1. K-means Clustering and SSE Calculation

a) Creating Clusters

First Set of Centroids (15, 32):

1. Calculate distances:

- Distance from each data point to 15: |5-15|, |10-15|, |15-15|, |20-15|, |25-15|, |30-15|, |35-15|

- Distance from each data point to 32: |5-32|, |10-32|, |15-32|, |20-32|, |25-32|, |30-32|, |35-32|

2. Assign data points to clusters:

- Cluster 1 (centroid = 15): Points closer to 15.

- Cluster 2 (centroid = 32): Points closer to 32.

Second Set of Centroids (12, 30):

1. Calculate distances:

- Distance from each data point to 12: |5-12|, |10-12|, |15-12|, |20-12|, |25-12|, |30-12|, |35-12|

- Distance from each data point to 30: |5-30|, |10-30|, |15-30|, |20-30|, |25-30|, |30-30|, |35-30|

2. Assign data points to clusters:

- Cluster 1 (centroid = 12): Points closer to 12.

- Cluster 2 (centroid = 30): Points closer to 30.

b) Calculating SSE (Sum of Squared Errors)

First Set of Centroids (15, 32):

1. Assign clusters:

- Cluster 1 (15): 5, 10, 15, 20, 25

- Cluster 2 (32): 30, 35

2. Calculate SSE:

- SSE = Σ(distance^2 of each point from its centroid)

Second Set of Centroids (12, 30):

1. Assign clusters:

- Cluster 1 (12): 5, 10, 15

- Cluster 2 (30): 20, 25, 30, 35

2. Calculate SSE:

- SSE = Σ(distance^2 of each point from its centroid)

2. Market Basket Research and Association Analysis

Market Basket Analysis uses association analysis to identify sets of items that frequently co-occur in transactions. This concept is used to understand customer purchasing patterns, discover relationships between products, and make recommendations. Association rules like "If a customer buys bread, they are likely to buy butter" help in cross-selling and promotions.

3. Example of Apriori Algorithm

Steps:

1. Identify frequent itemsets:

- Minimum support threshold.

- Count the occurrences of itemsets.

- Prune non-frequent itemsets.

2. Generate association rules:

- Minimum confidence threshold.

- Derive rules from frequent itemsets.

Example:

1. Transactions: {Milk, Bread}, {Milk, Diaper, Beer, Bread}, {Milk, Diaper, Beer, Coke}, {Bread, Butter}, {Milk, Diaper, Bread, Butter}

2. Frequent itemsets (min support = 50%): {Milk, Bread}, {Milk, Diaper}, {Bread, Butter}

3. Association rules (min confidence = 60%):

- If {Milk} then {Bread}

- If {Bread} then {Butter}

4. Hierarchical Clustering Distance Measurement

Distance Measurement:

- Single Linkage (minimum distance): Distance between the closest points of two clusters.

- Complete Linkage (maximum distance): Distance between the farthest points.

- Average Linkage: Average distance between all pairs of points.

- Centroid Linkage: Distance between the centroids of the clusters.

Iteration End:

- Continue merging clusters until the desired number of clusters is achieved or the distance between merged clusters exceeds a threshold.

5. Recomputing Cluster Centroids in K-means

After assigning each data point to the nearest centroid, recompute the centroid by calculating the mean of all points in the cluster:

\[ \text{New Centroid} = \frac{1}{n} \sum\_{i=1}^{n} x\_i \]

where \( n \) is the number of points in the cluster and \( x\_i \) are the data points.

6. Determining the Number of Clusters

Elbow Method:

- Plot SSE against the number of clusters.

- The "elbow" point where the rate of decrease sharply slows indicates an optimal number of clusters.

7. Advantages and Disadvantages of K-means Algorithm

Advantages:

- Simple and easy to implement.

- Scalable for large datasets.

- Works well for spherical clusters.

Disadvantages:

- Sensitive to initial centroids.

- Not suitable for non-spherical clusters.

- Requires pre-specifying the number of clusters.

8. Diagram Demonstrating Clustering

A simple diagram can include data points scattered on a 2D plane with different shapes/colors to represent different clusters, and centroids marked with a special symbol (e.g., a star).

9. K-means Clustering and SSE Calculation for Second Iteration

Given clusters:

- C1: (2,2), (4,4), (6,6)

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

- C3: (5,5), (9,9)

Recompute centroids:

\[ \text{Centroid C1} = \left( \frac{2+4+6}{3}, \frac{2+4+6}{3} \right) = (4,4) \]

\[ \text{Centroid C2} = \left( \frac{0+4+0+0+0+0+0+0}{8}, \frac{4+0+4+4+4+4+4+4}{8} \right) = (0.5, 3.5) \]

\[ \text{Centroid C3} = \left( \frac{5+9}{2}, \frac{5+9}{2} \right) = (7,7) \]

Calculate SSE:

\[ \text{SSE} = \sum\_{\text{each cluster}} \sum\_{\text{each point}} (\text{distance from centroid})^2 \]

10. Clustering Software Defects

To identify related software defects, text analytics is performed, and defects are grouped into clusters. Each new defect is compared to the centroids of existing clusters to determine its category.

Diagram:

- 20 defect points clustered into 5 clusters using k-means.

- New defects are assigned to the nearest cluster based on text similarity.