5

Clustering

IT 326: Data Mining

First semester 2021

Outline

- Cluster Analysis
- Partitioning Methods
- Hierarchical Methods
- Evaluation of Clustering
- Summary

What is Cluster Analysis?

- Clustering: the process of partitioning a set of data objects into subsets (clusters), where objects in a cluster are similar to one another, yet dissimilar to objects in other clusters.
  - Similar (or related) to one another within the same cluster.
  - Dissimilar (or unrelated) to the objects in other clusters.

- Cluster analysis (or clustering, data segmentation, automatic classification…).
  - Finding similarities between data according to the characteristics found in the data and grouping similar data objects into clusters.

- Considered as unsupervised learning: no predefined classes (i.e., learning by observations vs. learning by examples: supervised).
  - Descriptive data mining method, often for exploratory analysis.

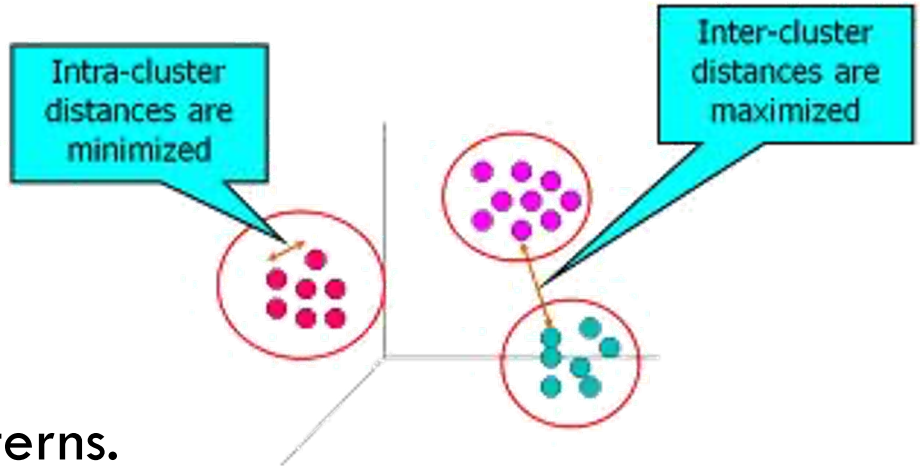
Examples of Clustering Applications

- Biology: Group genes that perform the same function.
- Group individuals with similar political view.
- Information Retrieval: Group documents of similar topic.
- Identify similar objects from pictures.
- Marketing: Help marketers discover distinct groups in their customer bases, and then use this knowledge to develop targeted marketing programs.
- City-planning: Identifying groups of houses according to their house type, value, and geographical location.

Quality: What Is Good Clustering?

- A good clustering method will produce high quality clusters:
  - **High** intra-class similarity: cohesive within clusters.
  - **Low** inter-class similarity: distinctive between clusters.

- The **quality** of a clustering method depends on:
  - The similarity measure used by the method.
  - Its implementation.
  - Its ability to discover some or all of the hidden patterns.



Intra-cluster vs. Inter-cluster

Measure the Quality of Clustering

Dissimilarity/Similarity metric:

- Similarity is expressed in terms of a distance function, typically metric: d(i, j).
- The definitions of distance functions are usually rather different for interval-scaled, Boolean, categorical, ordinal, ratio, and vector variables.
- Weights should be associated with different variables based on applications and data semantics.

Quality of clustering:

- There is usually a separate "quality" function that measures the "goodness" of a cluster.
- It is hard to define "similar enough" or "good enough" .
  - The answer is typically highly subjective.

Major Clustering Approaches

- Partitioning approach: Construct various partitions and then evaluate them by some criterion, e.g., minimizing the sum of square errors.
  - Typical methods: k-means, k-medoids…
- Hierarchical approach: Create a hierarchical decomposition of the set of data (or objects) using some criterion.
  - Typical methods: DIANA, AGNES…

Partitioning Methods

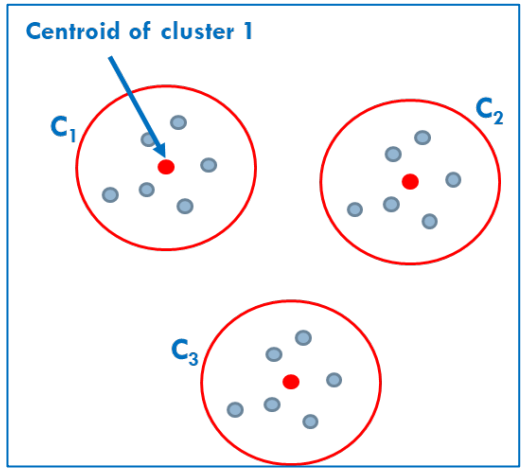
Partitioning method: Partitioning a database D of n objects into a set of k clusters, such that the sum of squared distances (E) is minimized (where c_i is the centroid or medoid of cluster C_i).

$$E = \sum_{i=1}^{k} \sum_{p \in C_i} dist(\boldsymbol{p}, \boldsymbol{c_i})^2,$$

$$E = \Sigma_{i=1}^{k} \Sigma_{p \in C_i} (p - c_i)^2$$

Centroid of cluster i

elements of cluster i

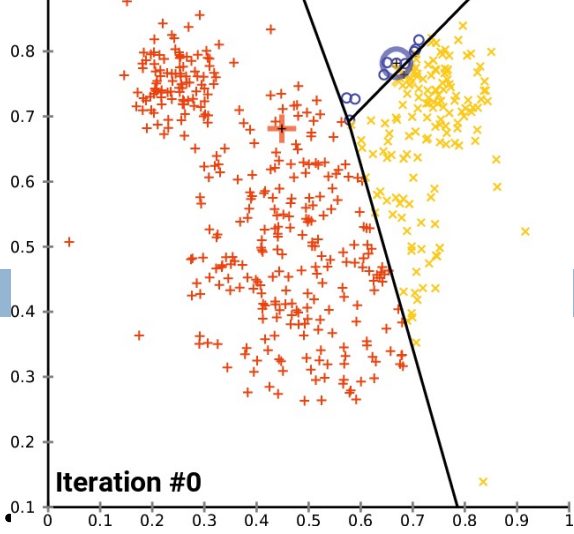


- Given k, find a partition of k clusters that optimizes the chosen partitioning criterion.

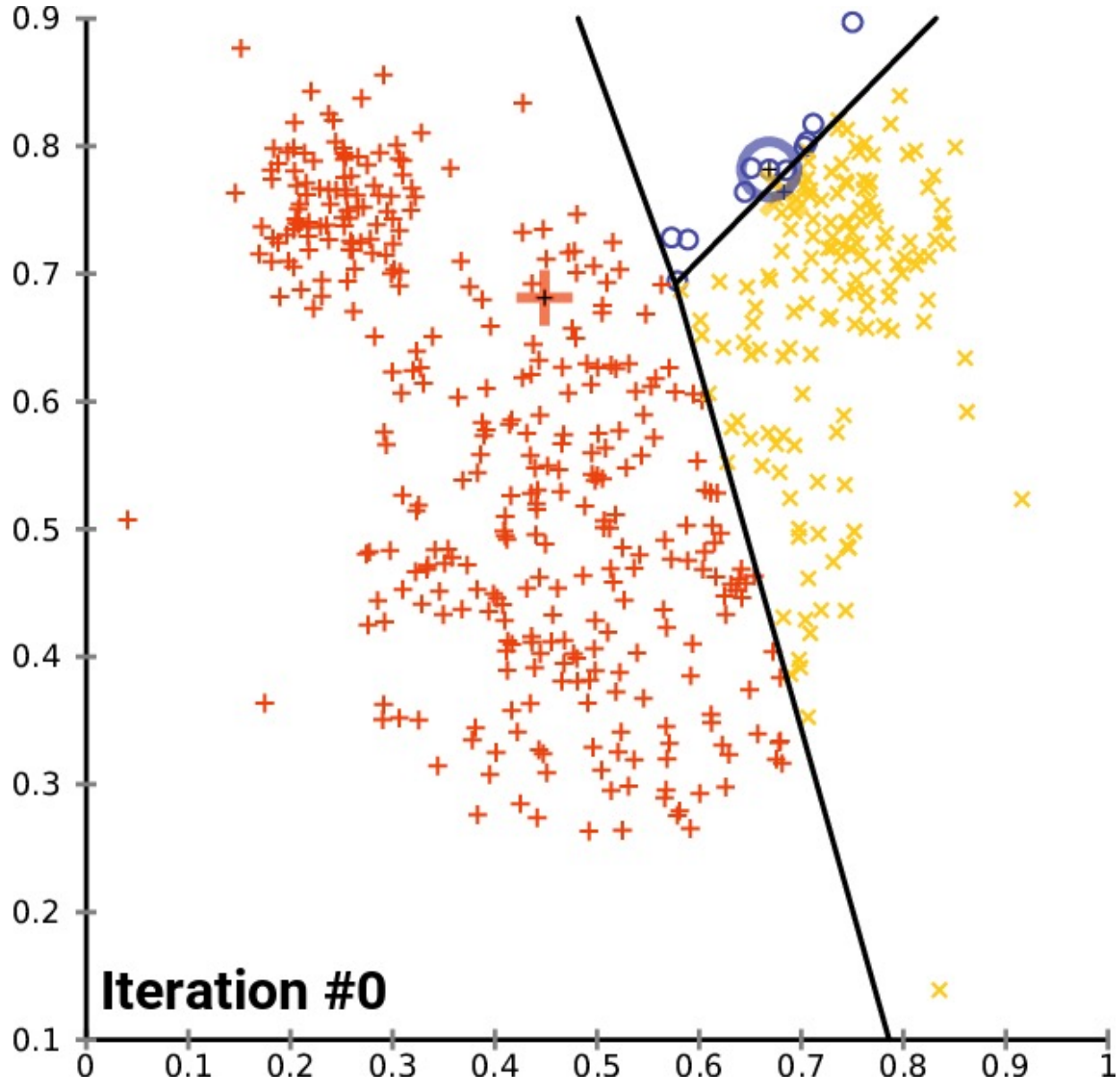
Partitioning Methods

- Heuristic methods:
  - k-means
    - Each cluster is represented by the center of the cluster.
  - k-medoids
    - or PAM (Partition around medoids)
    - Each cluster is represented by one of the objects in the cluster.

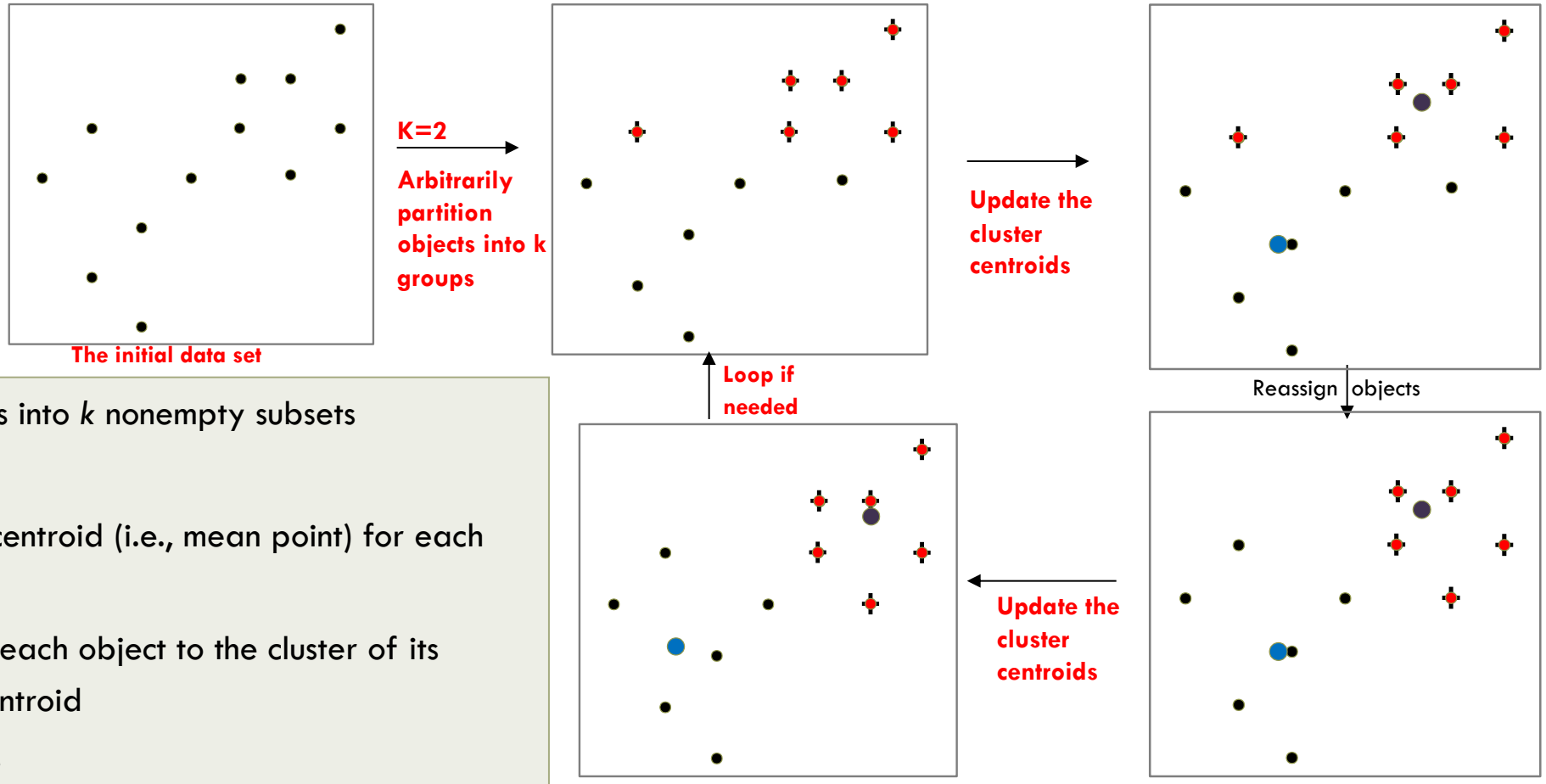
The K-Means Clustering Method



- Given k, the k-means algorithm is implemented in four steps:
  1. Choose k objects as the initial cluster centers (**seeds** or random).
  2. Assign each object to the cluster with the nearest center point based on Euclidean distance.
  3. Update the cluster centroids of the current partitioning (the centroid is the center, i.e., mean point, of the cluster).
  4. Go back to Step 2, or stop when the assignment does not change (centroid is the same).


0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
Iteration #0

Example 1: K-Means Clustering

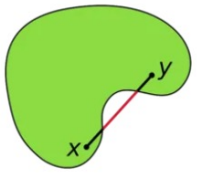


1. Partition objects into *k* nonempty subsets
2. Repeat:
   - Compute centroid (i.e., mean point) for each partition
   - (re)assign each object to the cluster of its nearest centroid
3. Until no change

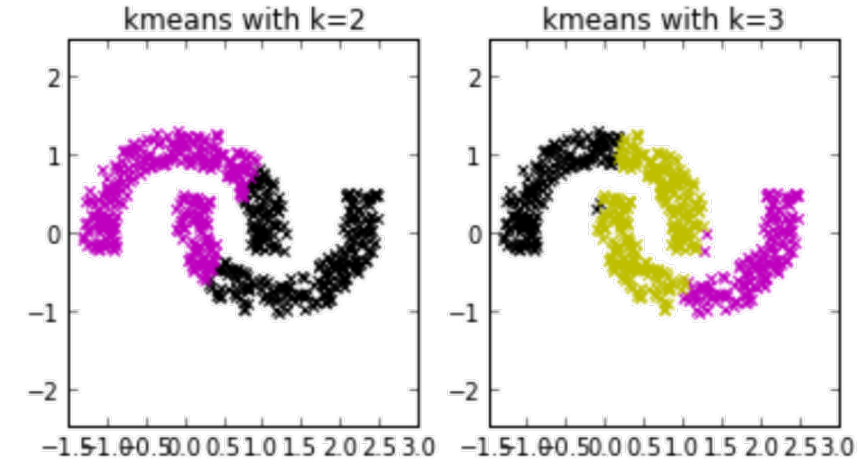
Online Example: http://www.onmyphd.com/?p=k-means.clustering&ckattempt=1, https://user.ceng.metu.edu.tr/~akifakkus/courses/ceng574/k-means

Comments on the K-Means Method

- **Strength:**
  - Efficient in processing large datasets.
    - Normally $k \,\&\, t << n$; where k is the number of clusters, t is the number of iterations, and n is the number of objects.
- **Weakness:**
  - Need to specify k, the number of clusters, in advance.
  - Sensitive to noisy data and outliers.
  - Not suitable to discover clusters with non-convex shapes.



A non-convex set



k-means performs poorly on the banana shapes

Example 2: K-Means Method

- Given:
  - The scores of two variables on each of seven individuals:

Subject	A	B
1	1.0	1.0
2	1.5	2.0
3	3.0	4.0
4	5.0	7.0
5	3.5	5.0
6	4.5	5.0
7	3.5	4.5

  - Number of clusters $K = 2$.
  - The distance measure is Euclidean distance .

$$d(i, j) = \sqrt{(x_{i1} - x_{j1})^2 + (x_{i2} - x_{j2})^2 + \cdots + (x_{ip} - x_{jp})^2}.$$

Example 2: K-Means Method cont.

Subject	A	B
1	1.0	1.0
2	1.5	2.0
3	3.0	4.0
4	5.0	7.0
5	3.5	5.0
6	4.5	5.0
7	3.5	4.5

Step 1: The initial clusters cluster centroids are m1 = (1.0,1.0) & m2=(5.0, 7.0).

	Individual	Mean Vector (centroid)
Group 1	1	(1.0, 1.0)
Group 2	4	(5.0, 7.0)

Step 2: Examine the remaining individuals and allocate them to the closest cluster.

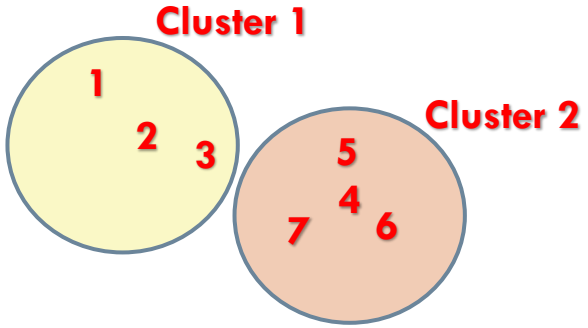
- For example, the distance between individual 2 and the two clusters can be calculated as follows;

$$d(m_1,2)=\sqrt{|1.0-1.5|^2+|1.0-2.0|^2}=1.12$$

$$d(m_2,2)=\sqrt{|5.0-1.5|^2+|7.0-2.0|^2}=6.10$$

which means that individual 2 is closer to cluster 1.

	Individual
Cluster 1	1, 2, 3
Cluster 2	4, 5, 6, 7



Example 2: K-Means Method cont.

Subject	A	B
1	1.0	1.0
2	1.5	2.0
3	3.0	4.0
4	5.0	7.0
5	3.5	5.0
6	4.5	5.0
7	3.5	4.5

Step 3: Recalculate the centroid of each cluster using the mean of the points.

	Individual
Cluster 1	1, 2, 3
Cluster 2	4, 5, 6, 7

$$m_1 = \left(\frac{1}{3}(1.0+1.5+3.0), \frac{1}{3}(1.0+2.0+4.0)\right) = (1.83, 2.33)$$

$$m_2 = \left(\frac{1}{4}(5.0+3.5+4.5+3.5), \frac{1}{4}(7.0+5.0+5.0+4.5)\right) = (4.12, 5.38)$$

- The new centroids are m1 = (1.8,2.3) & m2 =(4.1, 5.4).

	Individual	Mean Vector (centroid)
Cluster 1	1, 2, 3	(1.8, 2.3)
Cluster 2	4, 5, 6, 7	(4.1, 5.4)

Example 2: K-Means Method cont.

	Individual	Mean Vector (centroid)
Cluster 1	1, 2, 3	(1.8, 2.3)
Cluster 2	4, 5, 6, 7	(4.1, 5.4)

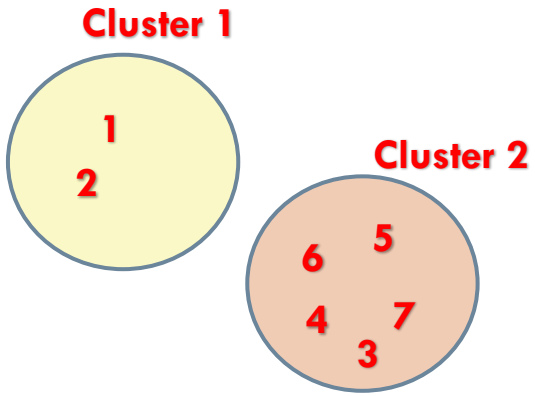
Subject	A	B
1	1.0	1.0
2	1.5	2.0
3	3.0	4.0
4	5.0	7.0
5	3.5	5.0
6	4.5	5.0
7	3.5	4.5

Step 4: Compare each individual's distance to its cluster mean and to that of the opposite cluster:

Individual	Distance to mean (centroid) of Cluster 1	Distance to mean (centroid) of Cluster 2
1	1.5	5.4
2	0.4	4.3
3	2.1	1.8
4	5.7	1.8
5	3.2	0.7
6	3.8	0.6
7	2.8	1.1

- Reassign to clusters, and then re-compute new centroids for clusters.
  - Individual **3** is **nearer** to the mean of cluster 2 than its own cluster (cluster 1).

	Individual	Mean Vector (centroid)
Cluster 1	1, 2	(1.3, 1.5)
Cluster 2	3, 4, 5, 6, 7	(3.9, 5.1)



Example 2: K-Means Method cont.

Step 5: Continue iterations until no more new cluster assignments occur.

- From the last results, each individual is nearer to its own cluster mean than that of the other cluster.

Activity#1: K-Means Method
Self Study

Consider A={6,7,8,11,12,13,18,20,22}; k=3 clusters,
Start with first three items, and the distance is the difference between two objects

1. Initialize each cluster.

   C1={6}, M1=6, C2={7}, M2=7, C3={8} and M3=8.

2. Round1: Assign each object to a cluster such that dist is minimal.
   - For example, object 11 is assigned to C3 because:
     - dist(M3,11)<dist(M2,11)
     - dist(M3,11)<dist(M1,11)
   - First Round cluster:
     - C1={6} M1=6, C2={7}, M2=7, C3={8,11,12,13,18,20,22} M3=104/7=14.86

3. Round2: Reassign each point to nearest cluster.
   - dist(8,M2)<dist(8,M3) 8 goes to C2.
   - dist(11,M3)<dist(11,M2) 11 goes to C3.
   - Second round cluster:
     - C1={6} M1=6, C2={7,8}, M2=7.5, C3={11,12,13,18,20,22} M3=16

4. Round3: Reassign each point to nearest cluster.
   - dist(11,M2)<dist(11,M3) 11 goes to C2.
   - Third round cluster:
     - C1={6} M1=6, C2={7,8,11}, M2=8.7, C3={12,13,18,20,22} M3=17

5. Round4: Reassign each point to nearest cluster.
   - dist(7,M1)<dist(7,M2) 7 goes to C1.
   - dist(12,M2)<dist(12,M3) 12 goes to C2.
   - Fourth round cluster:
     - C1={6,7} M1=6.5, C2={8,11,12}, M2=10.3, C3={13,18,20,22} M3=18.25

6. Round5: Reassign each point to nearest cluster.
   - dist(8,M1)<dist(8,M2)8 goes to C1.
   - dist(13,M2)<dist(13,M3)13 goes to C2
   - Fifth round cluster:
     - C1={6,7,8} M1=7, C2={11,12,13} M2=12, C3={18,20,22} M3=20

Hierarchical Clustering

- A hierarchical clustering works by grouping data objects into a tree of clusters (dendrogram).

- Hierarchical clustering is either:
  - Bottom-up (agglomerative).
  - Top-down (divisive).

- Use distance matrix as clustering criteria.
  - It does not require the number of clusters k as an input,
  - but needs a **termination condition.**

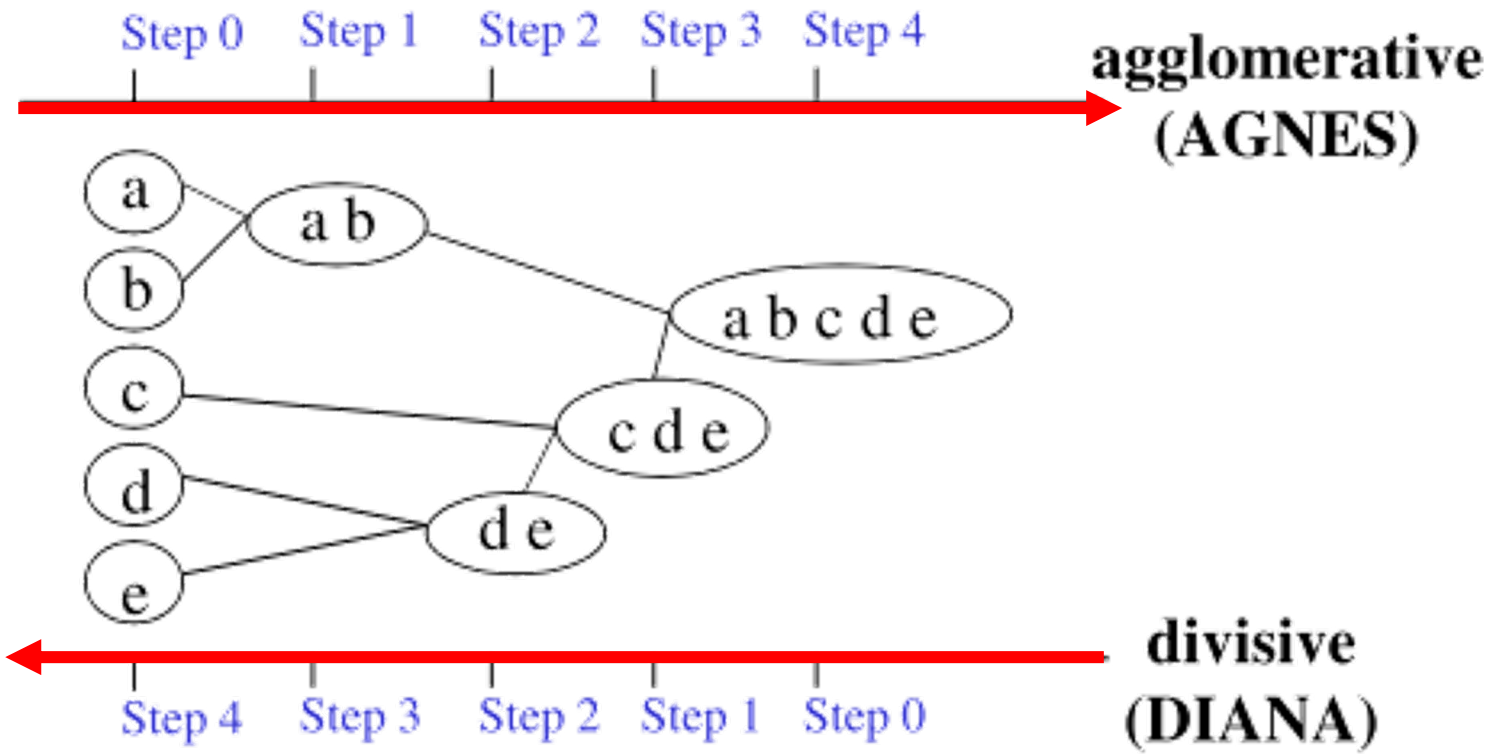
Hierarchical Clustering

- Bottom-up (Agglomerative) Algorithms:
  - Initially each item is in its own cluster.
  - Iteratively clusters are **merged** together.
  - Methods: AGNES (AGglomerative NESting).

- Top-down (Divisive) Algorithms:
  - Initially all items are in one cluster.
  - Iteratively **split** large clusters.
  - Methods: DIANA (DIvisive ANAlysis).

Hierarchical Clustering: Example

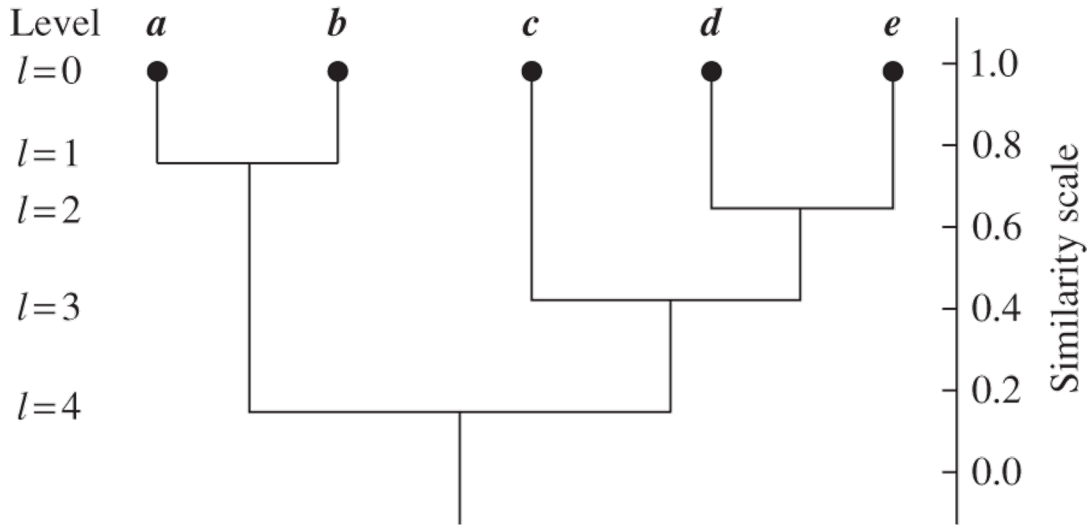
Dataset is {a, b, c, d, e}.



- A result of hierarchical clustering is represented by a dendrogram.

Dendrogram Structure

- Dendrogram shows how clusters are merged.
  - Decompose data objects into several levels of nested partitioning (tree of clusters), called a dendrogram.
  - A clustering of the data objects is obtained by cutting the dendrogram at the desired level, then each connected component forms a cluster.



Dendrogram representation for hierarchical clustering of data objects {*a*,*b*,*c*,*d*,*e*}.

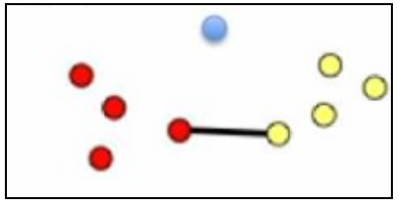
Distance Between Clusters

- **Single link:** smallest distance between an element in one cluster and an element in the other cluster, i.e., dist(Ki, Kj) = min(n_i, n_j) (Nearest neighbor clustering)
- **Complete link:** largest distance between an element in one cluster and an element in the other cluster, i.e., dist(Ki, Kj) = max(n_i, n_j) (Farthest neighbor clustering)
- **Average:** average distance between an element in one cluster and an element in the other cluster, i.e., dist(Ki, Kj) = avg(n_i, n_j)
- **Centroid:** distance between the centroids of two clusters,

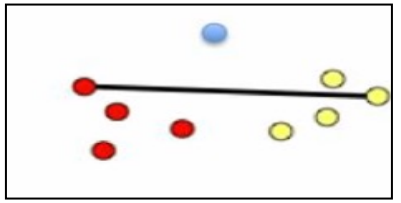
  i.e., dist(Ki, Kj) = dist(Ci, Cj)
- **Medoid:** distance between the medoids of two clusters,

  i.e., dist(Ki, Kj) = dist(Mi, Mj).

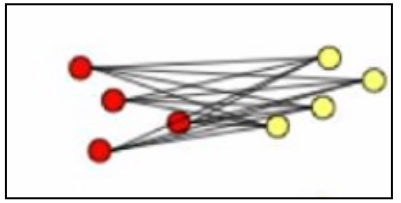
Medoid: a chosen, centrally located object in the cluster.



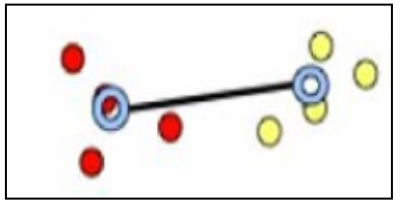
Single link



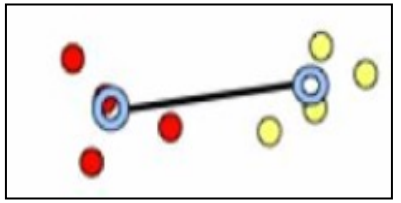
Complete link



Average

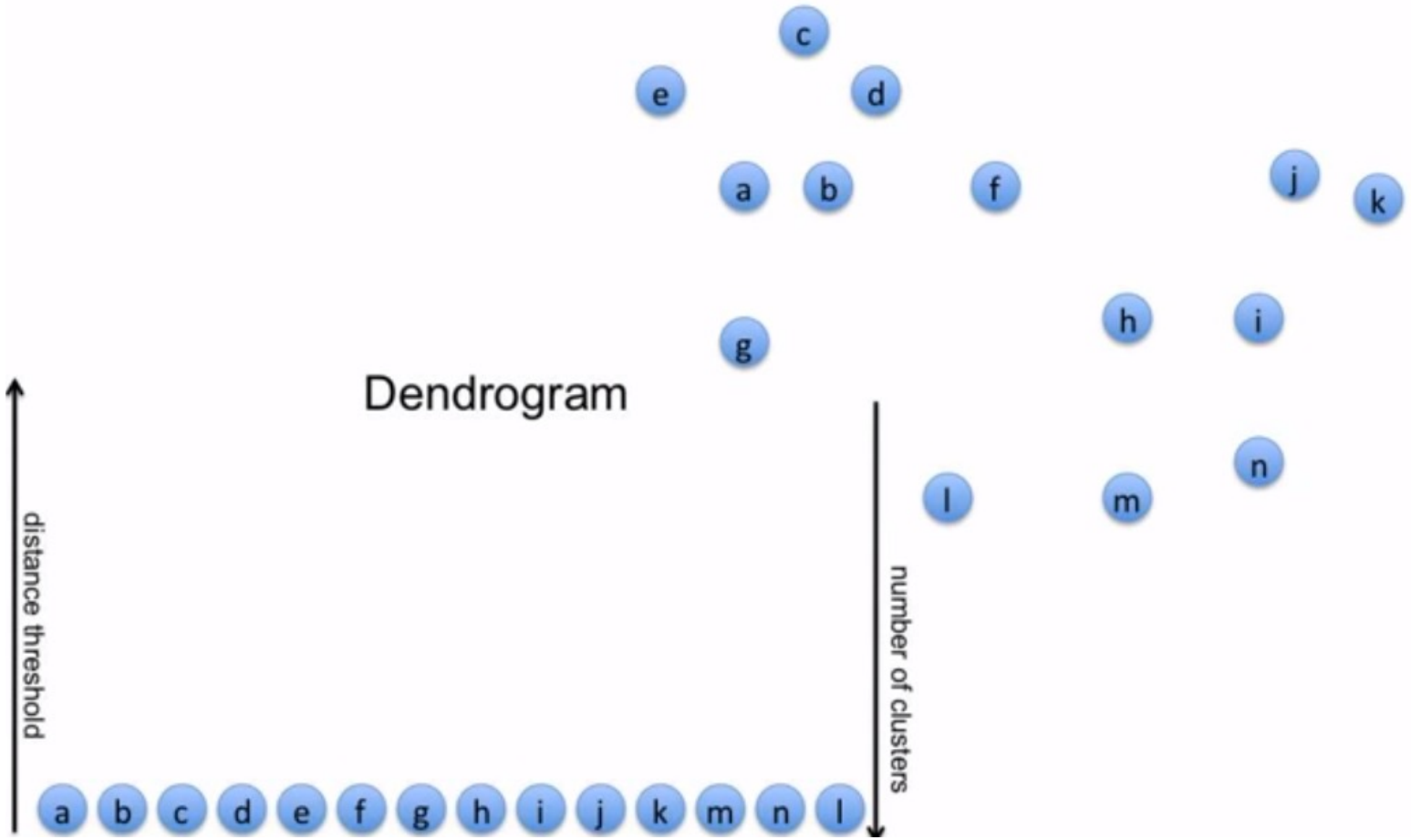


Centroid

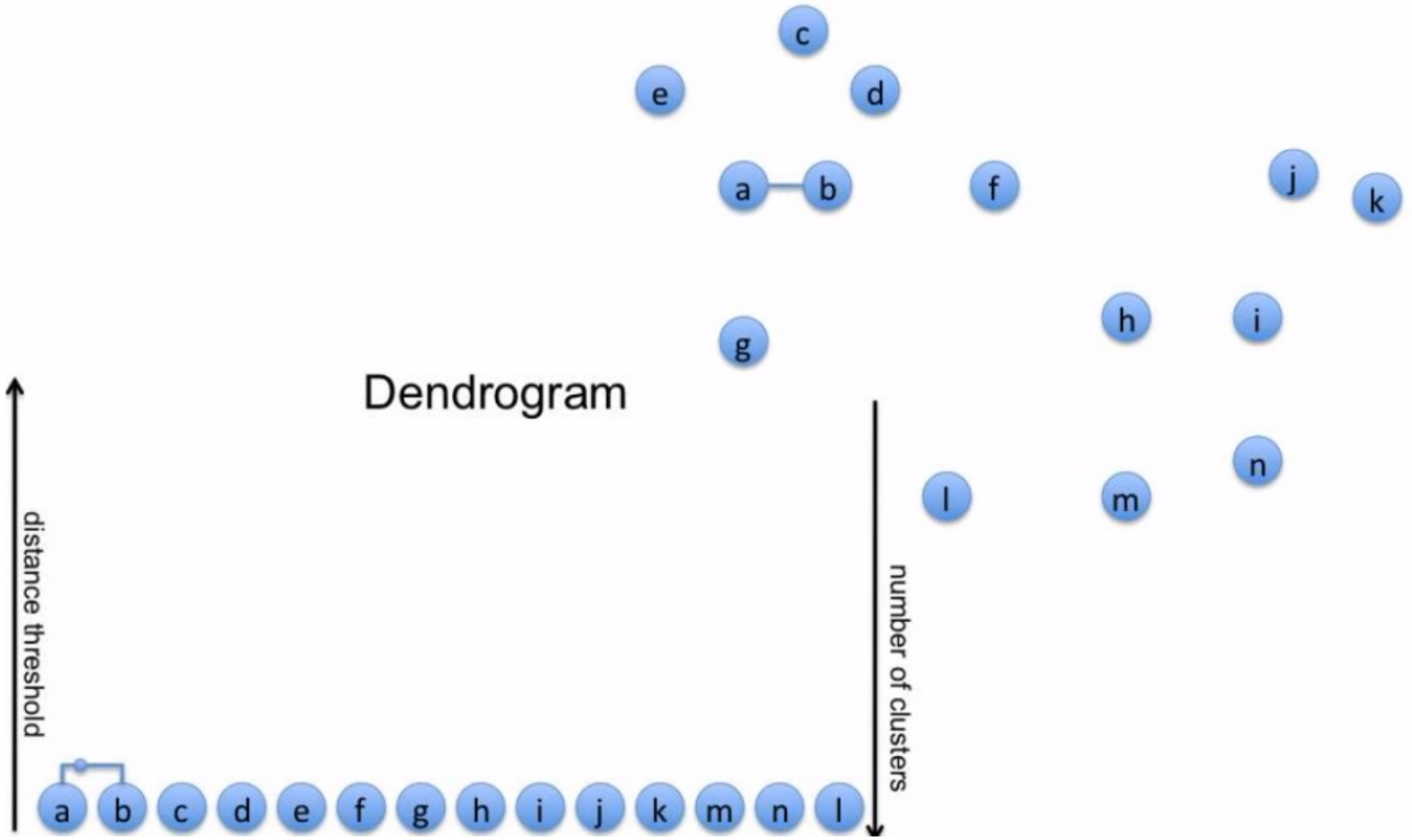


Medoid

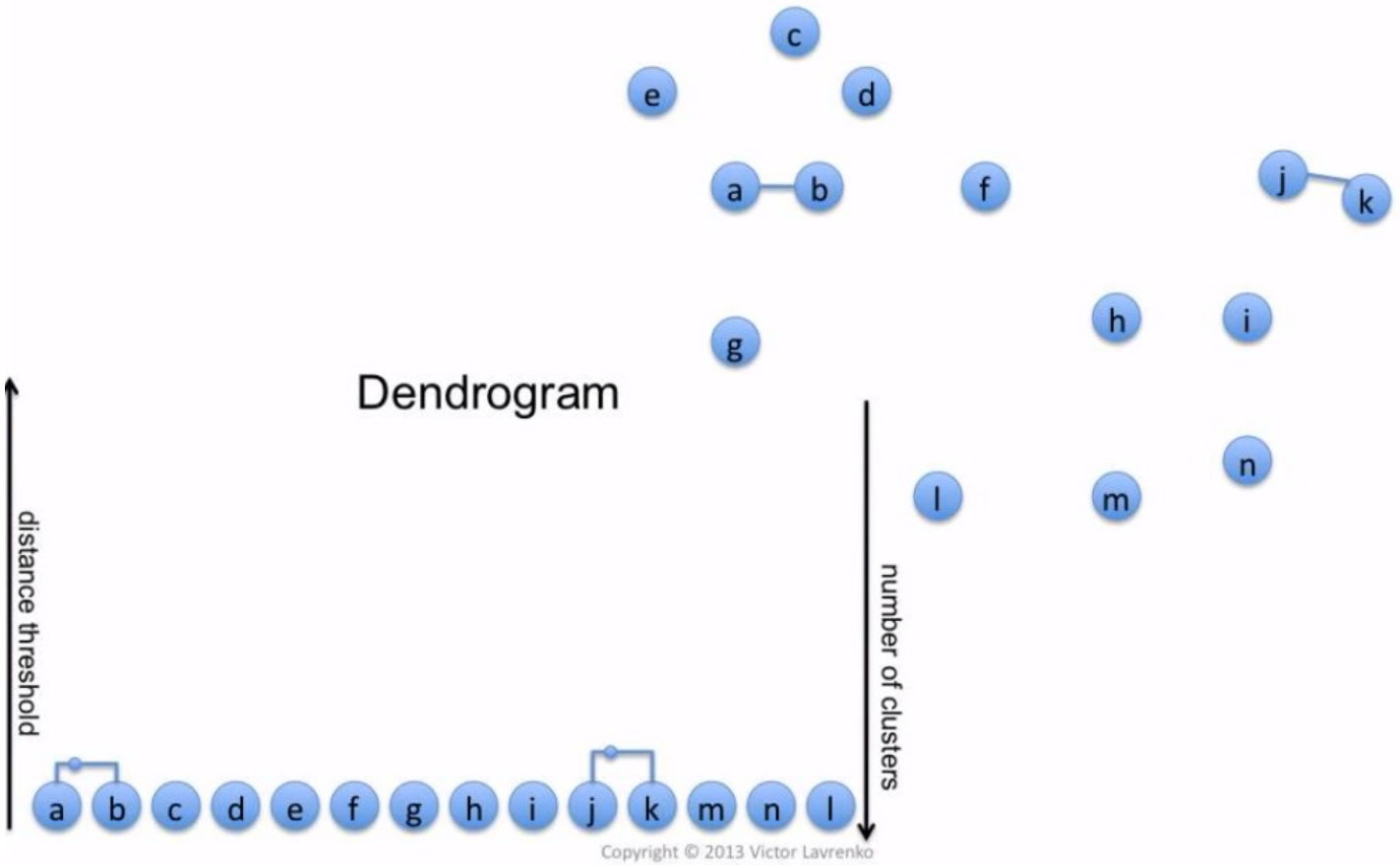
Example1 : Agglomerative Clustering -Dataset



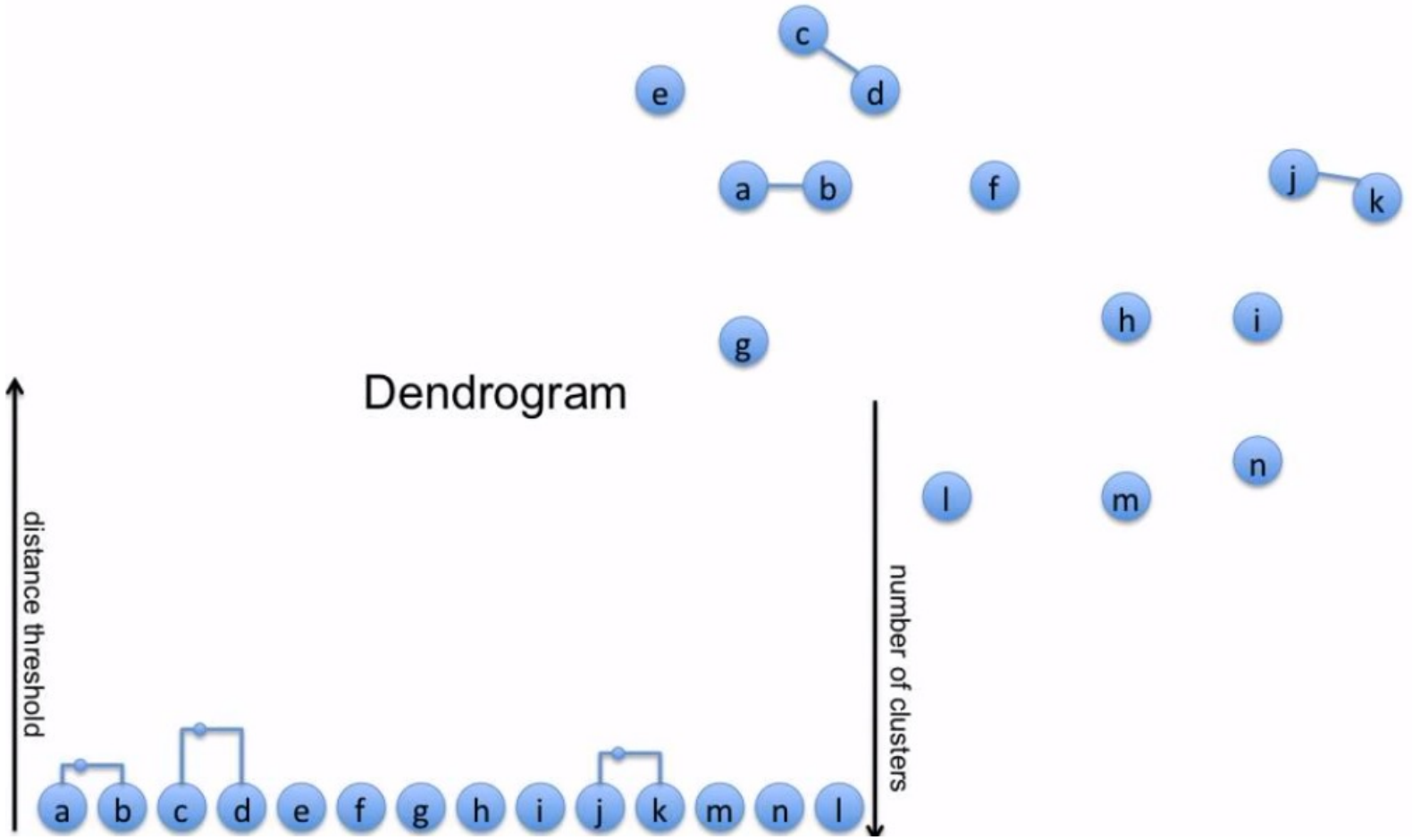
Example1 : Agglomerative Clustering - Step1



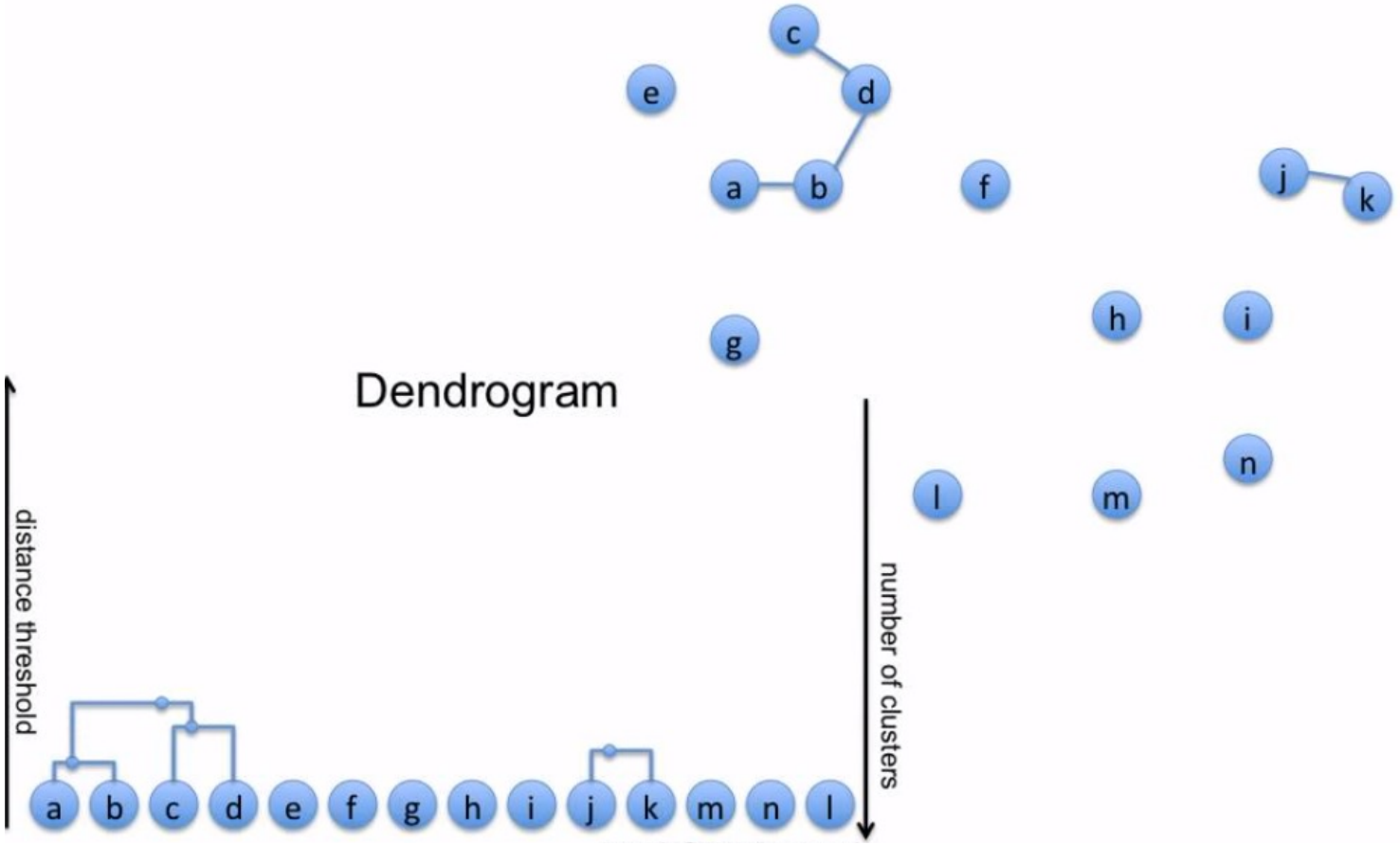
Example1 : Agglomerative Clustering - Step2



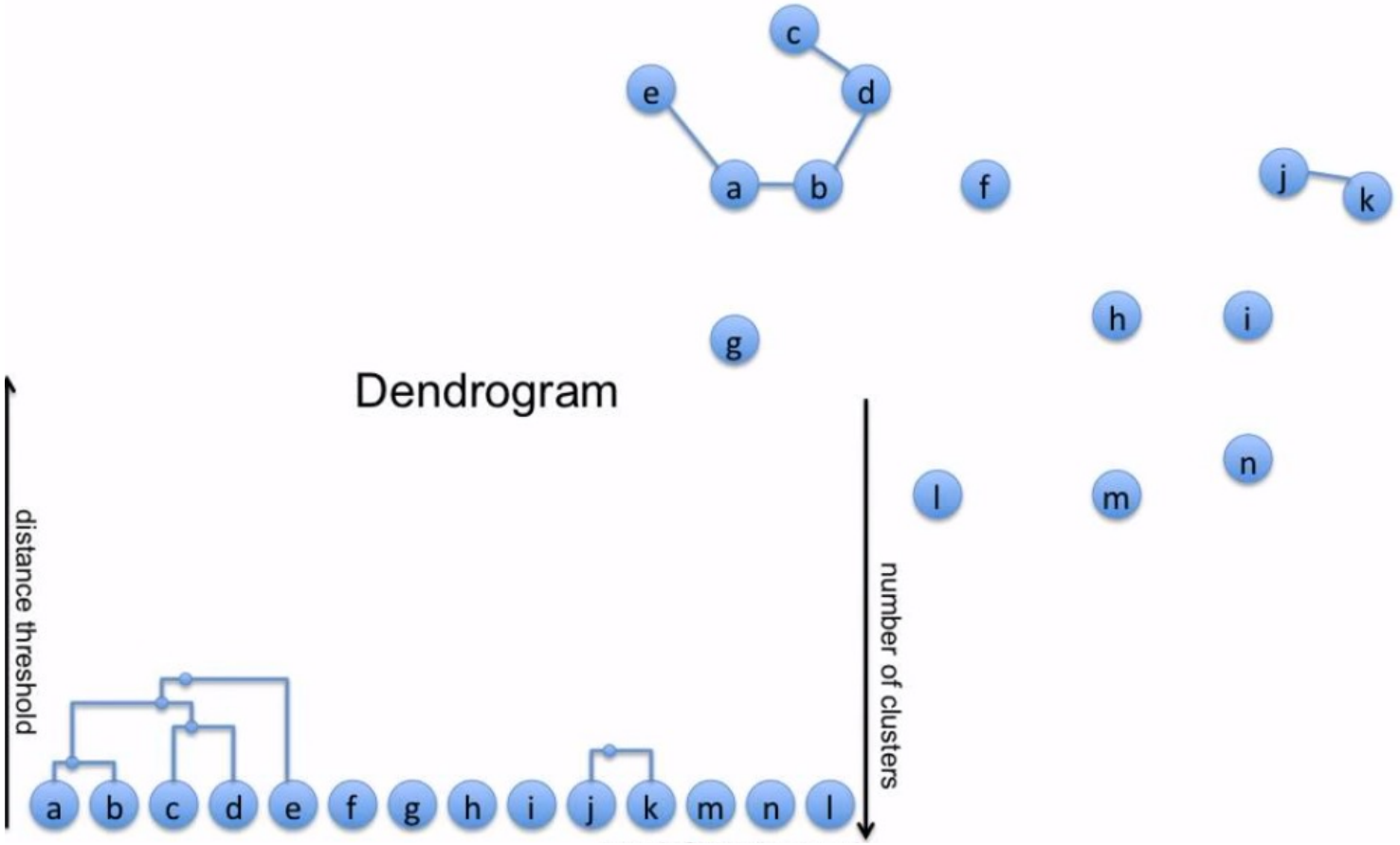
Example1 : Agglomerative Clustering - Step3



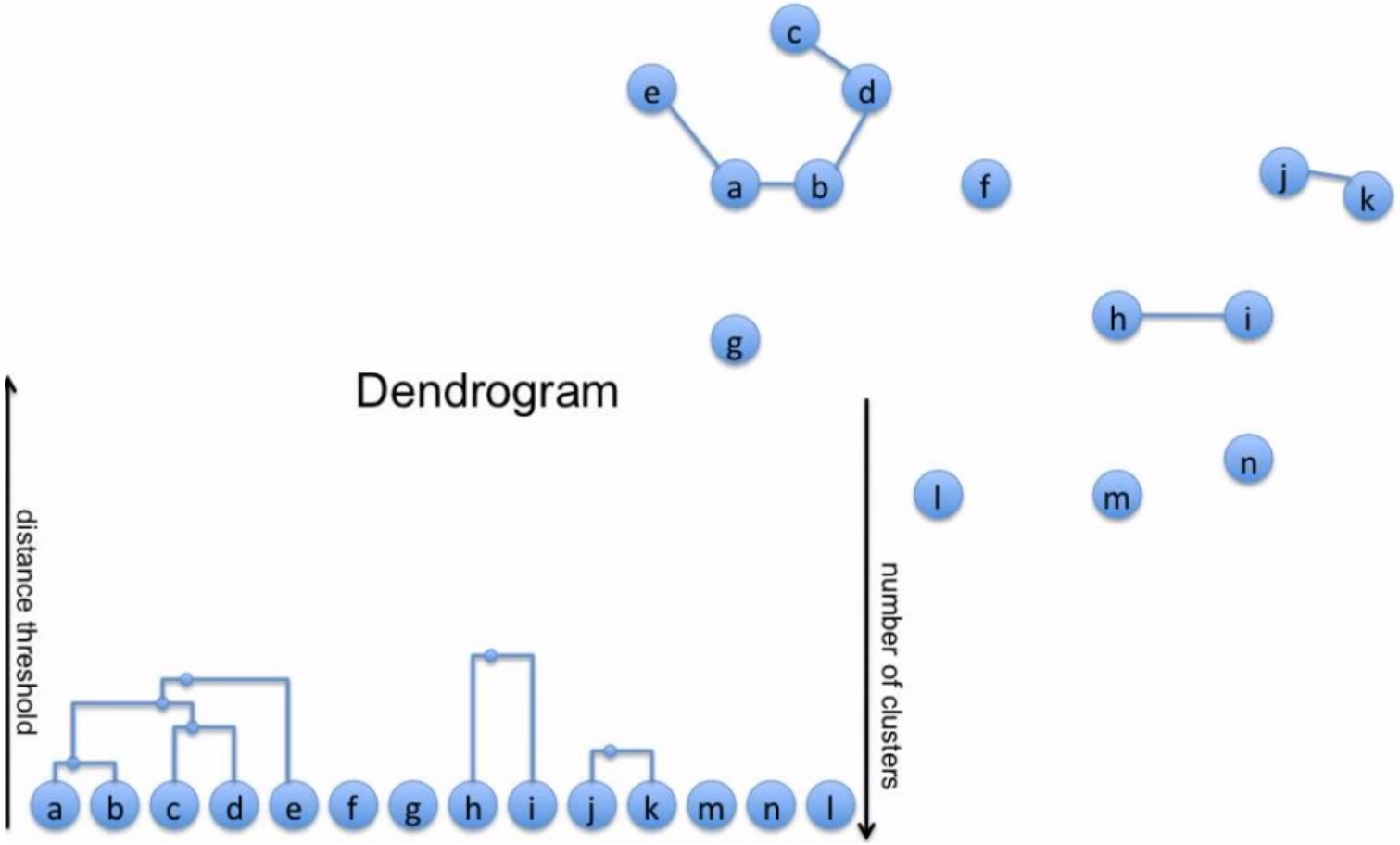
Example1 : Agglomerative Clustering - Step4



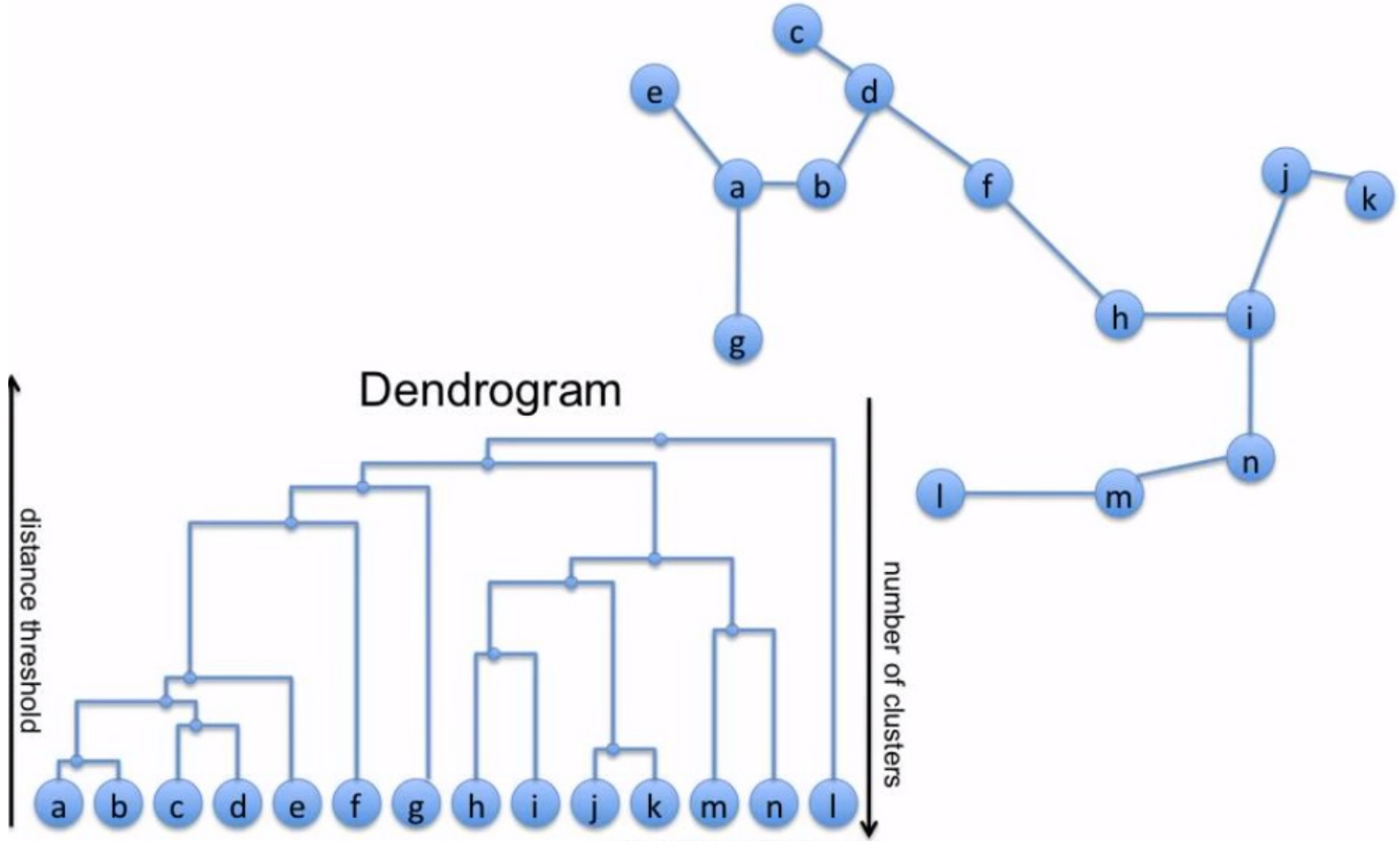
Example1 : Agglomerative Clustering - Step5



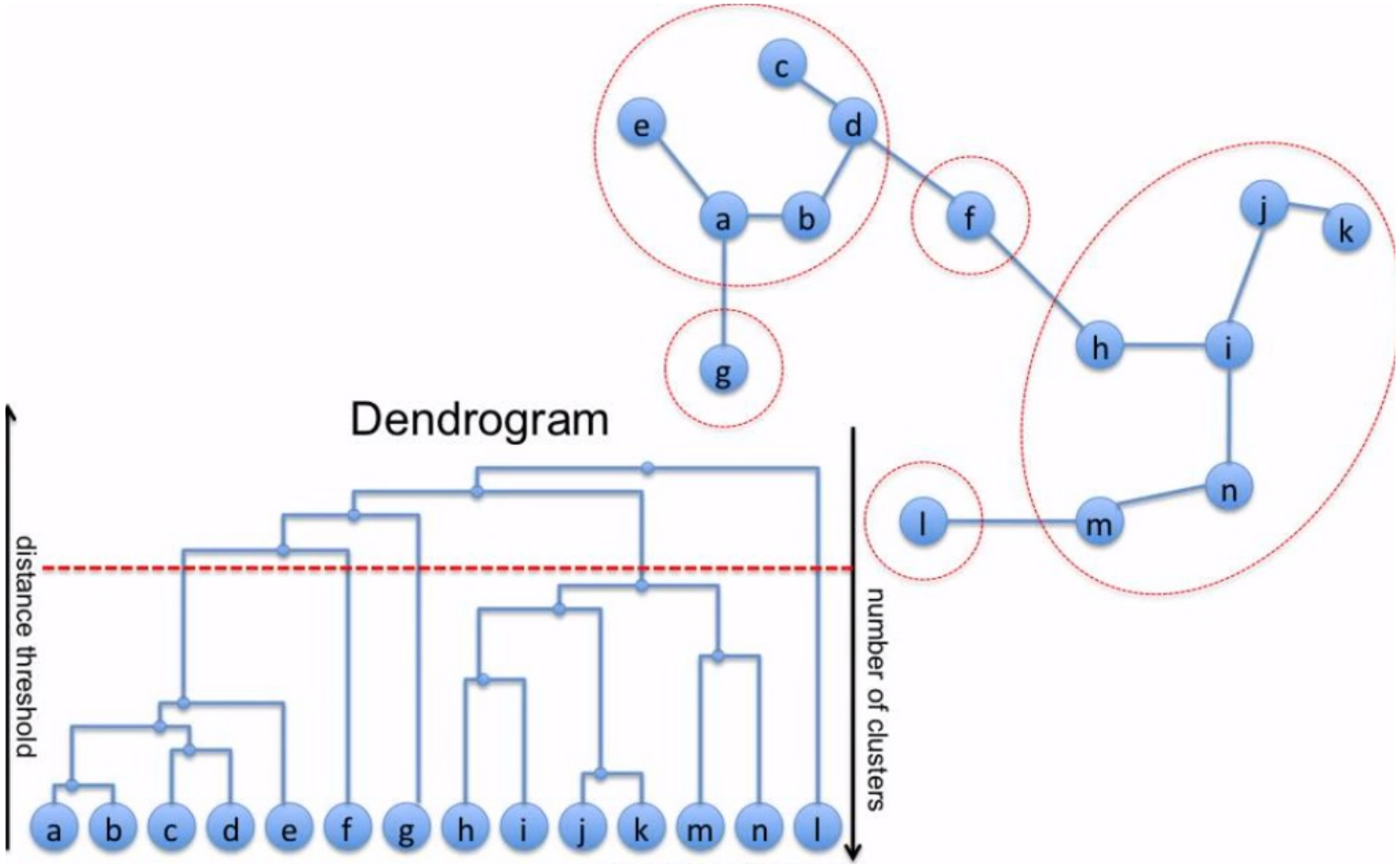
Example1 : Agglomerative Clustering - Step6



Example1 : Agglomerative Clustering - Final Step



Example1 : Agglomerative Clustering - Distance Cut

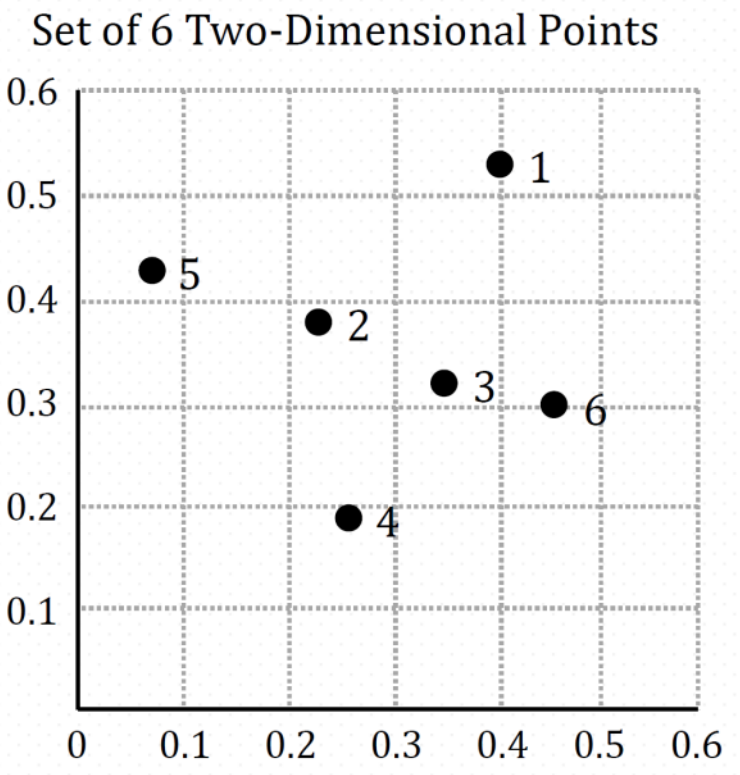


Flat clustering = threshold on distance to cut the tree

Example 2: Agglomerative clustering using single link

- Find the clusters using single link technique:
  - Use Euclidean distance.
  - Draw the dendrogram.

	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	0.3



Example2: Agglomerative clustering using single link cont.

- Step 1: calculate Euclidean distance and create the distance matrix.
- Euclidean distance $= \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2}$

	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	0.3

	P1	P2	P3	P4	P5	P6
P1	0					
P2	0.24	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

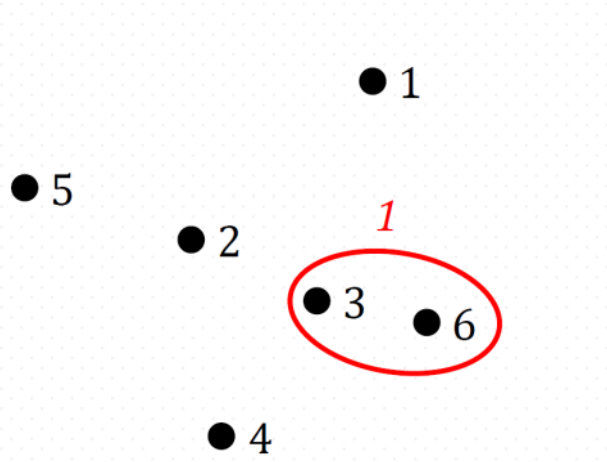
Distance(P1,P2)$= \sqrt{(0.40 - 0.22)^2 + (0.53 - 0.38)^2}$

$= \sqrt{(0.18)^2 + (0.15)^2} = 0.24$

Example2 : Agglomerative clustering using single link cont.

- Step 2: Find the smallest distance in this matrix and use the pair of elements to form a cluster.

	P1	P2	P3	P4	P5	P6
P1	0					
P2	0.24	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



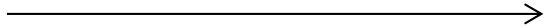
Points 3 and 6 have the smallest single link proximity distance.
Merge these points into one cluster and update the distances to this new cluster.

Example2 : Agglomerative clustering using single link cont.

- Step 3: Update the distance matrix.
  - Update the distance matrix after forming cluster(P3,P6).
  - When we calculate the distance between the newly created cluster(P3,P6) and an element (Px) we take the smallest distance between d(P3,Px) and d(P6,Px).

	P1	P2	P3	P4	P5	P6
P1	0					
P2	0.24	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

Distance between the new cluster (P3.P6) and P1
=Min[dist(P3,P1),dist(P6,P1)]
=Min [0.22,0.23]=0.22



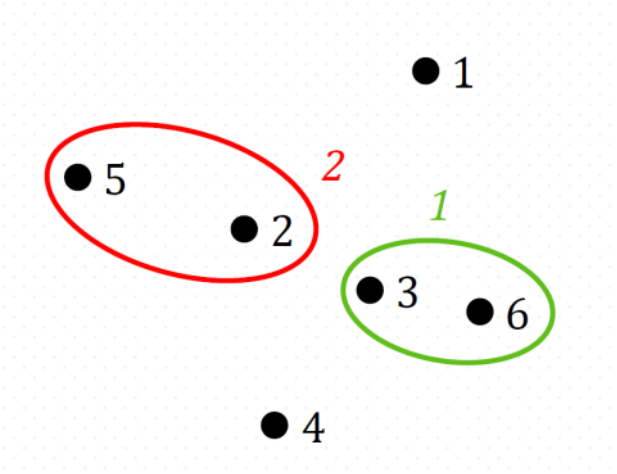
	P1	P2	P3,P6	P4	P5
P1	0				
P2	0.24	0			
P3,P6	0.22	0.15	0		
P4	0.37	0.20	0.15	0	
P5	0.34	0.14	0.28	0.29	0

As seen in the matrix, the only distances updates are those in the cells for the merged elements (P3 and P6)

Example2 : Agglomerative clustering using single link cont.

- Repeat steps 2 and 3 on the updated distance matrix. Iterate until all elements are grouped all together.

	P1	P2	P3,6	P4	P5
P1	0				
P2	0.24	0			
P3,6	0.22	0.15	0		
P4	0.37	0.20	0.15	0	
P5	0.34	0.14	0.28	0.29	0



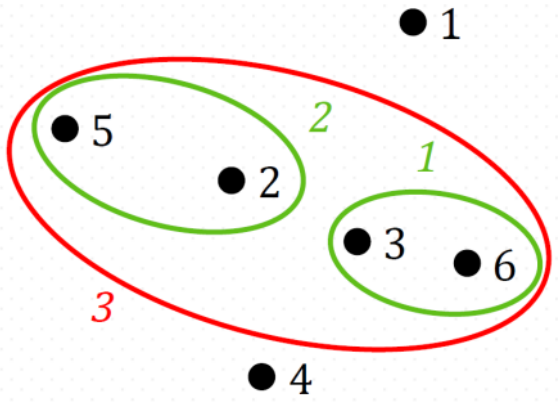
	P1	P2,5	P3,6	P4
P1	0			
P2,P5	0.24	0		
P3,P6	0.22	0.15	0	
P4	0.37	0.20	0.15	0

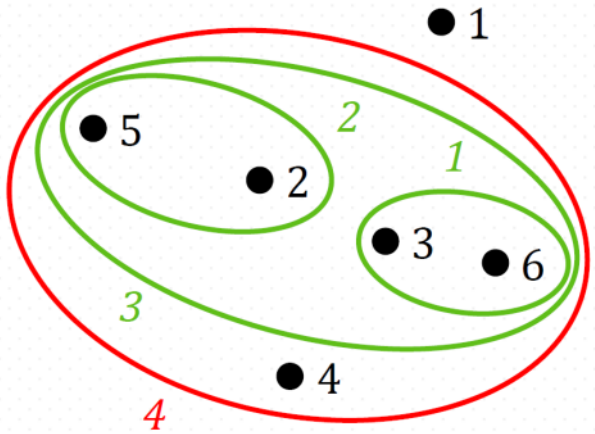
Example2 : Agglomerative clustering using single link cont.

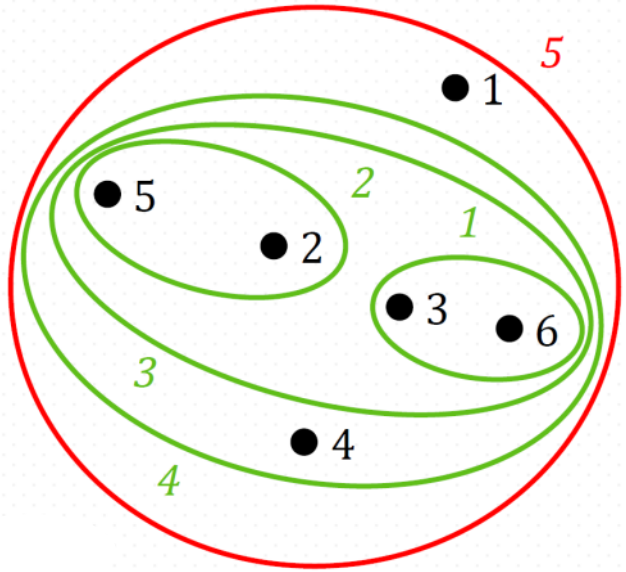
	P1	P2,5	P3,6	P4
P1	0			
P2,5	0.24	0		
P3,6	0.22	0.15	0	
P4	0.37	0.20	0.15	0

	P1	P2,5 ,3,6	P4
P1	0		
P2,5,3,6	0.22	0	
P4	0.37	0.15	0

	P1	P2,5 ,3,6.4
P1	0	
P2,5,3,6.4	0.22	0

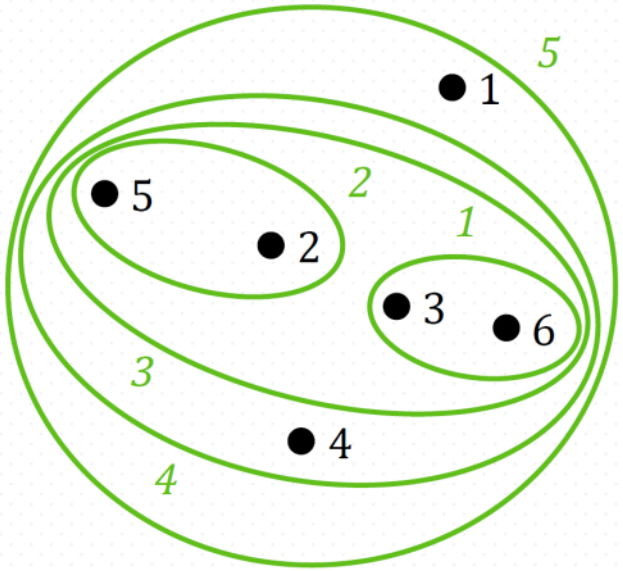




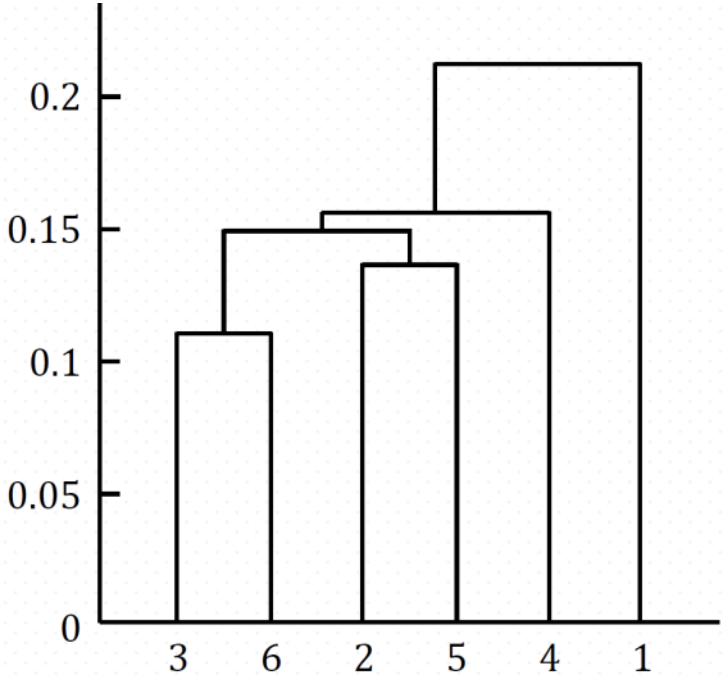


Example2: Agglomerative clustering using single link cont.

- The final result is:



Dendrogram



- We can apply a threshold on distance to cut the tree.

Activity#2: Agglomerative clustering using complete and average link. Self study

- Consider Example 2, apply agglomerative clustering:
  - Using complete link instead of single link.
    - Complete link is similar to single link BUT when computing distances between clusters we take the max distance of points in the clusters instead of min.
  - Using average link instead of single link.
    - Average link is similar to single link BUT when computing distances between clusters we take the average of distances between points in the clusters instead of min.

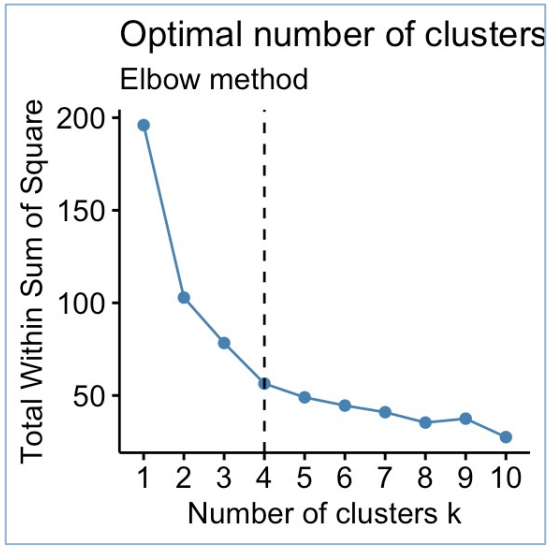
Evaluation of Clustering

- The major tasks of clustering evaluation include the following:
  - Assessing clustering tendency.
    - Assess whether a nonrandom structure exists in the data.
    - Clustering analysis on a data set is meaningful only when there is a nonrandom structure in the data.
  - Determining the number of clusters in a data set.
  - Measuring clustering quality.
    - assess how good the resulting clusters are?

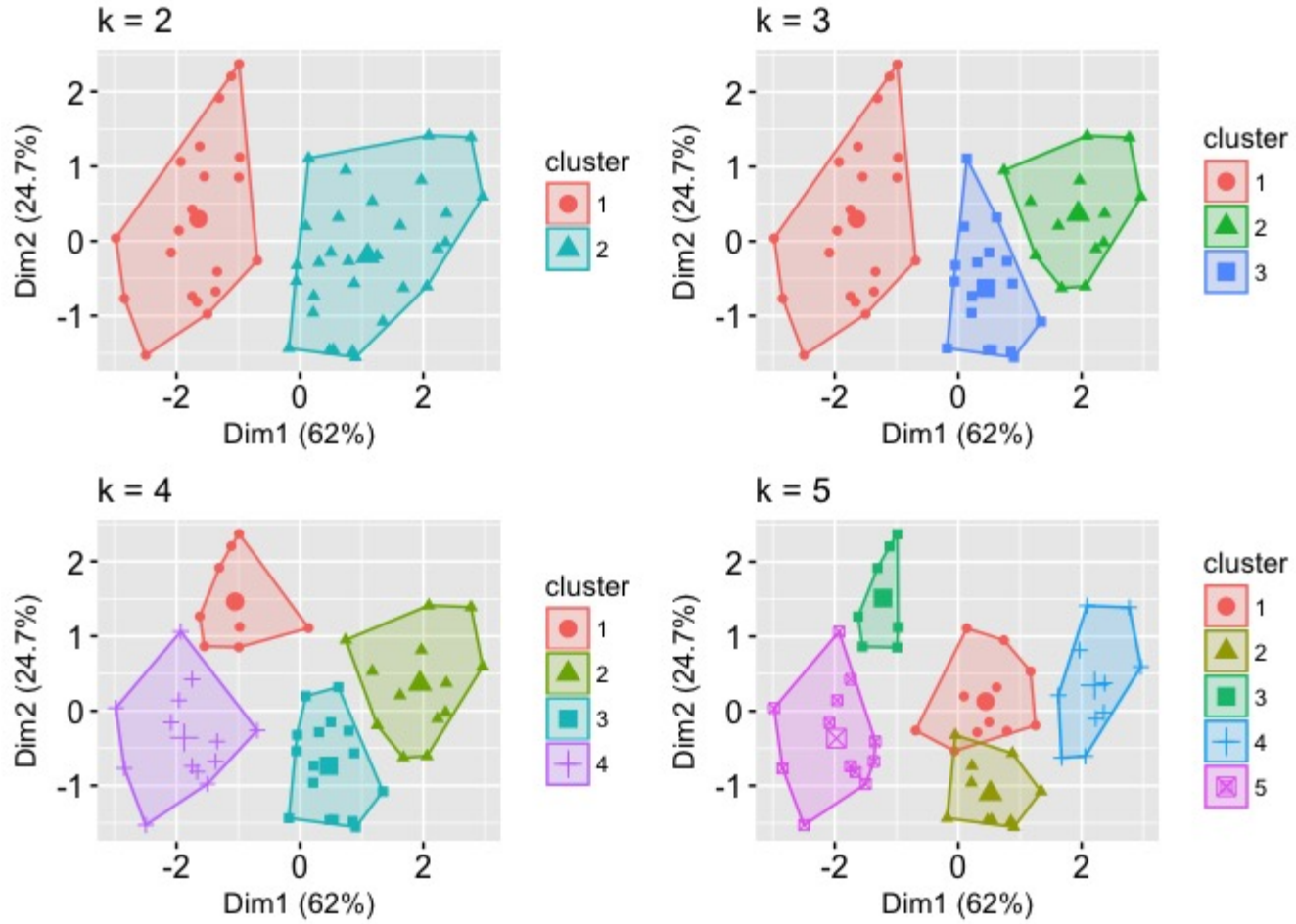
Determine the Number of Clusters

- Determining the "right" number of clusters in a data set is important:
  - some clustering algorithms, like *k*-means, require such a parameter.
  - appropriate number of clusters controls the proper granularity of cluster analysis.
- It depends on the dataset, as well as, the clustering resolution required by the user.
- Empirical method:
  - # of clusters ≈ $\sqrt{\frac{n}{2}}$ for a dataset of n points.
- Elbow method:
  - Use the turning point in the curve to determine the # of clusters.

Plot the curve of **total within-cluster sum of square (WSS)** with respect to **K: # of clusters**



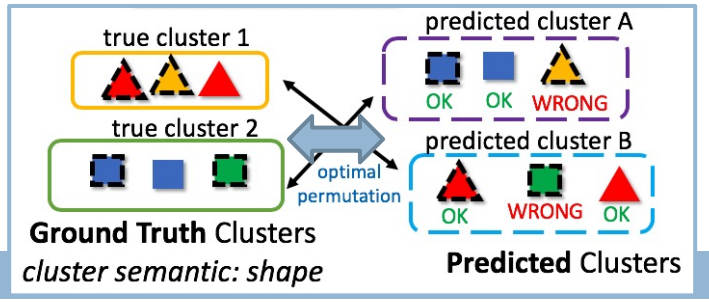
Determine the Number of Clusters cont.



Measuring Clustering Quality

- "*How good is the clustering generated by a method, and how can we compare the clusterings generated by different methods?*"
- Two methods: extrinsic vs. intrinsic
- Extrinsic: supervised (the ground truth is available)
  - Compare a clustering against the ground truth using certain clustering quality measure.
  - BCubed precision and recall metrics.
- Intrinsic: unsupervised (the ground truth is unavailable)
  - Evaluate the goodness of a clustering by considering how well the clusters are separated, and how compact the clusters are.
  - Silhouette coefficient.

Extrinsic Methods



- When the ground truth is available, we can compare it with a clustering to assess the clustering.
- Clustering quality measure: Q(C, Cg), for a clustering C given the ground truth Cg.
- Q is good if it satisfies the following 4 essential criteria:
  - Cluster homogeneity: the purer, the better.
  - Cluster completeness: should assign objects belong to the same category in the ground truth to the same cluster.
  - Rag bag: putting a heterogeneous object into a pure cluster should be penalized more than putting it into a rag bag (i.e., "miscellaneous" or "other" category).
  - Small cluster preservation: splitting a small category into pieces is more harmful than splitting a large category into pieces.

Clustering Quality: 4 Criteria

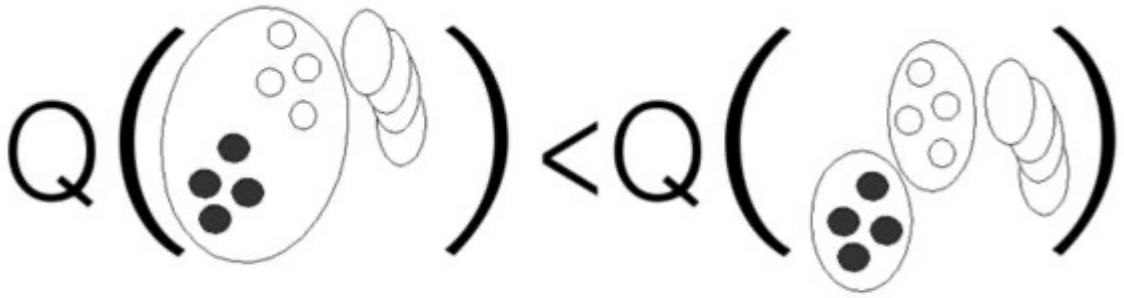


Figure 1: Constraint 1: Cluster Homogeneity

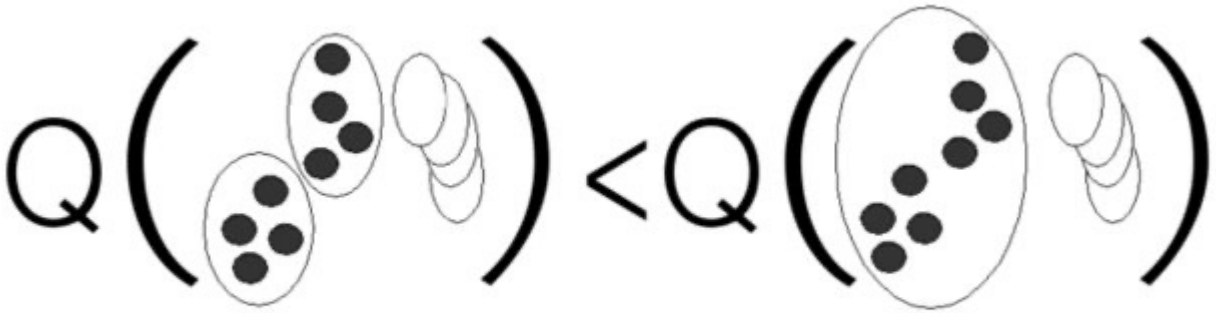


Figure 2: Constraint 2: cluster completeness

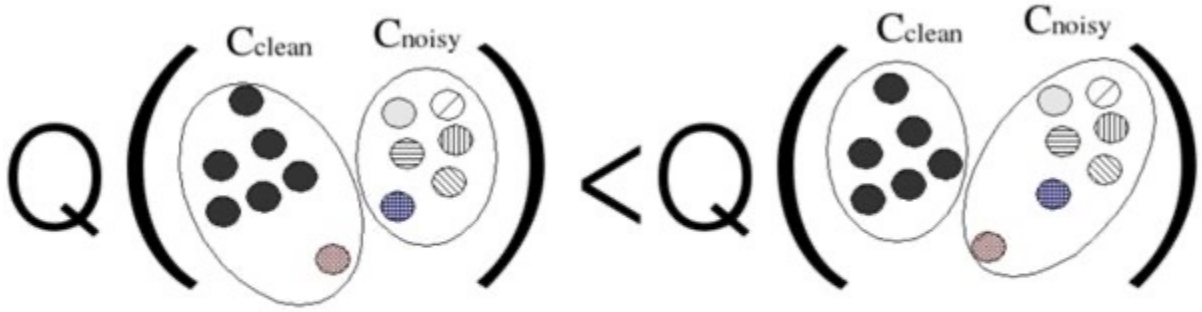


Figure 3: Constraint 3: Rag Bag

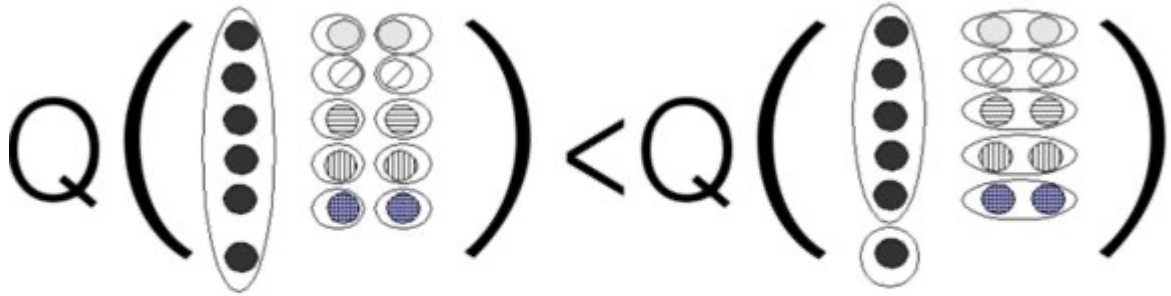


Figure 4: Clusters Size vs. Quantity

Extrinsic Methods

- The BCubed precision and recall metrics satisfy all four criteria.
- BCubed evaluates the precision and recall for every object in a clustering on a given data set according to ground truth.
  - The precision of an object indicates how many other objects in the same cluster belong to the same category as the object.

$$\text{Precision} = \frac{\#\text{ of items from same category in its cluster}}{\#\text{ items in its cluster}}$$

  - The recall of an object reflects how many objects of the same category are assigned to the same cluster.

$$\text{Recall} = \frac{\#\text{ of items from same category in its cluster}}{\#\text{ items in its category}}$$

Example: Precision and Recall

For each object o_i, we calculate precision and recall as follow:

$$\text{Precision}(o_i) = \frac{\#\text{ of items from same category in its cluster}}{\#\text{ items in its cluster}}$$

$$\text{Recall}(o_i) = \frac{\#\text{ of items from same category in its cluster}}{\#\text{ items in its category}}$$

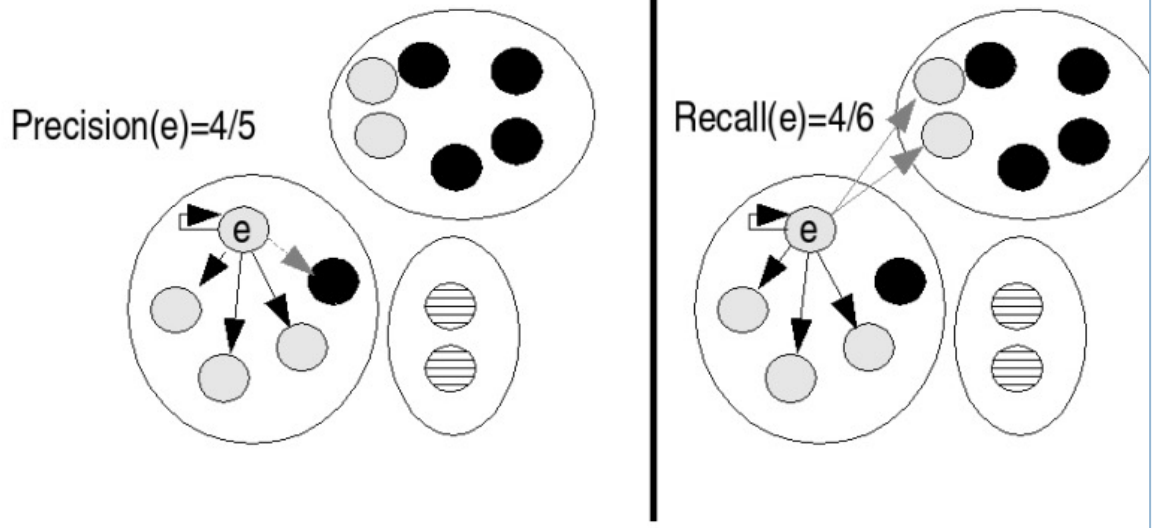
To measure the quality of a clustering:

the average BCubed precision and recall of all objects in the data set is computed, as follow:

$$\text{BCubed Precision} = \frac{1}{n}\sum_{i=1}^{n}\text{Precision}(o_i)$$

$$\text{BCubed Recall} = \frac{1}{n}\sum_{i=1}^{n}\text{Recall}(o_i)$$

Example:



Intrinsic Methods

- When the ground truth of a data set is not available, an intrinsic method is used to assess the clustering quality.
- In general, intrinsic methods evaluate a clustering by examining how well the clusters are separated and how compact the clusters are.
- Many intrinsic methods have the advantage of a similarity metric between objects in the data set.
  - The silhouette coefficient is such a measure.

Silhouette Coefficient

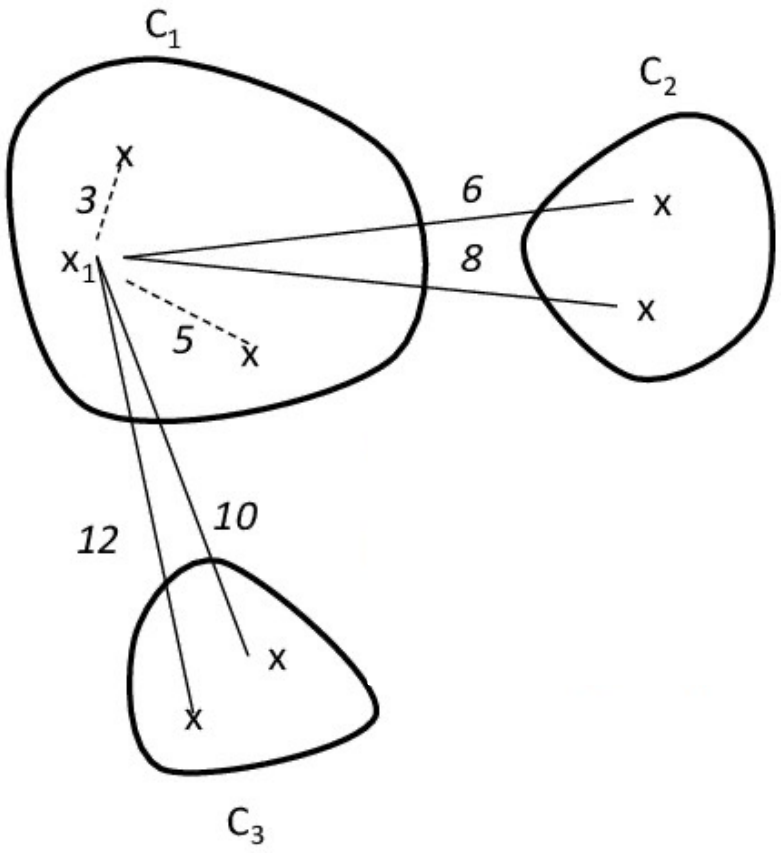
- The value of the silhouette coefficient is between -1 and 1.
- The silhouette coefficient of an object o is defined as:

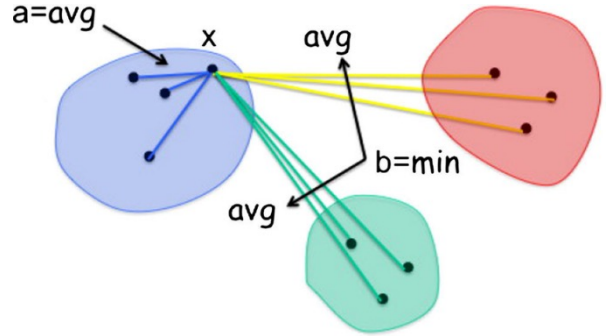
$$s(o) = \frac{b(o) - a(o)}{\max\{a(o), b(o)\}}.$$

- The value of $a(o)$ reflects the compactness of the cluster to which o belongs.
  - The smaller the value, the more compact the cluster.
- The value of $b(o)$ captures the degree to which o is separated from other clusters.
  - The larger $b(o)$ is, the more separated o is from other clusters.

Example: Silhouette Coefficient

$$s(o) = \frac{b(o) - a(o)}{\max\{a(o), b(o)\}}.$$





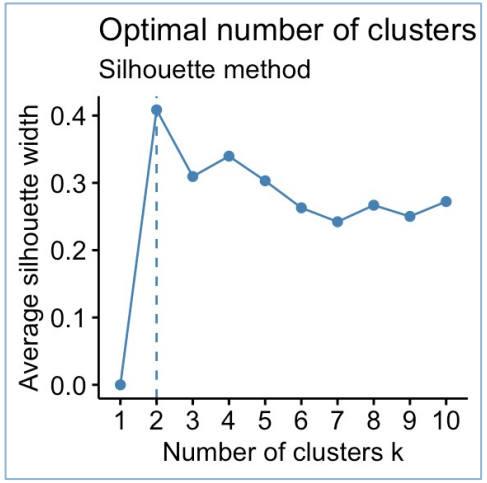
- $a\,(x_1) = \frac{3+5}{2} = 4$
- $b\,(x_1) = min\,(\frac{6+8}{2}, \frac{10+12}{2}) = 7$
- $s\,(x_1) = \frac{7-4}{7} = \frac{3}{7}$

Silhouette Coefficient

- When the silhouette coefficient value of o approaches 1:
  - the cluster containing o is compact and o is far away from other clusters, which is the preferable case.
- When the silhouette coefficient value is negative (i.e., $b(o) < a(o)$):
  - o is closer to the objects in another cluster than to the objects in the same cluster as o , which is a bad situation and should be avoided.
- To measure a cluster's fitness within a clustering:
  - the average silhouette coefficient value of all objects in the cluster is computed.
- To measure the quality of a clustering:
  - the average silhouette coefficient value of all objects in the data set is computed.
  - can be used to determine the optimal number of clusters *k*.

Plot the curve of **the average silhouette coefficient value** with respect to **K: # of clusters**

Optimal k : the one that maximizes the average silhouette.



Summary

- Cluster analysis groups objects based on their similarity and has wide applications.
- Clustering algorithms can be categorized into partitioning methods, and hierarchical methods.
- K-means and K-medoids algorithms are popular partitioning-based clustering algorithms.
- Agglomerative and Divisive algorithms are popular hierarchical-based clustering algorithms.
- Quality of clustering results can be evaluated in various ways.