HOMEWORK ASSIGNMENT: CLUSTERING

Course: Big Data Analytics

Topic: Clustering Algorithms (K-Means, Hierarchical, BFR, CURE)

PART 1: THEORY & CONCEPTS (40%)

Question 1: The Curse of Dimensionality

Based on the slides regarding the difficulty of high-dimensional clustering:

a. Explain why our geometric intuition in 2D/3D space becomes misleading when applied to a 10,000-dimensional space.

b. What happens to the distance between pairs of points as dimensionality increases significantly? Why does this make distance-based algorithms (like K-Means) less effective?

Question 2: K-Means Clustering

a. What is the Big-O time complexity of the K-Means algorithm? Explain the meaning of the variables in the formula (k, t, d, N).

b. "Picking k" is a challenge in K-Means. Describe one method to determine an appropriate value for k (e.g., the Elbow method).

Question 3: BFR Algorithm (Big Data)

The BFR algorithm is an extension of K-Means designed to handle very large datasets that cannot fit into memory (RAM). Explain the role of the following three sets of points in BFR:

1. **Discard Set (DS)**
2. **Compressed Set (CS)**
3. **Retained Set (RS)**

PART 2: CALCULATION EXERCISES (40%)

Question 4: Hierarchical Clustering

Given a dataset of 5 points in a 2D space:

$$P1(1,2), P2(2,2), P3(5,4), P4(5,5), P5(1,5)$$

Assume we are using Agglomerative Hierarchical Clustering.

a. Calculate the Euclidean distance matrix between all pairs of points.

b. Perform the first merge step and the second merge step. Which two points (or clusters) are merged at each step if using the Single Link (Min Distance) criterion?

c. Repeat question (b) using the Complete Link (Max Distance) criterion. Does the result change?

Question 5: Mahalanobis Distance (In BFR)

The BFR algorithm uses Mahalanobis distance instead of standard Euclidean distance.

a. Why is Mahalanobis distance more suitable than Euclidean distance when clusters have an ellipsoidal shape rather than a spherical one?

b. In the BFR algorithm, if a point has a Mahalanobis distance to a cluster centroid that is smaller than a certain threshold, how is that point processed?

PART 3: CRITICAL THINKING & COMPARISON (20%)

Question 6: CURE vs. K-Means

a. What is the major disadvantage of K-Means (and BFR) regarding cluster shapes?

b. How does the CURE algorithm overcome this disadvantage? (Hint: Mention how CURE uses "Representative points" instead of a single Centroid).

Question 7: Performance & Complexity

Why is Hierarchical Clustering (with complexity $O(N^3)$ or $O(N^2 logN)$) generally not used for Big Data, while K-Means (or BFR) is preferred?