## Assignment 18 Solutions

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

Ans: The main distinction between the two approaches is the use of labeled datasets. To put it simply, supervised learning uses labeled input and output data, while an unsupervised learning algorithm does not. Unsupervised learning models, in contrast, work on their own to discover the inherent structure of unlabeled data.

In Supervised learning, you train the machine using data which is well “labeled.” Unsupervised learning is a machine learning technique, where you do not need to supervise the model. For example, Baby can identify other dogs based on past supervised learning.

**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: The various types of clustering are:

- Connectivity-based Clustering (Hierarchical clustering): Hierarchical Clustering is a method of unsupervised machine learning clustering where it begins with a pre-defined top to bottom hierarchy of clusters. It then proceeds to perform a decomposition of the data objects based on this hierarchy, hence obtaining the clusters. This method follows two approaches based on the direction of progress, i.e., whether it is the top-down or bottom-up flow of creating clusters.

- Centroids-based Clustering (Partitioning methods): Centroid based clustering is considered as one of the most simplest clustering algorithms, yet the most effective way of creating clusters and assigning data points to it. The intuition behind centroid based clustering is that a cluster is characterized and represented by a central vector and data points that are in close proximity to these vectors are assigned to the respective clusters.

- Density-based Clustering (Model-based methods): If one looks into the previous two methods that we discussed, one would observe that both hierarchical and centroid based algorithms are dependent on a distance (similarity/proximity) metric. The very definition of a cluster is based on this metric. Density-based clustering methods take density into consideration instead of distances. Clusters are considered as the densest region in a data space, which is separated by regions of lower object density and it is defined as a maximal-set of connected points.

- Distribution-based Clustering: Until now, the clustering techniques as we know are based around either proximity (similarity/distance) or composition (density). There is a family of clustering algorithms that take a totally different metric into consideration – probability. Distribution-based clustering creates and groups data points based on their likely hood of belonging to the same probability distribution (Gaussian, Binomial etc.) in the data.

- Fuzzy Clustering: The general idea about clustering revolves around assigning data points to mutually exclusive clusters, meaning, a data point always resides uniquely inside a cluster and it cannot belong to more than one cluster. Fuzzy clustering methods change this paradigm by assigning a data-point to multiple clusters with a quantified degree of belongingness metric. The data-points that are in proximity to the center of a cluster, may also belong in the cluster that is at a higher degree than points in the edge of a cluster. The possibility of which an element belongs to a given cluster is measured by membership coefficient that vary from 0 to 1.

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

Ans: Calculate the Within-Cluster-Sum of Squared Errors (WSS) for different values of k, and choose the k for which WSS becomes first starts to diminish. In the plot of WSS-versus-k, this is visible as an elbow. Within-Cluster-Sum of Squared Errors sounds a bit complex.

**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 diagram that shows the attribute distances between each pair of sequentially merged classes After each merging, the distances between all pairs of classes are updated. The distances at which the signatures of classes are merged are used to construct a dendrogram.

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

Ans: I am going to refer to it as SSE, which stands for Sum of Squared Errors. The regression line is the line made using the function we defined above. To get the SSE we calculate the distance for each of the data points from the regression line then square the it, then we add to the sum.

The SSE is defined as the sum of the squared Euclidean distances of each point to its closest centroid. Since this is a measure of error, the objective of k-means is to try to minimize this value. The purpose of this figure is to show that the initialization of the centroids is an important step.

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

Ans: k-means is one of the simplest unsupervised learning algorithms that solve the well known clustering problem. The procedure follows a simple and easy way to classify a given data set through a certain number of clusters (assume k clusters) fixed apriori. The main idea is to define k centers, one for each cluster.

Step 1: Choose the number of clusters k.

Step 2: Select k random points from the data as centroids.

Step 3: Assign all the points to the closest cluster centroid.

Step 4: Recompute the centroids of newly formed clusters.

Step 5: Repeat steps 3 and 4.

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

Ans: 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 ?**

Ans: 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. Here is a dataset consisting of six transactions."