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

**Ans:** Supervised Learning: In supervised learning, the algorithm is trained on a labeled dataset, which means it is presented with input-output pairs. The algorithm learns to predict the output from the input data. Examples of supervised learning include regression and classification tasks. For example, predicting housing prices based on features like area, number of bedrooms, and location, or classifying emails as spam or not spam.

Unsupervised Learning: Unsupervised learning, on the other hand, deals with unlabeled data. The algorithm tries to learn the patterns and relationships within the dataset without any guidance. Clustering and association are examples of unsupervised learning. An example of unsupervised learning is clustering similar documents from a large text dataset without any predefined categories.

**2. Mention a few unsupervised learning applications.**

**Ans:** Market Basket Analysis

Customer Segmentation

Anomaly Detection

Pattern Recognition

Dimensionality Reduction

Density Estimation

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

**Ans:** Partitioning Methods: Divide the data into non-overlapping groups. Example: K-means clustering.

Hierarchical Methods: Create a tree of clusters where the root is the entire dataset and the leaves are individual data points. Example: Agglomerative Hierarchical clustering.

Density-Based Methods: Based on the idea that a cluster is a dense region of data points separated by low-density regions. Example: DBSCAN (Density-Based Spatial Clustering of Applications with Noise).

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

**Ans:** K-means algorithm determines the consistency of clustering by minimizing the sum of the squared distances between the data points and their respective cluster centroids. It calculates the centroid of each cluster and assigns data points to the nearest centroid. Then, it recalculates the centroids based on the newly formed clusters, repeating this process until the centroids no longer change significantly.

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

**Ans:** K-means uses the mean value of points in a cluster as the cluster center, whereas K-medoids uses the most centrally located point in a cluster as the cluster center. K-medoids is more robust to noise and outliers than K-means.

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

**Ans:** A dendrogram is a diagram that shows the hierarchical relationship between objects. It's commonly used in hierarchical clustering. It works by illustrating how each cluster is composed by linking the clusters together based on their similarity.

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

**Ans:** SSE is a metric used in the k-means algorithm to evaluate the quality of clustering. It measures the total variation within a cluster, calculated as the sum of the squared distance between each data point and the centroid of its assigned cluster. The goal of k-means is to minimize SSE.

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

**Ans:** Choose the number of clusters, K.

Randomly initialize K cluster centroids.

Assign each data point to the nearest centroid.

Recalculate the centroids based on the newly formed clusters.

Repeat steps 3 and 4 until the centroids no longer change significantly or a specified number of iterations is reached.

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

**Ans:** Single Link: It measures the similarity between two clusters as the similarity between their two most similar members.

Complete Link: It measures the similarity between two clusters as the similarity between their two least similar members.

**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.**

**Ans:** The Apriori principle helps reduce measurement overhead in business basket analysis by focusing on frequent item sets. It works on the principle that if an itemset is frequent, then all of its subsets must also be frequent. This reduces the number of itemsets that need to be examined, making the analysis more efficient.

For example, in a retail store, if a customer frequently buys bread and milk together, the Apriori algorithm can identify this as a frequent itemset, making it easier for the store to manage their inventory and place related products together for marketing purposes.