# UNIT 5:

# UNSUPERVISED LEARNING

# UNIT – V

**Unsupervised Learning :** Introduction, Clustering: Clustering as a machine learning task, Different types of clustering techniques, Partitioning methods, Hierarchical clustering, Finding Pattern using Association Rule: Definition of common terms, Association rule, The Apriority Algorithm for association rule learning.

---

## Unsupervised Learning

Unsupervised learning is a machine learning concept where the unlabeled and unclassified information is analysed to discover hidden knowledge. The algorithms work on the data without any prior training, but they are constructed in such a way that they can identify patterns, groupings, sorting order, and numerous other interesting knowledge from the set of data.

## Unsupervised VS Supervised Learning

| Category | Supervised Learning | Unsupervised Learning |
|---|---|---|
| Data | Labeled data is supplied | Unlabeled data is supplied |
| Training | Training will happen | No training will happen |
| Output | Try to learn the probability of outcome Y for particular input X and Predict the output | Find out the association between the features or their grouping to understand the nature of the data |
| Types of Algorithms | Classification and Regression | Clustering and Association Analysis |

## Applications of Unsupervised Learning

Because of its flexibility that it can work on uncategorized and unlabeled data, there are many domains where unsupervised learning finds its application. Few examples of such applications are as follows:

a. Segmentation of target consumer populations by an advertisement consulting agency on the basis of few dimensions such as demography, financial data, purchasing habits, etc. so that the advertisers can reach their target consumers efficiently.

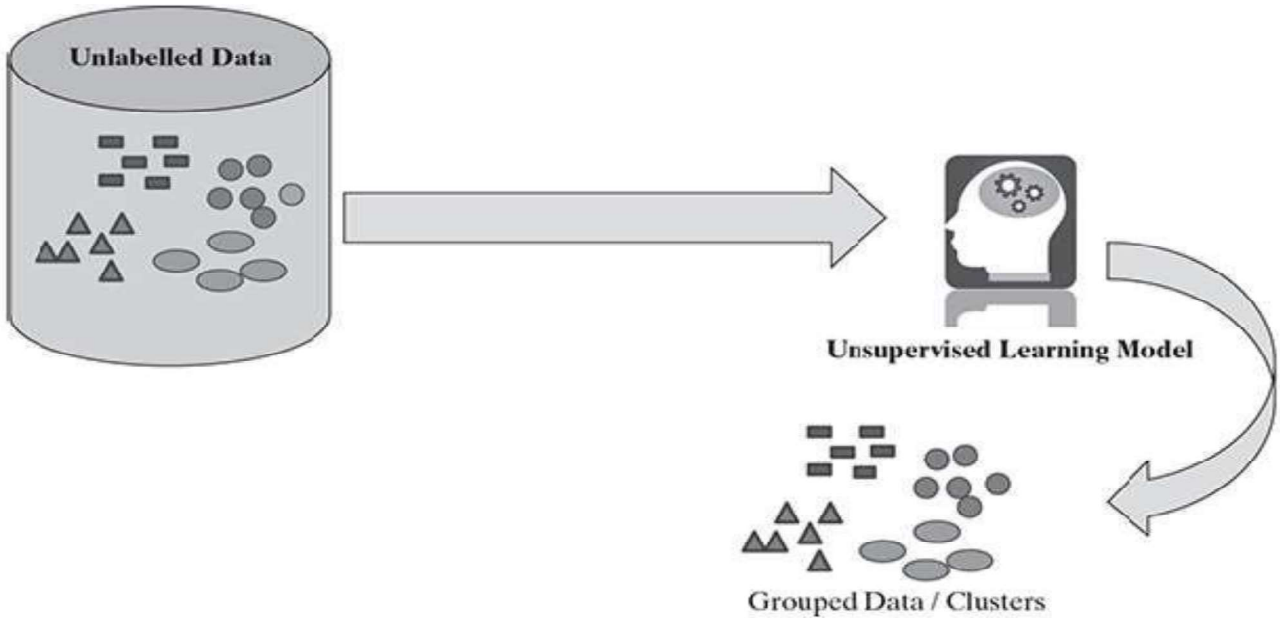
b. Anomaly or fraud detection in the banking sector by identifying the pattern of loan defaulters

c. Image processing and image segmentation such as face recognition, expression identification.

d. Grouping of important characteristics in genes to identify important influencers in new areas of genetics
e. Utilization by data scientists to reduce the dimensionalities in sample data to simplify modeling Document clustering and identifying potential labeling options

Chat bots, self-driven cars, and many more recent innovations are results of the combination of unsupervised and supervised learning.

## CLUSTERING:




Clustering refers to a broad set of techniques for finding subgroups, or clusters, in a data set on the basis of the characteristics of the objects within that data set in such a manner that the objects within the group are similar (or related to each other) but are different from (or unrelated to) the objects from the other groups.

Clustering is defined as an unsupervised machine learning task that automatically divides the data into **clusters** or groups of similar items.

The effectiveness of clustering depends on how similar or related the objects within a group are or how different or unrelated the objects in different groups are from each other.

**Uses of Cluster Analysis:** There are many different fields where cluster analysis is used effectively, such as —

a. **Text data mining:** this includes tasks such as text categorization, text clustering, document summarization, concept extraction, sentiment analysis, and entity relation modeling
b. **Customer segmentation:** creating clusters of customers on the basis of parameters such as demographics, financial conditions, buying habits, etc., which can be used by retailers and advertisers to promote their products in the correct segment

c. **Anomaly checking:** checking of anomalous behaviors such as fraudulent bank transaction, unauthorized computer intrusion, suspicious movements on a radar scanner, etc.

d. **Data mining:** simplify the data mining task by grouping a large number of features from an extremely large data set to make the analysis manageable

**Different types of clustering techniques**

The major clustering techniques are

1) Partitioning methods,

2) Hierarchical methods, and

3) Density-based methods.

**1) Partitioning methods**

Two of the most important algorithms for partitioning based clustering are

**k-means** and **k-medoid.**

**(a) K-means - A centroid-based technique**

✓ This is one of the oldest and most popularly used algorithm for clustering.

✓ The principle of the k-means algorithm is to assign each of the 'n' data points to one of the K clusters where 'K' is a user-defined parameter as the number of clusters desired.

✓ The objective is to maximize the homogeneity within the clusters and also to maximize the differences between the clusters.

**Simple algorithm of K-means:-**

**Step 1:** Select *K* points in the data space and mark them as initial centroids

**loop**

**Step 2:** Assign each point in the data space to the nearest centroid to form K clusters

**Step 3:** Measure the distance of each point in the cluster from the centroid

**Step 4:** Calculate the Sum of Squared Error (SSE) to measure the quality of the clusters *(described later in this chapter)*

**Step 5:** Identify the new centroid of each cluster on the basis of distance between points

**Step 6:** Repeat Steps 2 to 5 to refine until centroids do not change

**end loop**

**Example :-**

In the above figure, let's assume K=4 implying that we want to create four clusters out of this data set.

**Step1:** we assign four random points from the data set as the centroids, as represented by the * signs, and we assign the data points to the nearest centroid to create four clusters.

**Step2:** on the basis of the distance of the points from the corresponding centroids, the centroids are updated and points are reassigned to the updated centroids.

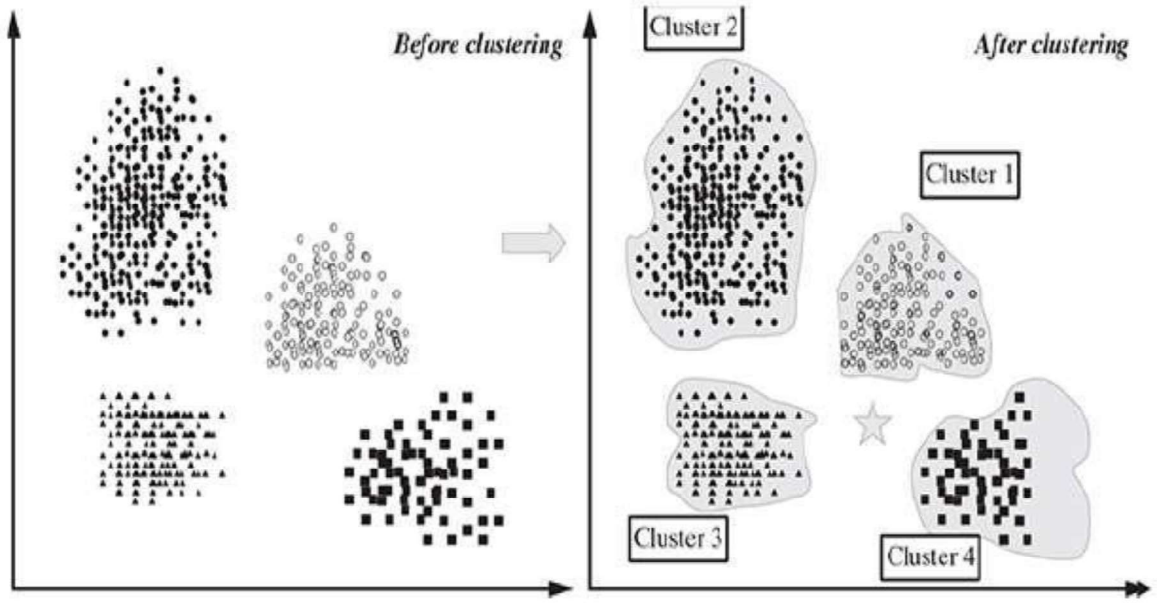
**Step3:** The iterative step is to recalculate the centroids of the data set after each iteration. The proximities of the data points from each other within a cluster is measured to minimize the distances. The measure of quality of clustering uses the SSE technique.

$$\text{SSE} = \sum_{i=1}^{k} \sum_{x \in C_i} \text{dist}\,(c_i, x)^2$$

Where, **dist()** calculates the ***Euclidean distance*** between the centroid c of the cluster C and the data points x in the cluster.

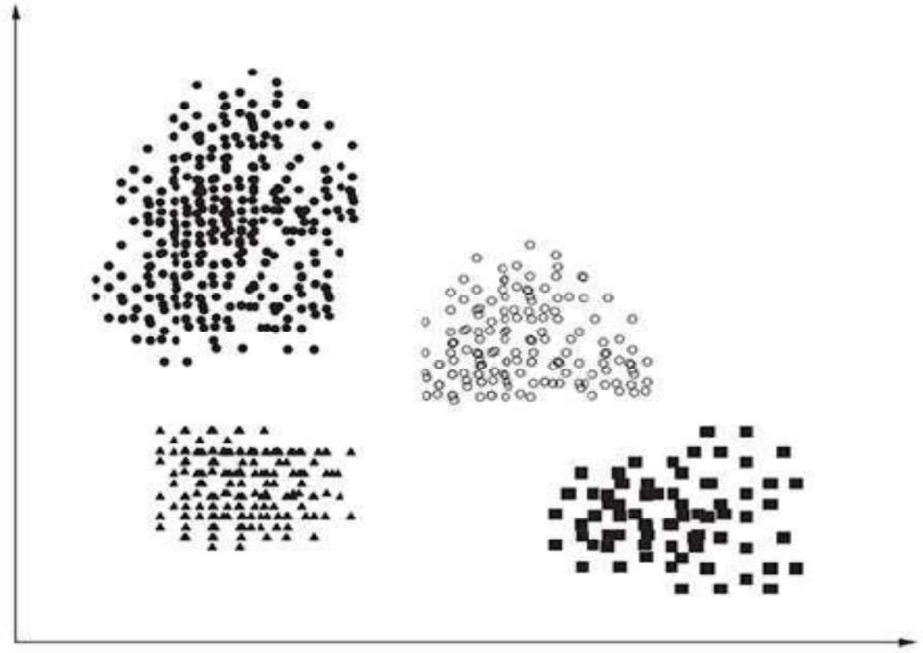
$$\text{dist}(x, y) = \sqrt{\sum_{1}^{n} (x_i - y_i)^2}$$

The lower the SSE for a clustering solution, the better is the representative position of the centroid.

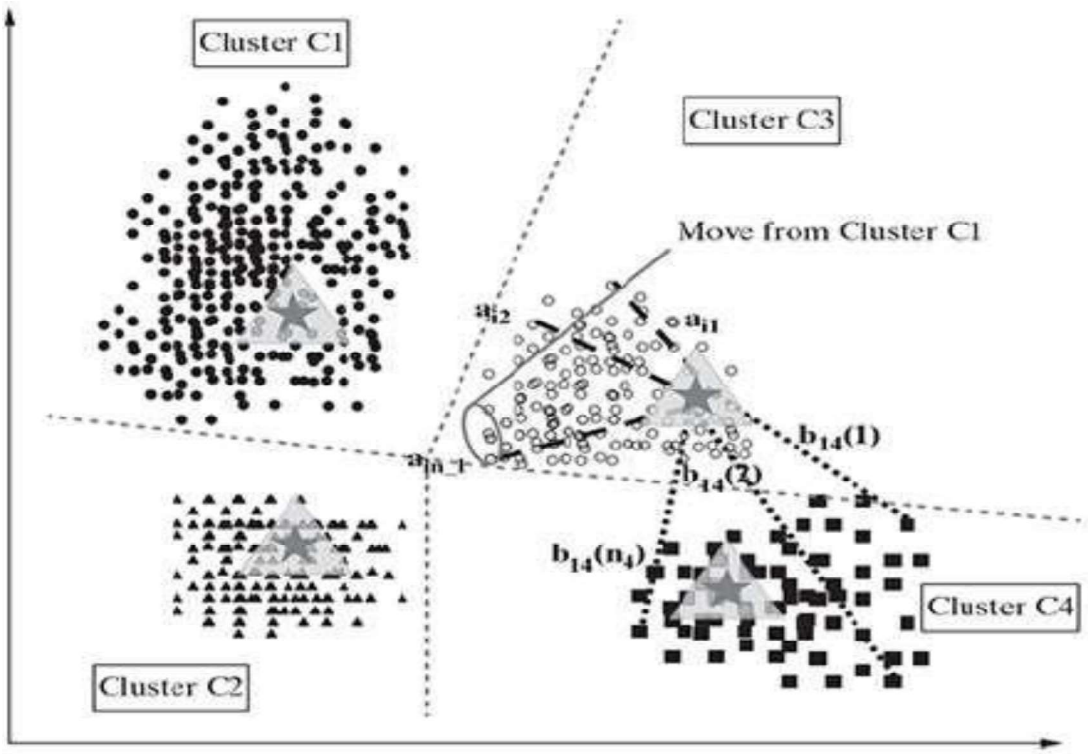



Clustering concept – before and after clustering

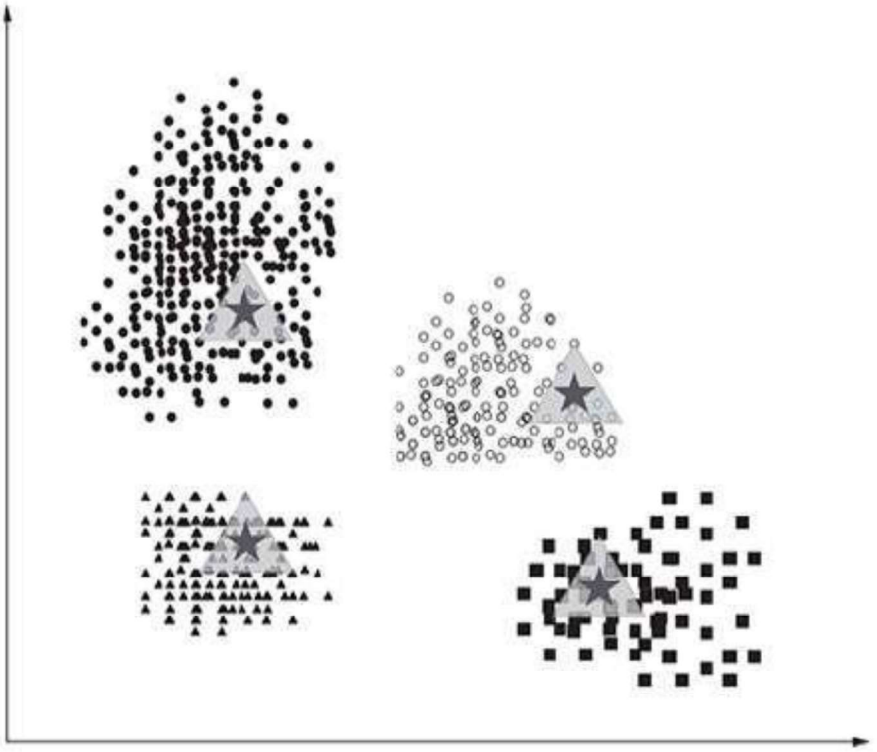
**Limitation of SSE:** One limitation of the squared error method is that in the case of presence of outliers in the data set, the squared error can distort the mean value of the clusters.



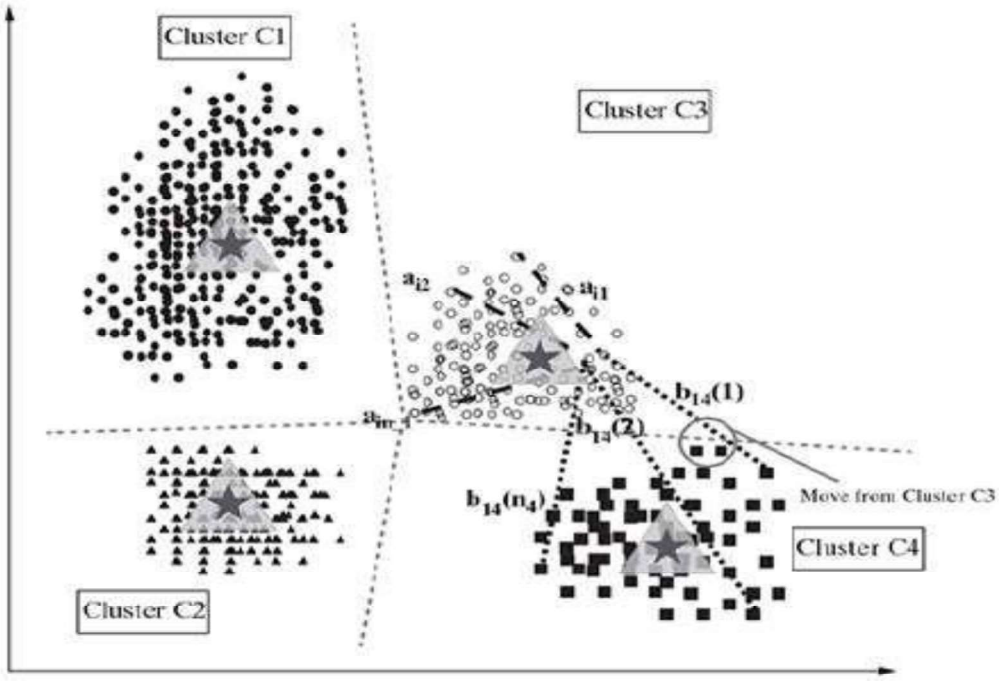
Clustering of data set



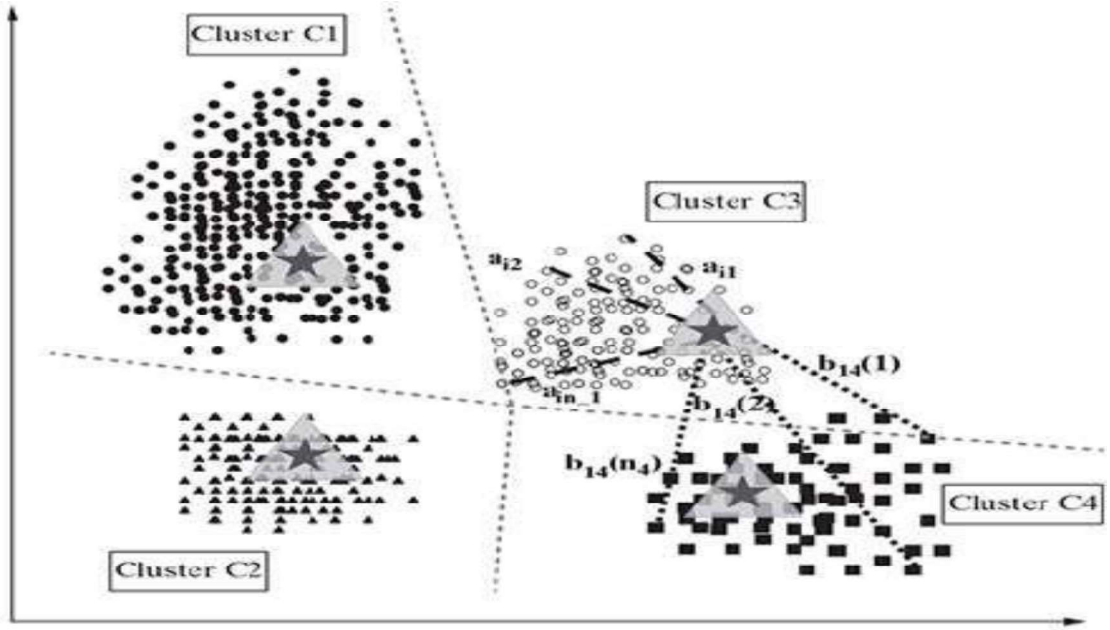

Iteration 2: Centroids recomputed and points redistributed among the clusters according to the nearest centroid



Clustering with initial centroids




Iteration 3: Final cluster arrangement: Centroids recomputed and points redistributed among the clusters according to the nearest centroid




Iteration 1: Four clusters and distance of points from the centroids

## K-means : Strengths and Weaknesses

| Strengths | Weaknesses |
|---|---|
| • The principle used for identifying the clusters is very simple and involves very less complexity of statistical terms<br>• The algorithm is very flexible and thus can be adjusted for most scenarios and complexities<br>• The performance and efficiency are very high and comparable to those of any sophisticated algorithm in term of dividing the data into useful clusters | • The algorithm involves an element of random chance and thus may not find the optimal set of cluster in some cases<br>• The starting point of guessing the number natural clusters within the data requires some experience of the user, so that the final outcome is efficient |

### (b) K-Medoids: a representative object-based technique

The k-means algorithm is sensitive to outliers in the data set and inadvertently produces skewed clusters when the means of the data points are used as centroids.

**Ex:-** Take an example of eight data points, and for simplicity, we can consider them to be 1-D data

with values 1, 2, 3, 5, 9, 10, 11, and 25. Point 25 is the outlier, and it affects the cluster formation negatively when the mean of the points is considered as centroids.

With K = 2, the initial clusters we arrived at are {1, 2,3, 6} and {9, 10, 11, 25}.

$$\text{The mean of the cluster } \{1, 2, 3, 6\} = \frac{12}{4} = 3,$$

and the mean of the cluster

$$\{9, 10, 12, 25\} = \frac{56}{4} = 14.$$

So, the SSE within the clusters is

$$(1-3)^2 + (2-3)^2 + (3-3)^2 + (6-3)^2 + (9-14)^2 + (10-14)^2 + (12-14)^2 + (25-14)^2 = 179$$

If we compare this with the cluster {1, 2, 3, 6, 9} and {10, 11, 25},

$$\text{the mean of the cluster } \{1, 2, 3, 6, 9\} = \frac{21}{5} = 4.2,$$

Because the SSE of the second clustering is lower, k-means tend to put point 9 in the same cluster with 1, 2, 3, and 6 though the point is logically nearer to points 10 and 11. This skewedness is introduced due to the outlier point 25, which shifts the mean away from the centre of the cluster.

k-medoids provides a solution to this problem. Instead of considering the mean of the data points

and the mean of the cluster

$$\{10, 12, 25\} = \frac{47}{3} = 15.67.$$

So, the SSE within the clusters is

$$(1-4.2)^2 + (2-4.2)^2 + (3-4.2)^2 + (6-4.2)^2 + (9-4.2)^2 + (10-15.67)^2 + (12-15.67)^2 + (25-15.67)^2 = 113.84$$

$$SSE = \sum_{i=1}^{k} \sum_{x \in C_i} dist\,(o_i, x)^2 \qquad (9.3)$$

where $o_i$ is the representative point or object of cluster $C_i$.

in the cluster, k-medoids considers k representative data points from the existing points in the data set as the centre of the clusters. It then assigns the data points according to their distance from these centres to form k clusters. Note that the medoids in this case are actual data points or objects from the data set and not an imaginary point as in the case when the mean of the data sets within cluster is used as the centroid in the k-means technique. The SSE is calculated as Thus, the k-medoids method groups n objects in k clusters by minimizing the SSE. Because of the use of medoids from the actual representative data points, k-medoids is less influenced by the outliers in the data.

One of the practical implementation of the k-medoids principle is the **Partitioning Around Medoids (PAM) algorithm.**

**Algorithm PAM**

**Step 1:** Randomly choose k points in the data set as the initial representative points loop

**Step 2:** Assign each of the remaining points to the cluster which has the nearest representative point

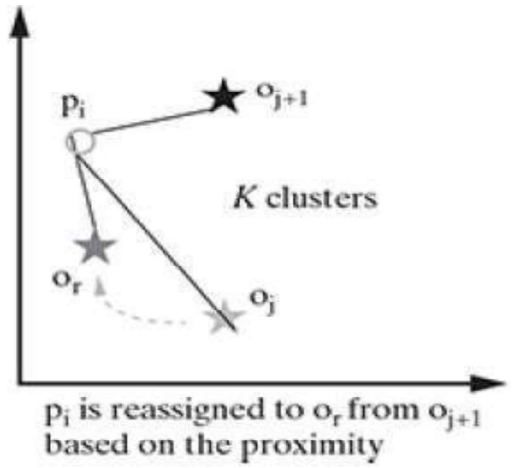
**Step 3:** Randomly select a non-representative point o in each cluster

**Step 4:** Swap the representative point o with o and compute the new SSE after swapping

**Step 5:** If SSE new < SSE old, then swap o with o to form the new set of k representative objects;

**Step 6:** Refine the k clusters on the basis of the nearest representative point. Logic continues until there is no change

end loop




. PAM algorithm: Reassignment of points to different clusters

## 2) Hierarchical methods

The hierarchical clustering methods are used to group the data into hierarchy or tree-like structure. For example, in a machine learning problem of organizing employees of a university in different departments, first the employees are grouped under the different departments in the university, and then within each department, the employees can be grouped according to their roles such as professors, assistant professors, supervisors, lab assistants, etc. This creates a hierarchical structure of the employee data and eases visualization and analysis.

### Types of Hierarchical Clustering Methods

There are two main techniques –

a) Agglomerative clustering (Bottom-up technique)

b) Divisive clustering (Top-down)

### a) Agglomerative clustering (Bottom-up technique)

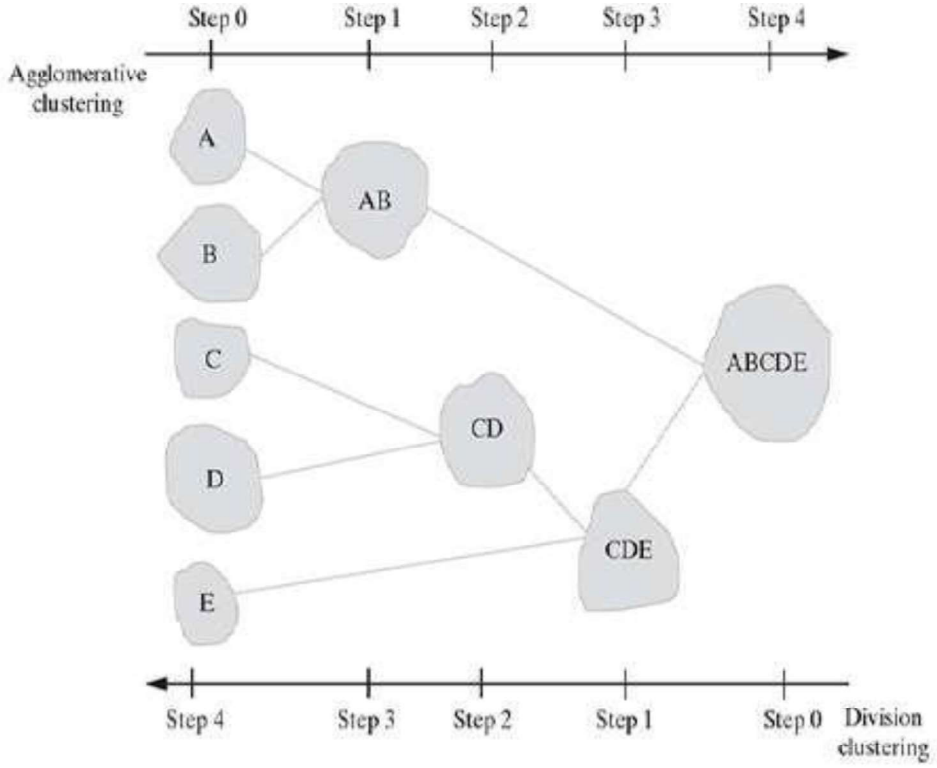
The agglomerative hierarchical clustering method uses the bottom-up strategy. It starts with each object forming its own cluster and then iteratively merges the clusters according to their similarity to form larger clusters. It terminates either when a certain clustering condition imposed by the user is achieved or all the clusters merge into a single cluster.

### b) Divisive clustering (Top-down)

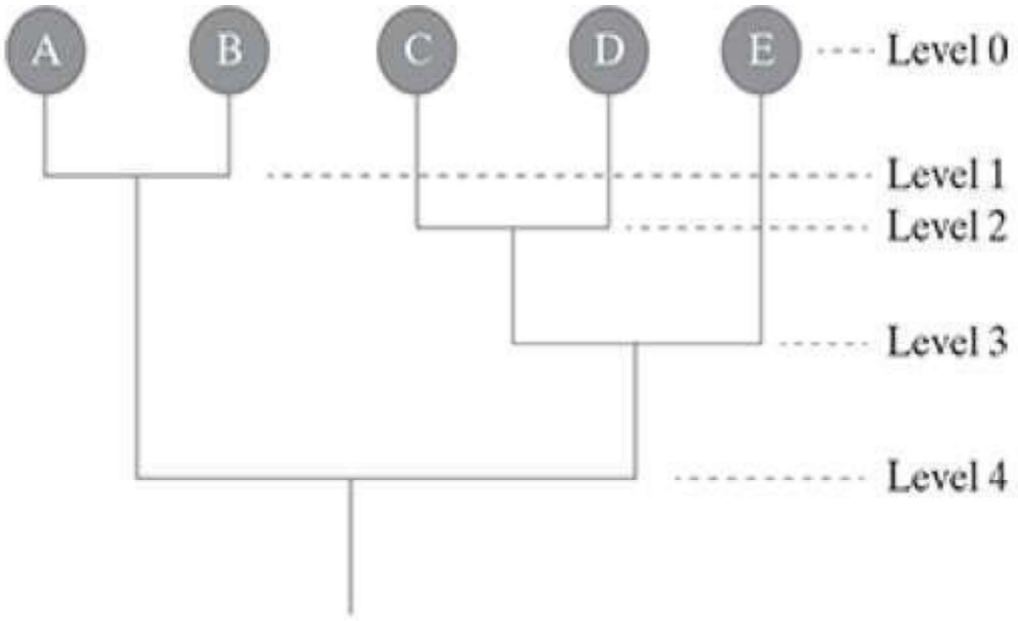
The divisive hierarchical clustering method uses a top-down strategy. The starting point is the largest cluster with all the objects in it, and then, it is split recursively to form smaller and smaller clusters, thus forming the hierarchy. The end of iterations is achieved when the objects in the final clusters are sufficiently homogeneous to each other or the final clusters contain only one object or the user-defined clustering condition is achieved.

### Dendogram technique :-

A dendrogram is a commonly used tree structure representation of step-by-step creation of hierarchical clustering. It shows how the clusters are merged iteratively (in the case of agglomerative clustering) or split iteratively (in the case of divisive clustering) to arrive at the optimal clustering solution.




Agglomerative and divisive hierarchical clustering




Dendrogram representation of hierarchical clustering

**Distance Measure**

There are four standard methods to measure the distance between clusters:

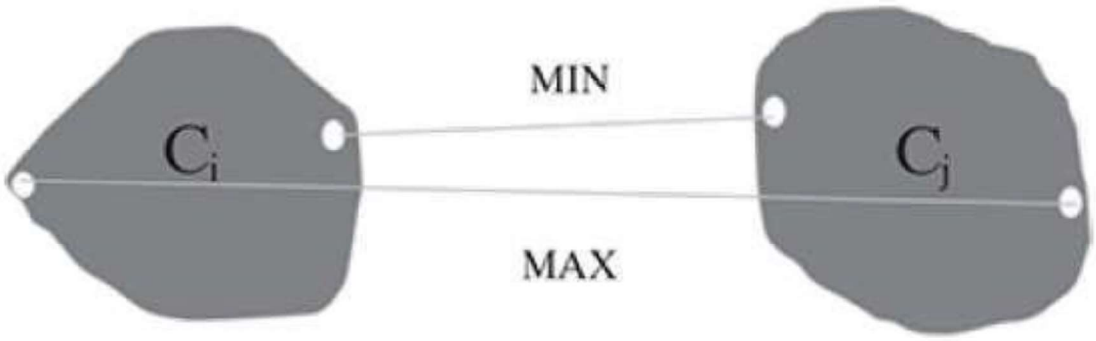
Let $C_i$ and $C_j$ be the two clusters with $n_i$ and $n_j$ respectively. $p_i$ and $p_j$ represents the points in clusters $C_i$ and $C_j$ respectively. We will denote the mean of cluster $C_i$ as $m_i$.

Minimum distance $D_{\min}(C_i, C_j) = \min_{p_i \in C_i,\ p_j \in C_j} \{|p_i - p_j|\}$

Maximum distance $D_{\max}(C_i, C_j) = \max_{p_i \in C_i,\ p_j \in C_j} \{|p_i - p_j|\}$

Mean distance $D_{\text{mean}}(C_i, C_j) = \{|m_i - m_j|\}$

Average distance $D_{\text{avg}}(C_i, C_j) = \frac{1}{n_i n_j} \sum_{p_i \in C_i,\ p_j \in C_j} |pi - pj|$




Distance measure in algorithmic methods

3) **Density-based methods – DBSCAN**

- In the above two approaches, the resulting clusters are spherical or nearly spherical in nature.
- The density-based clustering approach provides a solution to identify clusters of arbitrary shapes.
- The principle is based on identifying the dense area and sparse area within the data set and then run the clustering algorithm.

- DBSCAN is one of the popular density-based algorithm which creates clusters by using connected regions with high density.

**Differences between different CLUSTERING Methods**

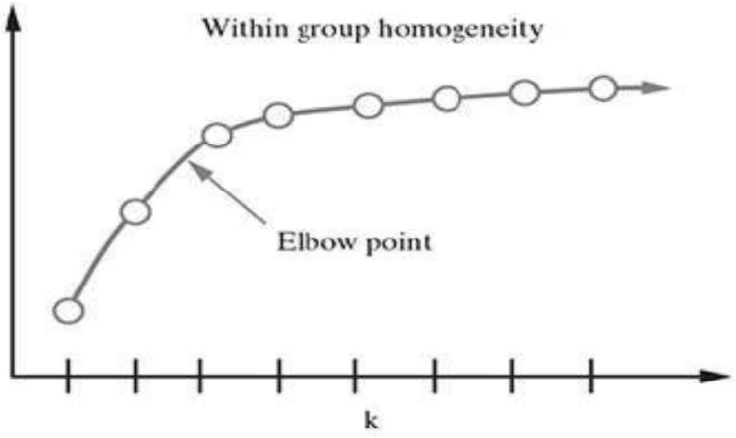
| Method | Characteristics |
|---|---|
| Partitioning methods | • Uses mean or medoid (etc.) to represent cluster centre<br>• Adopts distance-based approach to refine clusters<br>• Finds mutually exclusive clusters of spherical or nearly spherical shape<br>• Effective for data sets of small to medium size |
| Hierarchical methods | • Creates hierarchical or tree-like structure through decomposition or merger<br>• Uses distance between the nearest or furthest points in neighbouring clusters as a guideline for refinement<br>• Erroneous merges or splits cannot be corrected at subsequent levels |
| Density-based methods | • Useful for identifying arbitrarily shaped clusters<br>• Guiding principle of cluster creation is the identification of dense regions of objects in space which are separated by low-density regions<br>• May filter out outliers |

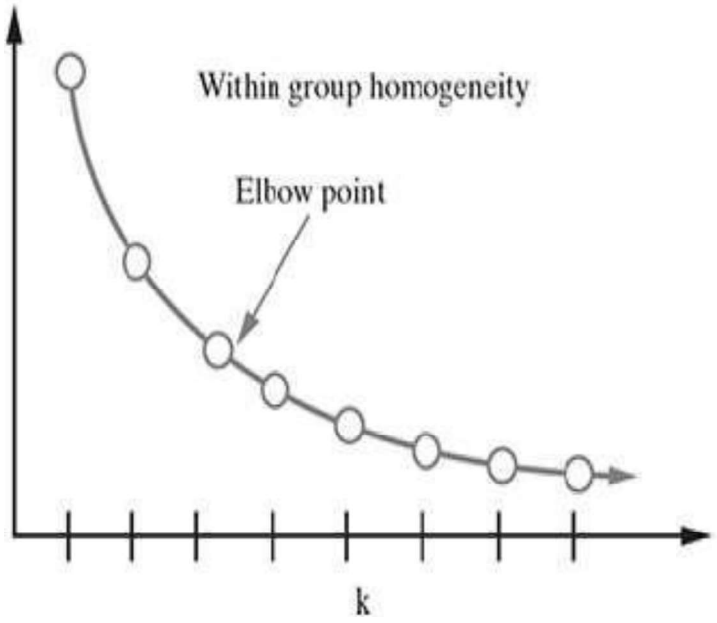
**Elbow method and Elbow point**

This method tries to measure the homogeneity or heterogeneity within the cluster and for various values of 'K' and helps in arriving at the optimal 'K'.

From the below figure, we can see the homogeneity will increase or heterogeneity will decrease with increasing 'K' as the number of data points inside each cluster reduces with this increase.

The point at which optimal clustering performance is produced is called as Elbow point.






## FINDING PATTERN USING ASSOCIATION RULE

Association rule presents a methodology that is useful for identifying interesting relationships hidden in large data sets. It is also known as association analysis, and the discovered relationships can be represented in the form of association rules comprising a set of frequent items.

A common application of this analysis is the **Market Basket Analysis** that retailers use for cross-selling of their products.

The application of association analysis is also widespread in other domains such as bioinformatics, medical diagnosis, scientific data analysis, and web data mining.

For example, by discovering the interesting relationship between food habit and patients developing breast cancer, a new cancer prevention mechanism can be found which will benefit thousands of people in the world.

### Few common terminologies used in association analysis

#### Item set

One or more items are grouped together and are surrounded by brackets to indicate that they form a set, or more specifically, an item set that appears in the data with some regularity.

{Bread, Milk, Egg} can be grouped together to form an item set as those are frequently bought together.

A collection of zero or more items is called an **item set**.

**A null item set** is the one which does not contain any item.

**k-Item set**: **In the association analysis**, an item set is called **k-item set** if it contains k number of items.

Thus, the item set {Bread, Milk, Egg} is a three-item set.

**Support Count**

**Support count** denotes the number of transactions in which a particular item set is present. This is a very important property of an item set as it denotes the frequency of occurrence for the item set. This is expressed as

$$\sigma(X) = |\{t_i | X \subseteq t_i, t_i \varepsilon T\}|$$

where |{}| denotes the number of elements in a set

*Market Basket Transaction Data*

| Transaction Number | Purchased Items |
|---|---|
| 1 | {Bread, Milk, Egg, Butter, Salt, Apple} |
| 2 | {Bread, Milk, Egg, Apple} |
| 3 | {Bread, Milk, Butter, Apple} |
| 4 | {Milk, Egg, Butter, Apple} |
| 5 | {Bread, Egg, Salt} |
| 6 | {Bread, Milk, Egg, Apple} |

The item set {Bread, Milk, Egg} occurs together three times and thus have a support count of 3.

**ASSOCIATION RULE**

- ✓ The result of the market basket analysis is expressed as a set of association rules that specify patterns of relationships among items.
- ✓ A typical rule might be expressed as {Bread, Milk}→{Egg}, which denotes that if Bread and Milk are purchased, then Egg is also likely to be purchased. Thus, association rules are learned from subsets of item sets.

  It should be noted that an association rule is an expression of X → Y where X and Y are disjoint item sets, i.e. X ∩ Y = 0.

**Measuring the strength of an association rule**

**Support** and **confidence** are the two concepts that are used for measuring the strength of an association rule.

**Support** denotes how often a rule is applicable to a given data set.

**Confidence** indicates how often the items in Y appear in transactions that contain X in a total transaction of N.

$$\text{Support}, s(X \rightarrow Y) = \frac{\sigma(X \bigcup Y)}{N}$$

$$\text{Confidence}, c(X \rightarrow Y) = \frac{\sigma(X \bigcup Y)}{\sigma(X)}$$

Support s({Bread, Milk} -> {Egg}) = support count of {Bread, Milk, Egg} / Total transactions

$$= 3/6$$

$$= 0.5$$

$$\text{Confidence } c(\{\text{Bread, Milk}\} \rightarrow \{\text{Egg}\}) = \frac{\text{support count of } \{\text{Bread, Milk, Egg}\}}{\text{support count of } \{\text{Bread, Milk}\}}$$

$$= \frac{3}{4}$$

$$= 0.75$$

**A low support** may indicate that the rule has occurred by chance. Also, from its application perspective, this rule may not be a very attractive business investment as the items are seldom bought together by the customers. Thus, support can provide the intelligence of identifying the most interesting rules for analysis.

Similarly, **confidence** provides the measurement for reliability of the inference of a rule. Higher confidence of a rule X → Y denotes more likelihood of to be present in transactions that contain X as it is the estimate of the conditional probability of Y given X.

**The APRIORI algorithm for association rule learning**

The main challenge of discovering an association rule and learning from it is the large volume of transactional data and the related complexity. Because of the variation of features in transactional data, the number of feature sets within a data set usually becomes very large. This leads to the problem of handling a very large number of item sets, which grows exponentially with the number of features.

Following steps are followed for generating association rules:

1. Decide the minimum support and minimum confidence of the association rules. From a set of transaction T, let us assume that we will find out all the rules that have support ≥ min S and confidence ≥ min C, where min S and min C are the support and confidence thresholds, respectively, for the rules to be considered acceptable.
2. Generate Frequent
3. Generate Rules

Build the a priori principle rules

One of the most widely used algorithm to reduce the number of item sets to search for the association rule is known as A priori.

a) If an item set is frequent, then all of its subsets must also be frequent.
b) If an item set is frequent, then all the supersets must be frequent too.

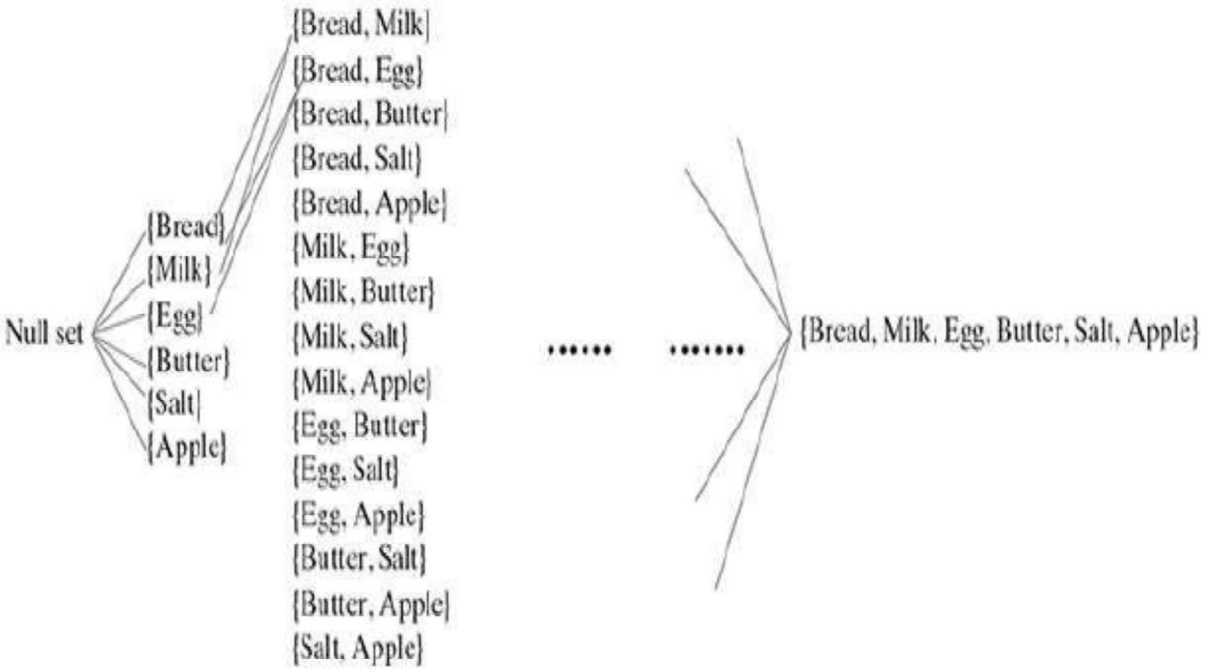
Demonstration of A priori principle

a. Consider the dataset given below

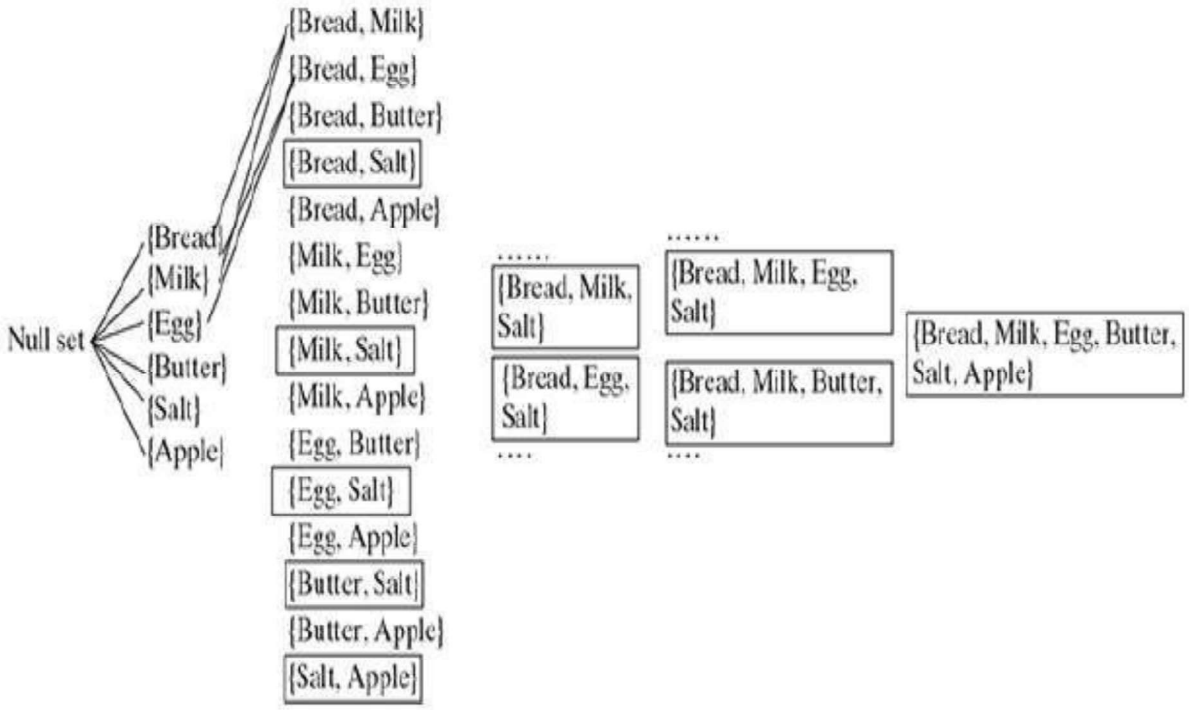
*Market Basket Transaction Data*

| Transaction Number | Purchased Items |
|---|---|
| 1 | {Bread, Milk, Egg, Butter, Salt, Apple} |
| 2 | {Bread, Milk, Egg, Apple} |
| 3 | {Bread, Milk, Butter, Apple} |
| 4 | {Milk, Egg, Butter, Apple} |
| 5 | {Bread, Egg, Salt} |
| 6 | {Bread, Milk, Egg, Apple} |

From the full item set of six items {Bread, Milk, Egg, Butter, Salt, Apple}, there are $2^6$ ways to create baskets or item sets (including the null item set) as shown in the below




Sixty-four ways to create itemsets from 6 items




Discarding the itemsets consisting of Salt

The actual process of creating rules involves two phases:

a. Identifying all item sets that meet a minimum support threshold set for the analysis
b. Creating rules from these item sets that meet a minimum confidence threshold which identifies the strong rules

**Strengths and Weaknesses of apriori algorithm**

| Strengths | Weaknesses |
|---|---|
| • Provides reasonable accuracy while working with very large amounts of transactional data<br>• Discovers rules that are easy to understand<br>• Provides valuable insight into the unexpected knowledge in data sets, which is a key aspect of learning | • Not very accurate in the case the data set is small as the smaller occurrences of itemsets may not be due to chance<br>• Some effort is involved to separate the insight from the common sense<br>• In the case of widespread presence of random patterns, the principle can draw spurious conclusions |

**----------END----------**

**TEXT BOOKS:**

1. Saikat Dutt, Subramanian Chandramouli, Amit Kumar Das, Machine Learning, Pearson, 2019.

**REFERENCE BOOKS:**

1. Ethern Alpaydin, ― Introduction to Machine Learning, MIT Press, 2004.
2. Stephen Marsland, ― Machine Learning - An Algorithmic Perspective, Second Edition, Chapman and Hall / CRC Machine Learning and Pattern Recognition Series, 2014