

DMDW – Module-5

By

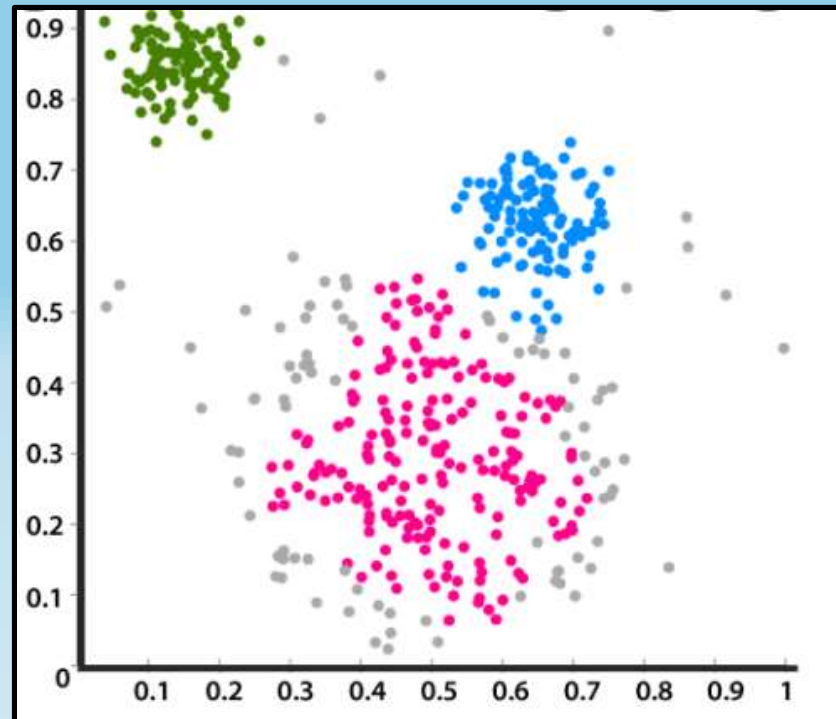
Dr. Pulak Sahoo

Associate Professor
Dept of CSE, SIT, BBSR

Module-5 Syllabus

Clustering: Overview, K-Means, K-Medoid, Agglomerative hierarchical clustering, DBSCAN, Cluster evaluation, Density-based clustering, Graph-based clustering, Scalable clustering algorithms.

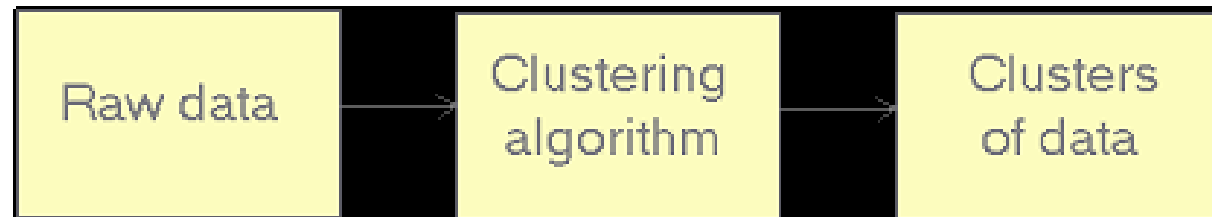
Cluster Analysis:



Basic Concepts & Methods

Cluster Analysis

- ▶ **Clustering** is the process of grouping a set of data objects into multiple groups or clusters so that objects within a cluster have high similarity, but are very dissimilar to objects in other clusters
- ▶ **Cluster analysis** is the process of partitioning a data set into subsets (cluster), such that objects in a cluster are similar to one another, yet dissimilar to objects in other clusters. The set of clusters resulting from a cluster analysis can be referred to as a clustering.



What is Clustering?

- ▶ **Clustering** is **unsupervised learning**
- ▶ **Clustering** is a method of grouping data that share similar patterns
- ▶ **Clustering** is a method by which a large data set is grouped into clusters of smaller sets of similar data

- **Example:**



After
clustering:



The usage of clustering

- ▶ **Cluster analysis is widely used in many applications :**
- ▶ **Ex:** Business intelligence, Web search, biology & security
- ▶ In **pattern recognition**, the concepts of cluster analysis are used
- ▶ **Ex:** handwritten characters, samples of speech, fingerprints & pictures
- ▶ In the **life sciences** (**Ex:** biology, botany, zoology, microbiology) the objects of analysis are **plants, animals & insects**. The clustering analysis may lead classification of the species into subspecies
- ▶ Widely used in **info, policy & decision sciences** including **surveys** on political issues, market analysis, survey of products, survey of sales programs.




Requirements for Cluster Analysis

- ▶ **Scalability**
- ▶ **Ability to deal with different types of attributes**
- ▶ **Discovery of clusters**
- ▶ **Requirements for domain knowledge**
- ▶ **Ability to deal with noisy data**
- ▶ **Incremental clustering & insensitivity to input order**
- ▶ **Capability of clustering high-dimensional data**
- ▶ **Constraint-based clustering**
- ▶ **Interpretability & usability**

Overview of Basic Clustering Methods

The major clustering methods can be classified into the following categories:

1. **Partitioning Methods**
2. **Hierarchical Method**
3. **Density-based Methods**
4. **Grid-based Methods** 

1. Partitioning Methods

- ▶ Given a set of ***n objects***, a *partitioning method* constructs ***k partitions of the data***, where each partition **represents a cluster & $k \leq n$**
- ▶ ***Each group*** must contain at ***least one object***
- ▶ Most partitioning methods are **distance-based**.
- ▶ Given the **no. of partitions k**, the partitioning method creates an **initial partitioning**.
- ▶ It then uses an **iterative relocation technique** that improves the partitioning by moving objects from one group to another
- ▶ **The criterion of a good partitioning** is that objects in the same cluster are “**close**” to each other, whereas objects in different clusters are “**far apart**”

2. Hierarchical Methods

- ▶ A **hierarchical method** creates a **hierarchical decomposition** of the given set of data objects
- ▶ It can be classified as **agglomerative** or **divisive**
- ▶ The **agglomerative approach** or the **bottom-up approach**, starts with each object forming a separate group.
- ▶ It **successively merges** the objects close to one another, until all the groups are merged into one or a termination condition holds
- ▶ The **divisive approach** or the **top-down approach**, starts with all the objects in the same cluster. In each successive iteration, a cluster is split into smaller clusters, until each object is in one cluster, or a termination condition holds
- ▶ Hierarchical clustering methods can be **distance-based** or **density-and continuity based**

3. Density-based Methods

- ▶ The idea is to **continue growing** a given cluster as long as the density of data points in the neighborhood **exceeds some threshold**
- ▶ **Ex:** for each data point within a given cluster, the **neighborhood of a given radius** has to contain at least a minimum no. of points
- ▶ Such a method can be used to **filter out noise or outliers** & **discover clusters of arbitrary shape**



Grid-based Methods

- ▶ **Grid-based methods** quantize the object space into a finite number of cells that form a grid structure
- ▶ All the clustering operations are performed on the **grid structure** (on the quantized space)
- ▶ The main advantage of this approach is its **fast processing time**, which is **independent** of the **no. of data objects** & **dependent** only on the no. of cells in each dimension in the quantized space
- ▶ Using grids is often an efficient approach to many spatial data mining problems, including clustering. Therefore, grid-based methods can be integrated with other clustering methods such as density-based methods & hierarchical methods



Overview of Clustering Methods

Method	General Characteristics
Partitioning methods	<ul style="list-style-type: none">– Find mutually exclusive clusters of spherical shape– Distance-based– May use mean or medoid (etc.) to represent cluster center– Effective for small- to medium-size data sets
Hierarchical methods	<ul style="list-style-type: none">– Clustering is a hierarchical decomposition (i.e., multiple levels)– Cannot correct erroneous merges or splits– May incorporate other techniques like microclustering or consider object “linkages”
Density-based methods	<ul style="list-style-type: none">– Can find arbitrarily shaped clusters– Clusters are dense regions of objects in space that are separated by low-density regions– Cluster density: Each point must have a minimum number of points within its “neighborhood”– May filter out outliers
Grid-based methods	<ul style="list-style-type: none">– Use a multiresolution grid data structure– Fast processing time (typically independent of the number of data objects, yet dependent on grid size)

1. Partitioning Methods

- ▶ The **simplest** & most fundamental version of **cluster analysis is partitioning**, which organizes the objects of a set into several exclusive groups
- ▶ Given a data set, D , of n objects & k no. of clusters to form, a **partitioning algorithm organizes the objects into k partitions ($k \leq n$)**, where each *partition* represents a **cluster**

k-Means: A Centroid-Based Technique

- ▶ The ***k-means algorithm*** defines the ***centroid*** of a cluster as the **mean value** of the points within the cluster
- ▶ **It proceeds as follows:**
 - 1) **Randomly select k of the objects**, *each of which initially represents a cluster mean or center*
 - 2) For each of the remaining objects, an object is assigned to a cluster, based on the **Euclidean distance** between the object & the cluster mean
 - 3) The ***k-means algorithm iteratively improves the clusters***
 - 4) For each cluster, it computes the **new mean** using the objects assigned to the cluster in the previous iteration.
 - 5) All the objects are then **reassigned using the updated means** as the new cluster centers
 - 6) The **iterations continue until the assignment is stable**

k-Means: A Centroid-Based Technique

- ▶ **Algorithm: *k-means*.** The k-means algorithm for partitioning, where each cluster's center is represented by the mean value of the objects in the cluster
- ▶ **Input:**
k: the number of clusters,
D: a data set containing n objects.
- ▶ **Output:** A set of k clusters

K-Means Algorithm

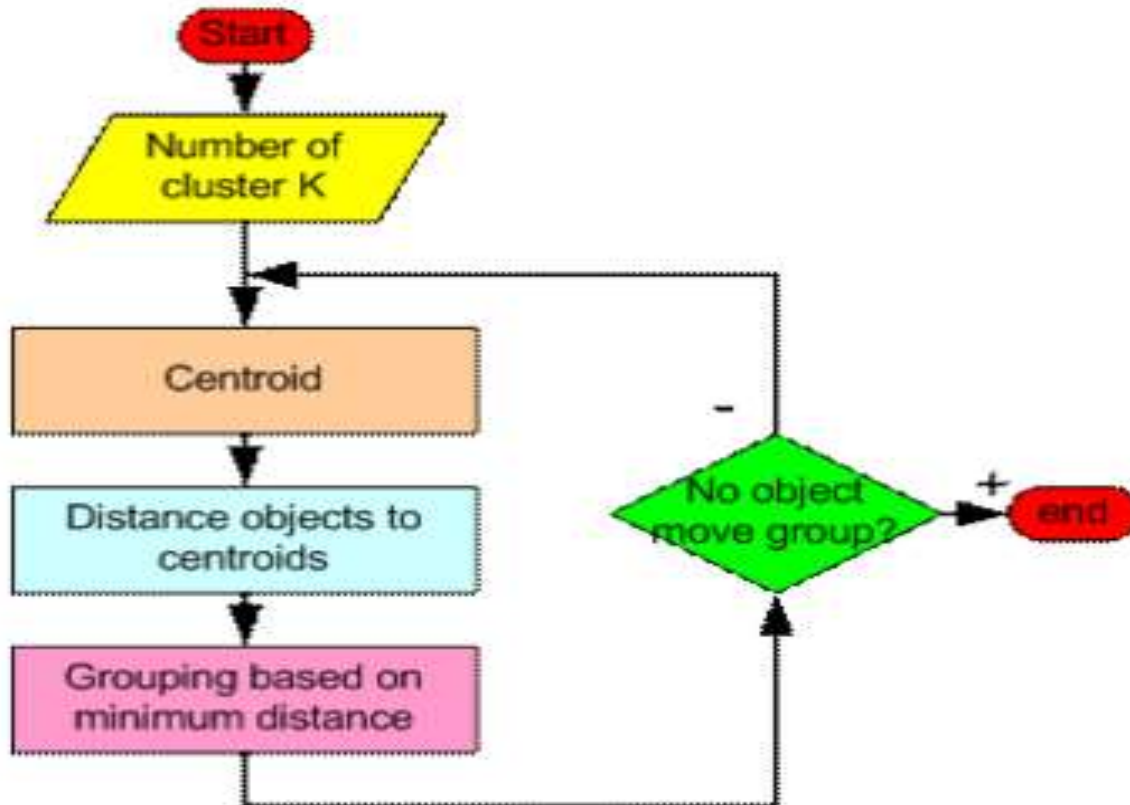
Method:

- (1) Arbitrarily **choose k objects from D as the initial cluster centers**
- (2) **Repeat:**
 - (3) **(re)assign** each object to the **cluster** to which the **object is the most similar**, based on the **mean** value of the objects in the cluster
 - (4) **Update** the **cluster means** *i.e., calculate the mean value of the objects for each cluster*
- (3) **until no change**

K-Means Algorithm

Iterate until *stable* (= no object move group):

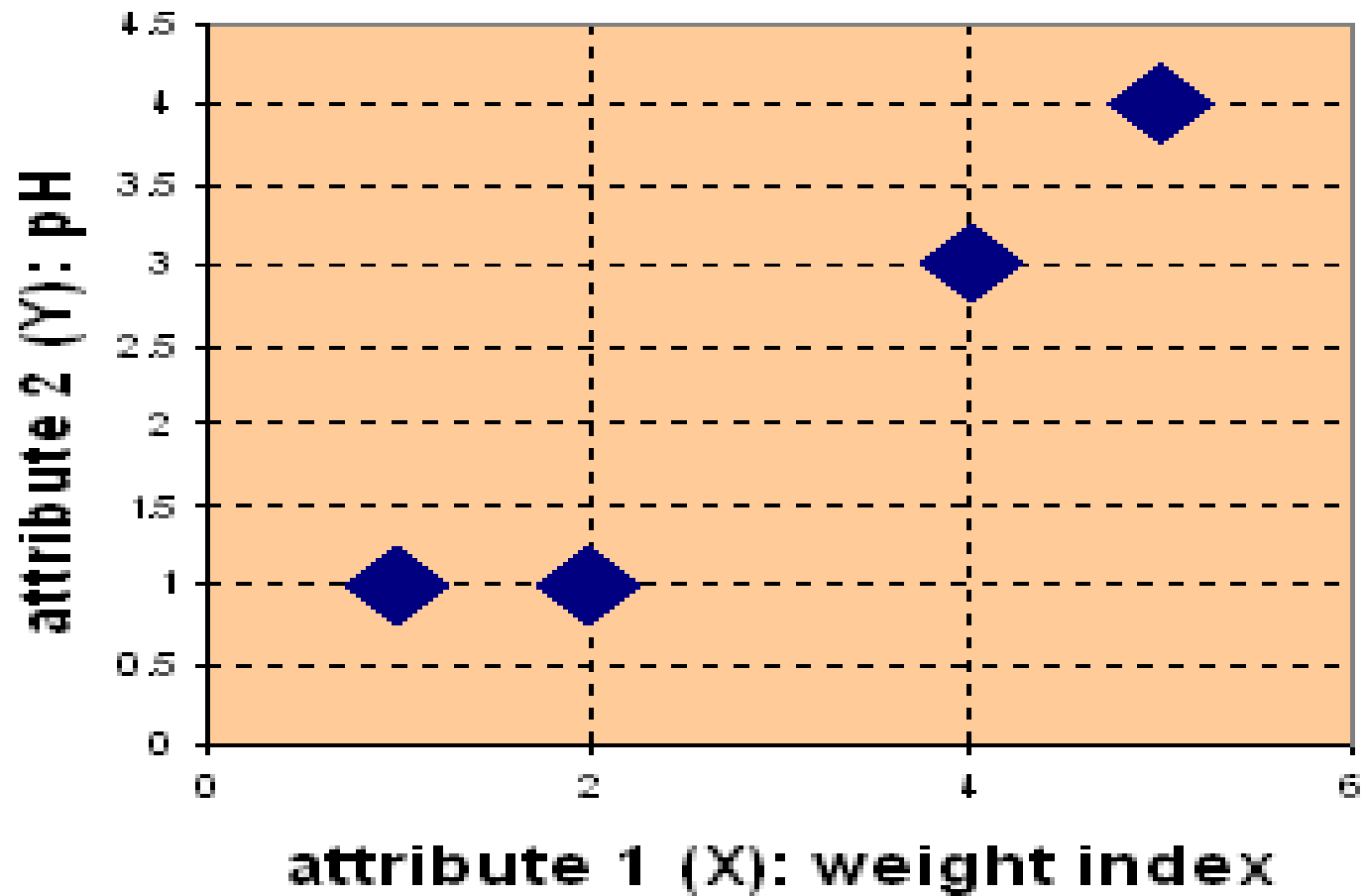
1. Determine the centroid coordinate
2. Determine the distance of each object to the centroids
3. Group the object based on minimum distance



Example: k-Means

Example: Suppose we have 4 objects as your training data points and each object have 2 attributes. Each attribute represents coordinate of the object.

Object	Attribute 1 (X):weight index	Attribute 2 (Y): pH
Medicine A	1	1
Medicine B	2	1
Medicine C	4	3
Medicine D	5	4



Example: k-Means Cont...

1. *Initial value of centroids*: Suppose we use medicine A and medicine B as the first centroids. Let \mathbf{c}_1 and \mathbf{c}_2 denote the coordinate of the centroids, then $\mathbf{c}_1 = (1, 1)$ and $\mathbf{c}_2 = (2, 1)$

Example: k-Means Cont...

2. Objects-Centroids distance: we calculate the distance between cluster centroid to each object. Let us use [Euclidean distance](#), then we have distance matrix at iteration 0 is

$$\mathbf{D}^0 = \begin{bmatrix} 0 & 1 & 3.61 & 5 \\ 1 & 0 & 2.83 & 4.24 \end{bmatrix} \quad \begin{array}{ll} \mathbf{c}_1 = (1,1) & \text{group - 1} \\ \mathbf{c}_2 = (2,1) & \text{group - 2} \end{array}$$

<i>A</i>	<i>B</i>	<i>C</i>	<i>D</i>	
1	2	4	5	<i>X</i>
1	1	3	4	<i>Y</i>

$\mathbf{c}_1 = (1,1)$ is $\sqrt{(4-1)^2 + (3-1)^2} = 3.61$, and its distance to the second centroid

$\mathbf{c}_2 = (2,1)$ is $\sqrt{(4-2)^2 + (3-1)^2} = 2.83$, etc.

Example: k-Means Cont...

3. *Objects clustering*: We assign each object based on the minimum distance. Thus, medicine A is assigned to group 1, medicine B to group 2, medicine C to group 2 and medicine D to group 2. The element of Group matrix below is 1 if and only if the object is assigned to that group.

$$\mathbf{G}^0 = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 1 & 1 \end{bmatrix} \begin{matrix} \text{group - 1} \\ \text{group - 2} \end{matrix}$$

A B C D

Example: k-Means Cont...

4. *Iteration-1, determine centroids*: Knowing the members of each group, now we compute the new centroid of each group based on these new memberships. Group 1 only has one member thus the centroid remains in $\mathbf{c}_1 = (1, 1)$. Group 2 now has three members, thus the centroid is the average coordinate among the three members:

$$\mathbf{c}_2 = \left(\frac{2+4+5}{3}, \frac{1+3+4}{3} \right) = \left(\frac{11}{3}, \frac{8}{3} \right).$$

Example: k-Means Cont...

5. *Iteration-1, Objects-Centroids distances*: The next step is to compute the distance of all objects to the new centroids. Similar to step 2, we have distance matrix at iteration 1 is

$$D^1 = \begin{bmatrix} 0 & 1 & 3.61 & 5 \\ 3.14 & 2.36 & 0.47 & 1.89 \end{bmatrix} \quad \begin{array}{l} \mathbf{c}_1 = (1,1) \text{ group-1} \\ \mathbf{c}_2 = (\frac{11}{3}, \frac{8}{3}) \text{ group-2} \end{array}$$

<i>A</i>	<i>B</i>	<i>C</i>	<i>D</i>	
$\begin{bmatrix} 1 & 2 & 4 & 5 \end{bmatrix}$				<i>X</i>
$\begin{bmatrix} 1 & 1 & 3 & 4 \end{bmatrix}$				<i>Y</i>

Example: k-Means Cont...

6. *Iteration-1, Objects clustering*: Similar to step 3, we assign each object based on the minimum distance. Based on the new distance matrix, we move the medicine B to Group 1 while all the other objects remain. The Group matrix is shown below

$$\mathbf{G}^1 = \begin{bmatrix} 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \end{bmatrix} \begin{matrix} \text{group} - 1 \\ \text{group} - 2 \end{matrix}$$

$A \quad B \quad C \quad D$

Example: k-Means Cont...

7. *Iteration 2, determine centroids:* Now we repeat step 4 to calculate the new centroids coordinate based on the clustering of previous iteration. Group1 and group 2 both has two members, thus the new centroids are

$$\mathbf{c}_1 = \left(\frac{1+2}{2}, \frac{1+1}{2} \right) = \left(1\frac{1}{2}, 1 \right) \quad \text{and} \quad \mathbf{c}_2 = \left(\frac{4+5}{2}, \frac{3+4}{2} \right) = \left(4\frac{1}{2}, 3\frac{1}{2} \right)$$

Example: k-Means Cont...

8. *Iteration-2, Objects-Centroids distances*: Repeat step 2 again, we have new distance matrix at iteration 2 as

$$\mathbf{D}^2 = \begin{bmatrix} 0.5 & 0.5 & 3.20 & 4.61 \\ 4.30 & 3.54 & 0.71 & 0.71 \end{bmatrix} \quad \begin{array}{l} \mathbf{c}_1 = (1\frac{1}{2}, 1) \text{ group-1} \\ \mathbf{c}_2 = (4\frac{1}{2}, 3\frac{1}{2}) \text{ group-2} \end{array}$$

A B C D

1	2	4	5	<i>X</i>
1	1	3	4	<i>Y</i>

Example: k-Means Cont...

9. *Iteration-2, Objects clustering:* Again, we assign each object based on the minimum distance.

$$\mathbf{G}^2 = \begin{bmatrix} 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \end{bmatrix} \begin{matrix} \text{group-1} \\ \text{group-2} \end{matrix}$$

$A \quad B \quad C \quad D$

We obtain result that $\mathbf{G}^2 = \mathbf{G}^1$.

Disadvantages

- ▶ **k-means** does not guarantee to find **the global optimum solution for clustering**
- ▶ The algorithm can be very **sensitive to outliers & noisy data**
- ▶ In other words, **k-means** will regularly **discover a local rather than global minimum**

K-Medoids: A Representative Object-based clustering technique

- ▶ The **K-medoids** clustering algorithm **Partitions Around Medoids**
- ▶ The **K-medoids algorithm is more robust to noise** than K-means algorithm
- ▶ In **K-medoids**, data points are chosen to be the **medoids**

▶ A **medoid** can be defined as that object of a cluster, whose **average dissimilarity** to all the objects in the cluster is **minimal**

Difference between k-means & k-medoids

- The difference between **k-means** & **k-medoids** is analogous to:
- The difference between **mean** & **median**:
 - Where **mean** indicates the **average value of all data items**
 - While **median** indicates all data items are **evenly distributed around it**

K-Medoids & PAM algorithm: Basic Idea

PAM algorithm

- ▶ The **PAM** (Partitioning Around Medoids) algorithm searches for **k representative objects (k medoids)** in a data set.
- ▶ After finding a set of **k medoids**, clusters are constructed by assigning each data point to the **nearest medoid**
- ▶ Next, each **selected medoid** & each **non-medoid data point** are **swapped** & the **objective function** is computed
- ▶ The **objective function** corresponds to the sum of the dissimilarities of all objects to their nearest medoid
- ▶ The **SWAP** step attempts to **improve the quality of the clustering** by exchanging **selected objects (medoids)** & **non-selected objects**

K-Medoids: Basic Idea

- ▶ If the **objective function** can be **reduced** by interchanging a selected object with an unselected object, then the **swap** is carried out. This continues till the objective function can **no longer be decreased**
- ▶ The **dissimilarity of the medoid(C_i) & object(P_i)** is calculated by using **$E = |P_i - C_i|$**

The cost in K-Medoids algorithm is given as

$$c = \sum_{C_i} \sum_{P_i \in C_i} |P_i - C_i|$$

K-Medoids Algorithm

- 1. Initialize:** select k random points out of the n data points as the **medoids**
- 2. Associate each data point to the **closest medoid** by using any common distance metric methods**
- 3. While the cost decreases:**
For each medoid m , for each data o point which is not a medoid:
 - a) **Swap m & o** , associate each data point to the closest medoid, **recompute the cost****
 - b) If the total cost $>$ that in the previous step, **undo the swap****

Steps of K-Medoids

The algorithm proceeds in two steps:

- ▶ **BUILD-step:** This step sequentially **selects k "centrally located" objects, to be used as initial medoids**
- ▶ **SWAP-step:** If the **objective function can be reduced** by **interchanging a selected object with an unselected object**, then the **swap is carried out**. This is continues till the objective function can no longer be decreased.

▶ Build phase:

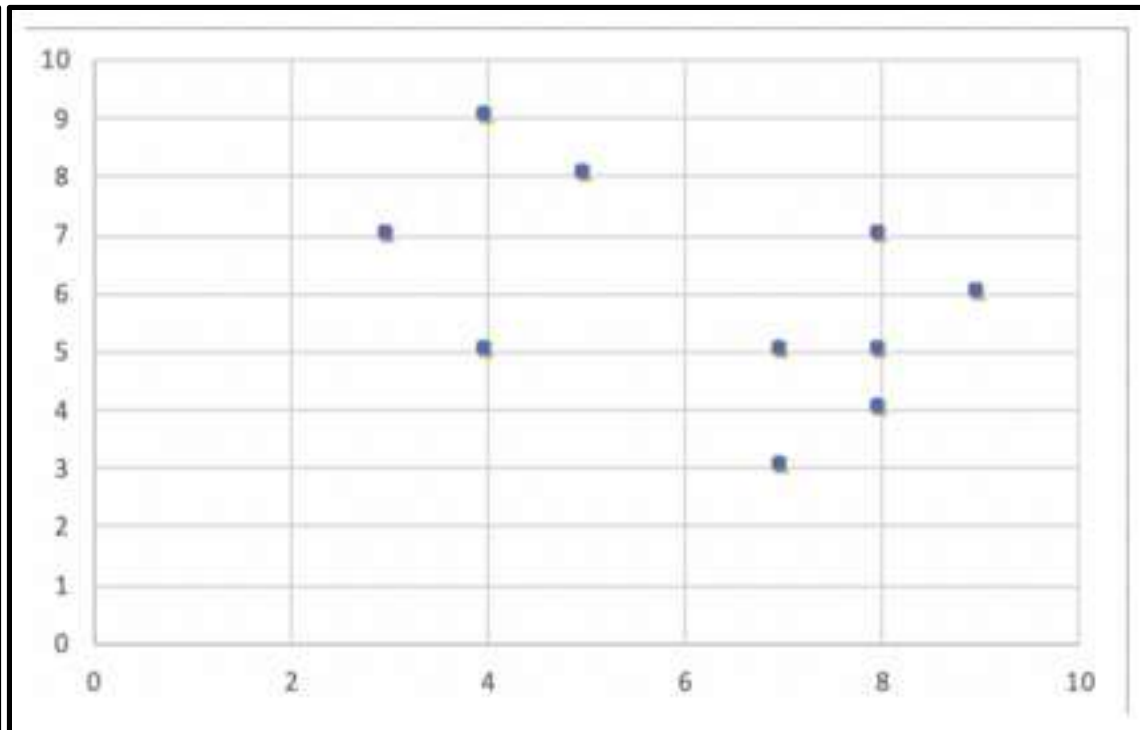
1. Select k objects to become the medoids, or in case these objects were provided use them as the medoids. Calculate the dissimilarity matrix if it was not provided
2. Assign every object to its closest medoid

Swap phase:

3. For each cluster search, check if any of the object of the cluster decreases the average dissimilarity coefficient; if it does, select the entity that decreases this coefficient the most as the medoid
4. If at least one medoid has changed go to (3), else end the algorithm.

Example: K-Medoids

	X	Y
0	8	7
1	3	7
2	4	9
3	9	6
4	8	5
5	5	8
6	7	3
7	8	4
8	7	5
9	4	5



Example: K-Medoids

Step 1:

Let the randomly selected 2 medoids, so select $k = 2$ and let

C1 - (4, 5) and **C2 - (8, 5)** are the two medoids.

Step 2: Calculating cost.

The dissimilarity of each non-medoid point with the medoids is calculated and tabulated:

	X	Y	Dissimilarity from C1	Dissimilarity from C2
0	8	7	6	2
1	3	7	3	7
2	4	9	4	8
3	9	6	6	2
4	8	5	-	-
5	5	8	4	6
6	7	3	5	3
7	8	4	5	1
8	7	5	3	1
9	4	5	-	-

Each point is assigned to the cluster of that medoid whose dissimilarity is less.

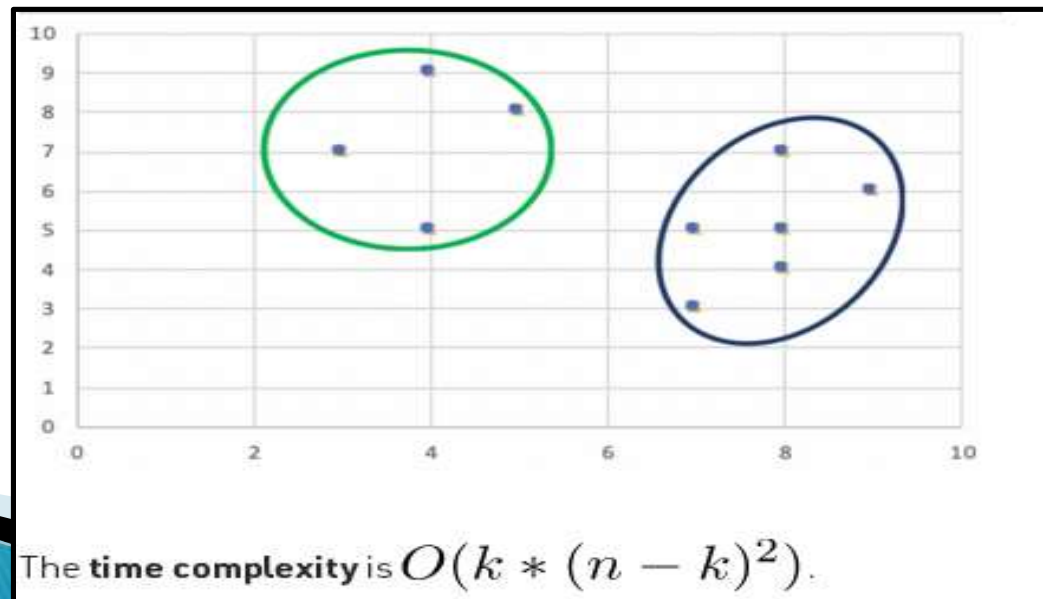
The points 1, 2, 5 go to cluster c1 and 0, 3, 6, 7, 8 go to cluster c2.

The cost = $(3 + 4 + 4) + (3 + 1 + 1 + 2 + 2) = 20$

- ▶ **Step 3: randomly select one non-medoid point & recalculate the cost**
- ▶ Let the randomly selected **point be (8, 4)**. The dissimilarity of each non-medoid point with the medoids : C1 (4, 5) & C2 (8, 4) is calculated & tabulated.

	X	Y	Dissimilarity from C1	Dissimilarity from C2
0	8	7	6	3
1	3	7	3	8
2	4	9	4	9
3	9	6	6	3
4	8	5	4	1
5	5	8	4	7
6	7	3	5	2
7	8	4	-	-
8	7	5	3	2
9	4	5	-	-

- ▶ Each point is assigned to that cluster whose dissimilarity is less. So, the **points 1, 2, 5 go to cluster C1** & **0, 3, 6, 7, 8 go to cluster C2**
- ▶ The New cost = $(3 + 4 + 4) + (2 + 2 + 1 + 3 + 3) = 22$
Swap Cost = New Cost – Previous Cost = $22 - 20$ & $2 > 0$
- ▶ As the swap cost is not less than zero, **we undo the swap**
- ▶ Hence, **previous (4, 5) & (8, 5)** are the final medoids. The clustering would be in the following way



Advantages:

- ▶ It is simple to understand & easy to implement
- ▶ K-Medoid Algorithm is fast & converges in a fixed no. of steps
- ▶ **PAM** is less sensitive to outliers

Disadvantages:

- The main disadvantage of K-Medoid algorithms is that it is not suitable for clustering non-spherical (arbitrary shaped) groups of objects
- k-Medoid works efficiently for small data sets but does not scale well for large data sets
- It may obtain different results for different runs on the same dataset because the k medoids are chosen randomly

Notes:

- ▶ The **PAM** algorithm works with a **matrix of dissimilarity** & to compute this matrix the algorithm can **use two metrics**
- ▶ The **Euclidean distances**, that are the **root sum-of-squares of differences**
- ▶ And, the **Manhattan distance** that are the **sum of absolute distances**
- ▶ We should **get similar results most of the time**, using either euclidean or Manhattan distance
- ▶ If your data contains **outliers**, Manhattan distance should give more robust results, whereas euclidean would be influenced by unusual values

Types of Distance Metrics

- ▶ Euclidean Distance
- ▶ Manhattan Distance
- ▶ Minkowski Distance
- ▶ Hamming Distance

1. Euclidean Distance

- ▶ **Euclidean Distance** represents the **shortest distance between two points**

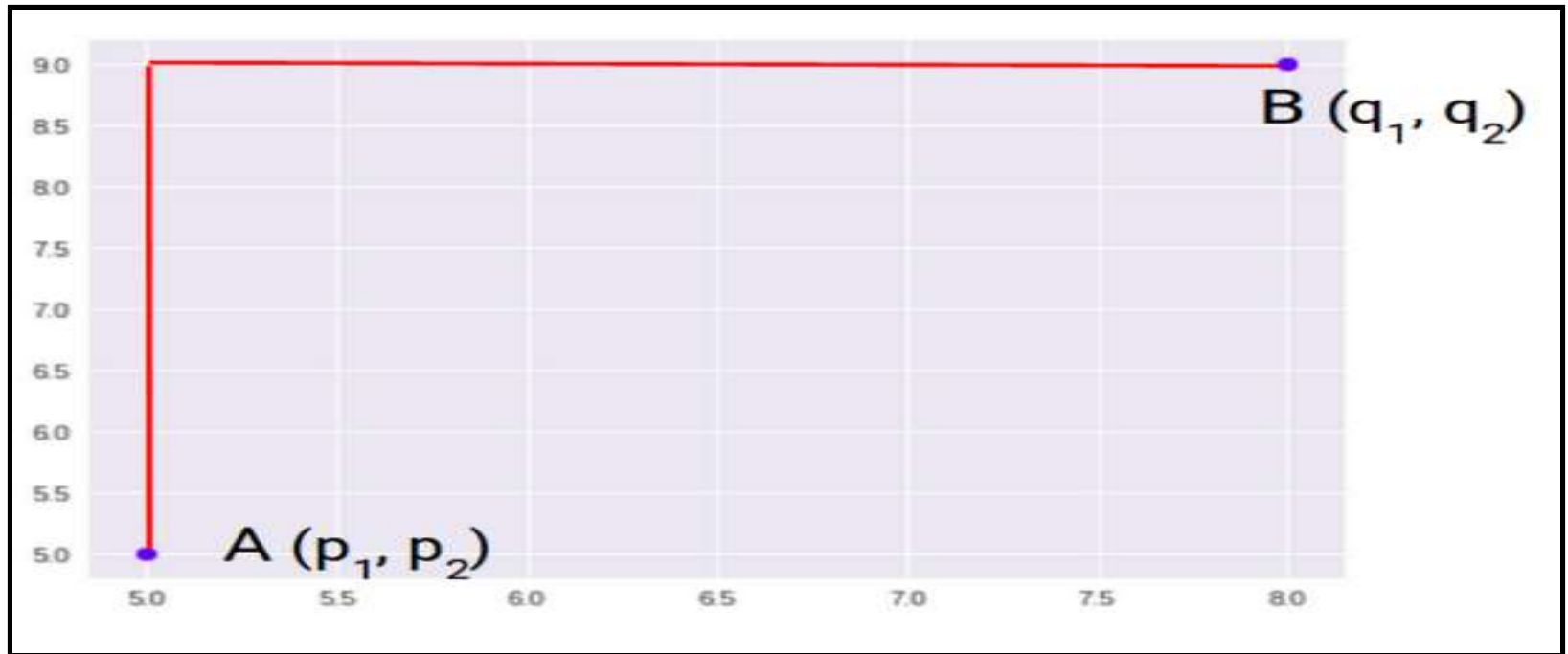
$$D_e = \left(\sum_{i=1}^n (p_i - q_i)^2 \right)^{1/2}$$

Where,

- n = number of dimensions
- p_i, q_i = data points

2. Manhattan Distance

- ▶ **Manhattan Distance** is the sum of absolute differences between points across all the dimensions



- Since the above representation is 2 dimensional, to calculate Manhattan Distance, we will take the **sum of absolute distances in both the x & y directions**. So, the Manhattan distance in a 2-dimensional space is given as:

$$d = |p_1 - q_1| + |p_2 - q_2|$$

And the generalized formula for an n-dimensional space is given as:

$$D_m = \sum_{i=1}^n |p_i - q_i|$$

Where,

- n = number of dimensions
- p_i, q_i = data points

Example:

Features	Coord1	Coord2	Coord3	Coord4
Object A	0	3	4	5
Object B	7	6	3	-1

The City Block Distance between point A and B is

$$\begin{aligned}d_{BA} &= |0 - 7| + |3 - 6| + |4 - 3| + |5 - (-1)| \\ &= 7 + 3 + 1 + 6 = 17\end{aligned}$$

3. Minkowski Distance

- ▶ **Minkowski Distance** is the generalized form of Euclidean & Manhattan Distance
- ▶ When the order is 1, both Minkowski & Manhattan Distance are the same
- ▶ When the order is 2, Minkowski & Euclidean distances are the same

The formula for Minkowski Distance is given as:

$$D = \left(\sum_{i=1}^n |p_i - q_i|^p \right)^{1/p}$$

► **Example:**

Features	Coord1	Coord2	Coord3	Coord4	Coord5	Coord6
Object A	0	3	4	5		
Object B	7	6	3	-1		

Input order parameter lambda =

	cost	time	weight	incentive
Object A	0	3	4	5
Object B	7	6	3	-1

Point A has coordinate (0, 3, 4, 5) and point B has coordinate (7, 6, 3, -1).

The Minkowski Distance of order 3 between point A and B is

$$d_{BA} = \left(|0-7|^3 + |3-6|^3 + |4-3|^3 + |5+1|^3 \right)^{\frac{1}{3}}$$

$$= \sqrt[3]{343+27+1+216} = \sqrt[3]{587} = 8.373$$

4. Hamming Distance

- ▶ **Hamming Distance** measures the **similarity between two strings of the same length**
- ▶ The **Hamming Distance** between two strings is the **no. of positions at which the corresponding characters are different**
- ▶ Let's say we have two strings: **“euclidean”** & **“manhattan”**
- ▶ We will **go character by character & match the strings & check the difference**
- ▶ **Seven characters** are **different** whereas two characters (the last two characters) are similar

euclidean and **manhattan**

Hence, the Hamming Distance here will be 7. Note that larger the Hamming Distance between two strings, more dissimilar will be those strings (and vice versa).

DMDW – Module-5

By

Dr. Pulak Sahoo

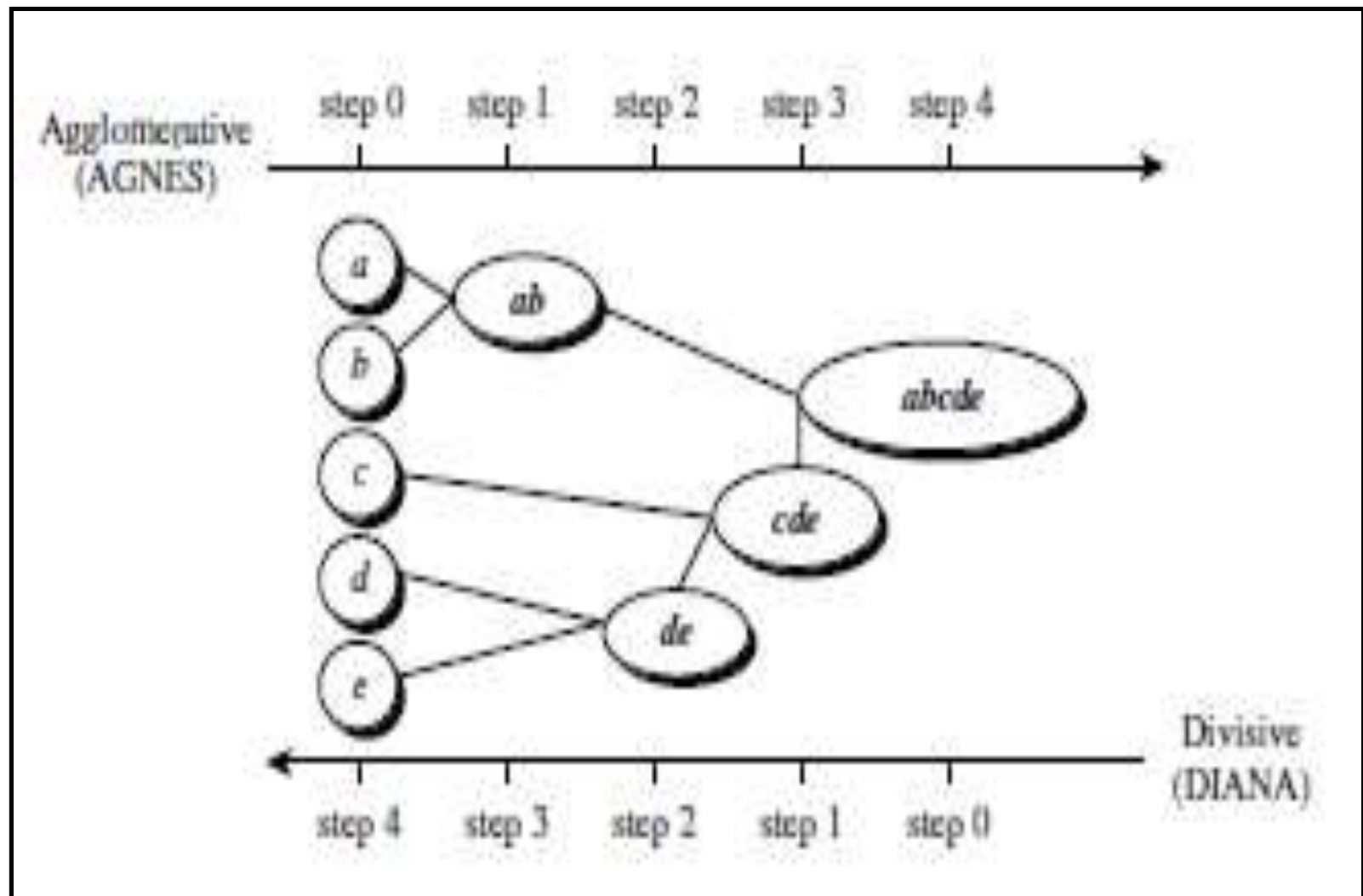
Associate Professor
Dept of CSE, SIT, BBSR

Hierarchical clustering

- ▶ A **hierarchical clustering** method works by **grouping data objects into a tree of clusters**
- ▶ A hierarchical method creates a **hierarchical decomposition** of the **given set of data objects**
- ▶ Two types of **hierarchical clustering** approaches are there:
 1. **Agglomerative approach**
 2. **Divisive approach**

2. Hierarchical Methods

- ▶ The **Agglomerative approach** or the **bottom-up approach**, starts with each object forming a separate group
 - ▶ It **successively merges** the objects close to one another, until **all the groups are merged into one** or a termination condition holds
-
- ▶ The **Divisive approach** or the **top-down approach**, starts with all the objects in the same cluster
 - ▶ In **each successive** iteration, a **cluster is split into smaller clusters**, until **each object is in one cluster**, or a termination condition holds
 - ▶ Hierarchical clustering methods can be **distance-based** or **density- and continuity based**



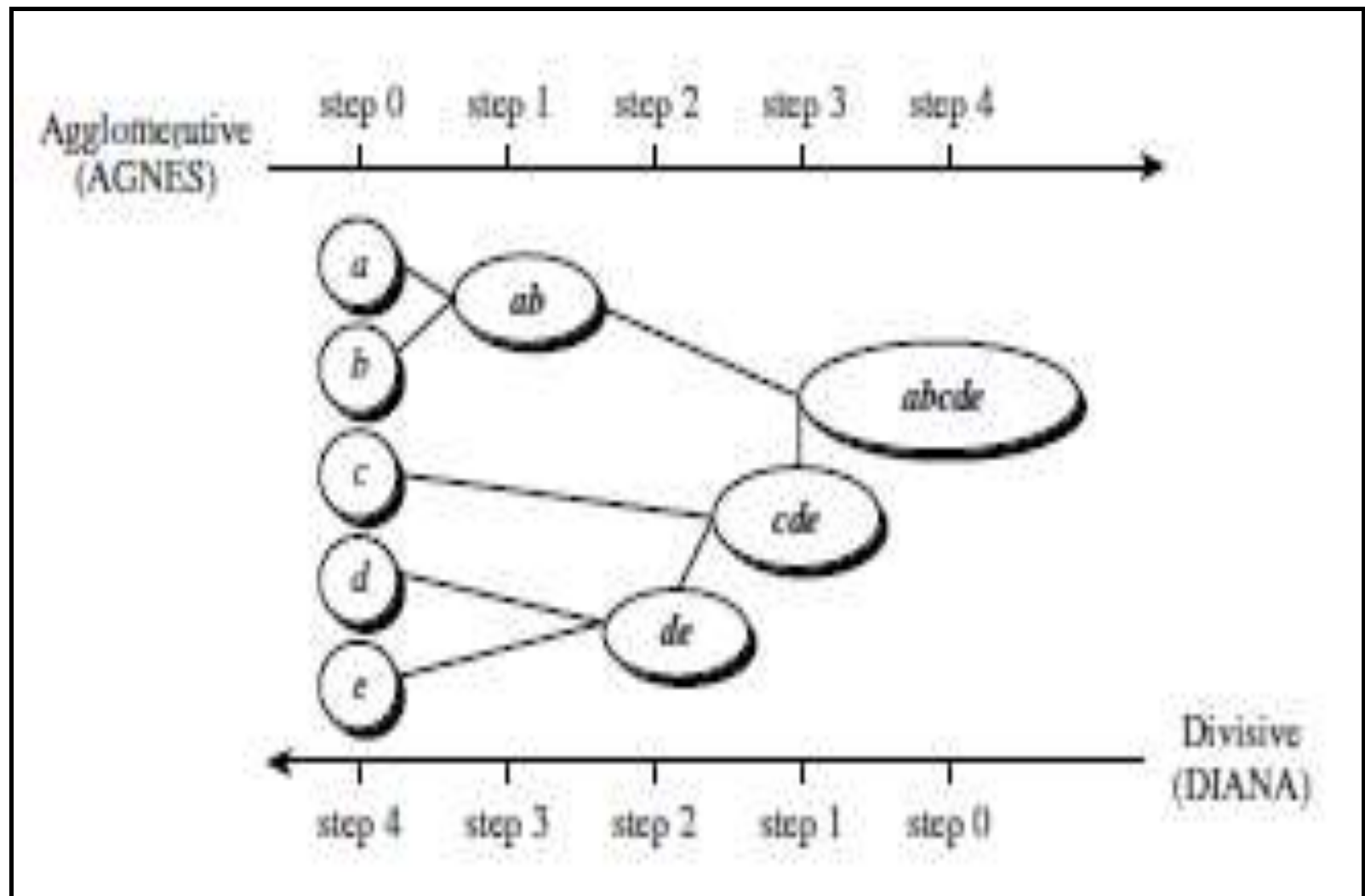


Density-based Methods

- ▶ The idea is to **continue growing** a **given cluster** as long as the **density of data points in the neighborhood exceeds some minimum threshold**
- ▶ **Ex: for each data point within a given cluster, the neighborhood of a given radius has to contain at least a minimum no. of points**
- ▶ Such a method can be used to **filter out noise or outliers & discover clusters of arbitrary shape**

★ Agglomerative Hierarchical Clustering

- ▶ The **agglomerative approach**, also called the **bottom-up approach**, starts with each object forming a separate group.
- ▶ It **successively merges** the objects or groups close to one another, until all the groups are merged into one (the topmost level of the hierarchy) or a **termination condition** holds Repeat
- ▶ Most **hierarchical clustering methods** belong to this category



Agglomerative Algorithm

Step by step algorithm of agglomerative approach to compute hierarchical clustering is as follow

1. Convert object features to distance matrix.
2. Set each object as a cluster (thus if we have 6 objects, we will have 6 clusters in the beginning)
3. Iterate until number of cluster is 1

1. Merge two closest clusters
2. Update distance matrix

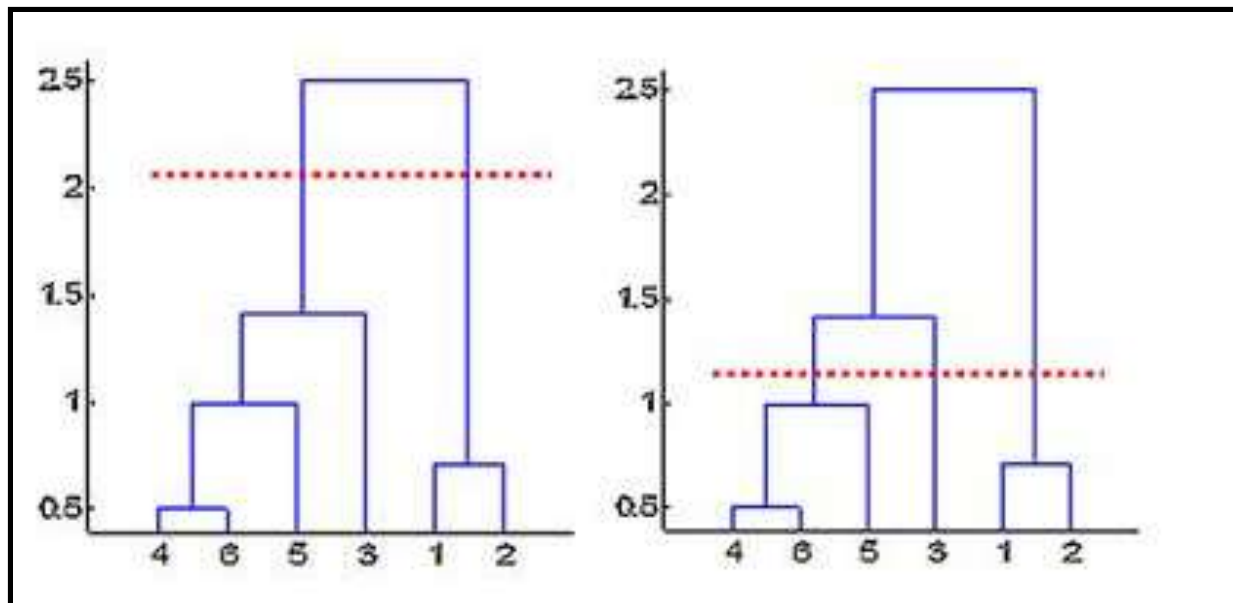


Divisive Hierarchical Clustering

- ▶ This top-down strategy does the reverse of agglomerative hierarchical clustering by starting with all objects in one cluster
- ▶ It subdivides the cluster into smaller & smaller pieces, until each object forms a cluster on its own or until it satisfies certain termination conditions, such as :
 - ▶ A desired number of clusters is obtained or
 - ▶ The diameter of each cluster is within a certain threshold

What is Dendrogram?

The standard output of hierarchical clustering is a dendrogram. A dendrogram is a cluster tree diagram where the distance of split or merge is recorded. Dendrogram is a visualization of hierarchical clustering.



Single Linkage Algorithm

Minimum distance : $d_{min}(C_i, C_j) = \min_{p \in C_i, p' \in C_j} |p - p'|$

Maximum distance : $d_{max}(C_i, C_j) = \max_{p \in C_i, p' \in C_j} |p - p'|$

Mean distance : $d_{mean}(C_i, C_j) = |m_i - m_j|$ Centroid Distance

Average distance : $d_{avg}(C_i, C_j) = \frac{1}{n_i n_j} \sum_{p \in C_i} \sum_{p' \in C_j} |p - p'|$

See
pic
after
2
slides

Single Linkage Algorithm Cont...

- ▶ When an algorithm uses the *minimum distance*, $d_{min}(C_i, C_j)$, to measure the **distance between clusters**, it is sometimes called a **nearest-neighbor clustering algorithm**
- ▶ If the clustering process is **terminated** when the **distance between nearest clusters exceeds a given threshold**, it is called a **single-linkage algorithm**

Linkages Between Objects

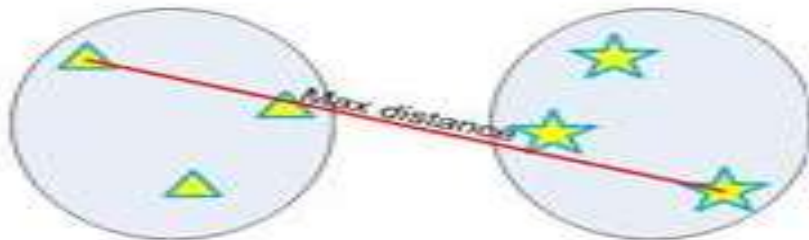
The rule of hierarchical clustering lie on how objects should be grouped into clusters. Given a distance matrix, linkages between objects can be computed through a criterion to compute distance between groups. Most common & basic criteria are

Single Linkage: minimum distance criterion



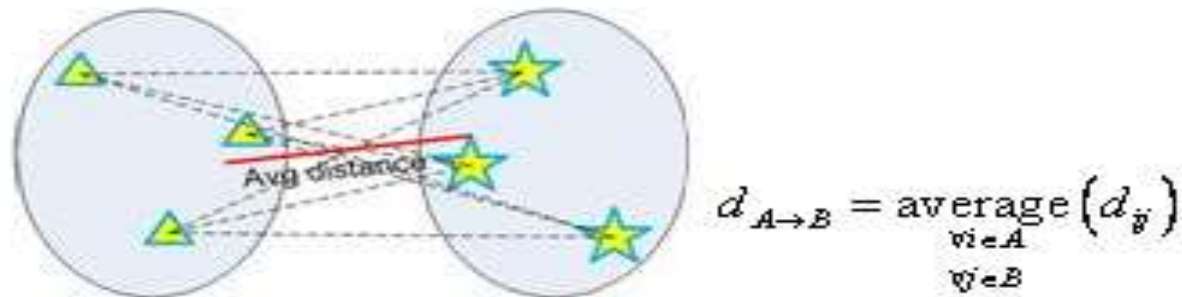
$$d_{A \rightarrow B} = \min_{\substack{v_i \in A \\ v_j \in B}} (d_{ij})$$

Complete Linkage: maximum distance criterion

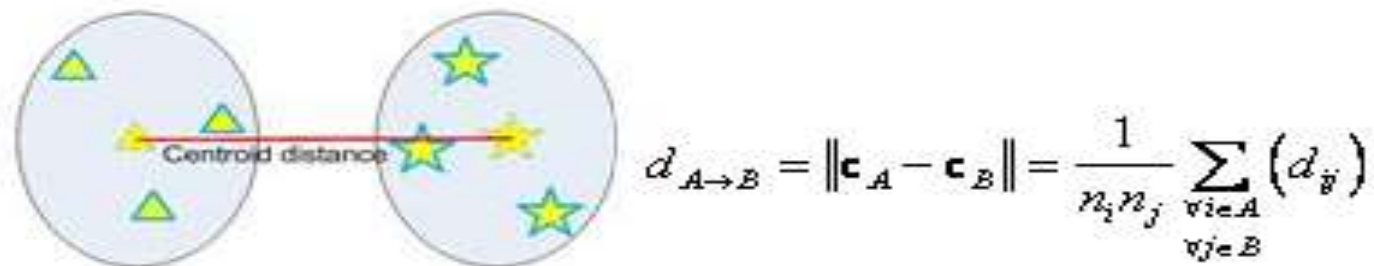


$$d_{A \rightarrow B} = \max_{\substack{v_i \in A \\ v_j \in B}} (d_{ij})$$

Average Group: average distance criterion

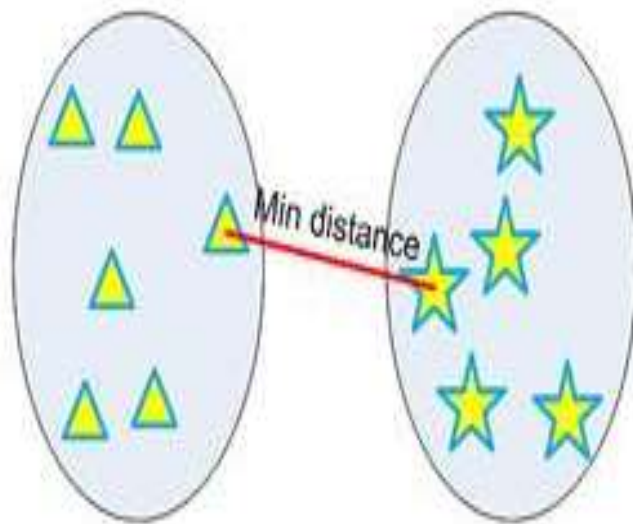


Centroid distance criterion



Ward: minimize variance of the merge cluster

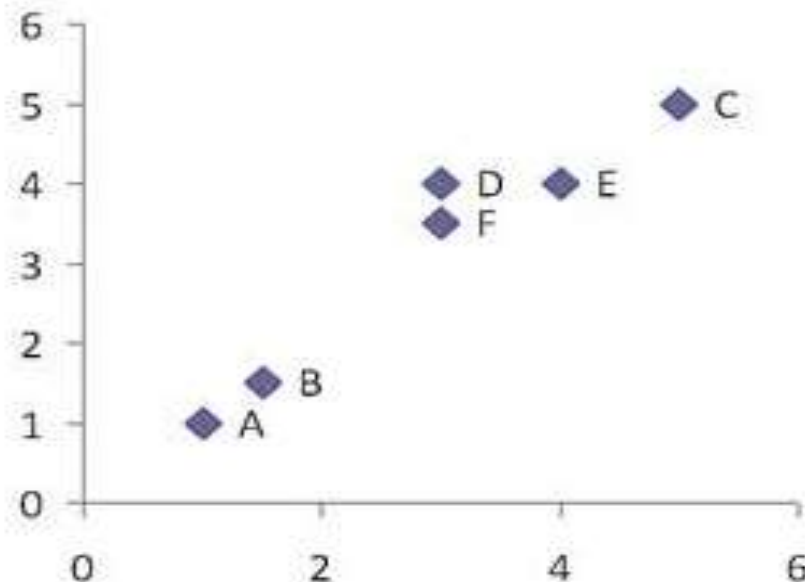
Minimum distance clustering is also called as single linkage hierarchical clustering or nearest neighbor clustering. Distance between two clusters is defined by the minimum distance between objects of the two clusters, as shown below.



Distance Matrix

To illustrate hierarchical clustering algorithm, let us use the following simple example. Suppose we have 6 objects (with name A, B, C, D, E and F) and each object have two measured features (X1 and X2). We can plot the features in a scattered plot to get the visualization of proximity between objects.

	X1	X2
A	1	1
B	1.5	1.5
C	5	5
D	3	4
E	4	4
F	3	3.5



The proximity between object can be measured as distance matrix. Suppose we use Euclidean distance , we can compute the distance between objects using the following formula

$$d_{ij} = \left(\sum_k (x_{ik} - x_{jk})^2 \right)^{\frac{1}{2}}$$

For example, distance between object A = (1, 1) and B = (1.5, 1.5) is computed as


$$d_{AB} = \left((1-1.5)^2 + (1-1.5)^2 \right)^{\frac{1}{2}} = \sqrt{\frac{1}{2}} = 0.7071$$

Dist	A	B	C	D	E	F
A	0.00	0.71	5.66	3.61	4.24	3.20
B	0.71	0.00	4.95	2.92	3.54	2.50
C	5.66	4.95	0.00	2.24	1.41	2.50
D	3.61	2.92	2.24	0.00	1.00	0.50
E	4.24	3.54	1.41	1.00	0.00	1.12
F	3.20	2.50	2.50	0.50	1.12	0.00



Minimum Distance

In general, if we have m objects, the number of distances on the lower triangular matrix (green part of the distance matrix) contain $\frac{1}{2}m(m-1)$ number of elements. In our example, we have 6 objects, thus the total distances that need to be computed is $\frac{1}{2}m(m-1) = \frac{1}{2} \cdot 6 \cdot 5 = 15$. We listed again below the 15 elements of distances as an array. Clearly the minimum distance is 0.5 (between object D and F).



0.71	5.66	3.61	4.24	3.20	4.95	2.92	3.54	2.50	2.24	1.41	2.50	1.00	0.50	1.12
------	------	------	------	------	------	------	------	------	------	------	------	------	------	------

Example: Agglomerative Clustering

Dist	A	B	C	D	E	F
A	0.00	0.71	5.66	3.61	4.24	3.20
B	0.71	0.00	4.95	2.92	3.54	2.50
C	5.66	4.95	0.00	2.24	1.41	2.50
D	3.61	2.92	2.24	0.00	1.00	0.50
E	4.24	3.54	1.41	1.00	0.00	1.12
F	3.20	2.50	2.50	0.50	1.12	0.00

We have 6 objects and we put each object into one cluster (analogue to put a ball into a basket). Instead of calling them as objects, now we call them clusters. Thus, in the beginning we have 6 clusters. Our goal is to group those 6 clusters such that at the end of the iterations, we will have only single cluster consists of the whole six original objects.

Example Cont...

In each step of the iteration, we find the closest pair clusters. In this case, the closest cluster is between cluster F and D with shortest distance of 0.5. Thus, we group cluster D and F into cluster (D, F). Then we update the distance matrix (see distance matrix below). Distance between ungrouped clusters will not change from the original distance matrix. Now the problem is how to calculate distance between newly grouped clusters (D, F) and other clusters?

Min Distance (Single Linkage)

Dist	A	B	C	D, F	E
A	0.00	0.71	5.66	?	4.24
B	0.71	0.00	4.95	?	3.54
C	5.66	4.95	0.00	?	1.41
D, F	?	?	?	0.00	?
E	4.24	3.54	1.41	?	0.00

Example Cont...

That is exactly where the linkage rule comes into effect. Using single linkage, we specify minimum distance between original objects of the two clusters.

Using the input distance matrix, distance between cluster (D, F) and cluster A is computed as

$$d_{(D,F) \rightarrow A} = \min(d_{DA}, d_{FA}) = \min(3.61, 3.20) = 3.20$$

Distance between cluster (D, F) and cluster B is

$$d_{(D,F) \rightarrow B} = \min(d_{DB}, d_{FB}) = \min(2.92, 2.50) = 2.50$$

Dist	A	B	C	D	E	F
A	0.00	0.71	5.66	3.61	4.24	3.20
B	0.71	0.00	4.95	2.92	3.54	2.50
C	5.66	4.95	0.00	2.24	1.41	2.50
D	3.61	2.92	2.24	0.00	1.00	0.50
E	4.24	3.54	1.41	1.00	0.00	1.12
F	3.20	2.50	2.50	0.50	1.12	0.00

Example Cont...

Similarly, distance between cluster (D, F) and cluster C is

$$d_{(D,F) \rightarrow C} = \min(d_{DC}, d_{FC}) = \min(2.24, 2.50) = 2.24$$

Finally, distance between cluster E and cluster (D, F) is calculated as

$$d_{E \rightarrow (D,F)} = \min(d_{ED}, d_{EF}) = \min(1.00, 1.12) = 1.00$$

Example Cont...

Then, the updated distance matrix becomes

Min Distance (Single Linkage)

A	B	C	D, F	E
0.00				
0.71	0.00			
1.66	4.95	0.00		
1.20	2.50	2.24	0.00	
1.24	3.54	1.41	1.00	0.00

Minimum
Distance
cluster (A,B)



Example Cont...

Looking at the lower triangular updated distance matrix, we found out that the closest distance between cluster B and cluster A is now 0.71. Thus, we group cluster A and cluster B into a single cluster name (A, B).

Now we update the distance matrix. Aside from the first row and first column, all the other elements of the new distance matrix are not changed.

Dist	A,B	C	(D, F)	E
A,B	0	?	?	?
C	?	0	2.24	1.41
(D, F)	?	2.24	0	1.00
E	?	1.41	1.00	0

Example Cont...

Using the input distance matrix (size 6 by 6), distance between cluster C and cluster (A,B) is computed as

$$d_{C \rightarrow (A,B)} = \min(d_{CA}, d_{CB}) = \min(5.66, 4.95) = 4.95$$

Distance between cluster (D, F) and cluster (A, B) is the minimum distance between all objects involves in the two clusters

$$d_{(D,F) \rightarrow (A,B)} = \min(d_{DA}, d_{DB}, d_{FA}, d_{FB}) = \min(3.61, 2.92, 3.20, 2.50) = 2.50$$

Similarly, distance between cluster E and (A, B) is

$$d_{E \rightarrow (A,B)} = \min(d_{EA}, d_{EB}) = \min(4.24, 3.54) = 3.54$$

Example Cont...

Then the updated distance matrix is

Min Distance (Single Linkage)

Dist	A,B	C	(D, F)	E
A,B	0	4.95	2.50	3.54
C	4.95	0	2.24	1.41
(D, F)	2.50	2.24	0	1.00
E	3.54	1.41	1.00	0

Observing the lower triangular of the updated distance matrix, we can see that the closest distance between clusters happens between cluster E and (D, F) at distance 1.00. Thus, we cluster them together into cluster ((D, F), E).

Example Cont...

The updated distance matrix is given below.

Min Distance (Single Linkage)

Dist	(A,B)	C	(D, F), E
(A,B)	0.00	4.95	2.50
C	4.95	0.00	1.41
(D, F), E	2.50	1.41	0.00

Minimum
Distance
cluster
 $\{((D,F),E),C\}$

Example Cont...

Distance between cluster ((D, F), E) and cluster (A, B) is calculated as

$$d_{((D,F),E) \rightarrow (A,B)} = \min(d_{DA}, d_{DB}, d_{FA}, d_{FB}, d_{EA}, d_{EB}) = \min(3.61, 2.92, 3.20, 2.50, 4.24, 3.54) = 2.50$$

Distance between cluster ((D, F), E) and cluster C yields the minimum distance of 1.41. This distance is computed as

$$d_{((D,F),E) \rightarrow C} = \min(d_{DC}, d_{FC}, d_{EC}) = \min(2.24, 2.50, 1.41) = 1.41$$

After that, we merge cluster ((D, F), E) and cluster C into a new cluster name (((D, F), E), C).

The updated distance matrix is shown in the figure below

Min Distance (Single Linkage)

Dist	(A,B)	((D, F), E),C
(A,B)	0.00	2.50
((D, F), E),C	2.50	0.00

**Minimum
Distance cluster
{(((D,F),E),C),(A,B)}**

Example Cont...

Min Distance (Single Linkage)		
Dist	(A,B)	(D, F), E),C
(A,B)	0.00	2.50
((D, F), E),C	2.50	0.00

The minimum distance of 2.5 is the result of the following computation

$$d_{(((D,F),E),C) \rightarrow (A,B)} = \min (d_{DA}, d_{DB}, d_{FA}, d_{FB}, d_{EA}, d_{EB}, d_{CA}, d_{CB})$$

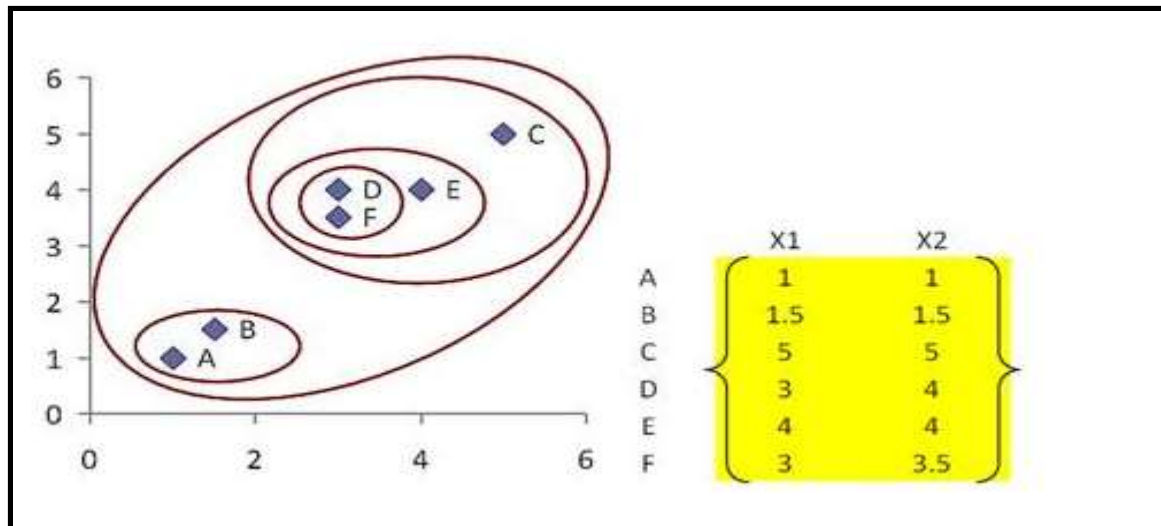
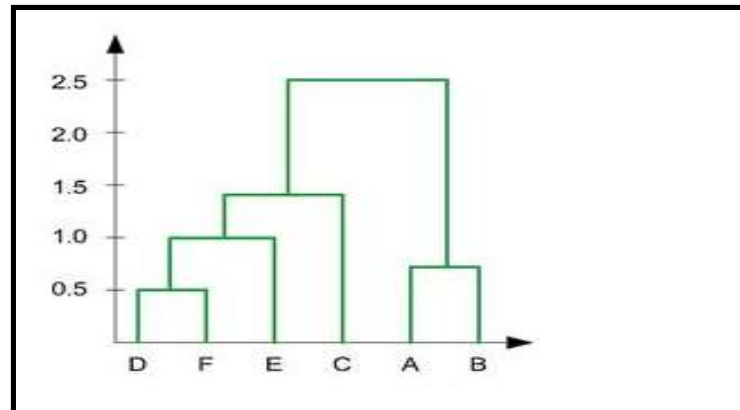
$$\underline{d_{(((D,F),E),C) \rightarrow (A,B)}} = \min (3.61, 2.92, 3.20, 2.50, 4.24, 3.54, 5.66, 4.95) = \underline{2.50}$$

Example Cont...

Now if we merge the remaining two clusters, we will get only single cluster contain the whole 6 objects. Thus, our computation is finished. We summarized the results of computation as follow:

1. In the beginning we have 6 clusters: A, B, C, D, E and F
2. We merge cluster D and F into cluster (D, F) at distance **0.50**
3. We merge cluster A and cluster B into (A, B) at distance **0.71**
4. We merge cluster E and (D, F) into ((D, F), E) at distance **1.00**
5. We merge cluster ((D, F), E) and C into (((D, F), E), C) at distance **1.41**
6. We merge cluster (((D, F), E), C) and (A, B) into ((((D, F), E), C), (A, B)) at distance **2.50**
7. The last cluster contain all the objects, thus conclude the computation

Example Cont...



DBSCAN with Neumericals

Density-Based Clustering Methods

- ▶ **Clustering based on density** (local cluster criterion), such as density-connected points
- ▶ **Major features:**
 - Discover clusters of arbitrary shape
 - Handle noise
 - One scan
 - Need density parameters as termination condition
- ▶ **Several interesting studies:**
 - DBSCAN: Ester, et al. (KDD'96)
 - OPTICS: Ankerst, et al (SIGMOD'99).
 - DENCLUE: Hinneburg & D. Keim (KDD'98)
 - CLIQUE: Agrawal, et al. (SIGMOD'98)

Density-Based Clustering: Basic Concepts

► Two parameters:

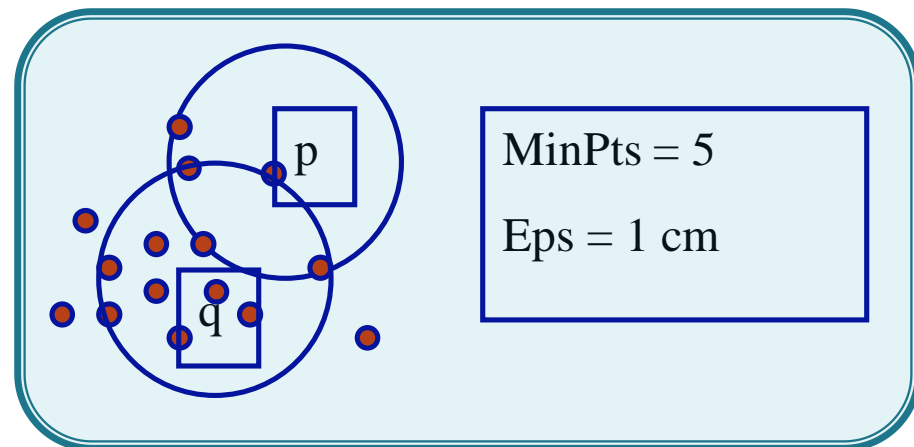
- **Eps**: Maximum radius of the neighbourhood (nbd)
- **MinPts**: Min. no. of points in an Eps-nbd of that point

► $N_{Eps}(p)$: $\{q \text{ belongs to } D \mid \text{dist}(p,q) \leq Eps\}$

► **Directly density-reachable**: A point p is directly density-reachable from a point q w.r.t. Eps , $MinPts$ if

- p belongs to $N_{Eps}(q)$
- core point condition:

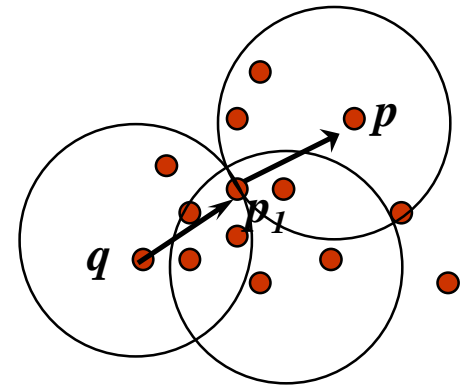
$$|N_{Eps}(q)| \geq MinPts$$



Density-Reachable and Density-Connected

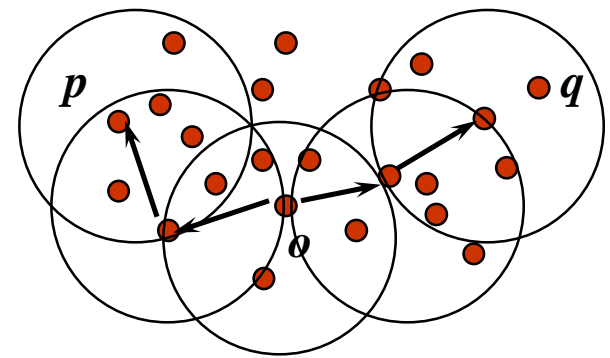
► Density-reachable:

- A point p is **density-reachable** from a point q w.r.t. Eps , $MinPts$ if there is a chain of points p_1, \dots, p_n , $p_1 = q$, $p_n = p$ such that p_{i+1} is directly density-reachable from p_i



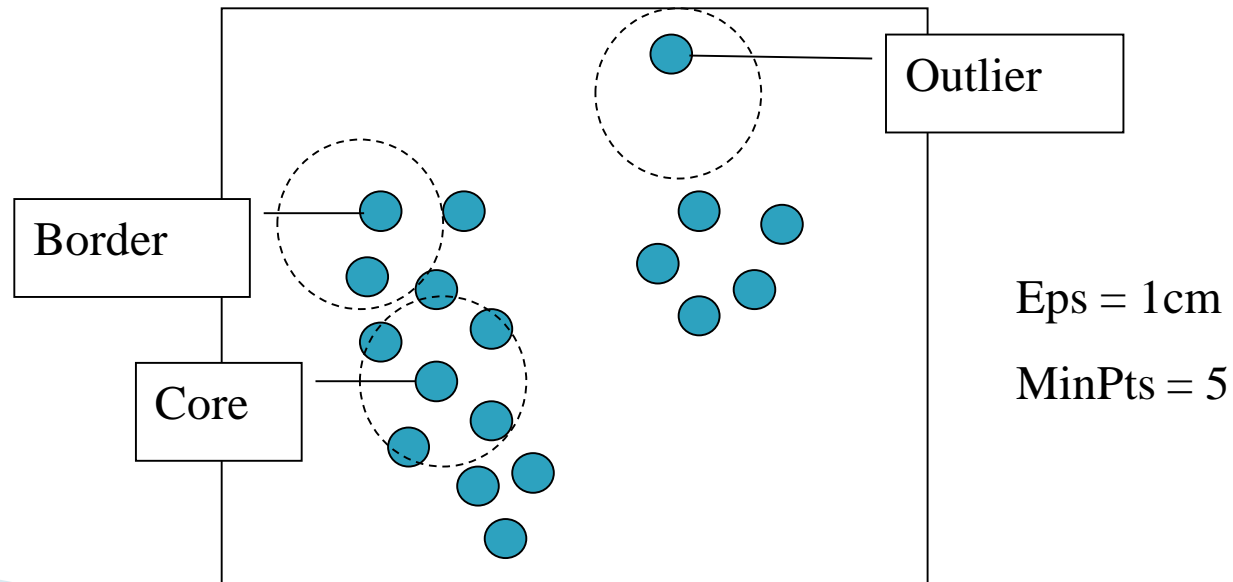
► Density-connected

- A point p is **density-connected** to a point q w.r.t. Eps , $MinPts$ if there is a point o such that both, p and q are density-reachable from o w.r.t. Eps and $MinPts$



DBSCAN: Density-Based Spatial Clustering of Applications with Noise

- ▶ Relies on a ***density-based*** notion of **cluster**: *A cluster is defined as a maximal set of density-connected points*
- ▶ Discovers clusters of **arbitrary shape** in **spatial databases with noise**



DBSCAN: The Algorithm

- ▶ Arbitrary select a point p
- ▶ Retrieve all points **density-reachable** from p w.r.t. Eps and $MinPts$
- ▶ If p is a **core point**, a cluster is formed
- ▶ If p is a **border point**, no points are density-reachable from p and DBSCAN visits the next point of the database
- ▶ Continue the process until all of the points have been processed

DBSCAN: Sensitive to Parameters

Figure 8. DBScan results for DS1 with MinPts at 4 and Eps at (a) 0.5 and (b) 0.4.

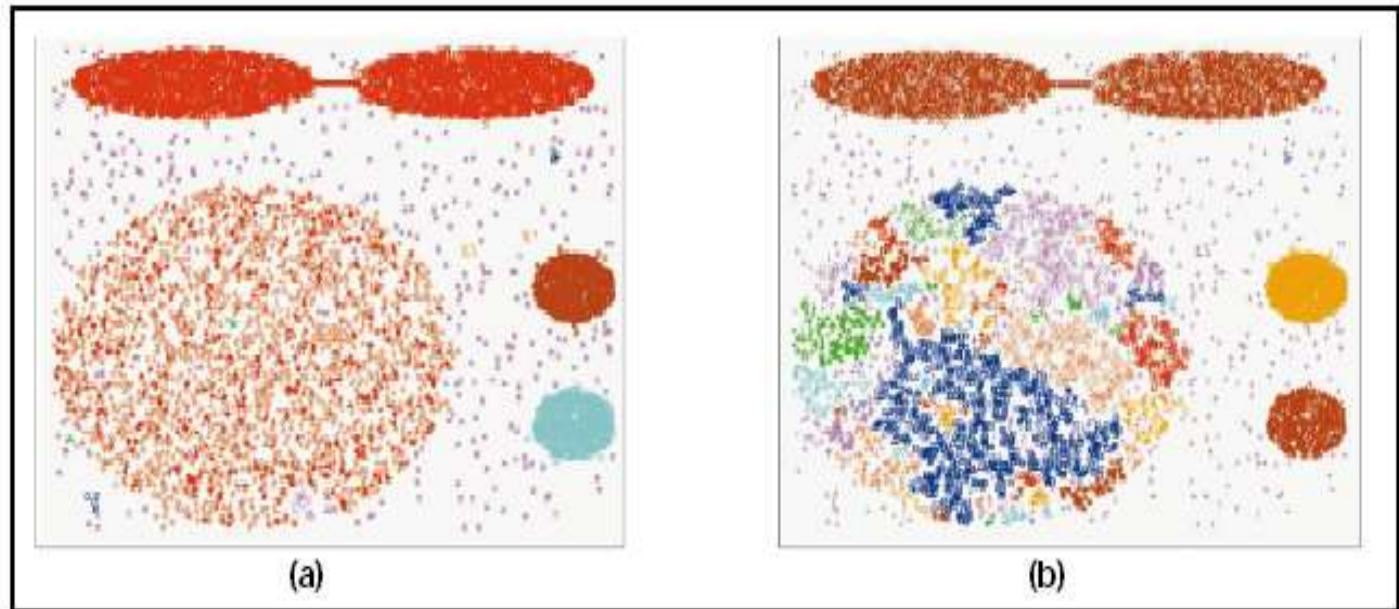
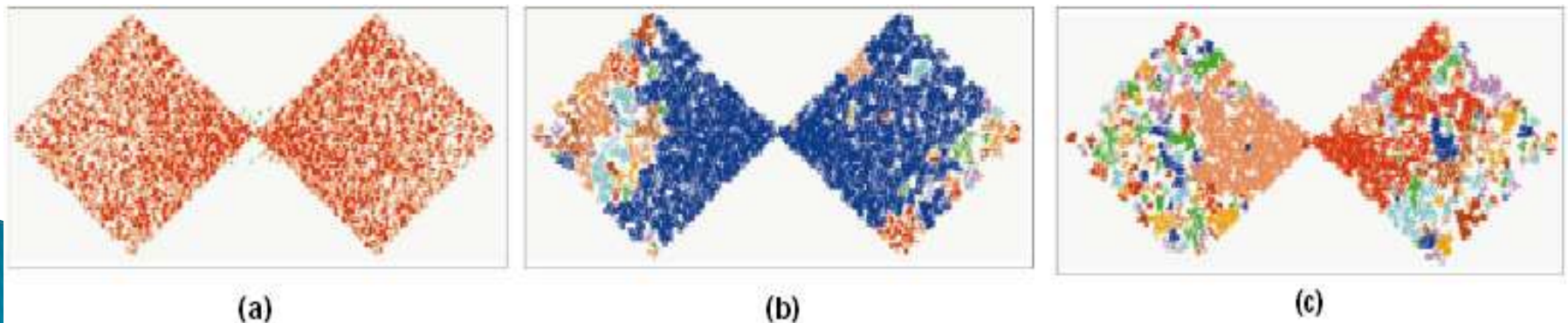


Figure 9. DBScan results for DS2 with MinPts at 4 and Eps at (a) 5.0, (b) 3.5, and (c) 3.0.



Center-Defined and Arbitrary

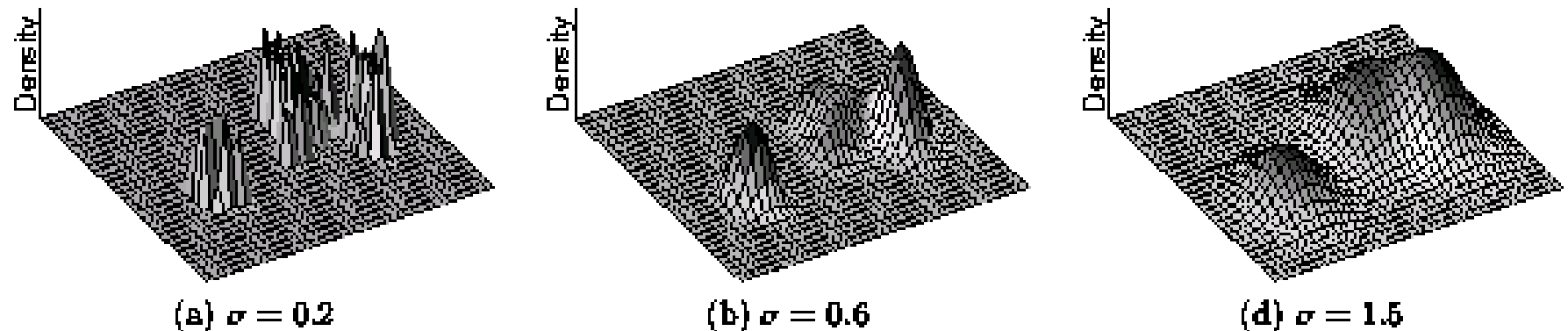


Figure 3: Example of Center-Defined Clusters for different σ

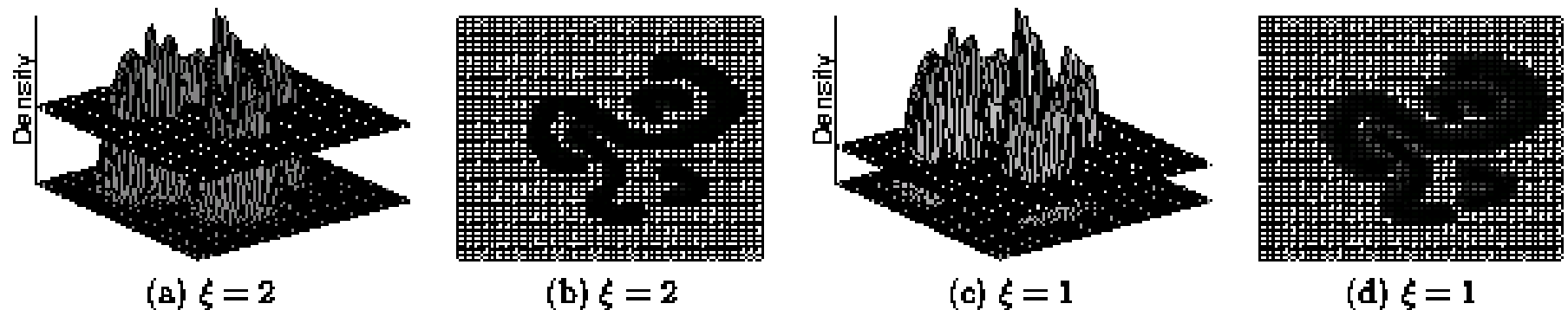


Figure 4: Example of Arbitrary-Shape Clusters for different ξ

Long Question– DBSCAN

Find no of core points ,border points and noise point where $\text{eps}=2$ and $\text{minpts}=3$.

Point	X	Y
P1	2	10
P2	2	5
P3	8	4
P4	5	8
P5	7	5
P6	6	4
P7	1	2
P8	4	9

Long Question– PAM (k-medoid) s-33

Suppose we have 10 objects as the training data points & each object has 2 attributes.

Each attribute represents coordinate of the object.

Considering $k=2$, find final two clusters using PAM algorithm.

x	y
8	7
3	7
4	9
9	6
8	5
5	8
7	3
8	4
7	5
4	5

Long Question – Dendrogram

Design a dendrogram using agglomerative clustering algorithm approach for following dataset.

	x1	x2
A	1	1
B	1.5	1.5
C	5	5
D	3	4
E	4	4
F	3	3.5

Long Question – K-Means

Consider the following set of two-dimensional records. Use the K-mean algorithm to cluster the data from given table.. We can use a value of 3 for K and can assume that the records with RIDs 1,3 and 5 are used for the initial cluster centroids (mean).

Rid	D1	D2
1	8	4
2	5	4
3	2	4
4	2	6
5	2	8
6	8	6

Long Question – K-Means

Suppose we have four objects as the training data points and each object has two attributes. Each attribute represents coordinate of the object. Considering $k=2$, find final two clusters using k-Means algorithm.

object	Attribute1	Attribute2
A	1	1
B	2	1
C	4	3
D	5	4

Long Question – K-Means

Suppose we have four objects as the training data points and each object has two attributes. Each attribute represents coordinate of the object. Considering $k=2$, find final two clusters using k-Means algorithm.

object	Attribute1	Attribute2
A	1	1
B	2	1
C	4	3
D	5	4

Short/Medium/Long Questions

Q: Write short notes on followings. (M)

- (a) Core point
- (b) Neighbour point
- (c) Noise point

Q: What are different types of linkage are used to find distance between two objects? (M)

Q: What is the disadvantage of hierarchical methods of clustering? (S)

Q: Describe a state under which the kmeans algorithm wont perform optimally. (S)

Q. What is the stopping criterion for the k-means clustering algorithm? (S)

Q. Describe in a systematic approach the different clustering techniques. (L)

Q. Explain the key differences between agglomerative and divisive hierarchical clustering. (L)

Short/Medium/Long Questions

Q. Briefly describe and give an example of each of the following clustering approaches: (L)

- a) partitioning methods
- b) hierarchical methods.

Q. Suppose that the data mining task is to cluster points (with (x, y) representing location) into three clusters. The points are $A1(2, 10)$, $A2(2, 5)$, $A3(8, 4)$, $B1(5, 8)$, $B2(7, 5)$, $B3(6, 4)$, $C1(1, 2)$, $C2(4, 9)$. The distance function is Euclidean distance. Suppose initially we assign $A1$, $B1$, and $C1$ as the center of each cluster respectively. Find out the final three clusters using k-Means algorithm. (L)

Q: Suppose that the data mining task is to cluster points (with (x, y) representing location) into three clusters. The points are $A1(2, 10)$, $A2(2, 5)$, $A3(8, 4)$, $B1(5, 8)$, $B2(7, 5)$, $B3(6, 4)$, $C1(1, 2)$, $C2(4, 9)$. The distance function is Euclidean distance. Suppose initially we assign $A1$, $B1$, and $C1$ as the center of each cluster respectively. Using k-Means algorithm find out the three cluster centers after the first round of execution. (L)

Short/Medium/Long Questions

Q. Find out the most suitable answer from the options given below.

Which clustering method is an efficient approach to many spatial data mining problems? (S)

- a) Partitioning Method
- b) Hierarchical Method
- c) Density-based Method
- d) Grid-based Method

Q. Why is clustering called as automatic classification? (S)

Q: Mention at least four applications of cluster analysis. (S)

Q. **Explain the different requirements for cluster analysis.(M)**

Q. Explain the difference between classification and clustering. (S)

Q. Discuss the different steps of k-Medoids clustering algorithm. (M)

Q. Explain the advantages and disadvantages of k-Means clustering algorithm.(M)

Short/Medium/Long Questions

Q. Explain the k-Means clustering algorithm. (S)

Q. Explain the different methods used for clustering. (L)

Q: Why is clustering known as unsupervised learning? (S)

Q. What is cluster analysis? (S)

Q. Explain the difference between classification and clustering. (S)

Q. Discuss the different steps of k-Medoids clustering algorithm. (M)

Q. Explain the advantages and disadvantages of k-Means clustering algorithm.(M)