sum_{k=1}^{p} x_{jk}^2}}$$

$$d_{ij} = 1 - Cosine\ Similarity(x_i, x_j)$$

- **Mahalanobis distance**

$$d_{ij} = d(\mathbf{x}_i, \mathbf{x}_j) = \sqrt{(\mathbf{x}_i - \mathbf{x}_j)^\top \mathbf{S}^{-1} (\mathbf{x}_i - \mathbf{x}_j)}, \quad \mathbf{S} \ \text{: var-cov matrix}$$

Distance Measures for Binary Variables

- **A contingency table for binary data**

		Object j		
		1	0	*sum*
	1	a	b	$a+b$
Object i	0	c	d	$c+d$
	sum	$a+c$	$b+d$	p

- **Distance = 1-Similarity**
  - Based on Simple matching coefficient

$$d_{ij} = d(\mathbf{x}_i, \mathbf{x}_j) = 1 - \frac{a+d}{a+b+c+d} = \frac{b+c}{p}$$

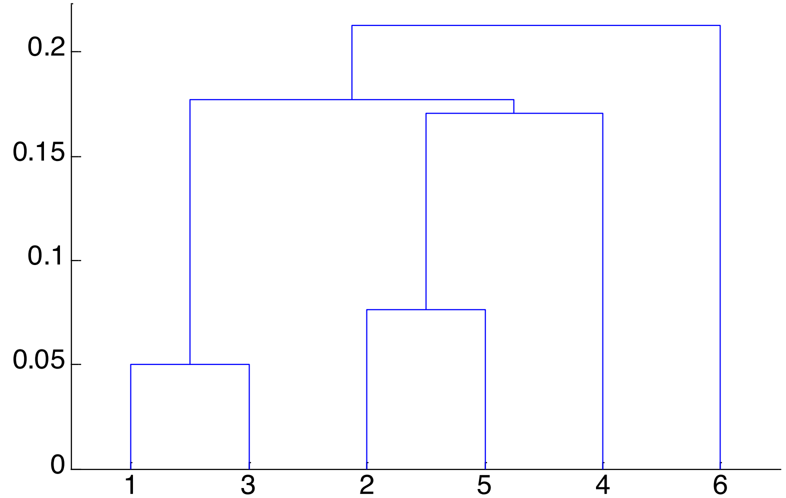
  - Based on Jaccard Coefficient

$$d_{ij} = 1 - \frac{a}{a+b+c} = \frac{b+c}{a+b+c}$$

Hierarchical Clustering

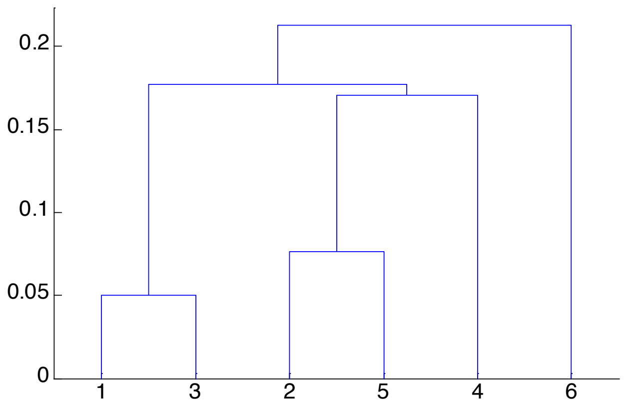
Hierarchical Clustering

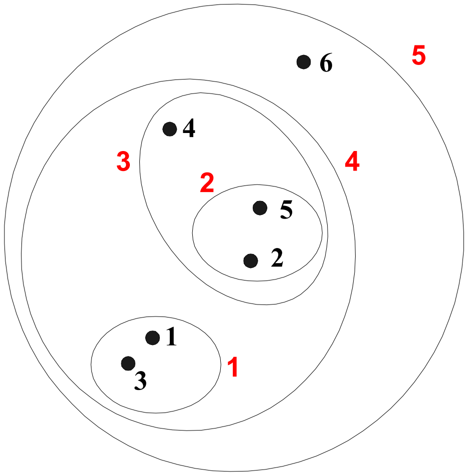
- **Hierarchical Clustering**
  - **Hierarchical clustering** is a general family of clustering algorithms that build nested clusters by **merging** or **splitting** them successively
  - This hierarchy of clusters is represented as a tree (or **dendrogram**)
    - A tree-like diagram that records the sequence of merges / splits
    - The root of the tree is the unique cluster that gathers all the samples, the leaves being the clusters with only one sample



Hierarchical Clustering

- **Two Types of Hierarchical Clustering**
  - **Agglomerative Clustering**
    - Start with the data points as individual clusters
    - At each step, **merge** the closest pair of clusters until only one cluster (or k clusters) left





  - **Divisive Clustering**
    - Start with one all-inclusive cluster (the entire dataset as a cluster)
    - At each step, **split** a cluster until each cluster contains an individual point (or there are k clusters)

Hierarchical Clustering

- **Agglomerative Clustering**
  - **Agglomerative clustering** performs a hierarchical clustering using a bottom up approach
  - **Step 0:** Start with the objects as individual clusters.
    - Consider each object as one cluster. $k = n$
  - **Step 1:** At each step, merge the closest pair of clusters until only one cluster (or k clusters) left.
    - Compute $d\left(C_i, C_j\right)$, $1 \le i \ne j \le k$, for every pair of clusters.
    - Find the minimum distance and combine two clusters as a single cluster.
    - $k \leftarrow k - 1$
  - **Step 2:** Stop or repeat.
    - If $k = 1$, stop
    - Otherwise, repeat Step 1.
  - The **linkage criteria** determines the metric used for the merge strategy

Linkage Strategies

- **Distance Measures between Clusters**

$$d\left(C_i, C_j\right): \text{distance between cluster } C_i \text{ and } C_j$$

  - Single Linkage:

$$d\left(C_i, C_j\right) = \min_{\mathbf{x} \in C_i, \mathbf{y} \in C_j} d(\mathbf{x}, \mathbf{y})$$

  - Complete Linkage:

$$d\left(C_i, C_j\right) = \max_{\mathbf{x} \in C_i, \mathbf{y} \in C_j} d(\mathbf{x}, \mathbf{y})$$

  - Average Linkage:

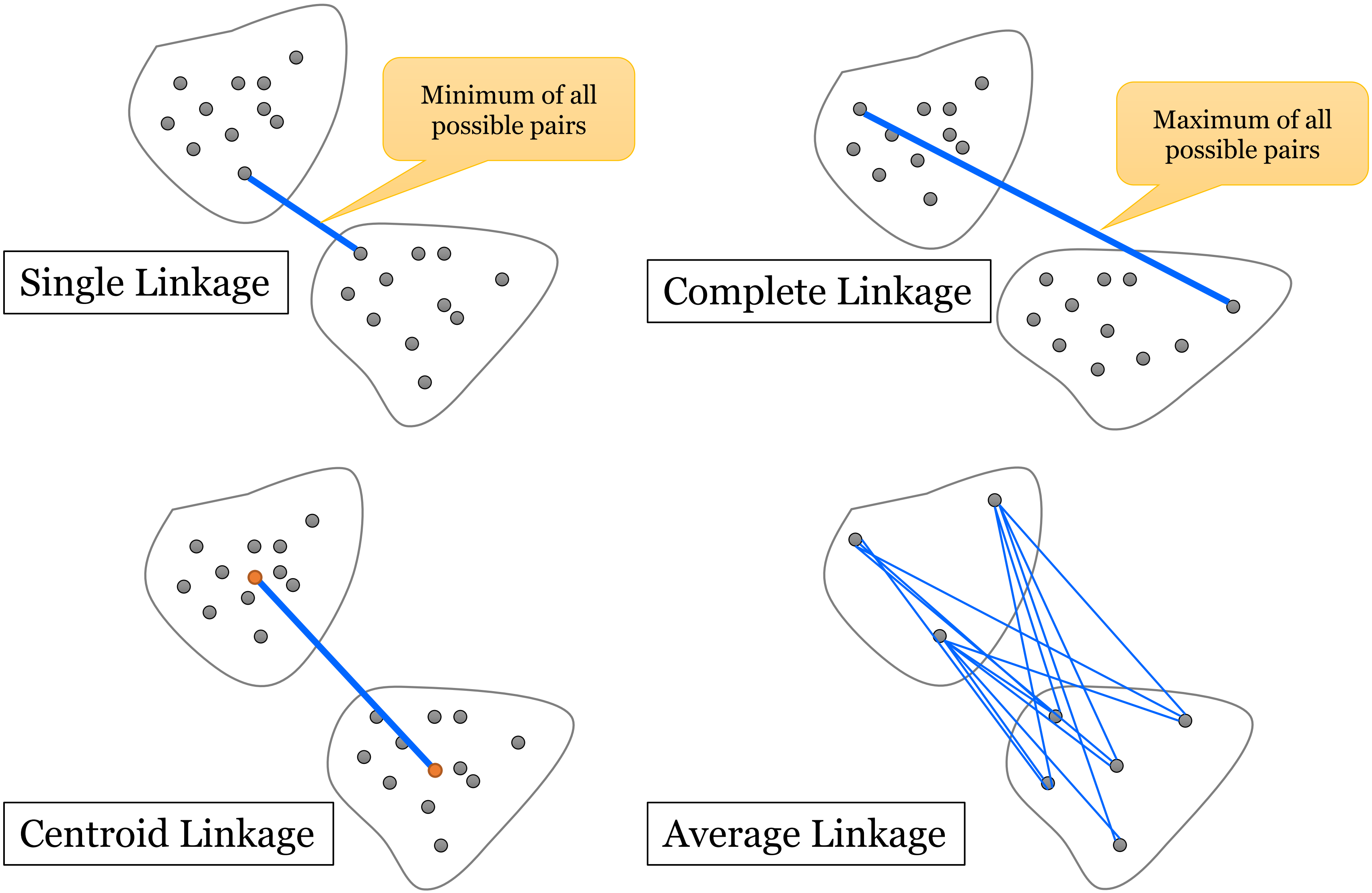
$$d\left(C_i, C_j\right) = \frac{1}{\left|C_i\right| \left|C_j\right|} \sum_{\mathbf{x} \in C_i, \mathbf{y} \in C_j} d(\mathbf{x}, \mathbf{y}) \ , \quad \left|C_i\right|: \text{ number of objects in } C_i$$

  - Centroid Linkage:

$$d\left(C_i, C_j\right) = d(\mathbf{c}_i, \mathbf{c}_j) \ , \quad \mathbf{c}_i = \left(\overline{x}_1^{(i)}, \overline{x}_2^{(i)}, \ldots, \overline{x}_p^{(i)}\right) \ , \ \overline{x}_h^{(i)} = \frac{1}{\left|C_i\right|} \sum_{a \in C_i} x_{ah}$$

  - Ward's Method: Will be introduced later.

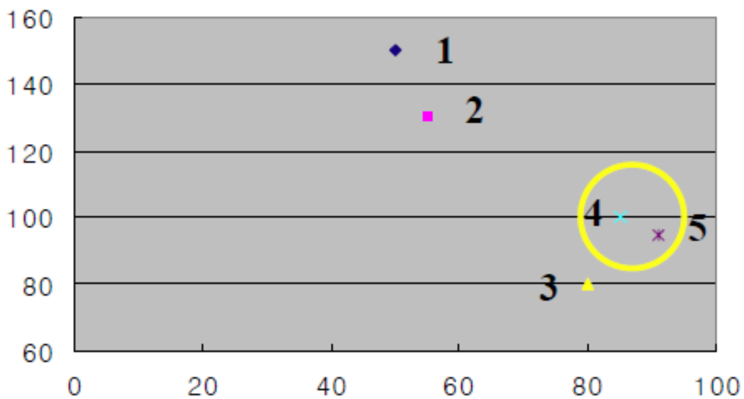
Linkage Strategies



Single Linkage Method (1/4)

ID	Income	Brand loyalty
1	150	50
2	130	55
3	80	80
4	100	85
5	95	91

Based on the minimum distance. Two objects separated by the shortest distance are placed in the first cluster. Then, next shortest distance is found, etc.



$$d\left(C_i, C_j\right) = \min_{\mathbf{x}\in C_i, \mathbf{y}\in C_j} d(\mathbf{x}, \mathbf{y})$$

ID	1	2	3	4	5
1	0.0				
2	20.6	0.0			
3	76.2	55.9	0.0		
4	61.0	42.4	20.6	0.0	
5	68.6	50.2	18.6	7.8	0.0

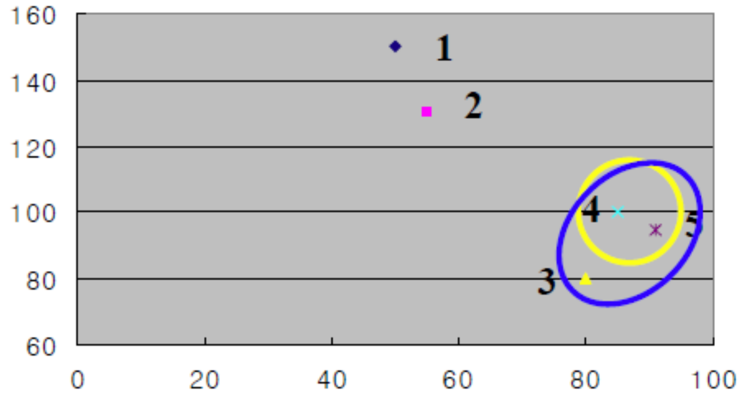
Single Linkage Method (2/4)

ID	(4,5)	1	2	3
(4,5)	0.0			
1	61.0	0.0		
2	42.4	20.6	0.0	
3	18.6	76.2	55.9	0.0

$$d\left((O_1),(O_4,O_5)\right)=\min\left\{d_{14},d_{15}\right\}=d_{14}=61.0$$

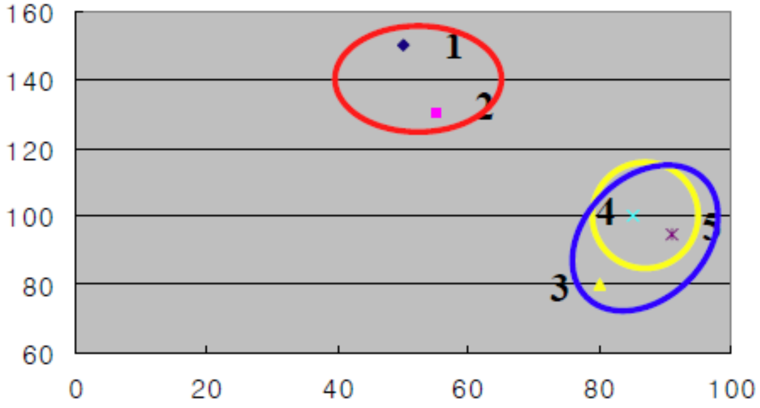
$$d\left((O_2),(O_4,O_5)\right)=\min\left\{d_{24},d_{25}\right\}=d_{24}=42.4$$

$$d\left((O_3),(O_4,O_5)\right)=\min\left\{d_{34},d_{35}\right\}=d_{35}=18.6$$



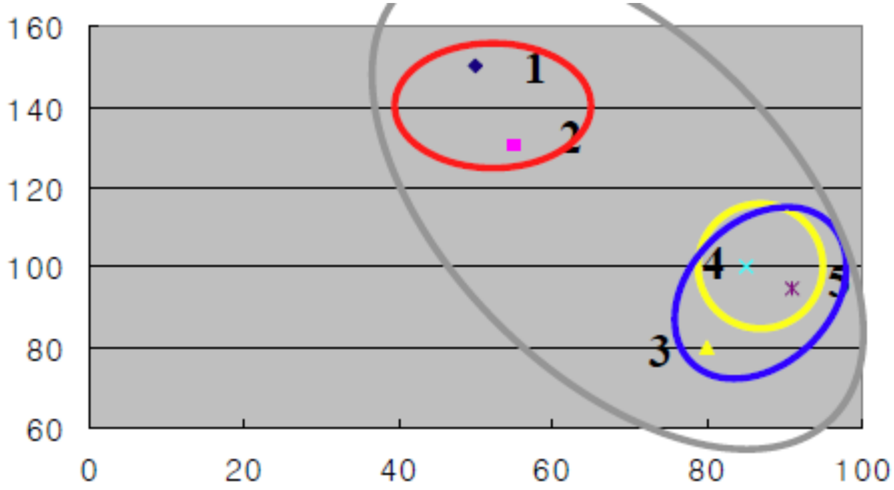
Single Linkage Method (3/4)

ID	(3,4,5)	1	2
(3,4,5)	0.0		
1	61.0	0.0	
2	42.4	20.6	0.0



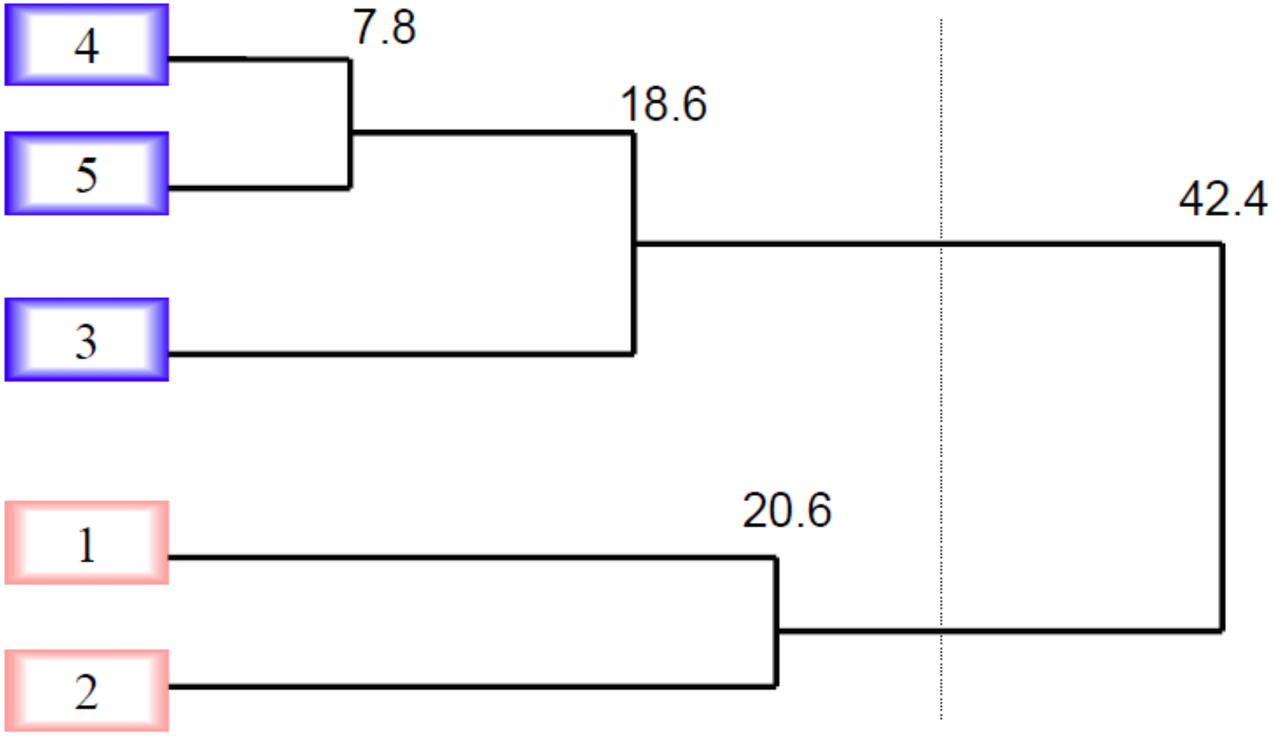
$$d\left(\left(O_1\right),\left(O_3,O_4,O_5\right)\right)=\min\left\{d_{13},d_{14},d_{15}\right\}=d_{14}=61.0$$

$$d\left(\left(O_2\right),\left(O_3,O_4,O_5\right)\right)=\min\left\{d_{23},d_{24},d_{25}\right\}=d_{24}=42.4$$



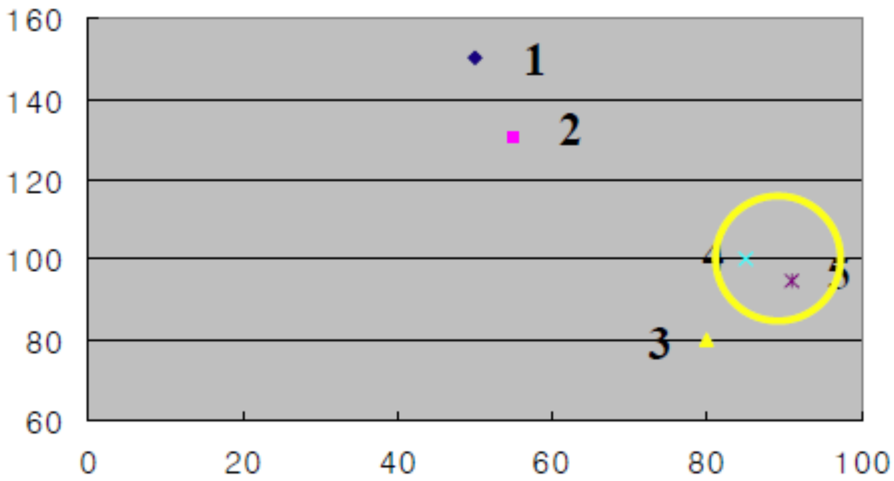
Single Linkage Method (4/4)

- **Dendrogram**



Complete Linkage Method (1/4)

$$d\left(C_i, C_j\right) = \max_{\mathbf{x}\in C_i, \mathbf{y}\in C_j} d(\mathbf{x}, \mathbf{y})$$



Equal to single linkage method, except that maximum distance is applied as cluster criterion.

ID	1	2	3	4	5
1	0.0				
2	20.6	0.0			
3	76.2	55.9	0.0		
4	61.0	42.4	20.6	0.0	
5	68.6	50.2	18.6	7.8	0.0

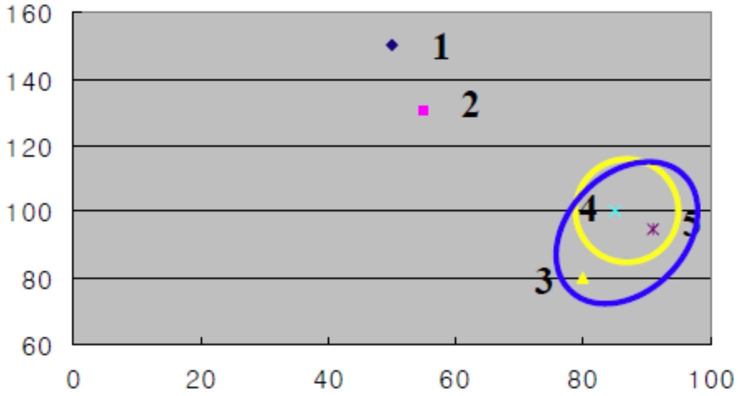
Complete Linkage Method (2/4)

ID	(4,5)	1	2	3
(4,5)	0.0			
1	68.6	0.0		
2	50.2	20.6	0.0	
3	20.6	76.2	55.9	0.0

$$d\left((O_1),(O_4,O_5)\right) = \max\left\{d_{14}, d_{15}\right\} = d_{15} = 68.6$$

$$d\left((O_2),(O_4,O_5)\right) = \max\left\{d_{24}, d_{25}\right\} = d_{25} = 50.2$$

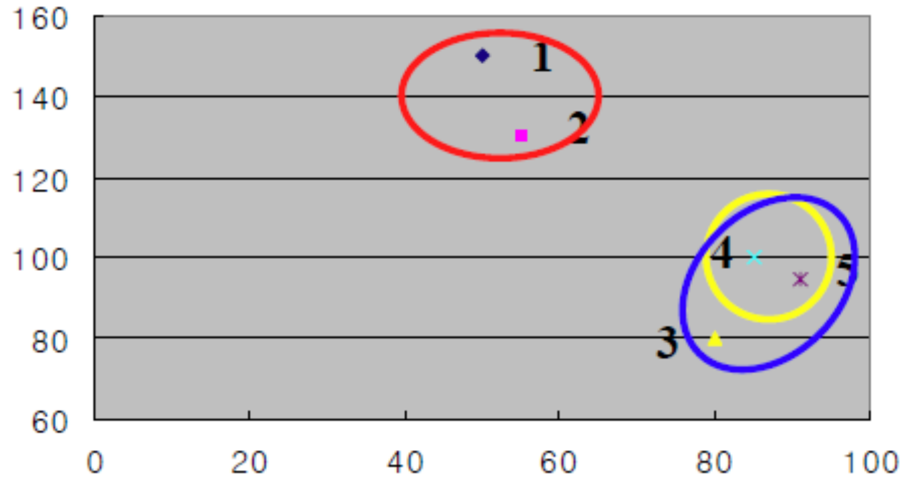
$$d\left((O_3),(O_4,O_5)\right) = \max\left\{d_{34}, d_{35}\right\} = d_{34} = 20.6$$

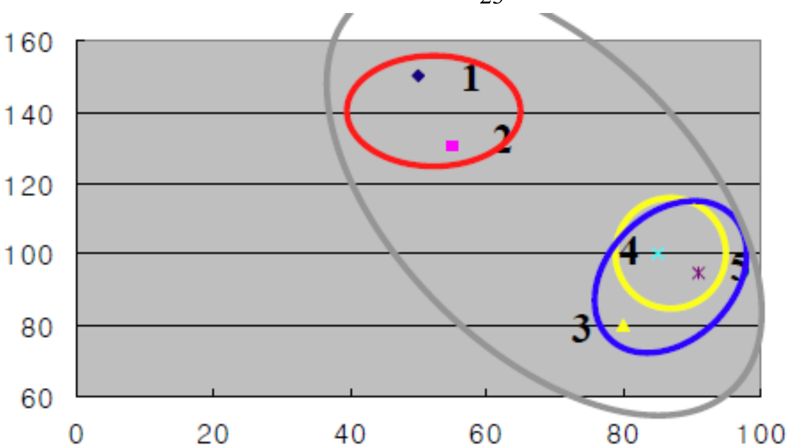


Complete Linkage Method (3/4)

ID	(3,4,5)	1	2
(3,4,5)	0.0		
1	76.2	0.0	
2	55.9	20.6	0.0

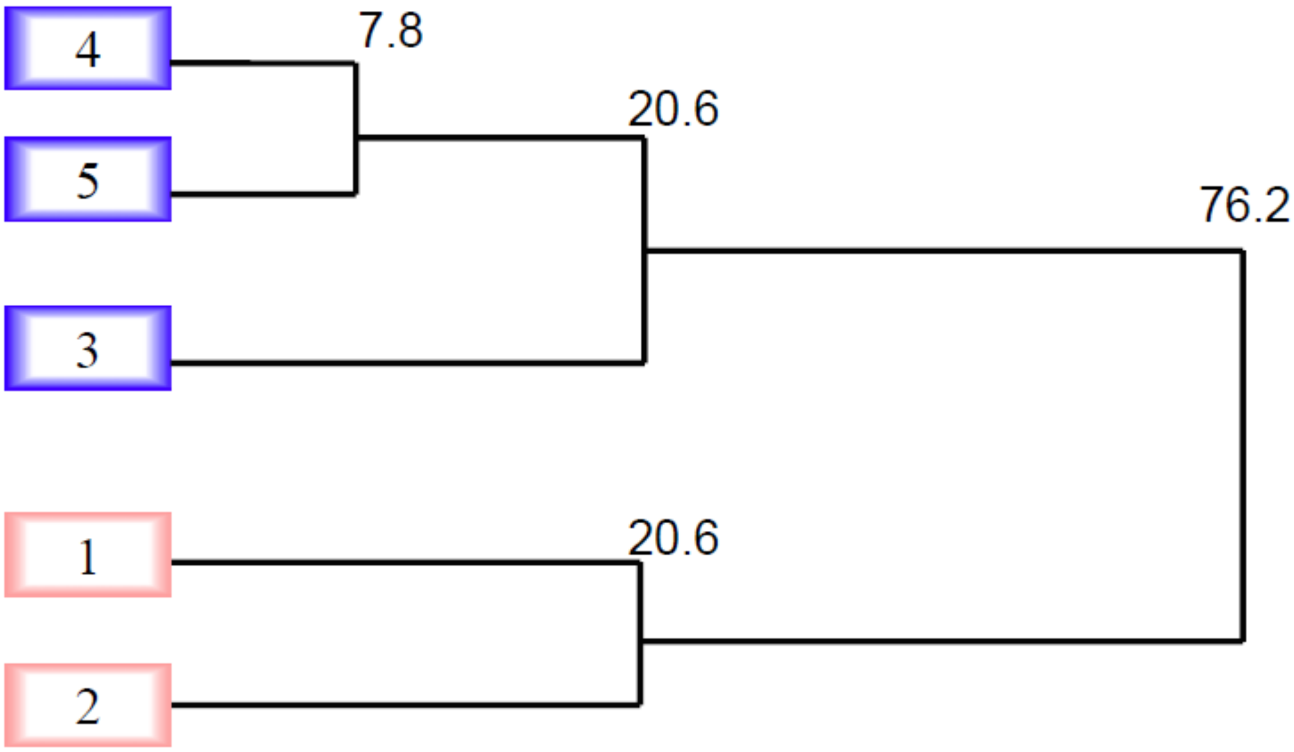
$$d\left(\left(O_1\right),\left(O_3,O_4,O_5\right)\right)=\max\left\{d_{13},d_{14},d_{15}\right\}$$
$$= d_{13} = 76.2$$
$$d\left(\left(O_2\right),\left(O_3,O_4,O_5\right)\right)=\max\left\{d_{23},d_{24},d_{25}\right\}$$
$$= d_{23} = 55.9$$





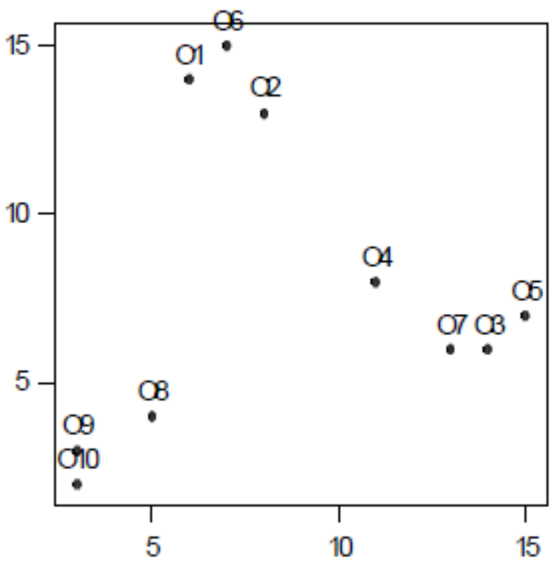
Complete Linkage Method (4/4)

- **Dendrogram**



Average Linkage Method (1/4)

	O_1	O_2	O_3	O_4	O_5	O_6	O_7	O_8	O_9	O_{20}
X_1	6	8	14	11	15	7	13	5	3	3
X_2	14	13	6	8	7	15	6	4	3	2



Iteration 0.

- Consider each object (O_i) as a cluster (C_i).
- Clustering result:

$$C_1 = \{O_1\}, C_2 = \{O_2\}, \ldots, C_{10} = \{O_{10}\}$$

- $k = 10$

Average Linkage Method (2/4)

	C1	C2	C3	C4	C5	C6	C7	C8	C9
C2	2.24								
C3	11.31	9.22							
C4	7.81	5.83	3.61						
C5	11.40	9.22	1.41	4.12					
C6	1.41	2.24	11.40	8.06	11.31				
C7	10.63	8.60	1.00	2.83	2.24	10.82			
C8	10.05	9.49	9.22	7.21	10.44	11.18	8.25		
C9	11.40	11.18	11.40	9.43	12.65	12.65	10.44	2.24	
C10	12.37	12.08	11.70	10.00	13.00	13.60	10.77	2.83	1.00

Iteration 1.

- Merge C_9 and C_{10}, having the closest distance.
- Let the merged cluster be C_9 .
- Clustering result:

$$k = 9: \; C_1 = \{O_1\}, C_2 = \{O_2\}, \ldots, C_9 = \{O_9, O_{10}\}$$

Average Linkage Method (3/4)

	C1	C2	C3	C4	C5	C6	C7	C8
C2	2.24							
C3	11.31	9.22						
C4	7.81	5.83	3.61					
C5	11.40	9.22	1.41	4.12				
C6	1.41	2.24	11.40	8.06	11.31			
C7	10.63	8.60	1.00	2.83	2.24	10.82		
C8	10.05	9.49	9.22	7.21	10.44	11.18	8.25	
C9	11.89	11.63	11.55	9.72	12.82	13.13	10.61	2.53

Iteration 2.

- Merge C_3 and C_7 , having the closest distance.
- Let the merged cluster be C_3 .
- Clustering result:

$$k=8:\ C_1=\{O_1\}, C_2=\{O_2\}, C_3=\{O_3,O_7\}, C_4=\{O_4\},$$
$$C_5=\{O_5\}, C_6=\{O_6\}, C_7=\{O_8\}, C_8=\{O_9,O_{10}\}$$

	C1	C2	C3	C4	C5	C6	C7
C2	2.24						
C3	11.31	9.22					
C4	7.81	5.83	3.22				
C5	11.40	9.22	1.83	4.12			
C6	1.41	2.24	11.11	8.06	11.31		
C7	10.05	9.49	8.73	7.21	10.44	10.82	
C8	11.89	11.63	11.08	9.72	12.82	13.13	2.53

Iteration 3.

- Merge C_1 and C_6 , having the closest distance.
- Let the merged cluster be C_1 .
- Clustering result:

$$k=7:\ C_1=\{O_1,O_6\}, C_2=\{O_2\}, C_3=\{O_3,O_7\}, C_4=\{O_4\},$$
$$C_5=\{O_5\}, C_6=\{O_8\}, C_7=\{O_9,O_{10}\}$$

Average Linkage Method (4/4)

Iteration 4.

$$k=6:\ C_1=\{O_1,O_6\},C_2=\{O_2\},C_3=\{O_3,O_5,O_7\},C_4=\{O_4\},C_5=\{O_8\},C_6=\{O_9,O_{10}\}$$

Iteration 5.

$$k=5:\ C_1=\{O_1,O_2,O_6\},C_2=\{O_3,O_5,O_7\},C_3=\{O_4\},C_4=\{O_8\},C_5=\{O_9,O_{10}\}$$

Iteration 6.

$$k=4:\ C_1=\{O_1,O_2,O_6\},C_2=\{O_3,O_5,O_7\},C_3=\{O_4\},C_5=\{O_8,O_9,O_{10}\}$$

Iteration 7.

$$k=3:\ C_1=\{O_1,O_2,O_6\},C_2=\{O_3,O_4,O_5,O_7\},C_3=\{O_8,O_9,O_{10}\}$$

Iteration 8.

	C1	C3
C3	9.64	
C8	11.56	10.38

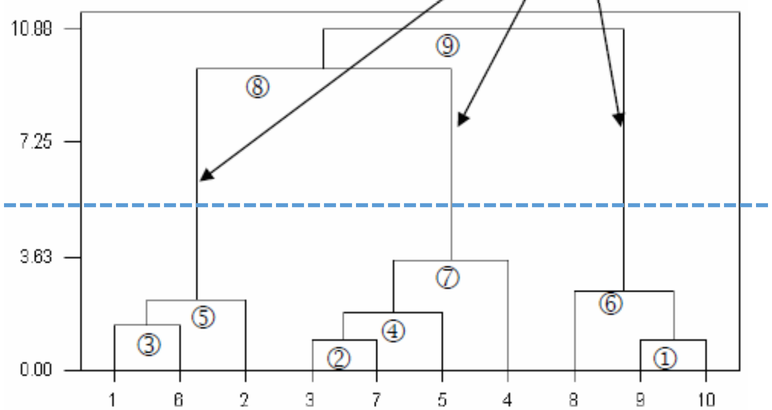
$$k=2:\ C_1=\{O_1,O_2,O_3,O_4,O_5,O_6,O_7\},C_2=\{O_8,O_9,O_{10}\}$$

Iteration 9.

$$k=1:\ C_1=\{O_1,O_2,O_3,O_4,O_5,O_6,O_7,O_8,O_9,O_{10}\}$$

3 clusters seem good.

Distance between clusters



Dendrogram

Ward's Method (1/6)

- **Under a clustering result** $P=\{C_1,C_2,\ldots,C_k\}$
  - **Within-cluster** sum of squares for cluster C_i

$$SSW(C_i)=\sum_{u\in C_i}\sum_{h=1}^{p}\left(x_{uh}-\bar{x}_h^{(i)}\right)^2=\sum_{u\in C_i}(\mathbf{x}_u-\mathbf{c}_i)^T(\mathbf{x}_u-\mathbf{c}_i)$$

  - Total within-cluster sum of squares

$$SSW(P)=\sum_{i=1}^{k}SSW(C_i)=\sum_{i=1}^{k}\sum_{u\in C_i}\sum_{h=1}^{p}\left(x_{uh}-\bar{x}_h^{(i)}\right)^2$$

$$=\sum_{i=1}^{k}\sum_{u\in C_i}(\mathbf{x}_u-\mathbf{c}_i)^T(\mathbf{x}_u-\mathbf{c}_i)$$

  - When combining C_i and C_j, let the clustering result be $\tilde{P}$.

$$SSW(\tilde{P})=\sum_{r\neq i,j}SSW(C_r)+SSW(C_i\cup C_j)$$

$$\boxed{\text{* Note that } SSW(\tilde{P})>SSW(P)}$$

Ward's Method (2/6)

- **Agglomerative Clustering with Ward's Method**
  - **Step 0:** Start with the objects as individual clusters.
    - Consider each object as one cluster. $k = n$
  - **Step 1:** At each step, merge the closest pair of clusters until only one cluster (or k clusters) left.
    - Compute new SSW for every pair of clusters assuming that they are combined.
    - Find **the minimum SSW** and combine two clusters as a single cluster.
    - Update the clustering result.
    - $k \leftarrow k - 1$
  - **Step 2:** Stop or repeat.
    - If $k = 1$, stop
    - Otherwise, repeat Step 1.

Ward's Method (3/6)

	O_1	O_2	O_3	O_4	O_5	O_6	O_7	O_8
X_1	4	20	3	19	17	8	19	18
X_2	15	13	13	4	17	11	12	6

Iteration 0.

$$k = 8: \; C_1 = \{O_1\}, C_2 = \{O_2\}, \ldots, C_8 = \{O_8\}$$

Iteration 1.

Table of total within-cluster sum of squares
when combining two clusters

	C1	C2	C3	C4	C5	C6	C7
C2	130.0	•					
C3	2.5	144.5	•				
C4	173.0	41.0	168.5	•			
C5	86.5	12.5	106	86.5	•		
C6	16.0	74.0	14.5	85	58.5	•	
C7	117.0	1.0	128.5	32	14.5	61	•
C8	138.5	26.5	137	2.5	61	62.5	18.5

$$k = 7: \; C_1 = \{O_1\}, C_2 = \{O_2, O_7\}, C_3 = \{O_3\}, C_4 = \{O_4\}$$
$$C_5 = \{O_5\}, C_6 = \{O_6\}, C_7 = \{O_8\}$$

Ward's Method (4/6)

Iteration 2.

	C1	C2	C3	C4	C5	C6
C2	165.33	•				
C3	3.50	182.66	•			
C4	174.00	49.33	169.50	•		
C5	87.50	18.67	107.00	87.50	•	
C6	17.00	90.67	15.50	86.00	59.50	•
C7	139.50	30.67	138.00	3.50	62.00	63.50

$$k = 6: \; C_1 = \{O_1\}, C_2 = \{O_2, O_7\}, C_3 = \{O_3\}, \\ C_4 = \{O_4, O_8\}, C_5 = \{O_5\}, C_6 = \{O_6\}$$

Iteration 3.

	C1	C2	C3	C4	C5
C2	167.83	•			
C3	6.00	185.16	•		
C4	210.33	60.75	206.33	•	
C5	90.00	21.17	109.50	101.00	•
C6	19.50	93.17	18.00	101.00	62.00

$$k = 5: \; C_1 = \{O_1, O_3\}, C_2 = \{O_2, O_7\}, \\ C_3 = \{O_4, O_8\}, C_4 = \{O_5\}, C_5 = \{O_6\}$$

Ward's Method (5/6)

Iteration 4.

	C1	C2	C3	C4
C2	264.25	•		
C3	312.00	63.25	•	
C4	133.50	23.67	103.50	•
C5	25.50	95.67	103.50	64.50

$$k = 4:\ C_1 = \{O_1, O_3\}, C_2 = \{O_2, O_5, O_7\},$$
$$C_3 = \{O_4, O_8\}, C_4 = \{O_6\}$$

Iteration 5.

	C1	C2	C3
C2	299.70	•	
C3	329.67	120.9	•
C4	43.17	115.75	121.17

$$k = 3:\ C_1 = \{O_1, O_3, O_6\}, C_2 = \{O_2, O_5, O_7\},$$
$$C_3 = \{O_4, O_8\}$$

Ward's Method (6/6)

Iteration 6.

	C1	C2
C2	324.83	•
C3	338.67	140.40

$k = 2:\ C_1 = \{O_1, O_3, O_6\}, C_2 = \{O_2, O_4, O_5, O_7, O_8\}$

Iteration 7.

	C1
C2	499.88

$k = 1:\ C_1 = \{O_1, O_2, O_3, O_4, O_5, O_6, O_7, O_8\}$