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

|  |  |
| --- | --- |
| **Supervised Learning** | **Unsupervised Learning** |
| Supervised learning algorithms are trained using labeled data. | Unsupervised learning algorithms are trained using unlabeled data. |
| Supervised learning model takes direct feedback to check if it is predicting correct output or not. | Unsupervised learning model does not take any feedback. |
| Supervised learning model predicts the output. | Unsupervised learning model finds the hidden patterns in data. |
| In supervised learning, input data is provided to the model along with the output. | In unsupervised learning, only input data is provided to the model. |
| The goal of supervised learning is to train the model so that it can predict the output when it is given new data. | The goal of unsupervised learning is to find the hidden patterns and useful insights from the unknown dataset. |
| Supervised learning needs supervision to train the model. | Unsupervised learning does not need any supervision to train the model. |
| Supervised learning can be categorized in **Classification** and **Regression** problems. | Unsupervised Learning can be classified in **Clustering** and **Associations** problems. |
| Supervised learning can be used for those cases where we know the input as well as corresponding outputs. | Unsupervised learning can be used for those cases where we have only input data and no corresponding output data. |
| Supervised learning model produces an accurate result. | Unsupervised learning model may give less accurate result as compared to supervised learning. |
| Supervised learning is not close to true Artificial intelligence as in this, we first train the model for each data, and then only it can predict the correct output. | Unsupervised learning is more close to the true Artificial Intelligence as it learns similarly as a child learns daily routine things by his experiences. |
| It includes various algorithms such as Linear Regression, Logistic Regression, Support Vector Machine, Multi-class Classification, Decision tree, Bayesian Logic, etc. | It includes various algorithms such as Clustering, KNN, and Apriori algorithm. |

2. Mention a few unsupervised learning applications.

The main applications of unsupervised learning include **clustering, visualization, dimensionality reduction, finding association rules, and anomaly detection**.

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

* Partitioning Method.
* Hierarchical Method.
* Density-based Method.
* Grid-Based Method.
* Model-Based Method.
* Constraint-based Method.

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.

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).

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

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.

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

rror Sum of Squares (SSE) is **the sum of the squared differences between each observation and its group's mean**. It can be used as a measure of variation within a cluster. If all cases within a cluster are identical the SSE would then be equal to 0.

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

1. Step 1: Choose the number of clusters k. ...
2. Step 2: Select k random points from the data as centroids. ...
3. Step 3: Assign all the points to the closest cluster centroid. ...
4. Step 4: Recompute the centroids of newly formed clusters. ...
5. Step 5: Repeat steps 3 and 4.

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

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.

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.

Apriori algorithm assumes **that any subset of a frequent itemset must be frequent**. Its the algorithm behind Market Basket Analysis. So, according to the principle of Apriori, if {Grapes, Apple, Mango} is frequent, then {Grapes, Mango} must also be frequent.

Market Basket Analysis is **a type of frequent itemset mining** which analyzes customer buying habits by finding associations between the different items that customers place in their “shopping baskets”.