| ML\_ASSISMENT\_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.**

**Sol:**

**For 15 and 32**

| **Data point** | **Distance to centroid (15)** | **Distance to centroid (32)** | **Assigned Cluster** |
| --- | --- | --- | --- |
| 5 | 10 | 27 | Centroid 15 |
| 10 | 5 | 22 | Centroid 15 |
| 15 | 0 | 17 | Centroid 15 |
| 20 | 5 | 12 | Centroid 15 |
| 25 | 10 | 7 | Centroid 32 |
| 30 | 15 | 2 | Centroid 32 |
| 35 | 20 | 3 | Centroid 32 |

New Centroid for Cluster 1: (5 + 10 + 15 + 20) / 4 = 12.5

New Centroid for Cluster 2: (25 + 30 + 35) / 3 = 30

Calculate the squared distances of each point from its cluster centroid:

For Cluster 1 (centroid 12.5):

(5 - 12.5)2 = 56.25

(10 - 12.5)2 = 6.25

(15 - 12.5)2 = 6.25

(20 - 12.5)2 = 56.25

For Cluster 2 (centroid 30):

(25 - 30)2 = 25

(30 - 30)2 = 0

(35 - 30)2 = 25

Sum up the squared distances for each point:

SSE = 56.25 + 6.25 + 6.25 + 56.25 + 25 + 0 + 25 = 175.25

**For 12 and 30**

| **Data point** | **Distance to centroid (12)** | **Distance to centroid (30)** | **Assigned Cluster** |
| --- | --- | --- | --- |
| 5 | 7 | 25 | Centroid 12 |
| 10 | 2 | 20 | Centroid 12 |
| 15 | 3 | 15 | Centroid 12 |
| 20 | 8 | 10 | Centroid 12 |
| 25 | 13 | 5 | Centroid 30 |
| 30 | 18 | 0 | Centroid 30 |
| 35 | 23 | 5 | Centroid 30 |

New Centroid for Cluster 1: (5 + 10 + 15 + 20) / 4 = 12.5

New Centroid for Cluster 2: (25 + 30 + 35) / 3 = 30

or Cluster 1 (centroid 12.5):

(5 - 12.5)2 = 56.25

(10 - 12.5)2 = 6.25

(15 - 12.5)2 = 6.25

(20 - 12.5)2 = 56.25

For Cluster 2 (centroid 30):

(25 - 30)2 = 25

(30 - 30)2 = 0

(35 - 30)2 = 25

Sum up the squared distances for each point:

SSE = 56.25 + 6.25 + 6.25 + 56.25 + 25 + 0 + 25 = 175.25

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

Sol:

* Market Basket Analysis is a technique used in retail and marketing to discover associations or relationships between products that customers tend to purchase together.
* Association analysis concepts, particularly the Apriori algorithm, play a crucial role in this process.
* The main idea is to identify frequent item sets and generate association rules that describe the patterns in customer purchasing behavior.

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

**Sol:**

Let's say we have a transaction dataset from a grocery store. Each transaction consists of a list of items purchased. The Apriori algorithm works in the following steps:

* Find all single items that occur above a minimum support threshold (e.g., items bought by at least 100 customers).
* Use these frequent single items to generate pairs of items and calculate their support.
* Remove pairs that do not meet the minimum support.
* Use the frequent pairs to generate triples and calculate their support.
* Continue this process until no more frequent item sets can be generated.

For instance, if the minimum support threshold is met, the algorithm could find that "bread" and "milk" are frequently purchased together. This information can lead to strategies like placing these items closer in the store to encourage further sales.

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

**Sol:**

* In hierarchical clustering, the distance between clusters is typically measured using various distance metrics, such as Euclidean distance or Manhattan distance.
* These metrics quantify the dissimilarity between two clusters based on the distances between their member points. The linkage criterion (how to measure the distance between clusters) affects the outcome of the clustering. Common linkage methods include single linkage, complete linkage, and average linkage.
* The iteration ends in hierarchical clustering when the desired number of clusters is achieved.
* This is often determined by interpreting a dendrogram, a tree-like diagram that illustrates the arrangement of clusters at different levels of the hierarchy.
* The height on the dendrogram represents the distance at which clusters merge. Cutting the dendrogram at a certain height results in a specific number of clusters.

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

Sol:

In the k-means algorithm:

* Assign each data point to the nearest cluster centroid.
* Recompute the cluster centroids by taking the mean of all data points assigned to each cluster.
* Repeat the above steps until convergence (when cluster assignments no longer change significantly).

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

**Sol:**

* One method for determining the required number of clusters is the "Elbow Method."
* In this approach, we plot the sum of squared distances (inertia) between data points and their assigned cluster centroids for a range of cluster numbers.
* The point where the plot starts to show diminishing returns (the "elbow" point) is often considered a good estimate for the optimal number of clusters.

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

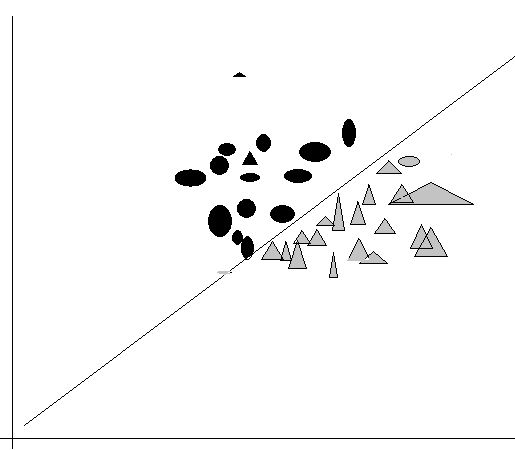
**Sol:**

Advantages:

* Simple and easy to implement.
* Scales well to large datasets.
* Fast convergence due to its iterative nature.

Disadvantages:

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

**8. Draw a diagram to demonstrate the principle of clustering.Sol:**

* Clustering aims to group similar points together based on some

similarity metric (distance).

* Points within the same cluster are closer to each other than to

points in other clusters.

* This creates distinct groups or clusters of points in the plot, each

potentially representing a meaningful pattern or category.

**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: (0,4), (4,0), (0,4), (0,4), (0,4), (0,4), (0,4), (0,4)

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?**

**Sol:**Calculate the centroids for the second iteration:

**For Cluster C1:** Centroid = ((2 + 4 + 6) / 3, (2 + 4 + 6) / 3) = (4, 4)

**For Cluster C2:** Centroid = ((0 + 4 + 0 + 0 + 0 + 0 + 0 + 0) / 8, (4 + 0 + 4 + 4 + 4 + 4 + 4 + 4) / 8) = (0.5, 3.5)

**For Cluster C3:** Centroid = ((5 + 9) / 2, (5 + 9) / 2) = (7, 7)

So, after the second iteration, the updated cluster centroids would be:

C1: (4, 4)

C2: (0.5, 3.5)

C3: (7, 7)

**Cluster C1: (2,2), (4,4), (6,6)**

**Centroid: (4, 4)**

**Squared distances from centroid for each point:**

(2,2): (2-4)^2 + (2-4)^2 = 4 + 4 = 8

(4,4): (4-4)^2 + (4-4)^2 = 0 + 0 = 0

(6,6): (6-4)^2 + (6-4)^2 = 4 + 4 = 8

**Total SSE for C1:**

**= 8 + 0 + 8**

**= 16**

**Cluster C2: (0,4), (4,0), (0,4), (0,4), (0,4), (0,4), (0,4), (0,4)**

**Centroid: ((0+4+0+0+0+0+0+0)/8, (4+0+4+4+4+4+4+4)/8) = (0.5, 3.5)**

**Squared distances from centroid for each point:**

**(**0,4): (0-0.5)^2 + (4-3.5)^2 = 0.25 + 0.25 = 0.5

(4,0): (4-0.5)^2 + (0-3.5)^2 = 13.0625 + 12.25 = 25.3125

(0,4): (0-0.5)^2 + (4-3.5)^2 = 0.25 + 0.25 = 0.5

(0,4): (0-0.5)^2 + (4-3.5)^2 = 0.25 + 0.25 = 0.5

(0,4): (0-0.5)^2 + (4-3.5)^2 = 0.25 + 0.25 = 0.5

(0,4): (0-0.5)^2 + (4-3.5)^2 = 0.25 + 0.25 = 0.5

(0,4): (0-0.5)^2 + (4-3.5)^2 = 0.25 + 0.25 = 0.5

(0,4): (0-0.5)^2 + (4-3.5)^2 = 0.25 + 0.25 = 0.5

**Total SSE for C2 :**

**= 0.5 + 25.3125 + 0.5 + 0.5 + 0.5 + 0.5 + 0.5 + 0.5**

**= 29.7125**

**Cluster C3: (5,5), (9,9)**

**Centroid:= (7, 7)**

**Squared distances from centroid for each point:**

(5,5): (5-7)^2 + (5-7)^2 = 4 + 4 = 8

(9,9): (9-7)^2 + (9-7)^2 = 4 + 4 = 8

**Total SSE for C3 = 8 + 8 = 16**

**Adding up the SSE values for all clusters:**

**SSE = SSE(C1) + SSE(C2) + SSE(C3)**

= 16 + 29.7125 + 16 = 61.7125

So, the SSE for this clustering is 61.7125.

**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.**

Cluster 1: ● ● ●

| | |

D1 D2 D3

Cluster 2: ● ●

| |

D4 D5

Cluster 3: ● ● ●

| | |

D6 D7 D8

Cluster 4: ● ●

| |

D9 D10

Cluster 5: ● ● ●

| | |

D11 D12 D13

D14 D15 D16 D17 D18 D19 D20

**Legend:**

**● - Data point (defect) Di - Defect i**

In this diagram, each cluster is represented by a group of data points (defects) enclosed within a cluster boundary. The number labels (D1, D2, etc.) indicate individual defect data points.

The clusters are formed based on the similarity of the defect details using the k-means algorithm.

Any new defect that arises after the initial clustering would need to be assigned to one of the existing clusters based on its similarity to the defects within those clusters.