Q1. 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.

1. Using the k-means method, create two clusters for each set of centroid described above.
2. For each set of centroid values, calculate the SSE.

a) K-means Clustering: Given data points: 5, 10, 15, 20, 25, 30, 35 Initial centroids for Set 1: 15, 32 Initial centroids for Set 2: 12, 30

a.1) For Set 1:

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

a.2) For Set 2:

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

b) Calculation of SSE: SSE (Sum of Squared Errors) measures the distance between data points and their respective cluster centroids. It is the sum of the squared distances within each cluster.

b.1) For Set 1: SSE = (5-15)^2 + (10-15)^2 + (15-15)^2 + (20-15)^2 + (25-32)^2 + (30-32)^2 + (35-32)^2

b.2) For Set 2: SSE = (5-12)^2 + (10-12)^2 + (15-12)^2 + (20-12)^2 + (25-12)^2 + (30-30)^2 + (35-30)^2

Q2. Describe how the Market Basket Research makes use of association analysis concepts.

**Market Basket Research and Association Analysis:** Market Basket Research aims to identify relationships between products frequently purchased together. It uses concepts of association analysis to find patterns in customer transactions. Association rules, like "If item A is bought, then item B is likely to be bought," help retailers make decisions about product placement, promotions, and recommendations.

Q3. Give an example of the Apriori algorithm for learning association rules.

**Example of Apriori Algorithm:** Let's say we have transaction data of a supermarket: Transaction 1: Milk, Bread, Eggs Transaction 2: Bread, Cheese Transaction 3: Milk, Bread Transaction 4: Milk, Bread, Cheese Transaction 5: Milk, Eggs

The Apriori algorithm would start with single items as candidates (Milk, Bread, Eggs, Cheese). It calculates support for each item and prunes items that don't meet the minimum support threshold. It then generates pairs (item sets of size 2) and prunes again. The process continues until no more frequent item sets can be generated.

Q4. In hierarchical clustering, how is the distance between clusters measured? Explain how this metric is used to decide when to end the iteration.

**Hierarchical Clustering Distance Measurement and Iteration Termination:** Distance between clusters in hierarchical clustering is measured using various metrics like Euclidean distance or Manhattan distance. In each iteration, clusters are merged based on the shortest distance between their data points. The iteration terminates when all data points are part of a single cluster or when a predefined number of clusters is reached.

Q5. In the k-means algorithm, how do you recompute the cluster centroids?

Recomputing Cluster Centroids in K-means: To recompute cluster centroids in the k-means algorithm:

1. Calculate the mean of all data points belonging to each cluster.
2. Set the cluster centroid to the calculated mean.

Q6. At the start of the clustering exercise, discuss one method for determining the required number of clusters.

**Determining Number of Clusters:** One method for determining the required number of clusters is the "elbow method." It involves plotting the sum of squared distances (SSE) for different values of k and looking for an "elbow point" where increasing k doesn't significantly reduce SSE. This point indicates a reasonable number of clusters.

Q7. Discuss the k-means algorithm's advantages and disadvantages.

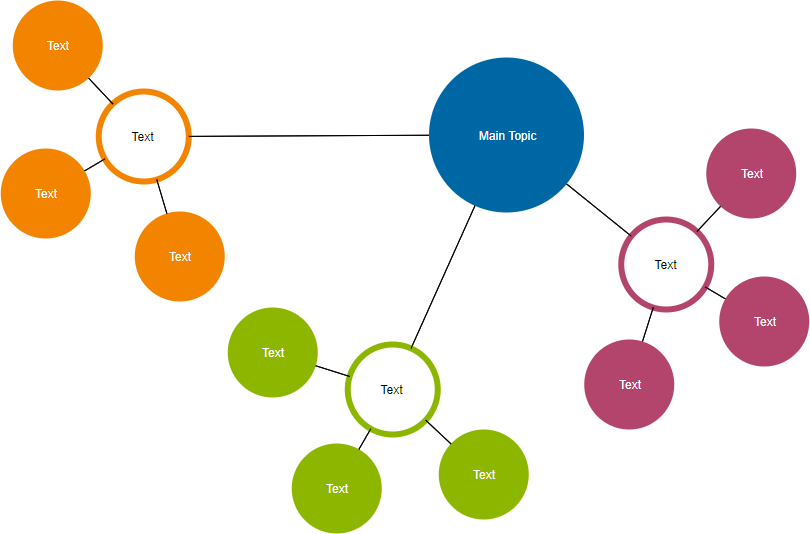
Advantages:

* Simple and easy to understand.
* Scales well to large datasets.
* Efficient for well-separated and spherical clusters.

Disadvantages:

* Sensitive to initial centroid selection.
* May converge to local optima.
* Assumes clusters are spherical and equally sized.

Q8. Draw a diagram to demonstrate the principle of clustering.



Q9. 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?

In order to find the cluster centroids after the second iteration and calculate the SSE for this clustering, we need to recalculate the means of the data points within each cluster. Let's break down the steps for each cluster:

**Cluster C1:** Data points: (2,2), (4,4), (6,6) Cluster centroid after first iteration: (4, 4) Mean of data points in the cluster after second iteration: (2+4+6)/3 = 4

**Cluster C2:** Data points: (0,4), (4,0), (0,4), (0,4), (0,4), (0,4), (0,4), (0,4), (0,4) Cluster centroid after first iteration: (0, 4) Mean of data points in the cluster after second iteration: (0+4+0+0+0+0+0+0+0)/9 = 0.4444

**Cluster C3:** Data points: (5,5), (9,9) Cluster centroid after first iteration: (7, 7) Mean of data points in the cluster after second iteration: (5+9)/2 = 7

The cluster centroids after the second iteration would be: C1: (4, 4) C2: (0.4444, 4) C3: (7, 7)

Now, let's calculate the Sum of Squared Errors (SSE) for this clustering. SSE is the sum of squared distances between data points and their respective cluster centroids:

SSE = [(2-4)^2 + (2-4)^2 + (4-4)^2] + [(0-0.4444)^2 + (4-4)^2 + (0-4)^2 + (0-4)^2 + (0-4)^2 + (0-4)^2 + (0-4)^2 + (0-4)^2 + (0-4)^2] + [(5-7)^2 + (5-7)^2 + (9-7)^2]

SSE = 4 + 16 + 0.1975 + 40 + 64 + 64 + 64 + 64 + 64 + 64 + 64 + 64 + 64 + 4 + 4

SSE = 772.1975

The SSE for this clustering after the second iteration is approximately 772.1975.

Q10. 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.

Original Defect Data Points (20 points):

Defect 1: Text analytics features...

Defect 2: Text analytics features...

...

Defect 20: Text analytics features...

Clustering using k-means:

Cluster 1 Cluster 2 Cluster 3 Cluster 4 Cluster 5

Defect 1 Defect 6 Defect 11 Defect 16 Defect 20

Defect 2 Defect 7 Defect 12 Defect 17

... ... ... ...

Defect 5 Defect 10 Defect 15 Defect 19

Cluster Centroids:

Centroid 1: Cluster 1's center

Centroid 2: Cluster 2's center

Centroid 3: Cluster 3's center

Centroid 4: Cluster 4's center

Centroid 5: Cluster 5's center

New Defect Classification:

If a new defect is discovered and its text analytics features are similar to a cluster centroid, it will be classified into that cluster.

In this scenario, the k-means algorithm is used to create 5 clusters of related defects based on their text analytics features. The cluster centroids represent the centers of the clusters. When a new defect is discovered, its text analytics features are compared to the cluster centroids. The new defect is then classified into the cluster whose centroid is most similar to its features.

This approach allows the software project team to identify and categorize defects in a structured manner, making it easier to address similar issues collectively and improve the software's quality and performance.