Data Mining – 19th April 2024

Consider the following set of one-dimensional points: {5, 7, 16, 18, 24, 26, 34, 38}.

(a) Suppose we apply k-means clustering to obtain three clusters, A, B, and C. If the three initial centroids are located at (15, 25, 31), respectively, show the clustering results after assigning each point to their closest centroid. (**3 points**)

* **Cluster A (centroid 15):** Points = {5, 7, 16, 18}
* **Cluster B (centroid 25):** Points = {24, 26}
* **Cluster C (centroid 31):** Points = {34, 38}

(b) Based on your answer in part (a), recompute the new locations of the centroids for A, B, and C. Compute also their overall SSE. (**3 points**)

Based on the assignments from part (a), we need to compute the new centroids for each cluster by taking the mean (average) of the points in each cluster. Then, we'll calculate the Sum of Squared Errors (SSE) for each cluster, and the overall SSE for all clusters combined.

**Clusters from Part (A)**

* **Cluster A:** {5, 7, 16, 18}
* **Cluster B:** {24, 26}
* **Cluster C:** {34, 38}

**Calculating New Centroids**

1. **Cluster A:** The new centroid is the average of 5, 7, 16, and 18.
2. **Cluster B:** The new centroid is the average of 24 and 26.
3. **Cluster C:** The new centroid is the average of 34 and 38.

**Calculating SSE**

The Sum of Squared Errors (SSE) for each cluster is calculated by summing up the squared differences between each point and the centroid of the cluster they belong to. Let's calculate these:

1. For **Cluster A**, calculate SSE using its new centroid.
2. For **Cluster B**, calculate SSE using its new centroid.
3. For **Cluster C**, calculate SSE using its new centroid.

Let's perform these calculations.

The recalculated centroids and their corresponding SSE values for the clusters are as follows:

**New Centroids**

* **Cluster A:** New centroid is 11.511.5
* **Cluster B:** New centroid is 25.025.0
* **Cluster C:** New centroid is 36.036.0

**SSE for Each Cluster**

* **SSE for Cluster A:** 125.0125.0
* **SSE for Cluster B:** 2.02.0
* **SSE for Cluster C:** 8.08.0

**Overall SSE**

The overall SSE, which is the sum of the SSE for all clusters, is 135.0135.0.

These results provide the new center points for the clusters after the initial assignment, and the SSE values show the variability of points within each cluster relative to their new centroids