**Assignment\_18**

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

**Ans: Supervised uses labelled data and semi-supervised learning uses unlabelled data to give necessary output.**

**Supervised Learning algorithms are decision tree, logistic regression, linear regression, support vector machine.**

**Unsupervised Learning algorithms are k-means clustering, hierarchical clustering, and apriori algorithm.**

2. Mention a few unsupervised learning applications.

**Ans: 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.

**Ans: Types of Clustering**

**Centroid-based Clustering.**

**Density-based Clustering.**

**Distribution-based Clustering.**

**Hierarchical Clustering.**

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

**Ans: K-means allocates every data point in the dataset to the nearest centroid (minimizing Euclidean distances between them), meaning that a data point is considered to be in a particular cluster if it is closer to that cluster's centroid than any other centroid.**

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

**Ans: 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.

**Ans: A dendrogram is a tree-like structure that explains the relationship between all the data points in the system. However, like a regular family tree, a dendrogram need not branch out at regular intervals from top to bottom as the vertical direction (y-axis) in it represents the distance between clusters in some metric.**

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

**Ans: SSE is defined as the sum of the squared distance between centroid and each member of the cluster. Then plot a K against SSE graph. We will observe that as K increases SSE decreases as disortation will be small. So the idea of this algorithm is to choose the value of K at which the graph decrease abruptly**.

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

**Ans: Step-1: Select the number K to decide the number of clusters.**

**Step-2: Select random K points or centroids. ...**

**Step-3: Assign each data point to their closest centroid, which will form the predefined K clusters.**

**Step-4: Calculate the variance and place a new centroid of each cluster.**

**Step-5:** **Repeat the third steps, which means reassign each datapoint to the new closest centroid of each cluster.**

**Step-6:** **If any reassignment occurs, then go to step-4 else go to FINISH.**

**Step-7: The model is ready.**

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

**Ans: Single-link clusters at step are maximal sets of points that are linked via at least one link (a single link) of similarity ; complete-link clusters at step are maximal sets of points that are completely linked with each other via links of similarity**

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: Apriori algorithm is a sequence of steps to be followed to find the most frequent itemset in the given database. This data mining technique follows the join and the prune steps iteratively until the most frequent itemset is achieved. A minimum support threshold is given in the problem or it is assumed by the user.**