1. What is the difference between supervised and unsupervised learning? Give some examples to illustrate your point.

Answer:---->Supervised Learning includes various algorithms such as Bayesian Logic, Decision Tree, Logistic Regression, Linear Regression, Multi-class Classification, Support Vector Machine etc. Unsupervised Learning includes various algorithms like KNN, Apriori Algorithm, and Clustering.supervised learning uses labeled input and output data, while an unsupervised learning algorithm does not. In supervised learning, the algorithm “learns” from the training dataset by iteratively making predictions on the data and adjusting for the correct answer.Which means some data is already tagged with the correct answer. After that, the machine is provided with a new set of examples(data) so that the supervised learning algorithm analyses the training data(set of training examples) and produces a correct outcome from labelled data.Supervised machine learning is generally used to classify data or make predictions, whereas unsupervised learning is generally used to understand relationships within datasets. Supervised machine learning is much more resource-intensive because of the need for labelled data.

2. Mention a few unsupervised learning applications.

Answer:---->Unsupervised learning finds a myriad of real-life applications, including: data exploration, customer segmentation, recommender systems, target marketing campaigns, and. data preparation and visualization, etc. Some use cases for unsupervised learning — more specifically, clustering — include: Customer segmentation, or understanding different customer groups around which to build marketing or other business strategies. Genetics, for example clustering DNA patterns to analyze evolutionary biology. The most common unsupervised learning method is cluster analysis, which applies clustering methods to explore data and find hidden patterns or groupings in data. With MATLAB you can apply many popular clustering algorithms: Hierarchical clustering: Builds a multilevel hierarchy of clusters by creating a cluster tree.

3. What are the three main types of clustering methods? Briefly describe the characteristics of each.

Answer:---->Partitioning based, hierarchical based, density-based-, grid-based-, and model-based clustering are the clustering methods. Characteristics of Clusters. Characteristics include: Anglo - competitive and result-oriented. Confucian Asia - result-driven, encourage group working together over individual goals Types of Clustering Centroid-based Clustering. Density-based Clustering. Distribution-based Clustering. Hierarchical Clustering. The various types of clustering are: Connectivity-based Clustering (Hierarchical clustering) Centroids-based Clustering (Partitioning methods) Distribution-based Clustering. Density-based Clustering (Model-based methods) Fuzzy Clustering. Constraint-based (Supervised Clustering)

4. Explain how the k-means algorithm determines the consistency of clustering.

Answer:--->It is an iterative process of assigning each data point to the groups and slowly data points get clustered based on similar features. The objective is to minimize the sum of distances between the data points and the cluster centroid, to identify the correct group each data point should belong to.in k-means clustering, the number of clusters that you want to divide your data points into i.e., the value of K has to be pre-determined whereas in Hierarchical clustering data is automatically formed into a tree shape form (dendrogram)Elbow method gives us an idea on what a good k number of clusters would be based on the sum of squared distance (SSE) between data points and their assigned clusters' centroids. We pick k at the spot where SSE starts to flatten out and forming an elbow.

5. With a simple illustration, explain the key difference between the k-means and k-medoids algorithms.

Answer:--->K-means attempts to minimize the total squared error, while k-medoids minimizes the sum of dissimilarities between points labeled to be in a cluster and a point designated as the center of that cluster. In contrast to the k -means algorithm, k -medoids chooses datapoints as centers (medoids or exemplars)k-means minimizes within-cluster variance, which equals squared Euclidean distances. In general, the arithmetic mean does this. It does not optimize distances, but squared deviations from the mean. k-medians minimizes absolute deviations, which equals Manhattan distance.k -medoids is a classical partitioning technique of clustering that splits the data set of n objects into k clusters, where the number k of clusters assumed known a priori (which implies that the programmer must specify k before the execution of a k -medoids algorithm)

6. What is a dendrogram, and how does it work? Explain how to do it.

Answer:---->A dendrogram is a diagram that shows the hierarchical relationship between objects. It is most commonly created as an output from hierarchical clustering. A dendrogram is a diagram that shows the attribute distances between each pair of sequentially merged classes. To avoid crossing lines, the diagram is graphically arranged so that members of each pair of classes to be merged are neighbors in the diagram. The Dendrogram tool uses a hierarchical clustering algorithm.Steps to Create Dendrogram Set the Y-axis max value. Choose the "Set Max and Min Value" option. Set X-axis and Y-axis tick number, click the "Appearance Options", and from the pop-up dialogue, change the tick numbers. Double click on the number label to edit numbers on X-axis.

7. What exactly is SSE? What role does it play in the k-means algorithm?

Answer:---->The quality of the cluster assignments is determined by computing the sum of the squared error (SSE) after the centroids converge, or match the previous iteration's assignment. The SSE is defined as the sum of the squared Euclidean distances of each point to its closest centroid.Intra-cluster variance (a.k.a., the squared error function or sum of squares within (SSW) or sum of squares error (SSE)) is used to quantify internal cohesion. It is defined as the sum of the squared distance between the average point (called Centroid) and each point of the clusterSSE is a good measure if we are trying to find spherically shaped clusters. Internal Measure: This is the more general one when the class label is not available. The silhouette coefficient is one such popular measure.

Double-click (or enter) to edit

8. With a step-by-step algorithm, explain the k-means procedure.

Answer:---->1) Randomly select 'c' cluster centers. 2) Calculate the distance between each data point and cluster centers. 3) Assign the data point to the cluster center whose distance from the cluster center is minimum of all the cluster centers..Use K means clustering to generate groups comprised of observations with similar characteristics. For example, if you have customer data, you might want to create sets of similar customers and then target each group with different types of marketing. K means clustering is a popular machine learning algorithm.The K-means clustering algorithm computes centroids and repeats until the optimal centroid is found. It is presumptively known how many clusters there are. It is also known as the flat clustering algorithm. The number of clusters found from data by the method is denoted by the letter 'K' in K-means

9. In the sense of hierarchical clustering, define the terms single link and complete link.

Answer:--->In single-link (or single linkage) hierarchical clustering, we merge in each step the two clusters whose two closest members have the smallest distance (or: the two clusters with the smallest minimum pairwise distance). Complete-link clustering can also be described using the concept of clique.Complete-linkage clustering is one of several methods of agglomerative hierarchical clustering. At the beginning of the

process, each element is in a cluster of its own. The clusters are then sequentially combined into larger clusters until all elements end up being in the same cluster.Single link - You link two clusters based on the minimum distance between 2 elements. A drawback of this method is that it tends to produce long thin clusters since you make the link based on only 2 points. Complete link - You link two clusters based on the max distance between 2 elements.

10. How does the apriori concept aid in the reduction of measurement overhead in a business basket analysis? Give an example to demonstrate your point.

Answer:---->

The lifeblood of retail businesses has always been sales. A retailer can never assume that his customers know all of his offerings. But rather, he must make the effort to present all applicable options in way which increases customer engagement and increase sales.

Association Rule Mining

Association Rule Mining is used when you want to find an association between different objects in a set, find frequent patterns in a transaction database, relational databases or any other information repository. The applications of Association Rule Mining are found in Marketing, Basket Data Analysis (or Market Basket Analysis) in retailing, clustering and classification.

The most common approach to find these patterns is Market Basket Analysis, which is a key technique used by large retailers like Amazon, Flipkart, etc to analyze customer buying habits by finding associations between the different items that customers place in their “shopping baskets”. The discovery of these associations can help retailers develop marketing strategies by gaining insight into which items are frequently purchased together by customers. The strategies may include:

Changing the store layout according to trends Customer behavior analysis Catalog design Cross marketing on online stores What are the trending items customers buy Customized emails with add-on sales etc.. Online retailers and publishers can use this type of analysis to:

Inform the placement of content items on their media sites, or products in their catalog Deliver targeted marketing (e.g. emailing customers who bought products specific products with other products and offers on those products that are likely to be interesting to them.