# UNIT IV

**Unsupervised Learning:** Clustering Introduction, Similarity and Distance Measures, Agglomerative Algorithms, Divisive Clustering, Minimum Spanning Tree [Reference 2]

**Association Rules:** Introduction, large Itemsets, Apriori Algorithm.

## 4.1 Introduction to Clustering

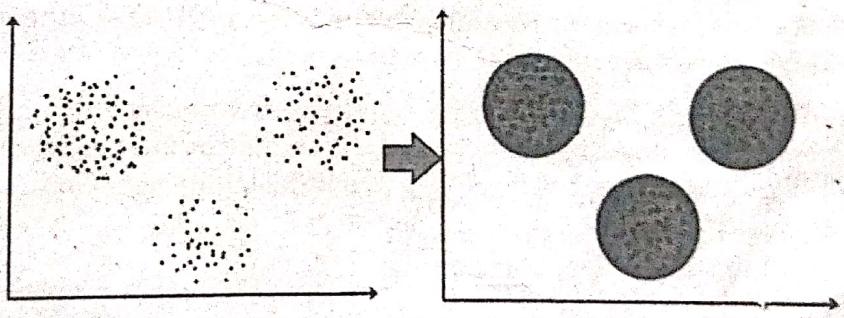
**Q1. Explain about clustering algorithms.**

*Ans :*

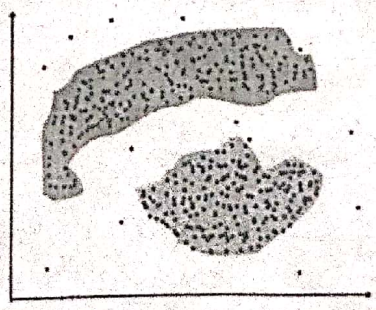
It is basically a type of unsupervised learning method. An unsupervised learning method is a method in which we draw references from datasets consisting of input data without labeled responses. Generally, it is used as a process to find meaningful structure, explanatory underlying processes, generative features, and groupings inherent in a set of examples.

Clustering is the task of dividing the population or data points into a number of groups such that data points in the same groups are more similar to other data points in the same group and dissimilar to the data points in other groups. It is basically a collection of objects on the basis of similarity and dissimilarity between them.

**For example,** The data points in the graph below clustered together can be classified into one single group. We can distinguish the clusters, and we can identify that there are 3 clusters in the below picture.



It is not necessary for clusters to be spherical. Such as:



**DBSCAN:** Density-based Spatial Clustering of Applications with Noise

These data points are clustered by using the basic concept that the data point lies within the given constraint from the cluster center. Various distance methods and techniques are used for the calculation of the outliers.

### Need of Clustering

Clustering is very much important as it determines the intrinsic grouping among the unlabelled data present. There are no criteria for good clustering. It depends on the user, what is the criteria they may use which satisfy their need. For instance, we could be interested in finding representatives for homogeneous groups (data reduction), in finding "natural clusters" and describe their unknown properties ("natural" data types), in finding useful and suitable groupings ("useful" data classes) or in finding unusual data objects (outlier detection). This algorithm must make some assumptions that constitute the similarity of points and each assumption make different and equally valid clusters.

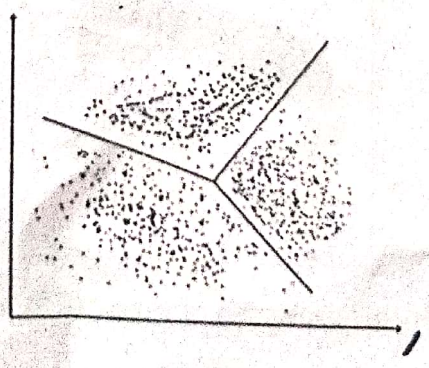
### Clustering Methods

- **Density-Based Methods:** These methods consider the clusters as the dense region having some similarities and differences from the lower dense region of the space. These methods have good accuracy and the ability to merge two clusters. Example DBSCAN (Density-Based Spatial Clustering of Applications with Noise), OPTICS (Ordering Points to Identify Clustering Structure), etc.
- **Hierarchical Based Methods:** The clusters formed in this method form a tree-type structure based on the hierarchy. New clusters are formed using the previously formed one. It is divided into two category
- Agglomerative (bottom-up approach)
- Divisive (top-down approach)

  **Examples:** CURE (Clustering Using Representatives), BIRCH (Balanced Iterative Reducing Clustering and using Hierarchies), etc.
- **Partitioning Methods:** These methods partition the objects into k clusters and each partition forms one cluster. This method is used to optimize an objective criterion similarity function such as when the distance is a major parameter example K-means, CLARANS (Clustering Large Applications based upon Randomized Search), etc.
- **Grid-based Methods:** In this method, the data space is formulated into a finite number of cells that form a grid-like structure. All the clustering operations done on these grids are fast and independent of the number of data objects example STING (Statistical Information Grid), wave cluster, CLIQUE (CLustering In Quest), etc.

### Clustering Algorithms

K-means clustering algorithm – It is the simplest unsupervised learning algorithm that solves clustering problem.K-means algorithm partitions n observations into k clusters where each observation belongs to the cluster with the nearest mean serving as a prototype of the cluster.



**Applications of Clustering in different fields:**

- **Marketing:** It can be used to characterize & discover customer segments for marketing purposes.
- **Biology:** It can be used for classification among different species of plants and animals.
- **Libraries:** It is used in clustering different books on the basis of topics and information.
- **Insurance:** It is used to acknowledge the customers, their policies and identifying the frauds.

City Planning: It is used to make groups of houses and to study their values based on their geographical locations and other factors present.

Earthquake studies: By learning the earthquake-affected areas we can determine the dangerous zones.

## 4.2 Similarity and Distance Measures, Agglomerative Algorithms

**Q2. Explain about smilarity and distance measures, Agglomerative Algorithms.**

*Ans :*

Before diving into the Agglomerative algorithms, we must understand the different concepts in clustering techniques. So, first, look at the concept of Clustering in Machine Learning:

Clustering is the broad set of techniques for finding subgroups or clusters on the basis of characterization of objects within dataset such that objects with groups are similar but different from the object of other groups. Primary guideline of clustering is that data inside a cluster should be very similar to each other but very different from those outside clusters. There are different types of clustering techniques like Partitioning Methods, Hierarchical Methods and Partitioning Methods.

| Method | Characteristics |
|---|---|
| **Partitioning Method** | • Uses mean/mediod to represent cluster centre<br>• Adopts distance-based approach to refine clusters<br>• Finds mutually exclusive clusters of spherical / nearly spherical shape<br>• Effective for datasets of small – medium age |
| **Hierarchical Method** | • Creates tree-like structure through decomposition. Uses distance between nearest / furthest points in neighboring clusters for refinement. Error can't be corrected at subsequent levels |
| **Density Based Method** | • Useful for identifying arbitarily shaped clusters<br>• May filter out outliers |

In Partitioning methods, there are 2 techniques namely, k-means and k-mediods technique (partitioning around mediods algorithm). But in order to learn about the Agglomerative Methods, we have to discuss the hierarchical methods.

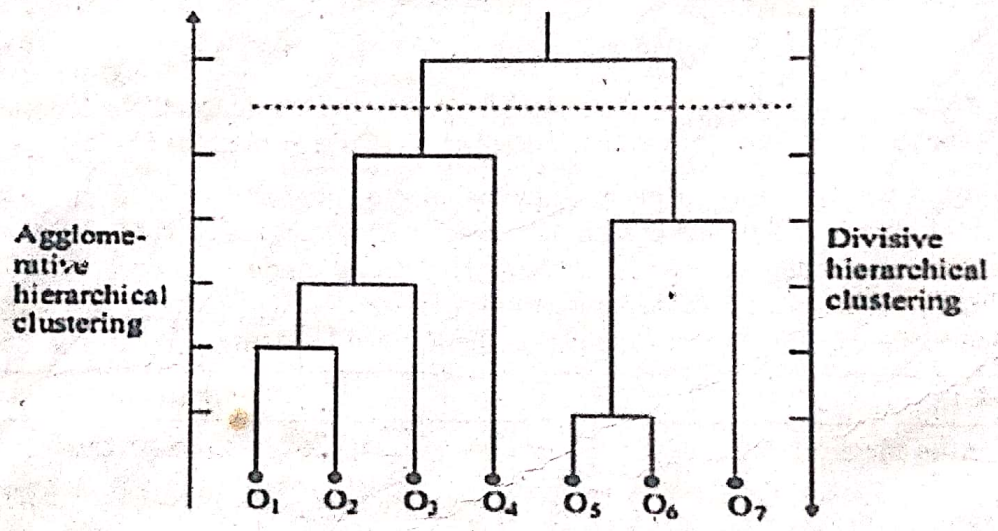
- **Hierarchical Methods:** Data is grouped into a tree like structure. There are two main clustering algorithms in this method:
  - **A. Divisive Clustering:** It uses the top-down strategy, the starting point is the largest cluster with all objects in it and then split recursively to form smaller and smaller clusters. It terminates when the user-defined condition is achieved or final clusters contain only one object.

- **B. Agglomerative Clustering:** It uses a bottom-up approach. It starts with each object forming its own cluster and then iteratively merges the clusters according to their similarity to form large clusters. It terminates either

➢ When certain clustering condition imposed by user is achieved or

➢ All clusters merge into a single cluster

A dendrogram, which is a tree like structure, is used to represent hierarchical clustering. Individual objects are represented by leaf nodes and the clusters are represented by root nodes. A representation of dendrogram is shown in this figure:




## 1. Agglomerative Algorithm: Single Link

Single-nearest distance or single linkage is the agglomerative method that uses the distance between the closest members of the two clusters. We will now solve a problem to understand it better:

Use Euclidean distance and draw the dendrogram.

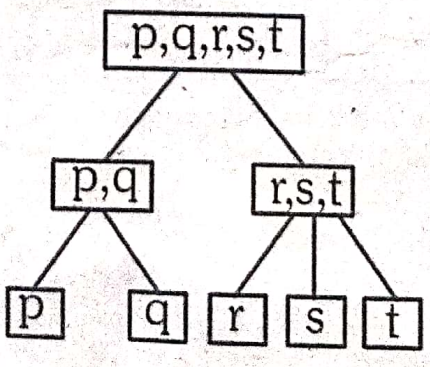



Fig.: Single link of Agglomerative Algorithm

## 4.3 Divisive Clustering

**Q3. Write about Divisive Clustering.**

*Ans :*

In data mining and statistics, hierarchical clustering analysis is a method of cluster analysis that seeks to build a hierarchy of clusters i.e. tree-type structure based on the hierarchy. Basically, there are two types of hierarchical cluster analysis strategies –1. Agglomerative Clustering: Also known as bottom-up approach or hierarchical agglomerative clustering (HAC). A structure that is more informative than the unstructured set of clusters returned by flat clustering. This clustering algorithm does not require us to prespecify the number of clusters. Bottom-up algorithms treat each data as a singleton cluster at the outset and then successively agglomerates pairs of clusters until all clusters have been merged into a single cluster that contains all data.

Algorithm : given a dataset ($d_1$, $d_2$, $d_3$, ....$d_N$) of size N

# compute the distance matrix

for i=1 to N:

# as the distance matrix is symmetric about

# the primary diagonal so we compute only lower

# part of the primary diagonal

for j=1 to i:

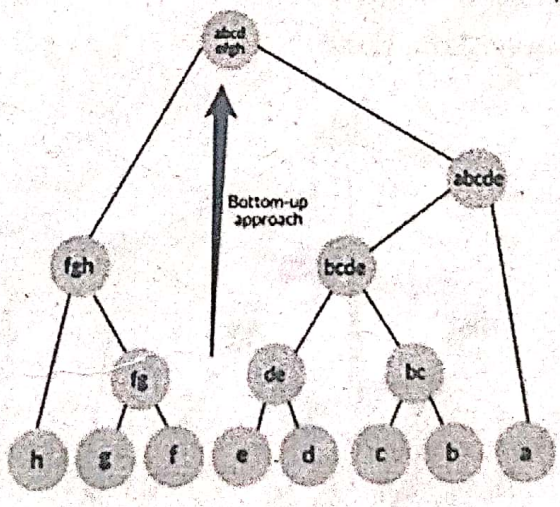
dis_mat[i][j] = distance[$d_i$, $d_j$]

each data point is a singleton cluster

**repeat**

merge the two cluster having minimum distance

update the distance matrixuntil only a single cluster remains




Hierarchical agglomerative clustering

Python implementation of the above algorithm using the scikit-learn library:

```
from sklearn.cluster import Agglomerative Clustering
import numpy as np
# randomly chosen dataset
```

```
X = np.array([[1, 2], [1, 4], [1, 0],
[4, 2], [4, 4], [4, 0]])
# here we need to mention the number of clusters
# otherwise the result will be a single cluster
# containing all the data
clustering = AgglomerativeClustering(n_clusters = 2).fit(X)
# print the class labels
print(clustering.labels_)
```

**Output :**

[1, 1, 1, 0, 0, 0]

**Divisive clustering**

Also known as a top-down approach. This algorithm also does not require to prespecify the number of clusters. Top-down clustering requires a method for splitting a cluster that contains the whole data and proceeds by splitting clusters recursively until individual data have been split into singleton clusters.

**Algorithm :**

given a dataset ($d_1$, $d_2$, $d_3$, ....$d_N$) of size N

at the top we have all data in one cluster

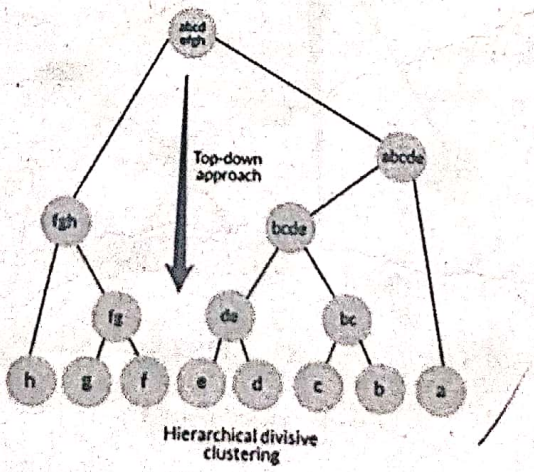
the cluster is split-using a flat clustering method eg. K-Means etc

**repeat**

choose the best cluster among all the clusters to

splitsplit that cluster by the flat clustering algorithm

until each data is in its own singleton cluster




Hierarchical divisive clustering

**Hierarchical Agglomerative vs Divisive clustering:**

Divisive clustering is more complex as compared to agglomerative clustering, as in the case of divisive clustering we need a flat clustering method as "subroutine" to split each cluster until we have each data having its own singleton cluster.

Divisive clustering is more efficient if we do not generate a complete hierarchy all the way down to individual data leaves. The time complexity of a naive agglomerative clustering is $O(n^3)$ because we exhaustively scan the N x N matrix dist_mat for the lowest distance in each of N-1 iterations. Using priority queue data structure we can reduce this complexity to $O(n^2 \log n)$. By using some more optimizations it can be brought down to $O(n^2)$. Whereas for divisive clustering given a fixed number of top levels, using an efficient flat algorithm like K-Means, divisive algorithms are linear in the number of patterns and clusters.

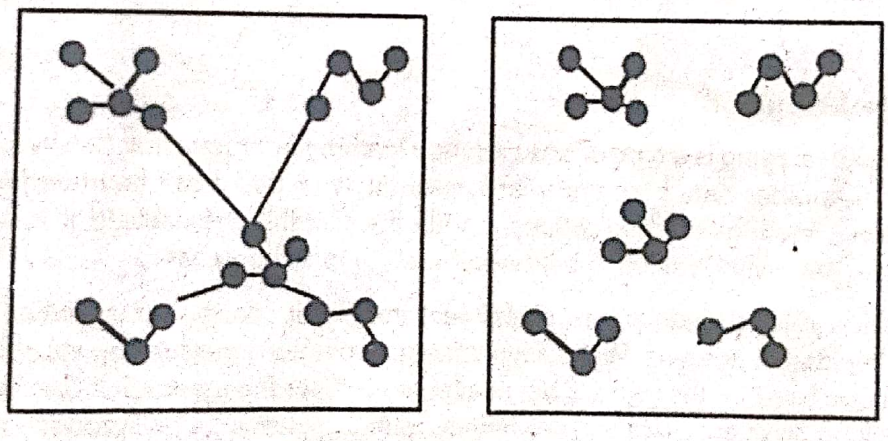
A divisive algorithm is also more accurate. Agglomerative clustering makes decisions by considering the local patterns or neighbor points without initially taking into account the global distribution of data. These early decisions cannot be undone. whereas divisive clustering takes into consideration the global distribution of data when making top-level partitioning decisions.

## 4.4 Minimum Spanning Tree Clustering

**Q4. Write about Minimum Spanning tree Clustering ?**

*Ans :*

Building MST (Minimum Spanning Tree) is a method for constructing hierarchy of clusters.It starts with a tree that consists of a point pp. In successive steps, look for the closest pair of points (p,q)(p;q) such that pp is in the current tree but qq is not. With this closest pair of points (p,q)(p,q), add qq to the tree and put an edge between pp and qq.



(a) An MST connecting all the data points

(b) Clusters after removal of longest edges

The procedure of constructing hierarchy of clusters using MST would be as follows:

Construct a MST as a proximity graph

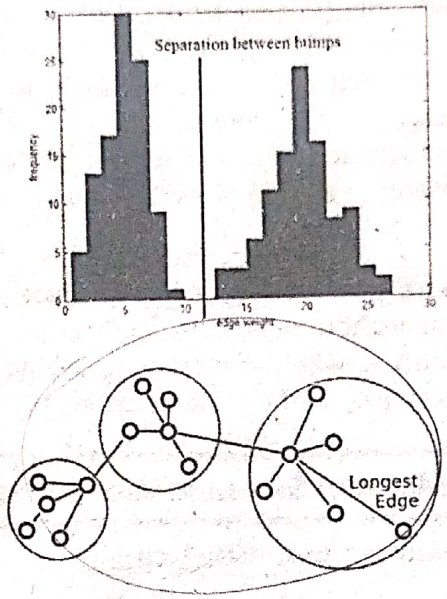
repeat

Split a cluster by breaking the inconsistent edge:

until Only singleton clusters remain

Note that the inconsistent edge is the link of the largest distance (smallest similarity).

The definition of inconsistency varies. For example, we can also use local inconsistency remove edges significantly larger than their neighborhood edges




## 4.5 Association Rules

### 4.5.1 Introduction, large Itemsets, Apriori Algorithm.

**Q5. Explain the following association rules learning, large intemsets and aprise algorithm.**

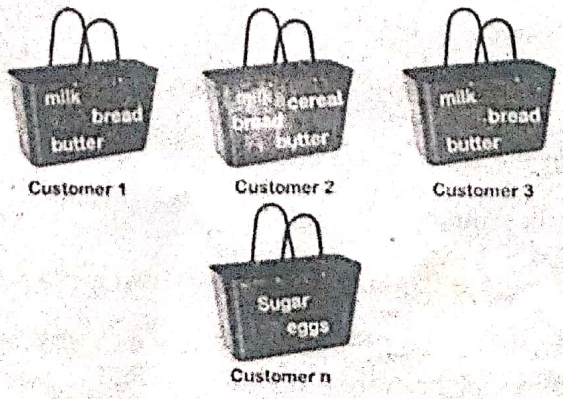
*Ans :*

**Association Rule Learning:**

Association rule learning is a type of unsupervised learning technique that checks for the dependency of one data item on another data item and maps accordingly so that it can be more profitable. It tries to find some interesting relations or associations among the variables of dataset. It is based on different rules to discover the interesting relations between variables in the database.

The association rule learning is one of the very important concepts of machine learning and it is employed in Market Basket analysis, Web usage mining, continuous production, etc. Here market basket analysis is a technique used by the various big retailer to discover the associations between items. We can understand it by taking an example of a supermarket, as in a supermarket, all products that are purchased together are put together.

For example, if a customer buys bread, he most likely can also buy butter, eggs, or milk, so these products are stored within a shelf or mostly nearby. Consider the below diagram:



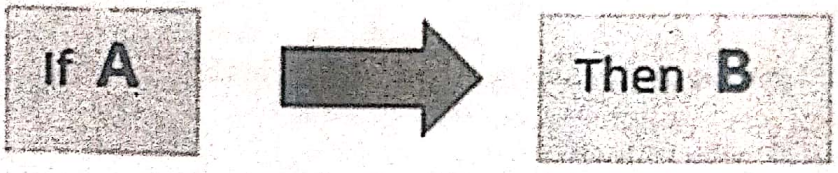
Association rule learning can be divided into three types of algorithms:

1. Apriori
2. Eclat
3. F-P Growth Algorithm

We will understand these algorithms in later chapters.

How does Association Rule Learning work?

Association rule learning works on the concept of If and Else Statement, such as if A then B.




Here the If element is called antecedent, and then statement is called as Consequent. These types of relationships where we can find out some association or relation between two items is known as single cardinality. It is all about creating rules, and if the number of items increases, then cardinality also increases accordingly. So, to measure the associations between thousands of data items, there are several metrics. These metrics are given below:

- Support
- Confidence
- Lift

**Let's understand each of them:**

**Support**

Support is the frequency of A or how frequently an item appears in the dataset. It is defined as the fraction of the transaction T that contains the itemset X. If there are X datasets, then for transactions T, it can be written as:

$$\text{Supp}(X) = \frac{\text{Freq}(X)}{T}$$

**Confidence**

Confidence indicates how often the rule has been found to be true. Or how often the items X and Y occur together in the dataset when the occurrence of X is already given. It is the ratio of the transaction that contains X and Y to the number of records that contain X.

$$\text{Confidence} = \frac{\text{Freq}(X, Y)}{\text{Freq}(X)}$$

**Lift**

It is the strength of any rule, which can be defined as below formula:

$$\text{Lift} = \frac{\text{Supp}(X, Y)}{\text{Supp}(X) \times \text{Supp}(Y)}$$

It is the ratio of the observed support measure and expected support if X and Y are independent of each other. It has three possible values:

- **If Lift= 1:** The probability of occurrence of antecedent and consequent is independent of each other.
- **Lift>1:** It determines the degree to which the two itemsets are dependent to each other.
- **Lift<1:** It tells us that one item is a substitute for other items, which means one item has a negative effect on another.

**Q6. What are the types of association rule learning?**

*Ans :*

Association rule learning can be divided into three algorithms:

**1. Apriori Algorithm**

This algorithm uses frequent datasets to generate association rules. It is designed to work on the databases that contain transactions.

This algorithm uses frequent datasets to generate association rules. It is designed to work on the databases that contain transactions. This algorithm uses a breadth-first search and Hash Tree to calculate the item set efficiently.

It is mainly used for market basket analysis and helps to understand the products that can be bought together. It can also be used in the healthcare field to find drug reactions for patients.

**2. Eclat Algorithm**

Eclat algorithm stands for Equivalence Class Transformation. This algorithm uses a depth-first search technique to find frequent itemsets in a transaction database. It performs faster execution than Apriori Algorithm.

**3. F-P Growth Algorithm**

The F-P growth algorithm stands for Frequent Pattern, and it is the improved version of the Apriori Algorithm. It represents the database in the form of a tree structure that is known as a frequent pattern/or tree. The purpose of this frequent tree is to extract the most frequent patterns.

Applications of Association Rule Learning

It has various applications in machine learning and data mining. Below are some popular applications of association rule learning:

- **Market Basket Analysis:** It is one of the popular examples and applications of association rule mining. This technique is commonly used by big retailers to determine the association between items.
- **Medical Diagnosis:** With the help of association rules, patients can be cured easily, as it helps in identifying the probability of illness for a particular disease.
- **Protein Sequence:** The association rules help in determining the synthesis of artificial Proteins.
- It is also used for the Catalog Design and Loss-leader Analysis and many more other applications.

# Short Question and Answers

**1. What is clustering algorithms?**

*Ans :*

It is basically a type of unsupervised learning method. An unsupervised learning method is a method in which we draw references from datasets consisting of input data without labeled responses. Generally, it is used as a process to find meaningful structure, explanatory underlying processes, generative features, and groupings inherent in a set of examples.

Clustering is the task of dividing the population or data points into a number of groups such that data points in the same groups are more similar to other data points in the same group and dissimilar to the data points in other groups. It is basically a collection of objects on the basis of similarity and dissimilarity between them.

**2. Why clustering in Machine Learning?**

*Ans :*

Clustering is very much important as it determines the intrinsic grouping among the unlabelled data present. There are no criteria for good clustering. It depends on the user, what is the criteria they may use which satisfy their need. For instance, we could be interested in finding representatives for homogeneous groups (data reduction), in finding "natural clusters" and describe their unknown properties ("natural" data types), in finding useful and suitable groupings ("useful" data classes) or in finding unusual data objects (outlier detection). This algorithm must make some assumptions that constitute the similarity of points and each assumption make different and equally valid clusters.

**Clustering Methods**

- **Density-Based Methods:** These methods consider the clusters as the dense region having some similarities and differences from the lower dense region of the space. These methods have good accuracy and the ability to merge two clusters. Example DBSCAN (Density-Based Spatial Clustering of Applications with Noise), OPTICS (Ordering Points to Identify Clustering Structure), etc.
- **Hierarchical Based Methods:** The clusters formed in this method form a tree-type structure based on the hierarchy. New clusters are formed using the previously formed one. It is divided into two category
- Agglomerative (bottom-up approach)
- Divisive (top-down approach)

  **Examples:** CURE (Clustering Using Representatives), BIRCH (Balanced Iterative Reducing Clustering and using Hierarchies), etc.
- **Partitioning Methods:** These methods partition the objects into k clusters and each partition forms one cluster. This method is used to optimize an objective criterion similarity function such as when the distance is a major parameter example K-means, CLARANS (Clustering Large Applications based upon Randomized Search), etc.

- **Grid-based Methods:** In this method, the data space is formulated into a finite number of cells that form a grid-like structure. All the clustering operations done on these grids are fast and independent of the number of data objects example STING (Statistical Information Grid), wave cluster, CLIQUE (CLustering In Quest), etc.

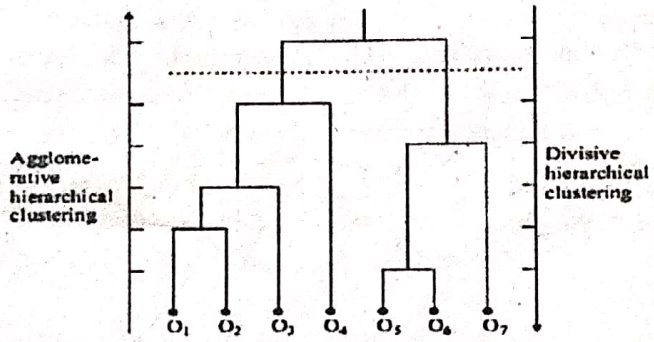
**3. Write about Agglomerative Clustering .**

*Ans :*

It uses a bottom-up approach. It starts with each object forming its own cluster and then iteratively merges the clusters according to their similarity to form large clusters. It terminates either

- When certain clustering condition imposed by user is achieved or
- All clusters merge into a single cluster

A dendrogram, which is a tree like structure, is used to represent hierarchical clustering. Individual objects are represented by leaf nodes and the clusters are represented by root nodes. A representation of dendrogram is shown in this figure:




Now we will look into the variants of Agglomerative methods:

**Agglomerative Algorithm: Single Link**

Single-nearest distance or single linkage is the agglomerative method that uses the distance between the closest members of the two clusters. We will now solve a problem to understand it better:

Question. Find the clusters using a single link technique. Use Euclidean distance and draw the dendrogram.

| Sample No. | 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 | 0.30 |

**4. Write about Divisive Clustering.**

*Ans :*

Divisive clustering also known as a top-down approach. This algorithm also does not require to prespecify the number of clusters. Top-down clustering requires a method for splitting a cluster that contains the whole data and proceeds by splitting clusters recursively until individual data have been split into singleton clusters.

**Algorithm :**

given a dataset ($d_1$, $d_2$, $d_3$, ....$d_N$) of size N

at the top we have all data in one cluster

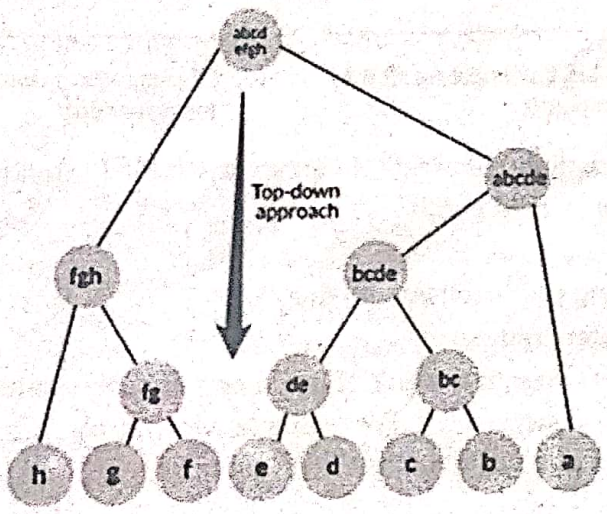
the cluster is split using a flat clustering method eg. K-Means etc

**repeat**

choose the best cluster among all the clusters to

splitsplit that cluster by the flat clustering algorithm

until each data is in its own singleton cluster




Hierarchical divisive clustering

**5. Hierarchical Agglomerative vs Divisive clustering**

*Ans :*

Divisive clustering is more complex as compared to agglomerative clustering, as in the case of divisive clustering we need a flat clustering method as "subroutine" to split each cluster until we have each data having its own singleton cluster.

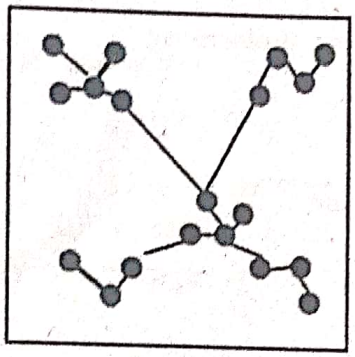
Divisive clustering is more efficient if we do not generate a complete hierarchy all the way down to individual data leaves. The time complexity of a naive agglomerative clustering is $O(n^3)$ because we exhaustively scan the N x N matrix dist_mat for the lowest distance in each of N-1 iterations. Using priority queue data structure we can reduce this complexity to $O(n^2 \log n)$. By using some more optimizations it can be brought down to $O(n^2)$. Whereas for divisive clustering given a fixed number of top levels, using an efficient flat algorithm like K-Means, divisive algorithms are linear in the number of patterns and clusters.

A divisive algorithm is also more accurate. Agglomerative clustering makes decisions by considering the local patterns or neighbor points without initially taking into account the global distribution of data. These early decisions cannot be undone, whereas divisive clustering takes into consideration the global distribution of data when making top-level partitioning decisions.

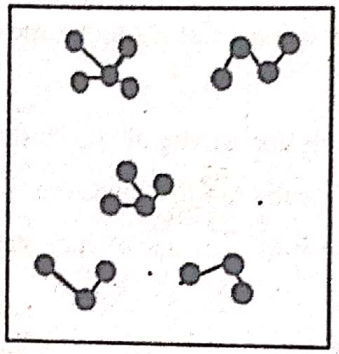
**6. Write short notes on Minimum Spanning Tree Clustering.**

*Ans :*

Building MST (Minimum Spanning Tree) is a method for constructing hierarchy of clusters.It starts with a tree that consists of a point pp. In successive steps, look for the closest pair of points (p,q)(p,q) such that pp is in the current tree but qq is not. With this closest pair of points (p,q)(p,q), add qq to the tree and put an edge between pp and qq.



(a) An MST connecting all the data points



(b) Clusters after removal of longest edges

The procedure of constructing hierarchy of clusters using MST would be as follows:

Construct a MST as a proximity graph

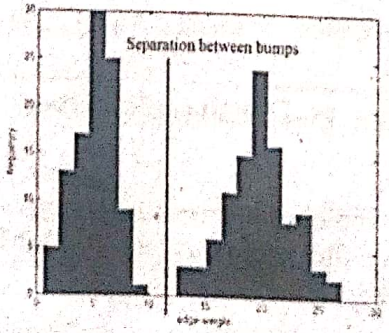
repeat

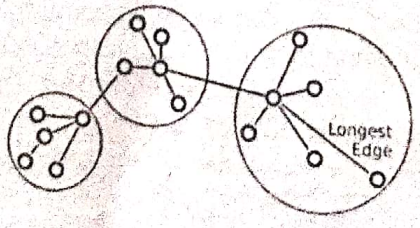
Split a cluster by breaking the inconsistent edge.

until Only singleton clusters remain

Note that the inconsistent edge is the link of the largest distance (smallest similarity).

The definition of inconsistency varies. For example, we can also use local inconsistency remove edges significantly larger than their neighborhood edges



**7. What are the types of association rule learning?**

*Ans :*

Association rule learning can be divided into three algorithms:

**1. Apriori Algorithm**

This algorithm uses frequent datasets to generate association rules. It is designed to work on the databases that contain transactions.

This algorithm uses frequent datasets to generate association rules. It is designed to work on the databases that contain transactions. This algorithm uses a breadth-first search and Hash Tree to calculate the item set efficiently.

It is mainly used for market basket analysis and helps to understand the products that can be bought together. It can also be used in the healthcare field to find drug reactions for patients.

**2. Eclat Algorithm**

Eclat algorithm stands for Equivalence Class Transformation. This algorithm uses a depth-first search technique to find frequent itemsets in a transaction database. It performs faster execution than Apriori Algorithm.

**3. F-P Growth Algorithm**

The F-P growth algorithm stands for Frequent Pattern, and it is the improved version of the Apriori Algorithm. It represents the database in the form of a tree structure that is known as a frequent pattern or tree. The purpose of this frequent tree is to extract the most frequent patterns.

Applications of Association Rule Learning

It has various applications in machine learning and data mining. Below are some popular applications of association rule learning:

- **Market Basket Analysis:** It is one of the popular examples and applications of association rule mining. This technique is commonly used by big retailers to determine the association between items.
- **Medical Diagnosis:** With the help of association rules, patients can be cured easily, as it helps in identifying the probability of illness for a particular disease.
- **Protein Sequence:** The association rules help in determining the synthesis of artificial Proteins.
- It is also used for the Catalog Design and Loss-leader Analysis and many more other applications.

# One Mark Answers

**1. Define Apriori Algorithm?**

*Ans :*

This algorithm uses frequent datasets to generate association rules. It is designed to work on the databases that contain transactions. This algorithm uses a breadth-first search and Hash Tree to calculate the item set efficiently.

**2. Define Association Rule Learning?**

*Ans :*

Association rule learning is a type of unsupervised learning technique that checks for the dependency of one data item on another data item and maps accordingly so that it can be more profitable.

**3. What is clustering?**

*Ans :*

Clustering is the task of dividing the population or data points into a number of groups such that data points in the same groups are more similar to other data points in the same group and dissimilar to the data points in other groups.

**4. What is Apriori Algorithm?**

*Ans :*

This algorithm uses frequent datasets to generate association rules. It is designed to work on the databases that contain transactions. This algorithm uses a breadth-first search and Hash Tree to calculate the item set efficiently.

**5. What is Probability density function?**

*Ans :*

Probability density is the relationship between observations and their probability. Some outcomes of a random variable will have low probability density and other outcomes will have a high probability density.