CS246 Exam 2025 — Question 3: Clustering (12 points)

Part 1 (4 points) — Hierarchical Clustering

Task: Apply agglomerative hierarchical clustering with **complete linkage** (distance between clusters = maximum Euclidean distance between any pair of points, one from each cluster) until 2 clusters remain.

Distance Matrix

	A	B	C	D	E	F
A	0.00	3.64	3.91	3.64	1.58	3.16
B	3.64	0.00	4.12	2.00	3.04	4.72
C	3.91	4.12	0.00	2.24	4.92	1.50
D	3.64	2.00	2.24	0.00	3.91	3.20
E	1.58	3.04	4.92	3.91	0.00	4.53
F	3.16	4.72	1.50	3.20	4.53	0.00

Merging Process

Iteration 1 → 2: Smallest distance is C–F = 1.50. Merge {C} and {F}.

Iteration 2 → 3: Recompute {C,F} distances using max: {C,F}–A = max(3.91, 3.16) = 3.91, etc. Smallest overall is A–E = 1.58. Merge {A} and {E}.

Iteration 3 → 4: Recompute {A,E} distances using max: {A,E}–B = max(3.64, 3.04) = 3.64, etc. Smallest overall is B–D = 2.00. Merge {B} and {D}.

Iteration 4 → 5: Three clusters remain. Distances:

- {A,E}–{B,D} = max(3.64, 3.64, 3.04, 3.91) = **3.91** ← smallest
- {B,D}–{C,F} = max(4.12, 4.72, 2.24, 3.20) = 4.72
- {A,E}–{C,F} = max(3.91, 3.16, 4.92, 4.53) = 4.92

Merge {A,E} and {B,D}.

Answer

Iteration	Clusters
Iteration 1	{A}, {B}, {C}, {D}, {E}, {F}
Iteration 2	{A}, {B}, **{C, F}**, {D}, {E}
Iteration 3	**{A, E}**, {B}, {C, F}, {D}
Iteration 4	{A, E}, **{B, D}**, {C, F}
Iteration 5	**{A, B, D, E}**, **{C, F}**

Part 2 (4 points) — K-Means Clustering

Task: Run K-Means on points {1, 5, 8, 13, 19, 28} with initial centroids $\mu_1 = 8$, $\mu_2 = 28$.

Iteration 1 — Assignment

Point	$\lvert x - 8 \rvert$	$\lvert x - 28 \rvert$	Cluster
1	7	27	C_1
5	3	23	C_1
8	0	20	C_1
13	5	15	C_1
19	11	9	C_2
28	20	0	C_2

Update centroids: $\mu_1 = (1+5+8+13)/4 = \mathbf{6.75}$, $\mu_2 = (19+28)/2 = \mathbf{23.5}$

Iteration 2 — Assignment

Point	$\lvert x - 6.75 \rvert$	$\lvert x - 23.5 \rvert$	Cluster
1	5.75	22.5	C_1
5	1.75	18.5	C_1
8	1.25	15.5	C_1

Point	|x − 6.75|	|x − 23.5|	Cluster
13	6.25	10.5	C_1
19	12.25	4.5	C_2
28	21.25	4.5	C_2

Assignments unchanged → **converged!**

Answer

- **Clusters:** {1, 5, 8, 13} and {19, 28}
- **Centroids:** 6.75 and 23.5

Part 3 (2 points) — Reducing K-Means' Sensitivity to Initialization

Answer: (b) — Choosing initial centroids that are spread out far across the space.

This is the idea behind **K-Means++**: the first center is chosen randomly, then each subsequent center is sampled with probability proportional to $D(p)^2$ (squared distance to the nearest existing center). This ensures centroids are spread out, avoiding the worst-case scenarios where multiple centroids start in the same cluster.

Why the others fail: (a) random points in empty space don't help and may clump together; (c) increasing k changes the problem entirely and doesn't fix initialization sensitivity.

Part 4 (2 points) — BFR: Discard Set vs. Compressed Set

Answer: (b) — DS is for points sufficiently close to an existing cluster centroid; CS is for subclusters of points that are not near any existing centroid but are close to each other.

The three BFR data structures:

Set	What goes here	Storage
Discard Set (DS)	Points confidently assigned to a known cluster	Summarized as (N, SUM, SUMSQ)
Compressed Set (CS)	Tight groups of points that are near *each other* but far from all centroids	Same (N, SUM, SUMSQ) summary
Retained Set (RS)	Isolated points that don't fit anywhere yet	Stored individually

Why the others fail: (a) reverses the storage — DS summarizes, it doesn't store individually; (c) DS is continuously updated with new batches; (d) both DS and CS use the same compact (N, SUM, SUMSQ) representation.