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

2. Mention a few unsupervised learning applications.

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

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

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

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

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

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

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

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:

1. Supervised learning is a type of machine learning in which the algorithm is trained on labeled data to predict or classify new data. Examples include spam filtering, sentiment analysis, and image recognition. Unsupervised learning, on the other hand, is a type of machine learning in which the algorithm is trained on unlabeled data to discover patterns or relationships within the data. Examples include clustering, anomaly detection, and dimensionality reduction.
2. Some examples of unsupervised learning applications include:

* Customer segmentation in marketing
* Fraud detection in finance
* Topic modeling in natural language processing
* Image segmentation in computer vision
* Anomaly detection in cybersecurity

1. The three main types of clustering methods are:

* Hierarchical clustering: This method creates a tree-like structure called a dendrogram that shows how the data points are grouped together. There are two types of hierarchical clustering: agglomerative and divisive. Agglomerative clustering starts with each data point as its own cluster and then merges the closest clusters until all points belong to a single cluster. Divisive clustering starts with all points in a single cluster and then recursively splits the cluster into smaller clusters until each point is in its own cluster.
* Partitioning clustering: This method divides the data into non-overlapping clusters, where each point belongs to only one cluster. The most popular partitioning clustering algorithm is k-means.
* Density-based clustering: This method clusters data points based on the density of the points. It is suitable for datasets with irregular shapes and different densities. Examples of density-based clustering algorithms include DBSCAN and OPTICS.

1. The k-means algorithm determines the consistency of clustering by minimizing the sum of squared distances between each data point and its assigned cluster center. This metric is called the within-cluster sum of squares (WCSS). The algorithm iteratively assigns data points to the nearest cluster center and updates the center based on the mean of the assigned points until the WCSS no longer decreases.
2. The key difference between k-means and k-medoids is the method used to determine the cluster center. K-means uses the mean of the assigned data points as the center, while k-medoids uses the data point itself that minimizes the sum of dissimilarities between itself and all other points in the cluster. This makes k-medoids more robust to outliers and non-normal distributions.
3. A dendrogram is a tree-like diagram that shows the hierarchical relationship between data points in a cluster analysis. It works by starting with all data points as individual clusters and then iteratively merging clusters based on a chosen distance metric. The distance between clusters can be calculated using different methods, such as single linkage, complete linkage, or average linkage.
4. SSE stands for sum of squared errors, which is a measure of how spread out the data points are within a cluster. In the k-means algorithm, the goal is to minimize the SSE by finding the optimal number of clusters and the optimal cluster centers that minimize the total SSE across all clusters.
5. The k-means algorithm can be described by the following steps:
   1. Choose the number of clusters "k"
   2. Initialize k cluster centers randomly
   3. Assign each data point to the nearest cluster center
   4. Recalculate the cluster centers as the mean of the assigned data points
   5. Repeat steps 3 and 4 until the clusters no longer change or a stopping criterion is met
6. In hierarchical clustering, single link and complete link are methods used to calculate the distance between two clusters. Single link measures the distance between the closest points in the two clusters, while complete link measures the distance between the farthest points in the two clusters. The choice of linkage method can affect the resulting dendrogram and the interpretation of the clusters.
7. The Apriori algorithm is a popular method used in market basket analysis to identify patterns of co-occurrence of items in transactions, and it can be used to reduce measurement overhead. The basic idea is that if a set of items is frequent, then all of its subsets must also be frequent. In other words, if people tend to buy bread, milk, and eggs together, then they are likely to buy bread and milk together, bread and eggs together, and milk and eggs together.

By applying the Apriori concept, we can identify the frequent itemsets in a dataset by gradually increasing the size of itemsets until we reach a minimum support threshold. Once the frequent itemsets are identified, we can then use them to identify association rules that describe the relationships between the items. This can help businesses to understand the buying habits of their customers and to make decisions about how to stock their stores, how to promote their products, and how to price their products.

For example, suppose we have a dataset that contains information about the items that customers have purchased from a grocery store. We might want to know which items are frequently purchased together so that we can arrange the store layout to make shopping easier. Using the Apriori algorithm, we can identify the frequent itemsets, such as {bread, milk, eggs}, {bread, butter, cheese}, {milk, butter, cheese}, and so on. Once we have identified the frequent itemsets, we can then use them to identify association rules, such as "if a customer buys bread, they are likely to buy milk and eggs as well." This information can help the store to stock these items near each other, making shopping more convenient for the customer and potentially increasing sales for the store.