CLUSTER ANALYSIS INTRODUCTION

Objectives

Student should able to

- To know the various methodologies for clustering

Syllabus:

Overview of Basic Clustering Methods, Partitioning Methods:k-Means: A Centroid-Based Technique, k-Mededis: A Representative Object-Based Technique, Hirearchical Methods :Agglomerative versus Divisive Hierarchical Clustering, Density-Based Methods : DBSCAN: Density-Based Clustering Based on Connected Regions with High Density.

Outcomes:

Student should able to

- Apply the different methods of clustering to identify distinct groups.
- Summarize various types of clustering techniques

6.0 Cluster Analysis Introduction

What is Cluster Analysis?

Cluster: Is a collection of data objects that are

– Similar to one another within the same cluster

– Dissimilar to the objects in other clusters

Clustering: The process of

– Grouping a set of data objects into clusters.

  - Clustering is an example of **unsupervised learning**, it do not rely on predefined classes and class-labeled training samples.

- It is a form of learning by observation rather than learning by examples.
- Typical applications
  - As a **stand-alone tool** to get insight into data distribution
  - As a **preprocessing step** for other algorithms
  - **General Applications** : market research, pattern recognition, data analysis, and image processing.
- A **good clustering** method will produce high quality clusters with
  - high intra-class similarity
  - low inter-class similarity
- The quality of a clustering result depends on both the similarity measure used by the method and its implementation

Applications of Clustering

- By clustering we can identify dense and sparse regions in object space and, discover overall distribution patterns and interesting correlations among data attributes
- Marketing: Help marketers discover distinct groups in their customer bases, and then use this knowledge to develop targeted marketing programs
- Land use: Identification of areas of similar land use in an earth observation database

- Insurance: Identifying groups of motor insurance policy holders with a high average claim cost
- City-planning: Identifying groups of houses according to their house type, value, and geographical location
- Earth-quake studies: Observed earth quake epicenters should be clustered along continent faults

6.1. Overview of Basic Clustering Methods

There exist a large number of clustering algorithms, the choice of clustering algorithm depends both on the type of data available and on the particular purpose and application.

Major clustering methods can be classified into the following categories,

- Partitioning methods
- Hierarchy methods
- Density-based methods
- Grid-based methods
- Model-based methods

Partitioning methods:

Given a database of n objects and k , the number of clusters to form , a partitioning algorithm organizes the objects into k partitions(k<=n), where each partition represents a cluster.

It satisfy the following requirements:

(1) each group must contain at least one object, and

(2) each object must belong to exactly one group.

It uses iterative relocation technique that attempts to improve the partitioning by moving objects from one group to another.

Popular methods are

(1) **k-means algorithm**, where each cluster is represented by the mean value of the objects in the cluster, and

(2) **k-medoids algorithm**, where each cluster is represented by one of the objects located near the centre of the cluster.

Hierarchical methods :

A hierarchical method creates a hierarchical decomposition of the given set of data objects. hierarchical method can be classified as being either agglomerative or divisive, based on how the hierarchical decomposition is formed.

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.

The divisive approach, also called 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 eventually each object is in one cluster, or a termination condition holds.

Density-based methods :

The clustering methods based on density. Continue growing a given cluster as long as the density (number of objects or data points) in the "neighborhood" exceeds some threshold. i.e each data point within a given cluster, the neighborhood of a given radius has to contain at least a minimum number of points.

Ex: OPTICS.

Density-based methods can be used to filter out noise and 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 (i.e., on the quantized space). The main advantage of this approach is its fast processing time, which is typically independent of the

number of data objects and dependent only on the number of cells in each dimension in the quantized space.

Ex : STING, CLIQUE

Model-based methods :

Model-based methods hypothesize a model for each of the clusters and find the best fit of the data to the give model. A model based algorithm may locate clusters by constructing a density function that reflect the spatial distribution of the data points.

Method	General Characteristics
Partitioning Methods	• 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	• Clustering is a hierarchical decomposition • Cannot correct erroneous merges or splits • May incorporate other techniques like microclustering or consider object “linkages”
Density-Based Methods	• 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	• Use a multi resolution grid data structure • Fact processing time

Table : Overview of Clustering Methods

6.2. Partitioning Methods

– Given a database of n objects and k , the number of clusters to form , a partitioning algorithm organizes the objects into k partitions(k<=n), where each partition represents a cluster.

– The clusters are formed to optimize an objective-partitioning criterion, often called a similarity function, such as distance, so that the objects within a cluster are "similar," whereas the objects of different clusters are "dissimilar“ in terms of the database attributes.

6.2.1. k-Means: A Centroid-Based Technique

The most well known and commonly used partitioning methods are k-means, k- medoids, and their variations.

– The k-means algorithm takes the input parameter, k, and partitions a set of n objects into k clusters so that the resulting intra cluster similarity is high but the inter cluster similarity is low.

– Cluster similarity is measured in regard to the mean value of the objects in a cluster, which can be viewed as the cluster's center of gravity.

The k-means algorithm proceeds as follows:

First, it randomly selects k of the objects, each of which initially represents a cluster mean or center . For each of the remaining objects, an object is assigned to the cluster to which it is the most similar, based on the distance between the object and the cluster mean.

– It then computes the new mean for each cluster. This process iterates until the criterion function converges.

– Typically, the squared-error criterion is used, defined as

$$E = \sum_{i=1}^{k} \sum_{p \in C_i} |p - m_i|^2,$$

– Where E is the sum of the square error for all objects in the data set; p is the point in space representing a given object; and m_i is the mean of cluster C_i (both p and m_i are multidimensional).

– In other words, for each object in each cluster, the distance from the object to its cluster center is squared, and the distances are summed.

– This criterion tries to make the resulting k clusters as compact and as separate as possible. The k-means procedure is follows as

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 :

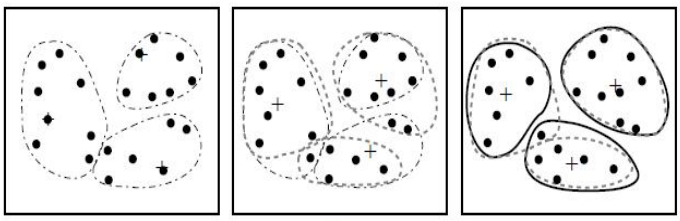
k: the number of clusters.

D: a data set containing n objects.

Output : A set of k clusters.

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 objects is the most similar, based on the mean value of the objects in the cluster.
4. Update the cluster means, that is, calculate the mean value of the object for each cluster
5. **Until** no change



7.3 Clustering of a set of objects based on the *k*-means method. (The mean of each cluster is marked by a "+".)

- The algorithm attempts to determine *k partitions that minimize the square-error* function. It works well when the clusters are compact clouds that are rather well separated from one another.

- **Strength:**

  Efficient and scalable: O(tkn), where n is number of objects, k is number of clusters, and t is number of iterations. Normally, k, t << n.

- **Weakness**
  - Applicable only to objects in a continuous n-dimensional space
  - Need to specify the number of clusters k , in advance.
  - Sensitive to noisy data and outliers.
  - Not suitable to discover clusters with non-convex shapes and clusters of very different size.
- There are several variants of *K*-means to overcome its weaknesses
  - *K*-Medoids: resistance to noise and/or outliers
  - *K*-Modes: extension to categorical data clustering analysis
  - CLARA: extension to deal with large data sets
  - Mixture models (EM-Expectation Maximization algorithm): handling uncertainty of clusters

6.2.2. k-Medodis: A Representative Object-Based Technique

k-Means algorithm is sensitive to outliers . Since an object with an extremely large value may substantially distort the distribution of the data.

- Medoid – the most centrally located point in a cluster, as a representative point of the cluster.

- The partitioning method is then performed based on the principle of minimizing the sum of the dissimilarities between each object and its corresponding reference point.

- That is, an absolute-error criterion is used, defined as

$$E = \sum_{j=1}^{k} \sum_{p \in C_j} |p - o_j|,$$

- Where *E is the sum of the absolute error for all objects in the data set; p is the point in* space representing a given object in cluster C_j; *and* o_j *is the representative object of* C_j.

- *T*he algorithm iterates until, eventually, each representative object is actually the medoid, or most centrally located object, of its cluster. This is the basis of the *k-medoids* method for grouping *n objects into k clusters.*

- The initial representative objects (or seeds) are chosen arbitrarily. The iterative process of replacing representative objects (the medoids) by non medoid objects continues as long as the quality of the resulting clustering is improved.

– This quality is estimated using a cost function that measures the average dissimilarity between an object and the representative object of its cluster. To determine whether a non representative object, o_{random}, *is a good replacement for a current representative* object, o_j, *the following four cases are examined for each of the non medoid objects, p, as illustrated in Figure.*

–

- Case 1: p currently belongs to representative object, o_j. If o_j is replaced by o_{random} as a representative object and p is closest to one of the other representative objects, o_i, $i \neq j$, then p is reassigned to o_i.
- Case 2: p currently belongs to representative object, o_j. If o_j is replaced by o_{random} as a representative object and p is closest to o_{random}, then p is reassigned to o_{random}.
- Case 3: p currently belongs to representative object, o_i, $i \neq j$. If o_j is replaced by o_{random} as a representative object and p is still closest to o_i, then the assignment does not change.
- Case 4: p currently belongs to representative object, o_i, $i \neq j$. If o_j is replaced by o_{random} as a representative object and p is closest to o_{random}, then p is reassigned to o_{random}.

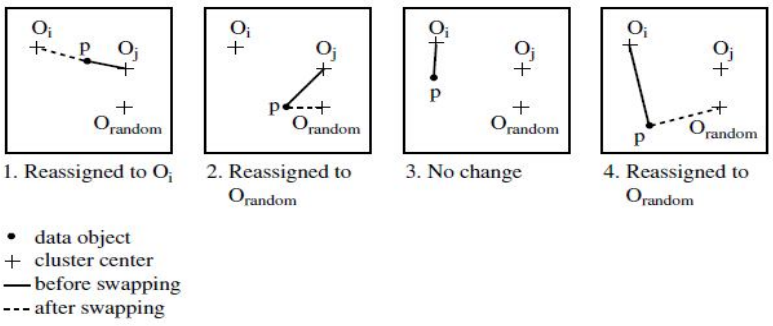


Figure 7.4 Four cases of the cost function for k-medoids clustering.

Algorithm k-medoids. PAM, a k-medoids algorithm for partitioning, or central objects.

Input :

k: the number of clusters.

D: a data set containing n objects.

Output : A set of k clusters.

Method :

1. Arbitrarily choose k objects from D as the initial representative objects or seeds.
2. **Repeat**
3. assign each remaining object to the cluster with the nearest representative objects
4. randomly select a non representative object o_{random}
5. Compute the total cost, S, of swapping representative object , O_j,with o_{random}
6. If S<0 then swap O_j with o_{random} to form the new set of k representative objects
7. **Until** no change

6.3. Hierarchical Methods

A **hierarchical clustering method** works by grouping data objects into a hierarchy or “tree” of clusters. Representing data objects in the form of a hierarchy is useful for data summarization and visualization. For example, as the manager of human resources at *AllElectronics* you may organize your employees into major groups such as executives, managers, and staff. You can further partition these groups into smaller subgroups. For instance, the general group of staff can be further divided into subgroups of senior officers, officers, and trainees. All these groups forma hierarchy.We can easily summarize or characterize the data that are organized into a hierarchy, which can be used to find, say, the average salary of managers and of officers.

Consider handwritten character recognition as another example. A set of handwriting samples may be first partitioned into general groups where each group corresponds to a unique character. Some groups can be further

partitioned into subgroups since a character may be written in multiple substantially different ways. If necessary, the hierarchical partitioning can be continued recursively until a desired granularity is reached.

a hierarchy for *AllElectronics* employees structured on, say, salary. In the study of evolution, hierarchical clustering may group animals according to their biological features to uncover evolutionary paths, which are a hierarchy of species. As another example, grouping configurations of a strategic game (e.g., chess or checkers) in a hierarchical way may help to develop game strategies that can be used to train players.

Hierarchical clustering methods can encounter difficulties regarding the selection of merge or split points. Such a decision is critical, because once a group of objects is merged or split, the process at the next step will operate on the newly generated clusters. It will neither undo what was done previously, nor perform object swapping between clusters. Thus, merge or split decisions, if not well chosen, may lead to low-quality clusters. Moreover, the methods do not scale well because each decision of merge or split needs to examine and evaluate many objects or clusters.

6.3.1. Agglomerative versus Divisive Hierarchical Clustering

A hierarchical clustering method can be either *agglomerative* or *divisive*, depending on whether the hierarchical decomposition is formed in a bottom-up (merging) or topdown (splitting) fashion.

An **agglomerative hierarchical clustering method** uses a bottom-up strategy. It typically starts by letting each object form its own cluster and iteratively merges clusters into larger and larger clusters, until all the objects are in a single cluster or certain termination conditions are satisfied. The single cluster becomes the hierarchy's root. For the merging step, it finds the two clusters that are closest to each other (according to some similarity measure), and combines the two to form one cluster. Because two clusters are merged per

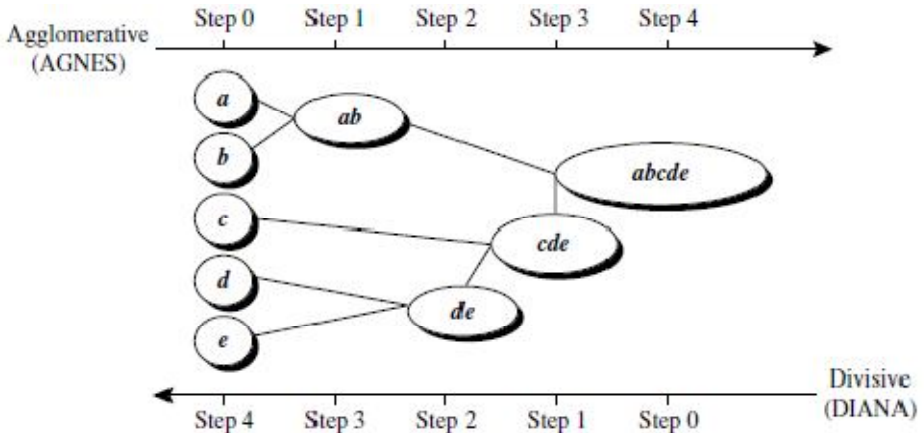
iteration, where each cluster contains at least one object, an agglomerative method requires at most *n* iterations.

A **divisive hierarchical clustering method** employs a top-down strategy. It starts by placing all objects in one cluster, which is the hierarchy's root. It then divides the root cluster into several smaller subclusters, and recursively partitions those clusters into smaller ones. The partitioning process continues until each cluster at the lowest level is coherent enough—either containing only one object, or the objects within a cluster are sufficiently similar to each other. In either agglomerative or divisive hierarchical clustering, a user can specify the desired number of clusters as a termination condition.

Example : **Agglomerative versus divisive hierarchical clustering.**

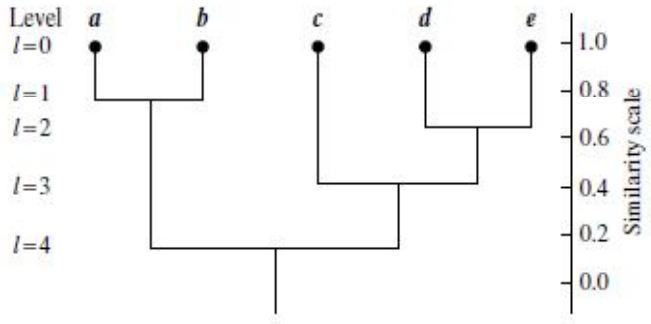
AGNES (AGglomerative NESting), an agglomerative hierarchical clustering method, and **DIANA** (DIvisive ANAlysis), a divisive hierarchical clustering method, on a data set of five objects, f***a***,***b***, ***c***,***d***, ***e***g. Initially, AGNES, the agglomerative method, places each object into a cluster of its own. The clusters are then merged step-by-step according to some criterion. For example, clusters *C*1 and *C*2 may be merged if an object in *C*1 and an object in *C*2 form the minimum Euclidean distance between any two objects from different clusters. This is a **single-linkage** approach in that each cluster is represented by all the objects in the cluster, and the similarity between two clusters is measured by the similarity of the *closest* pair of data points belonging to different clusters. The cluster-merging process repeats until all the objects are eventually merged to form one cluster.

DIANA, the divisive method, proceeds in the contrasting way. All the objects are used to form one initial cluster. The cluster is split according to some rinciple such as the maximum Euclidean distance between the closest neighboring objects in the cluster. The cluster-splitting process repeats until, eventually, each new cluster contains only a single object.



Agglomerative and divisive hierarchical clustering on data objects $\{a, b, c, d, e\}$.

A tree structure called a **dendrogram** is commonly used to represent the process of hierarchical clustering. It shows how objects are grouped together (in an agglomerative method) or partitioned (in a divisive method) step-by-step. a dendrogram for the five objects presented, where *l* D 0 shows the five objects as singleton clusters at level 0. At *l* D 1, objects ***a*** and ***b*** are grouped together to form the first cluster, and they stay together at all subsequent levels.We can also use a vertical axis to show the similarity scale between clusters. For example, when the similarity of two groups of objects, f***a***,***b***g and f***c***,***d***, ***e***g, is roughly 0.16, they are merged together to form a single cluster.



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

A challenge with divisive methods is how to partition a large cluster into several smaller ones. For example, there are $2^{n-1}-1$ possible ways to partition a set of *n*

objectsinto two exclusive subsets, where *n* is the number of objects. When *n* is large, it is computationally prohibitive to examine all possibilities.

6.4. Density-Based Methods

– To discover the clusters with arbitrary shape density based clusters are used.

– These typically regard clusters as dense regions or objects in the data space that are separated by regions of low denisty (representing noise).

– Various Density-Based clustering methods are
  - DBSCAN
  - OPTICS
  - DENCLUE

6.4.1. DBSCAN: Density-Based Clustering Based on Connected Regions with High Density:

A Density-Based Clustering Method Based on Connected Regions with Sufficiently High Density.

– DBSCAN (Density-Based Spatial Clustering of Applications with Noise) is a density based clustering algorithm.

– The algorithm grows regions with sufficiently high density into clusters and discovers clusters of arbitrary shape in spatial databases with noise.

– It defines a cluster as a maximal set of *density-connected points.*

– The basic ideas of density-based clustering involve a number of new definitions.

- The neighborhood within a radius ε of a given object is called the **ε-neighborhood** of the object.
- If the ε-neighborhood of an object contains at least a minimum number, *MinPts*, of objects, then the object is called a **core object**.
- Given a set of objects, D, we say that an object p is **directly density-reachable** from object q if p is within the ε-neighborhood of q, and q is a core object.
- An object p is **density-reachable** from object q with respect to ε and *MinPts* in a set of objects, D, if there is a chain of objects $p_1, \ldots, p_n$, where $p_1 = q$ and $p_n = p$ such that p_{i+1} is directly density-reachable from p_i with respect to ε and *MinPts*, for $1 \leq i \leq n$, $p_i \in D$.
- An object p is **density-connected** to object q with respect to ε and *MinPts* in a set of objects, D, if there is an object $o \in D$ such that both p and q are density-reachable from o with respect to ε and *MinPts*.

– Density- reachability and density connectivity. Consider Figure for a given e (epsilon) represented by the radius of the circles, and, say, let MinPts = 3. Based on the above definitions:

– Of the labeled points , m, p, o, and r are core objects because each is in an e-neighborhood containing at least three points.

– q is directly density-reachable fromm.mis directly density-reachable from p and vice versa.

– q is (indirectly) density-reachable from p because q is directly density-reachable from m and m is directly density-reachable from p. However, p is not density-reachable from q because q is not a core object. Similarly, r and s are density-reachable from o,and o is density-reachable from r.

– *o, r, and s are all density-connected.*

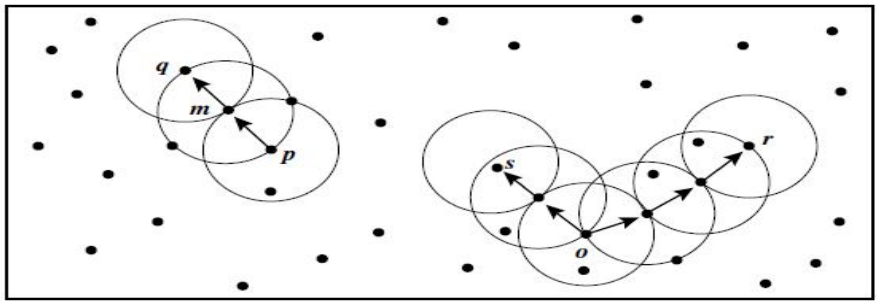


Figure 7.10 Density reachability and density connectivity in density-based clustering. Based on [EKSX96].

DBSCAN: Algorithm

Algorithm: DBSCAN: a density-based clustering algorithm

Input:

- D: a data set containing n objects
- € : the radius parameter, and
- Minpts: the neighborhood density threshold

Output: A set of density-based clusters.

Method:

1. Mark all objects as unvisited
2. **do**
3. Randomly select an unvisited object p
4. Mark p as visited
5. **if** the €-neighborhood of p has at least MinPts objects
6. Create a new cluster C, and add p to C
7. Let N be the set of objects in the €-neighborhood of P

8. **for** each point p1 in N
9. If p^1 is unvisited
10. Mark p1 as visited
11. If €-neighborhood of p^1 has at least MinPts points add those points to N
12. If p^1 is not yet a member of any cluster, add p^1 to C
13. **end for**
14. Output C
15. **else** mark p as noise
16. **until** no object is unvisited

UNIT-VI
Assignment-Cum-Tutorial Questions
SECTION-A

Objective Questions

1. The process of grouping a set of physical or abstract objects into classes of similar objects is called clustering. [T/F]
2. Clustering is a _____ type of learning. [ ]

   A) Supervised. B) Unsupervised. C) Both A & B. D) None of the above.
3. The formula for Ecludean distance d(i,j)= _____ with i =(xi1,xi2,....,xip) and j= (xj1,xj2,...,xjp) are p-Dimensional data objects.
4. The formula for manhattan distance d(i,j)= _____ with i =(xi1,xi2,....,xip) and j= (xj1,xj2,...,xjp) are p-Dimensional data objects.
5. The formula for Minkowski distance d(i,j)= _____ with i =(xi1,xi2,....,xip) and j= (xj1,xj2,...,xjp) are p-Dimensional data objects.
6. In _____ the class label of object/smaple is not known. [ ]

   A) Association rule mining. B) Classificationa.

   C) Clustering D) None of the above.
7. Main memory-base clustering algorithms uses following _____data structures. [ ]

   A) Data Matrix. B) Dissimilarity Matrix. C) Clustering Matrix. D) Both A and B.
8. Which of the following are the examples of Interval-scaled variables.

   A) Weight. B) Height. C) Weather Report. D) All of the above.[ ]
9. The Euclidean distance of an object to itself is _____ [ ]

   A) Zero. B) One. C) Two . D) Three.
10. ____ methods discovers the cluster with arbitrary shape. [ ]

    A) Partitioning. B) Hierarchical.

    C) Density-Based. D) All of the above.

SECTION-B

SUBJECTIVE QUESTIONS

1. What is cluster analysis? Explain any four requirements for clustering data.
2. Distinguish between the Binary, Nominal, Ordinal, and Ratio-Scaled variables.
3. Categorize major clustering methods.
4. Write a K-Means clustering algorithm.
5. suppose that the data mining task is to cluster the following 8 points (with (x,y) representing location) into three clusters.

   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 a center of each cluster, respectively. Use the K-Means algorithm to show only

   i. The 3 cluster centers after the first round execution.

   ii. The final 3 clusters.
6. Briefly Explain the k-Medoids clustering algorithm with an example.
7. With an example explain about the DBSCAN clustering method.