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.

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

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

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

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

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

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

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

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 clustering's SSE be?

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:

1.  
a) Using the k-means method with the first set of random centroids:

* First iteration:
  + Cluster 1: [5, 10, 15, 20], Centroid = 12.5
  + Cluster 2: [25, 30, 35], Centroid = 30
* Second iteration:
  + Cluster 1: [5, 10, 15, 20], Centroid = 12.5
  + Cluster 2: [25, 30, 35], Centroid = 30

Using the k-means method with the second set of random centroids:

* First iteration:
  + Cluster 1: [5, 10, 15], Centroid = 10
  + Cluster 2: [20, 25, 30, 35], Centroid = 27.5
* Second iteration:
  + Cluster 1: [5, 10, 15], Centroid = 10
  + Cluster 2: [20, 25, 30, 35], Centroid = 27.5

b) SSE for the first set of random centroids:

* SSE for Cluster 1 = [(5-12.5)^2 + (10-12.5)^2 + (15-12.5)^2 + (20-12.5)^2] = 125
* SSE for Cluster 2 = [(25-30)^2 + (30-30)^2 + (35-30)^2] = 50
* Total SSE = 125 + 50 = 175

SSE for the second set of random centroids:

* SSE for Cluster 1 = [(5-10)^2 + (10-10)^2 + (15-10)^2] = 50
* SSE for Cluster 2 = [(20-27.5)^2 + (25-27.5)^2 + (30-27.5)^2 + (35-27.5)^2] = 187.5
* Total SSE = 50 + 187.5 = 237.5

1. Market Basket Research utilizes association analysis concepts to identify relationships between products that are purchased together. It helps to uncover which items are most frequently bought together and which products are most likely to be purchased if the customer has already purchased a specific product. For instance, if a customer buys a computer, they are more likely to purchase a printer. By identifying these relationships, businesses can increase sales by suggesting complementary products or using cross-selling strategies.
2. Example of the Apriori algorithm for learning association rules:

Suppose we have a dataset with transactions, and we want to find association rules with a minimum support of 50% and a minimum confidence of 80%. The steps are as follows:

* Step 1: Identify all the frequent items in the dataset (items with a support greater than or equal to the minimum support threshold).
* Step 2: Generate all possible rules from the frequent itemsets.
* Step 3: Calculate the support and confidence for each rule.
* Step 4: Prune the rules that do not meet the minimum confidence threshold.
* Step 5: Return the remaining rules that meet the minimum support and minimum confidence thresholds.

1. In hierarchical clustering, the distance between clusters is measured using various distance metrics such as Euclidean distance, Manhattan distance, or cosine distance. The metric is used to decide when to end the iteration by measuring the distance between the clusters at each step. If the distance between clusters is below a certain threshold, or if the number of clusters is reduced to a predefined number, the iteration stops, and the final clusters are formed.
2. In the k-means algorithm, the cluster centroids are recomputed as follows:

* For each cluster, the centroid is recalculated as the mean of all the data points assigned to that cluster.
* This new set of centroids is used to reassign each data point to its nearest cluster.
* The process of reassigning data points and recalculating centroids is repeated until the cluster assignments no longer change or a maximum number of iterations is reached.

1. One method for determining the required number of clusters at the start of the clustering exercise is the elbow method. This method involves plotting the SSE (sum of squared errors) as a function of the number of clusters, and looking for the "elbow point" where the SSE begins to level off. This point indicates that the additional clusters are not significantly reducing the SSE.
2. Advantages of the k-means algorithm include its simplicity and efficiency, making it suitable for large datasets. It is also easy to interpret the results of k-means clustering. However, k-means is sensitive to initial cluster centroids, and the results may be affected by outliers or noise in the data. It also assumes that the clusters are roughly spherical and of equal size, which may not be appropriate for all datasets.
3. A diagram to demonstrate the principle of clustering might show a scatter plot of data points, with clusters indicated by different colors or symbols. The cluster centroids could be marked with larger symbols, and lines could be drawn to indicate the boundaries between clusters.
4. If you were to run a second iteration of k-means clustering on the data points listed, the new cluster centroids would be:

C1: (4,4), (6,6)

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

C3: (2,2), (4,4), (5,5), (6,6), (9,9)

The SSE for this clustering would be: SSE = 2\*(0^2 + 0^2 + 0^2) + 9\*(4^2) + 5\*(1^2 + 1^2 + 1^2 + 1^2 + 1^2) = 69

1. A simple diagram to explain this process might show the 20 data points as circles, with the five clusters indicated by different colors or symbols. The team could then label each cluster with a brief description of the types of defects it represents. Any new defect discovered during testing could be assigned to the appropriate cluster based on its characteristics, ensuring that similar defects are tracked and addressed consistently.