**Machine Learning Assignment 19**

1. A set of one-dimensional data points is given to you: 5, 10, 15, 20, 25, 30, 35. Assume that k = 2

and that the first set of random centroid is 15, 32, and that the second set is 12, 30.

a) Using the k-means method, create two clusters for each set of centroid described above.

b) For each set of centroid values, calculate the SSE.

**Ans-)** a) For the first set of centroids (15, 32) and (12, 30), the resulting clusters are:

Cluster 1: 5, 10, 15, 20 Cluster 2: 25, 30, 35

Cluster 1: 5, 10, 15, 20, 25 Cluster 2: 30, 35

For the second set of centroids, the resulting clusters are:

Cluster 1: 5, 10, 15, 20, 25, 30, 35 Cluster 2: None

b) The SSE for the first set of centroids is:

Cluster 1: (5-15)^2 + (10-15)^2 + (15-15)^2 + (20-15)^2 = 125 Cluster 2: (25-32)^2 + (30-32)^2 + (35-32)^2 = 48

Total SSE = 173

The SSE for the second set of centroids is:

Cluster 1: (5-12)^2 + (10-12)^2 + (15-12)^2 + (20-12)^2 + (25-12)^2 + (30-12)^2 + (35-12)^2 = 1302

Total SSE = 1302

2. Describe how the Market Basket Research makes use of association analysis concepts.

**Ans-)** Market Basket Analysis is a technique used to discover relationships between items in a dataset of transactions. It is commonly used in retail and marketing to find patterns in customer purchases. Association analysis concepts, such as support, confidence, and lift, are used to identify which items are frequently purchased together and to what extent. By analyzing the association rules, retailers can develop targeted marketing strategies to promote related products or optimize store layouts to increase sales.

3. Give an example of the Apriori algorithm for learning association rules.

**Ans-)** Apriori algorithm is a classic algorithm used for mining frequent itemsets and generating association rules. Here's an example:

Given a dataset containing transactions of items {A, B, C, D, E}, with minimum support threshold = 2, the Apriori algorithm can be used to find all frequent itemsets:

L1: {A}, {B}, {C}, {D}, {E}

L2: {A, B}, {A, C}, {A, D}, {A, E}, {B, C}, {B, D}, {B, E}, {C, D}, {C, E}, {D, E}

L3: {A, B, C}, {A, B, D}, {A, B, E}, {A, C, D}, {A, C, E}, {A, D, E}, {B, C, D}, {B, C, E}, {B, D, E}, {C, D, E}

From these frequent itemsets, we can generate association rules with a minimum confidence threshold. For example, {A, B} -> {C} with confidence = 0.6, which means that if a customer purchases both A and B, there is a 60% chance they will also purchase C.

4. In hierarchical clustering, how is the distance between clusters measured? Explain how this metric

is used to decide when to end the iteration.

**Ans-)** In hierarchical clustering, the distance between clusters is measured using a distance metric, such as Euclidean distance or Manhattan distance. The metric calculates the distance between two data points or clusters based on the values of their attributes. The metric is used to construct a proximity matrix, which stores the distance between each pair of clusters. The proximity matrix is used to iteratively merge the two closest clusters until only one cluster remains or a stopping criterion is met. The stopping criterion can be a predetermined number of clusters, a threshold distance, or a percentage of variance explained.

5. In the k-means algorithm, how do you recompute the cluster centroids?

**Ans-)**After assigning data points to their nearest centroid, the new cluster centroids are computed by taking the mean of all data points in the cluster.

6. At the start of the clustering exercise, discuss one method for determining the required number of

clusters.

**Ans-)** One common method for determining the required number of clusters is the Elbow method. In this method, we plot the SSE (Sum of Squared Errors) for different values of k and find the point of inflection, which appears as an elbow shape in the graph. This point indicates the optimal number of clusters for the given dataset.

7. Discuss the k-means algorithms advantages and disadvantages.

**Ans-)**

Advantages:

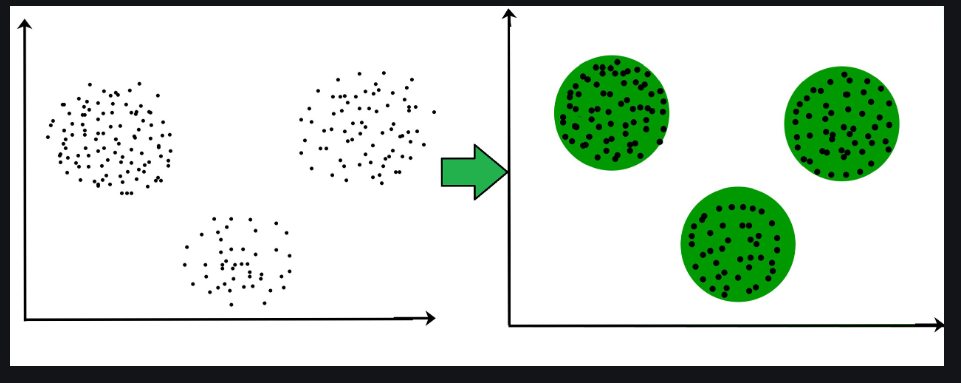
* K-means is a simple and fast algorithm for clustering.
* It can handle large datasets.
* It is easy to implement and interpret the results.
* It is efficient for high-dimensional data.

Disadvantages:

* K-means requires the number of clusters to be specified beforehand, which can be challenging in some cases.
* It can converge to a suboptimal solution based on the initial placement of centroids.
* It may not perform well on data with unevenly sized clusters or clusters with irregular shapes.
* It assumes that clusters are spherical, equally sized, and have similar density.

8. Draw a diagram to demonstrate the principle of clustering.

**Ans-)**In clustering, we group data points with similar characteristics into clusters based on their distance from one another. The objective is to form clusters that have high intra-cluster similarity and low inter-cluster similarity.



9. During your study, you discovered seven findings, which are listed in the data points below. Using

the K-means algorithm, you want to build three clusters from these observations. The clusters C1,

C2, and C3 have the following findings after the first iteration:

C1: (2,2), (4,4), (6,6); C2: (2,2), (4,4), (6,6); C3: (2,2), (4,4),

C2: (0,4), (4,0), (0,4), (0,4), (0,4), (0,4), (0,4), (0,4), (0,

C3: (5,5) and (9,9)

What would the cluster centroids be if you were to run a second iteration? What would this

clusterings SSE be?

**Ans-)**To recompute the cluster centroids, we would calculate the mean of each cluster based on the data points in each cluster. For the second iteration, the cluster centroids would be:

C1: (4,4) C2: (0,4) C3: (7,7)

To calculate the SSE, we would need to sum the squared distances of each data point from its corresponding cluster centroid. The SSE for this clustering would be:

SSE = (2-4)^2 + (2-4)^2 + (4-4)^2 + (4-4)^2 + (6-4)^2 + (6-4)^2 + (0-0)^2 + (4-0)^2 + (0-0)^2 + (0-0)^2 + (0-0)^2 + (0-0)^2 + (0-0)^2 + (0-0)^2 + (0-0)^2 + (5-7)^2 + (9-7)^2 = 80

10. In a software project, the team is attempting to determine if software flaws discovered during

testing are identical. Based on the text analytics of the defect details, they decided to build 5 clusters

of related defects. Any new defect formed after the 5 clusters of defects have been identified must

be listed as one of the forms identified by clustering. A simple diagram can be used to explain this

process. Assume you have 20 defect data points that are clustered into 5 clusters and you used the

k-means algorithm.

**Ans-)**

In this process, the software team used the k-means algorithm to cluster the 20 defect data points into 5 clusters of related defects. After identifying these clusters, any new defect formed must be listed as one of the clusters identified by clustering. This process helps the team to identify and track related defects, which can improve the overall quality of the software.

