Answer the following questions

1. How can we choose a number of cluster in clustering analysis? What is/are the common method(s) to determine an optimal number of clusters?
   Ans: Selecting the optimal number of clusters (k) in clustering analysis is crucial. In the lecture, the k-means algorithm iteratively groups data points into a predefined number of clusters based on their similarity. To determine the appropriate number of clusters, common methods include the elbow method, average silhouette method, and gap statistic method. These techniques assist in selecting the number of clusters that best captures the underlying structure of the data.

2. Explain how outliers can affect the performance of k-means clustering. What could be a way or ways to mitigate its/their effect?

   Ans: Centroids in k-means clustering are gravitational grabbed by outliers, which distorts cluster sizes and shapes and makes centroids misrepresent actual cluster centers. Determining the ideal number of clusters may be compromised by this distortion, which can create fictitious gaps or combine clusters. It is possible to mitigate the effect of outliers on k-clustering by implementing preventive measures such as outlier detection and removal before clustering. Furthermore, using robust clustering algorithms like DBSCAN or OPTICS—which are designed to deal with outliers—or switching to alternative distance metrics that are less prone to outliers are workable alternatives.

3. What is a stopping criteria? What could act as a stopping criteria in k-means clustering?

   Ans: In k-means clustering, the stopping criterion marks the moment to end the iterative process of assigning data points to clusters and updating centroids. We decide this based on factors like convergence, where centroids stabilize or change minimally. Alternatively, we might set a maximum iteration limit to prevent indefinite running. Also, halting when clustering quality doesn't improve much ensures efficient use of

resources. These decisions ensure the algorithm wraps up effectively, achieving good clustering outcomes without unnecessary computational burden.

4. Given an example of a situation where k-means clustering may not be suitable for clustering analysis.

   Ans: When it comes to k-means clustering, handling data with vastly different scales can pose a challenge. See, k-means works by measuring distances between data points. Now, if these scales vary a lot, distances in the larger-scale dimensions tend to have more influence, throwing off the clustering process. So, when your data has significant scale differences, it can mess with how well k-means performs. In simpler terms, k-means might not be the best choice for these kinds of situations.

5. What are the applications of clustering analysis? -- *those which are not in the lecture's slides.*

   Ans: Sensor data analytic from Iot and Language processing

65070503428

65070503442

65070503468