Task 8 - K-Means Clustering Interview Q&A (Extended)

1. How does K-Means clustering work?

K-Means groups data into K clusters based on similarity. It randomly initializes K centroids, assigns each point to the nearest centroid, then updates centroids based on the mean of the points in each cluster. This repeats until the centroids stabilize.

2. What is the Elbow method?

The Elbow Method is used to determine the optimal number of clusters (K) by plotting inertia (within-cluster sum of squares) against different K values. The point where the rate of decrease sharply slows down (forming an 'elbow') is considered the best K.

3. What are the limitations of K-Means?

K-Means requires specifying K in advance, is sensitive to the initial placement of centroids, doesn't perform well with non-spherical or overlapping clusters, and can be affected by outliers and noisy data.

4. How does initialization affect results?

Different initial centroids can lead to different final clusters. K-Means can get stuck in a local minimum. Using k-means++ or multiple initializations helps improve consistency.

5. What is inertia in K-Means?

Inertia is the sum of squared distances between each point and its assigned cluster centroid. Lower inertia indicates tighter clusters and is used as a performance metric in the Elbow Method.

6. What is Silhouette Score?

Silhouette Score measures how similar an object is to its own cluster versus other clusters. It ranges from -1 (poor) to 1 (good). Higher scores indicate well-separated, cohesive clusters.

7. How do you choose the right number of clusters?

Use the Elbow Method and Silhouette Score to choose K. Also consider domain knowledge and visualization of clustering results.

8. What's the difference between clustering and classification?

Clustering is unsupervised and groups data based on similarity without labels. Classification is supervised and assigns data to predefined classes based on labeled training data.