Unsupervised Learning

- Unsupervised learning is a machine learning concept where the unlabelled 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

- In the concept of unsupervised learning where the objective is to observe only the features $X_1 : X_2 : \ldots X_n$;
- It is to find out the association between the features or their grouping to understand the nature of the data.
- This analysis may reveal an interesting correlation between the features or a common behaviour within the subgroup of the data, which provides better understanding of the data.
- In terms of statistics, a supervised learning algorithm .the probability of outcome Y for a particular input X, which is called the posterior probability.
- Unsupervised learning is closely related to density estimation in statistics.
- Here, every input and the corresponding targets are concatenated to create a new set of input such as $\{(X_1 , Y_1), (X_2 , Y_2), \ldots, (X_n , Y_n)\}$, which leads to a better understanding of the correlation of X and Y; this probability notation is called the joint probability.
- For example unsupervised learning helps in pushing movie promotions to the correct group of people.
- In earlier days, movie promotions were blind push of the same data to all demography, such that everyone used to watch the same posters or trailers irrespective of their choice or preference.
- So, in most of the cases, the person watching the promotion or trailer would end up ignoring it, which leads to waste of effort and money on the promotion.
- But with the advent of smart devices and apps, there is now a huge database available to understand what type of movie is liked by what segment of the demography.
- Machine learning helps to find out the pattern or the repeated behaviour of the smaller groups/clusters within this database to provide the intelligence about liking or disliking of certain types of movies by different groups within the demography.
- So, by using this intelligence, the smart apps can push only the relevant movie promotions or trailers to the selected groups, which will significantly increase the chance of targeting the right interested person for the movie.

There are two method used for explaining the principle underlying unsupervised learning

1. Clustering and
2. Association Analysis.

Clustering is a broad class of methods used for discovering unknown subgroups in data, which is the most important concept in unsupervised learning. Another technique is **Association Analysis** which identifies a low-dimensional representation of the observations that can explain the variance and identify the association rule for the explanation.

APPLICATION OF UNSUPERVISED LEARNING

Because of its flexibility that it can work on uncategorized and unlabelled data, there are many domains where unsupervised learning finds its application.

Few examples of such applications are as follows:

- 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
- Anomaly or fraud detection in the banking sector by identifying the pattern of loan defaulters
- Image processing and image segmentation such as face recognition, expression identification, etc.
- Grouping of important characteristics in genes to identify important influencers in new areas of genetics
- Utilization by data scientists to reduce the dimensionalities in sample data to simplify modelling
- Document clustering and identifying potential labelling options

Today, unsupervised learning is used in many areas involving Artificial Intelligence (AI) and Machine Learning (ML). Chat bots, self-driven cars, and many more recent innovations are results of the combination of unsupervised and supervised learning.

There are two major aspects of unsupervised learning, namely

- Clustering which helps in segmentation of the set of objects into groups of similar objects and
- Association Analysis which is related to the identification of relationships among objects in a data set.

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.
- 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.

- It is often domain specific to define what is meant by two objects to be similar or dissimilar and thus is an important aspect of the unsupervised machine learning task.
- As an example, suppose we want to run some advertisements of a new movie for a countrywide promotional activity.
- We have data for the age, location, financial condition, and political stability of the people in different parts of the country.
- We may want to run a different type of campaign for the different parts grouped according to the data we have.
- Any logical grouping obtained by analysing the characteristics of the people will help us in driving the campaigns in a more targeted way.
- Clustering analysis can help in this activity by analysing different ways to group the set of people and arriving at different types of clusters.

There are many different fields where cluster analysis is used effectively, such as

Text data mining: this includes tasks such as text categorization, text clustering, document summarization, concept extraction, sentiment analysis, and entity relation modelling

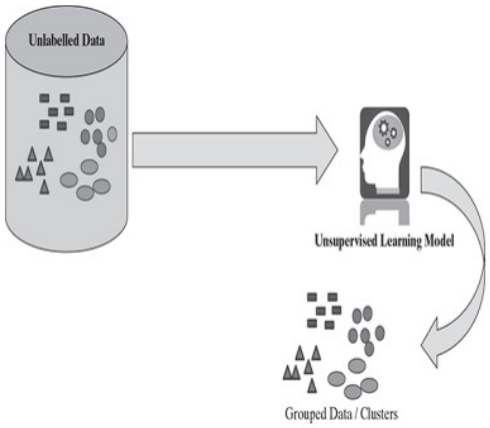
Customer segmentation: creating clusters of customers on the basis of parameters such as demographics, financial conditions, buying habits, etc., which can be used by retailers and advertisers to promote their products in the correct segment

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

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

CLUSTERING AS A MACHINE LEARNING TASK

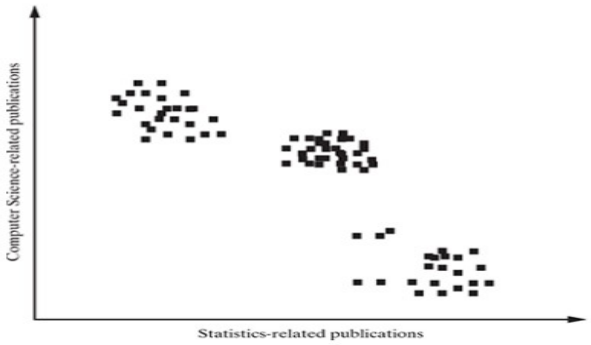
- Clustering is defined as an unsupervised machine learning task that automatically divides the data into clusters or groups of similar items.
- The analysis achieves this without prior knowledge of the types of groups required and thus can provide an insight into the natural groupings within the data set.
- The primary guideline of clustering task is that the data inside a cluster should be very similar to each other but very different from those outside the cluster.
- We can assume that the definition of similarity might vary across applications, but the basic idea is always the same, that is, to create the group such that related elements are placed together.
- Using this principle, whenever a large set of diverse and varied data is presented for analysis, clustering enables to represent the data in a smaller number of groups.
- It helps to reduce the complexity and provides insight into patterns of relationships to generate meaningful and actionable structures within the data.
- The effectiveness of clustering is measured by the homogeneity within a group as well as the difference between distinct groups.



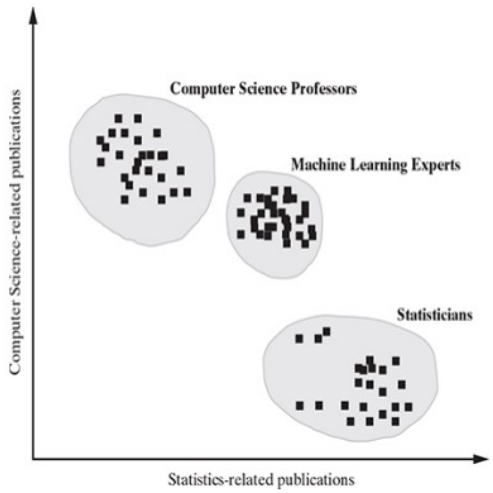
- Through clustering, we are trying to label the objects with class labels.
- But clustering is somewhat different from the classification and numeric prediction.
- In each of these cases, the goal was to create a model that relates features to an outcome or to other features and the model identifies patterns within the data.
- In contrast, clustering creates new data. Unlabelled objects are given a cluster label which is inferred entirely from the relationship of attributes within the data.

For an example.

- You were invited to take a session on Machine Learning in a reputed university for induction of their professors on the subject.
- Before you create the material for the session, you want to know the level of acquaintance of the professors on the subject so that the session is successful.
- But you do not want to ask the inviting university, but rather do some analysis on your own on the basis of the data available freely.
- As Machine Learning is the intersection of Statistics and Computer Science, you focused on identifying the professors from these two areas also.
- So, you searched the list of research publications of these professors from the internet, and by using the machine learning algorithm, you now want to group the papers and thus infer the expertise of the professors into three buckets – Statistics, Computer Science, and Machine Learning.
- After plotting the number of publications of these professors in the two core areas, namely Statistics and Computer Science, you obtain a scatter plot



- Some inferences that can be derived from the pattern analysis of the data is that there seems to be three groups or clusters emerging from the data.
- The pure statisticians have very less Computer Science-related papers, whereas the pure Computer Science professors have less number of statistics-related papers than Computer Science related papers.
- There is a third cluster of professors who have published papers on both these areas and thus can be assumed to be the persons knowledgeable in machine learning concepts,.
- Thus, in the above problem, we used visual indication of logical grouping of data to identify a pattern or cluster and labelled the data in three different clusters.
- The main driver for our clustering was the closeness of the points to each other to form a group.
- The clustering algorithm uses a very similar approach to measure how closely the data points are related and decides whether they can be labelled as a homogeneous group.



Different types of clustering techniques

The major clustering techniques are

1. Partitioning methods,
2. Hierarchical methods, and
3. Density-based methods.

Their approach towards creating the clusters, way to measure the quality of the clusters, and applicability are different.

Partitioning methods

- Two of the most important algorithms for partitioning based clustering are k-means and k-medoid.
- In the k means algorithm, the centroid of the prototype is identified for clustering, which is normally the mean of a group of points.
- Similarly, the k-medoid algorithm identifies the medoid which is the most representative point for a group of points.
- We can also infer that in most cases, the centroid does not correspond to an actual data point, whereas medoid is always an actual data point.

K-means – A centroid-based technique

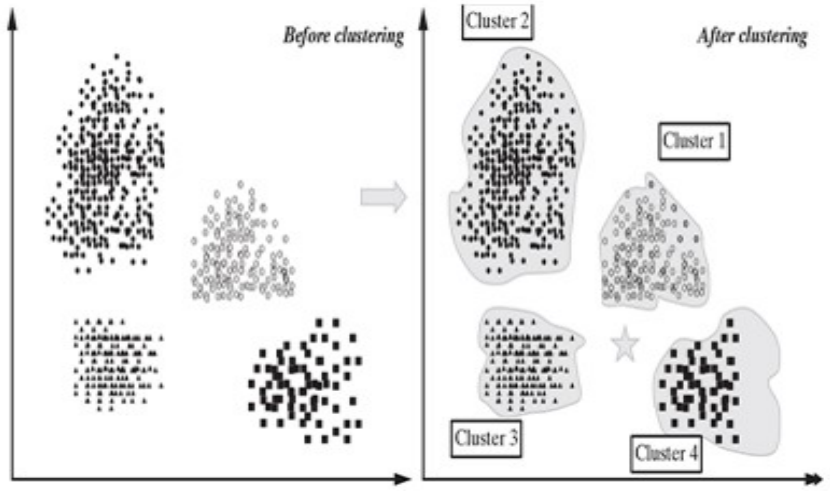
- This is one of the oldest and most popularly used algorithm for clustering.
- The basic principles used by this algorithm also serves as the basis for other more sophisticated and complex algorithms.
- The principle of the k-means algorithm is to assign each of the 'n' data points to one of the K clusters 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.
- The homogeneity and differences are measured in terms of the distance between the objects or points in the data set.

Strengths	Weaknesses
• The principle used for identifying the clusters is very simple and involves very less complexity of statistical terms • The algorithm is very flexible and thus can be adjusted for most scenarios and complexities • 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 • 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

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
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

- On a certain set of data points, and the k-means algorithm can be applied to find out the clusters generated from this data set.
- Let K = 4, implying that four clusters are created out of this data set.
- As the first step, assign four random points from the data set as the centroids, as represented by the * signs, and assign the data points to the nearest centroid to create four clusters.
- In the second step, 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.
- After three iterations,it is found that the centroids are not moving as there is no scope for refinement, and thus, the k-means algorithm will terminate.
- This provides the most logical four groupings or cluster of the data sets where the homogeneity within the groups is highest and difference between the groups is maximum.



Choosing appropriate number of clusters

- One of the most important success factors in arriving at correct clustering is to start with the correct number of cluster assumptions.

- Different numbers of starting cluster lead to completely different types of data split.
- So , there must be some prior knowledge about the number of clusters and k-means algorithm can be started with that prior knowledge.
- For example, if we are clustering the data of the students of a university, it is always better to start with the number of departments in that university.
- Sometimes, the business needs or resource limitations drive the number of required clusters.
- For example, if a movie maker wants to cluster the movies on the basis of combination of two parameters – budget of the movie: high or low, and casting of the movie: star or non-star, then there are 4 possible combinations, and thus, there can be four clusters to split the data.
- For a small data set, sometimes a rule of thumb that is followed is

$$K = \sqrt{\frac{n}{2}}$$

  - which means that K is set as the square root of n/2 for a data set of n examples.
- But unfortunately, this thumb rule does not work well for large data sets.
- There are several statistical methods to arrive at the suitable number of clusters.

Choosing the initial centroids

- Another key step for the k-means algorithm is to choose the initial centroids properly.
- One common practice is to choose random points in the data space on the basis of the number of cluster requirement and refine the points as we move into the iterations.
- But this often leads to higher squared error in the final clustering, thus resulting in sub-optimal clustering solution.
- The assumption for selecting random centroids is that multiple subsequent runs will minimize the SSE and identify the optimal clusters.
- One effective approach is to employ the hierarchical clustering technique on sample points from the data set and then arrive at sample K clusters.
- The centroids of these initial K clusters are used as the initial centroids.
- This approach is practical when the data set has small number of points and K is relatively small compared to the data points.
- In k-means algorithm, 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 distance of the data point from its nearest centroid can also be calculated to minimize the distances to arrive at the refined centroid.

- The Euclidean distance between two data points is measured as follows:

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

- Using this function, the distance between the example data and its nearest centroid and the objective is calculated to minimize this distance.
- The measure of quality of clustering uses the SSE technique.
- The formula used is as follows:

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

- where dist() calculates the Euclidean distance between the centroid c of the cluster C and the data points x in the cluster.
- The summation of such distances over all the 'K' clusters gives the total sum of squared error., The lower the SSE for a clustering solution, the better is the representative position of the centroid.
- Thus, in K-means clustering algorithm Algorithm the recomputation of the centroid involves calculating the SSE of each new centroid and arriving at the optimal centroid identification.
- After the centroids are repositioned, the data points nearest to the centroids are assigned to form the refined clusters.
- It is observed that the centroid that minimizes the SSE of the cluster is its mean.
- 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 is often used as the first step of identifying the subgroups within a unlabeled set of data which then is used for classifying the new observed data.
- At the beginning we are not clear about the pattern or classes that exist within the unlabeled data set.
- By using the clustering algorithm at that stage we find out the groups of similar objects within the data set and form sub-groups and classes.
- Later when a new object is observed, then using the classification algorithms we try to place that into one of the sub-groups identified in the earlier stage.

- Let's take an example. We are running a software testing activity and we identified a set of defects in the software.
- For easy allocation of these defects to different developer groups, the team is trying to identify similar groups of defects.
- Often text analytics is used as the guiding principle for identifying this similarity.
- Suppose there are 4 subgroups of defects identified, namely, GUI related defects, Business logic related defects, Missing requirement defects and Database related defects.
- Based on this grouping, the team identified the developers to whom the defects should be sent for fixing.
- As the testing continues, there are new defects getting created.
- We have the categories of defects identified now and thus the team can use classification algorithms to assign the new defect to one of the 4 identified groups or classes which will make it easy to identify the developer who should be fixing it

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.
- Let us 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}.

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

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},

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

and the mean of the cluster

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

So, the SSE within the clusters is

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

- 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 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.
- 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

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

- where o is the representative point or object of cluster C .
- 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, kmedoids 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

Partitioning Around Medoids (PAM) algorithm

- ✓ 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 SSEnew < SSEold, 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

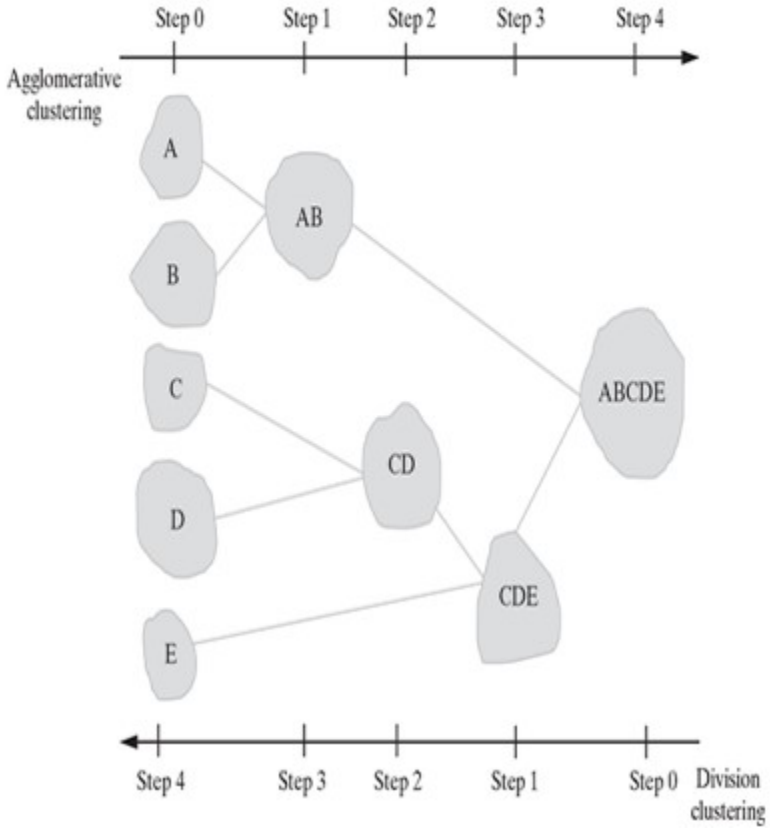
In this algorithm, we replaced the current representative object with a non-representative object and checked if it improves the quality of clustering.

In the iterative process, all possible replacements are attempted until the quality of clusters no longer improves.

Hierarchical clustering

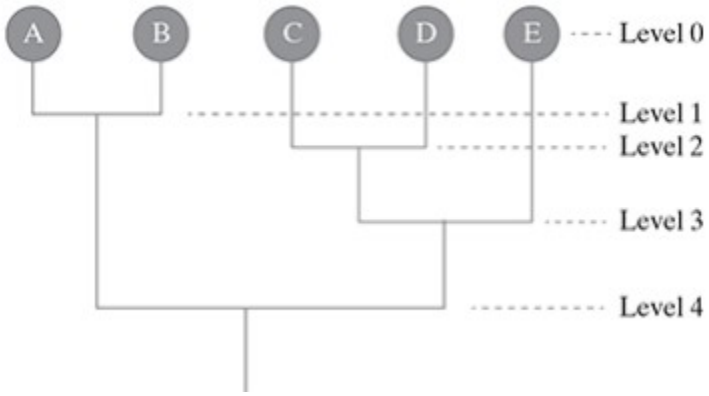
- There are situations when the data needs to be partitioned into groups at different levels such as in a hierarchy.
- The hierarchical clustering methods are used to group the data into hierarchy or tree-like structure.
- For example, in a machine learning problem of organizing employees of a university 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.
- Similarly, there may be a data set which has an underlying hierarchy structure that we want to discover and we can use the hierarchical clustering methods to achieve that.
- There are two main hierarchical clustering methods: agglomerative clustering and divisive clustering.

- **Agglomerative clustering** is a bottom-up technique which starts with individual objects as clusters and then iteratively merges them to form larger clusters.
- **divisive method** starts with one cluster with all given objects and then splits it iteratively to form smaller clusters..



- 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.
- The divisive hierarchical clustering method uses a topdown 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.
- In both these cases, it is important to select the split and merger points carefully, because the subsequent splits or mergers will use the result of the previous ones and there is no option to perform any object swapping between the clusters or rectify the decisions made in previous steps, which may result in poor clustering quality at the end.

- 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.
- The following figure shows a dendro-gram with four levels and how the objects are merged or split at each level to arrive at the hierarchical clustering.



- One of the core measures of proximities between clusters is the distance between them.
- 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.
- Denote the mean of cluster C_i as m_i

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

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

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

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

Distance measure in algorithmic methods

- Often the distance measure is used to decide when to terminate the clustering algorithm.
- For example, in an agglomerative clustering, the merging iterations may be stopped once the MIN distance between two neighbouring clusters becomes less than the userdefined threshold.
- So, when an algorithm uses the minimum distance D_{min} to measure the distance between the clusters, then it is referred to as nearest neighbour min clustering algorithm, and if the decision to stop the algorithm is based on a user-defined limit on D_{min} , then it is called single linkage algorithm.
- On the other hand, when an algorithm uses the maximum distance D_{max} to measure the distance between the clusters, then it is referred to as furthest neighbour clustering algorithm, and if the decision to stop the algorithm is based on a user-defined limit on D_{max} then it is called complete linkage algorithm.
- As minimum and maximum measures provide two extreme options to measure distance between the clusters, they are prone to the outliers and noisy data.
- Instead, the use of mean and average distance helps in avoiding such problem and provides more consistent results.

Density-based methods - DBSCAN

- when the partitioning and hierarchical clustering methods are used, the resulting clusters are spherical or nearly spherical in nature.
- In the case of the other shaped clusters such as S shaped or uneven shaped clusters, the above two types of method do not provide accurate results.
- 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

.

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.
- For example, every large grocery store accumulates a large volume of data about the buying pattern of the customers.

- On the basis of the items purchased together, the retailers can push some cross selling either by placing the items bought together in adjacent areas or creating some combo offer with those different product types.
- The below association rule signifies that people who have bought bread and milk have often bought egg also; so, for the retailer, it makes sense that these items are placed together for new opportunities for cross-selling.
    - {Bread, Milk} → {Egg}
- 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.

Definition of common terms

1.Itemset

- One or more items are grouped together and are surrounded by brackets to indicate that they form a set, or more specifically, an itemset that appears in the data with some regularity.
- For example, , {Bread, Milk, Egg} can be grouped together to form an itemset as those are frequently bought together.
- To generalize this concept, if I = $\{i_1 , i_2 ,\ldots, i_n \}$ are the items in a market basket data and T = $\{t_1 , t_2 ,\ldots, t_n \}$ are the set of all the transactions, then each transaction t contains a subset of items from I.
- A collection of zero or more items is called an itemset.
- A null itemset is the one which does not contain any item.
- In the association analysis, an itemset is called k-itemset if it contains k number of items.
- Thus, the itemset {Bread, Milk, Egg} is a three-itemset.

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}

2.Support count

- Support count denotes the number of transactions in which a particular itemset is present.

- This is a very important property of an itemset as it denotes the frequency of occurrence for the itemset.
- 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
- The itemset {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 itemsets.
- For example, the preceding rule was identified from the set of {Bread, Milk, Egg}. It should be noted that an association rule is an expression of X → Y where X and Y are disjoint itemsets, i.e. X ∩ Y = 0.
- 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.
- Confidence denotes the predictive power or accuracy of the rule. So, the mathematical expressions are

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

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

- if we consider the association rule {Bread, Milk} → {Egg}, then from the above formula

$$\text{Confidence } c(\{\text{Bread, Milk}\} \rightarrow \{\text{Egg}\}) = \frac{\text{support count of } \{\text{Bread, Milk, Egg}\}}{\text{support count of } \{\text{Bread, Milk}\}}$$
$$= \frac{3}{4}$$
$$= 0.75$$

$$\text{Confidence } c(\{\text{Bread, Milk}\} \rightarrow \{\text{Egg}\}) = \frac{\text{support count of } \{\text{Bread, Milk, Egg}\}}{\text{support count of } \{\text{Bread, Milk}\}}$$
$$= \frac{3}{4}$$
$$= 0.75$$

- It is important to understand the role of support and confidence in the association analysis.
- 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.
- Also, understand that the confidence of X leading to Y is not the same as the confidence of Y leading to X.
- For example, confidence of {Bread, Milk} → {Egg} = 0.75 but confidence of
- $\{\text{Egg}\} \rightarrow \{\text{Bread, Milk}\} = \frac{3}{5} = 0.6.1$
- Here, the rule {Bread, Milk} → {Egg} is the strong rule.
- As association rule learners are unsupervised, there is no need for the algorithm to be trained; this means that no prior labelling of the data is required.
- The programme is simply run on a data set in the hope that interesting associations are found. Association rule analysis is used to search for interesting connections among a very large number of variables.
- Though human beings are capable of such insight quite intuitively, sometimes it requires expert-level knowledge or a great deal of experience to achieve the performance of a rule-learning algorithm.

- Additionally, some data may be too large or complex for humans to decipher and analyse so easily

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 itemsets, which grows exponentially with the number of features.
- If there are k items which may or may not be part of an itemset, then there is 2^k ways of creating itemsets with those items.
- For example, if a seller is dealing with 100 different items, then the learner need to evaluate
- $2^{100} = 1 \times e^{30}$ itemsets for arriving at the rule, which is computationally impossible.
- So, it is important to filter out the most important (and thus manageable in size) itemsets and use the resources on those to arrive at the reasonably efficient association rules.
- The first step is to 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 ≥ minS and confidence ≥ minC, where minS and minC are the support and confidence thresholds, respectively, for the rules to be considered acceptable.
- Now, even if we put the minS = 20% and minC = 50%, it is seen that more than 80% of the rules are discarded; this means that a large portion of the computational efforts could have been avoided if the itemsets for consideration were first pruned and the itemsets which cannot generate association rules with reasonable support and confidence were removed.

Build the apriori principle rules

- One of the most widely used algorithm to reduce the number of itemsets to search for the association rule is known as Apriori.
- It has proven to be successful in simplifying the association rule learning to a great extent.
- The principle got its name from the fact that the algorithm utilizes a simple prior belief (i.e. a priori) about the properties of frequent itemsets: If an itemset is frequent, then all of its subsets must also be frequent.
- This principle significantly restricts the number of itemsets to be searched for rule generation.
- For example, if in a market basket analysis, it is found that an item like ‘Salt’ is not so frequently bought along with the breakfast items, then it is fine to remove all the itemsets

containing salt for rule generation as their contribution to the support and confidence of the rule will be insignificant.

- The converse also holds true: If an itemset is frequent, then all the supersets must be frequent too.
- These are very powerful principles which help in pruning the exponential search space based on the support measure and is known as support-based pruning.
- The key property of the support measure used here is that the support for an itemset never exceeds the support for its subsets.
- This is also known as the antimonotone property of the support measure.
- The actual process of creating rules involves two phases:
- Identifying all itemsets that meet a minimum support threshold set for the analysis
- Creating rules from these itemsets that meet a minimum confidence threshold which identifies the strong rules
- The first phase involves multiple iterations where each successive iteration requires evaluating the support of storing a set of increasingly large itemsets.
- For instance, iteration 1 evaluates the set of one-item itemsets (oneitemsets), iteration 2 involves evaluating the twoitemsets, etc..
- The result of each iteration N is a set of all N-itemsets that meet the minimum support threshold.
- Normally, all the itemsets from iteration N are combined in order to generate candidate itemsets for evaluation in iteration N + 1, but by applying the Apriori principle, we can eliminate some of them even before the next iteration starts.
- If {X}, {Y}, and {Z} are frequent in iteration 1 while {W} is not frequent, then iteration 2 will consider only {X, Y}, {X, Z}, and {Y, Z}.
- By continuing with the iterations, let us assume that during iteration 2, it is discovered that {X, Y} and {Y, Z} are frequent, but {X, Z} is not.
- Although iteration 3 would normally begin by evaluating the support for {X, Y, Z}, this step need not occur at all.
- The Apriori principle states that {X, Y, Z} cannot be frequent, because the subset {X, Z} is not. Therefore, in iteration 3, the algorithm may stop as no new itemset can be generated.
- Once we identify the qualifying itemsets for analysis, the second phase of the Apriori algorithm begins.
- For the given set of frequent itemsets, association rules are generated from all possible subsets.
- For example, {X, Y} would result in candidate rules for {X} → {Y} and {Y} → {X}.
- These rules are evaluated against a minimum confidence threshold, and any rule that does not meet the desired confidence level is eliminated, thus finally yielding the set of strong rules.

- Though the Apriori principle is widely used in the market basket analysis and other applications of association rule help in the discovery of new relationship among objects, there are certain strengths and weaknesses we need to keep in mind before employing it over the target data set:

Strengths	Weaknesses
• Provides reasonable accuracy while working with very large amounts of transactional data • Discovers rules that are easy to understand • 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 • Some effort is involved to separate the insight from the common sense • In the case of widespread presence of random patterns, the principle can draw spurious conclusions