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.

a--->To create two clusters using the k-means method for the first set of centroids (15, 32), we would follow these steps:

Calculate the distance between each data point and the two centroids. We can use the Euclidean distance formula to do this: sqrt((x1 - x2)^2 + (y1 - y2)^2)

For example, to calculate the distance between the data point 5 and the centroid 15, we would do:

sqrt((5 - 15)^2 + (0 - 0)^2) = 10

Assign each data point to the cluster with the nearest centroid. Using the distances calculated above, we would assign the following data points to each cluster: Cluster 1 (centroid 15): 5, 10, 15 Cluster 2 (centroid 32): 20, 25, 30, 35

Recalculate the centroids for each cluster by taking the mean of the data points in each cluster. For cluster 1, the new centroid would be: (5 + 10 + 15)/3 = 10

For cluster 2, the new centroid would be:

(20 + 25 + 30 + 35)/4 = 27.5

Repeat steps 1-3 until the centroids do not change or a maximum number of iterations is reached. To create two clusters using the k-means method for the second set of centroids (12, 30), we would follow the same steps as above, using the new centroids. The resulting clusters would be:

Cluster 1 (centroid 12): 5, 10 Cluster 2 (centroid 30): 15, 20, 25, 30, 35

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

Answer:--->In market basket analysis, association rules are used to predict the likelihood of products being purchased together. Association rules count the frequency of items that occur together, seeking to find associations that occur far more often than expected.The goal of Market Basket Analysis is to understand consumer behavior by identifying relationships between the items that people buy. For example, people who buy green tea are also likely to buy honey. So Market Basket Analysis would quantitatively establish that there is a relationship between Green Tea and Honey.The main objective of market basket analysis is to identify products that customers want to purchase. Market basket analysis enables sales and marketing teams to develop more effective product placement, pricing, cross-sell, and up-sell strategies.

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

Answer:--->Apriori Algorithm is one of the algorithm used for transaction data in Association Rule Learning. It allows us to mine the frequent itemset in order to generate association rule between them. Example: list of items purchased by customers, details of website which are frequently visited etc.Apriori algorithm refers to an algorithm that is used in mining frequent products sets and relevant association rules. Generally, the apriori algorithm operates on a database containing a huge number of transactions. For example, the items customers but at a Big Bazar.

4. In hierarchical clustering, how is the distance between clusters measured? Explain how this metric is used to decide when to end the iteration.

Answer:---->In complete linkage hierarchical clustering, the distance between two clusters is defined as the longest distance between two points in each cluster. For example, the distance between clusters “r” and “s” to the left is equal to the length of the arrow between their two furthest points. Average Linkage: In average linkage, we define the distance between two clusters to be the average distance between data points in the first cluster and data points in the second cluster.To get the optimal number of clusters for hierarchical clustering, we make use a dendrogram which is tree-like chart that shows the sequences of merges or splits of clusters. If two clusters are merged, the dendrogram will join them in a graph and the height of the join will be the distance between those clusters

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

Answer:--->Select k centroids. These will be the center point for each segment. Assign data points to nearest centroid.

Reassign centroid value to be the calculated mean value for each cluster.

Reassign data points to nearest centroid.

Repeat until data points stay in the same cluster. K-means clustering algorithm computes the centroids and iterates until we it finds optimal centroid. It assumes that the number of clusters are already known. It is also called flat clustering algorithm. The number of clusters identified from data by algorithm is represented by 'K' in K-means.Initial K cluster centroids are selected randomly from the observations. Distance between each observation and clusters' centroid is calculated and the observation is assigned to a cluster with minimal distance from the centroid of that cluster.

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

Answer:---->A simple method to calculate the number of clusters is to set the value to about √(n/2) for a dataset of 'n' points.Optimization techniques such as genetic algorithms are useful in determining the number of clusters that gives rise to the largest silhouette.The silhouette coefficient may provide a more objective means to determine the optimal number of clusters. This is done by simply calculating the silhouette coefficient over a range of k, and identifying the peak as the optimum K.To get the optimal number of clusters for hierarchical clustering, we make use a dendrogram which is tree-like chart that shows the sequences of merges or splits of clusters. If two clusters are merged, the dendrogram will join them in a graph and the height of the join will be the distance between those clusters

7. Discuss the k-means algorithm's advantages and disadvantages.

Answer:--->Advantages of k-means Guarantees convergence. Can warm-start the positions of centroids. Easily adapts to new examples. Generalizes to clusters of different shapes and sizes, such as elliptical clusters.The main advantage of a clustered solution is automatic recovery from failure, that is, recovery without user intervention. Disadvantages of clustering are complexity and inability to recover from database corruption.Guarantees convergence. Can warm-start the positions of centroids. Easily adapts to new examples. Generalizes to clusters of different shapes and sizes, such as elliptical clusters.With a large number of variables, K--Means may be computa onally faster than hierarchical clustering (if K is small). • k--Means may produce ghter clusters than hierarchical clustering. • An instance can change cluster (move to another cluster) when the centroids are re-- computed.

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

Answer:---->Hierarchical cluster analysis begins by separating each object into a cluster by itself. At each stage of the analysis, the criterion by which objects are separated is relaxed in order to link the two most similar clusters until all of the objects are joined in a complete classification tree.The k-means algorithm searches for a pre-determined number of clusters within an unlabeled multidimensional dataset. It accomplishes this using a simple conception of what the optimal clustering looks like: The "cluster center" is the arithmetic mean of all the points belonging to the clusterClustering is the task of dividing the population or data points into a number of groups such that data points in the same groups are more similar to other data points in the same group than those in other groups. In simple words, the aim is to segregate groups with similar traits and assign them into clusters.

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

Answer:--->In the second iteration of the K-means algorithm, the cluster centroids would be recalculated based on the current observations in each cluster.

For cluster C1, the centroid would be calculated as the average of the coordinates of the observations in the cluster. This would give a centroid of (4,4).

For cluster C2, the centroid would be calculated as the average of the coordinates of the observations in the cluster. This would give a centroid of (0,4).

For cluster C3, the centroid would be calculated as the average of the coordinates of the observations in the cluster. This would give a centroid of (7,7).

To calculate the SSE (sum of squared errors) for this clustering, you would sum the squared distance of each observation to its corresponding cluster centroid. For example, for the first observation in cluster C1 (2,2), the SSE would be calculated as (2-4)^2 + (2-4)^2 = 4. The SSE for the entire clustering would be the sum of the SSE for each observation.

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.

Answer:----->

The process of using the K-means algorithm to cluster the defects in a software project can be represented using the following diagram:

The team begins by selecting the number of clusters they want to create (in this case, 5 clusters).

The team then selects the initial centroids for each cluster. These can be chosen randomly or using some other method.

The team then assigns each defect data point to the cluster with the closest centroid, based on some distance metric (such as Euclidean distance).

The team then recalculates the centroids for each cluster based on the observations in the cluster.

The team repeats steps 3 and 4 until the centroids stabilize, or until a predetermined number of iterations has been reached.

Once the clustering process is complete, the team can examine the clusters and use the insights gained to identify and resolve common defects.

When a new defect is discovered, it can be added to one of the existing clusters or a new cluster can be created if the defect is significantly different from the existing clusters.