University Mohamed Boudiaf - M'sila
Faculty of MI, CS Department
Master1 AI

May 09, 2026
Duration : 2h

Final Exam

Data Mining: Techniques and Applications

Exercise 1 (14 pts)

Consider the dataset D = {-3.5; -2.75; 2; 3; 5}. We want to perform the clustering of D using two approaches: Hierarchical Ascending Clustering (HAC) and K-means.

Part A (4 pts):

1- Apply the K-means algorithm starting from the following initializations: $\mu1 = -3.5$; $\mu2 = 2$; $\mu3 = 5$
2- Find three clusters using K-means.

We will use the Manhattan Distance.

Part B (6 pts):

1- Apply HAC to the data D.
2- Draw the dendrogram.
3- Find three clusters from the dendrogram. At what cutting distances can you find these three clusters?

We will use the Manhattan Distance and the following metric between two clusters:

$$d\left(C_i, C_j\right) = \min_{x \in C_i, z \in C_j} |x - z|$$

Part C (4 pts): We aim to compare the quality of the clusters found by K-means and hierarchical clustering. To do this, we use the intra-class inertia defined by the following formula:

$$I_{intra} = \frac{1}{n} \sum_{i=1}^{k} \sum_{e \in C_i} d^2(e, g_i)$$

Where e is an element of the cluster C_i and g_i is the center of gravity of the cluster C_i

1- Calculate the intra-class inertia for the three clusters found using K-means (Part A).
2- Calculate the intra-class inertia for the three clusters found using hierarchical clustering (Part B, Question 3).
3- Compare the clusters found using K-means and hierarchical clustering. What are the common clusters found in both models?

Exercise 2 (6 pts)

Consider the following Boolean database with 5 items and 10 transactions.

Run the algorithm a priori with Minimal Support = 0.3, and find the rules with the following format:

Item1 → {Item2, Item3, Item4}.

	X1	X2	X3	X4	X5
t1	0	1	0	0	1
t2	0	0	1	0	1
t3	0	0	1	0	0
t4	1	1	1	1	1
t5	1	1	1	1	1
t6	1	1	1	1	0
t7	1	0	1	1	0
t8	1	0	1	1	0
t9	1	0	0	0	1
t10	1	0	0	0	1

Dr. Tahar Mehenni

Part A — K-means

Initializations : **μ_1 = -3.5, μ_2 = 2, μ_3 = 5**

Step 1 — Initial assignment:

Point	d(x, μ_1=-3.5)	d(x, μ_2=2)	d(x, μ_3=5)	Cluster
-3.5	0	5.5	8.5	**C_1**
-2.75	0.75	4.75	7.75	**C_1**
2	5.5	0	3	**C_2**
3	6.5	1	2	**C_2**
5	8.5	3	0	**C_3**

1.5 Pt

→ C_1 = {-3.5, -2.75}, C_2 = {2, 3}, C_3 = {5}

Step 2 — Recompute centroids:

- μ_1 = (-3.5 + -2.75)/2 = **-3.125**
- μ_2 = (2 + 3)/2 = **2.5**
- μ_3 = 5/1 = **5**

0.5 Pt

Step 3 — Reassign:

Point	d(x, -3.125)	d(x, 2.5)	d(x, 5)	Cluster
-3.5	0.375	6	8.5	**C_1**
-2.75	0.375	5.25	7.75	**C_1**
2	5.125	0.5	3	**C_2**
3	6.125	0.5	2	**C_2**
5	8.125	2.5	0	**C_3**

1.5 Pt

→ C_1 = {-3.5, -2.75}, C_2 = {2, 3}, C_3 = {5} — **no change → convergence!**

✅ **Final K-means clusters:**

C_1 = {-3.5, -2.75}, μ_1 = -3.125 **C_2 = {2, 3}**, μ_2 = 2.5 **C_3 = {5}**, μ_3 = 5

0.5 Pt

Part B — HAC (Single Linkage, Manhattan distance)

Initial distance matrix ($|x_i - x_j|$):

	-3.5	**-2.75**	**2**	**3**	**5**
-3.5	—	0.75	5.5	6.5	8.5
-2.75	0.75	—	4.75	5.75	7.75
2	5.5	4.75	—	1	3
3	6.5	5.75	1	—	2
5	8.5	7.75	3	2	—

1 Pt

Iteration 1 — Min: d(-3.5, -2.75) = **0.75** → Merge → **A = {-3.5, -2.75}**

0.5 Pt

Update (single linkage):

- d(2, A) = min(5.5, 4.75) = 4.75
- d(3, A) = min(6.5, 5.75) = 5.75
- d(5, A) = min(8.5, 7.75) = 7.75

Iteration 2 — Matrix:

	A={-3.5,-2.75}	**2**	**3**	**5**
A	—	4.75	5.75	7.75
2	4.75	—	1	3
3	5.75	1	—	2
5	7.75	3	2	—

1 Pt

Min: d(2, 3) = **1** → Merge → **B = {2, 3}**

0.5 Pt

Update:

- d(A, B) = min(4.75, 5.75) = 4.75
- d(5, B) = min(3, 2) = 2

Iteration 3 — Matrix:

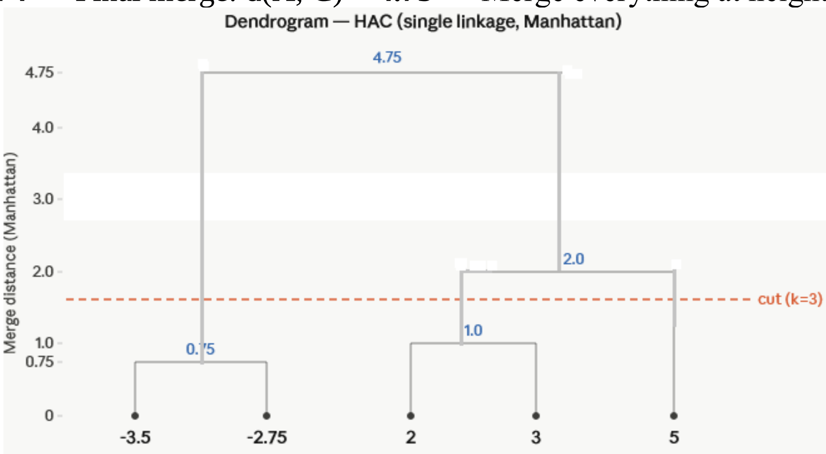
	A={-3.5,-2.75}	B={2,3}	5
A	—	4.75	7.75
B	4.75	—	2
5	7.75	2	—

1 Pt

Min: d(B, 5) = **2** → Merge {2, 3} and {5} → **C = {2, 3, 5}**
Update:

- d(A, C) = min(4.75, 7.75) = 4.75

Iteration 4 — Final merge: d(A, C) = **4.75** → Merge everything at height 4.75.



1.5 Pt

Now the dendrogram:**Reading for k=3:** Cutting between d=1.0 and d=2 (dashed orange line at d=3.0) gives: **H_1 = {-3.5, -2.75}**, **H_2 = {2, 3}**, **H_3 = {5}** 0.5 Pt

Part C — Intra-class Inertia

C.1 — K-means

C_1 = {-3.5, -2.75}, $\mu_1 = -3.125$

Point	**$(x_i - \mu_1)^2$**
-3.5	$(-0.375)^2 =$ **0.140625**
-2.75	$(0.375)^2 =$ **0.140625**

0.25 Pt

$W_1 =$ **0.28125**

C_2 = {2, 3}, $\mu_2 = 2.5$

Point	**$(x_i - \mu_2)^2$**
2	$(-0.5)^2 =$ **0.25**
3	$(0.5)^2 =$ **0.25**

0.25 Pt

$W_2 =$ **0.5**

C_3 = {5}, $\mu_3 = 5$ → $W_3 =$ **0**

W_Kmeans = 1/5(0.28125 + 0.5 + 0) = 0.15625 0.5 Pt

C.2 — HAC

HAC clusters: H_1 = {-3.5, -2.75}, H_2 = {2, 3}, H_3 = {5}

These are **identical** to K-means, so centroids and inertia are the same: 1 Pt

W_HAC = 1/5(0.28125 + 0.5 + 0) = 0.15625

C.3 — Comparison & Summary

Method	**C_1**	**C_2**	**C_3**	**Inertia W**
K-means	{-3.5, -2.75}	{2, 3}	{5}	**0.15625**
HAC	{-3.5, -2.75}	{2, 3}	{5}	**0.15625**

1 Pt

Both methods produce **exactly the same three clusters** with identical inertia.
The **common clusters across both methods** are: {-3.5, -2.75}, {2, 3}, and {5}. 1 Pt

1-itemset (0.5 pt)	Freq	Support
X1	7	0.7
X2	4	0.4
X3	7	0.7
X4	5	0.5
X5	6	0.6

2-itemset (0.5 pt)

	X1	X2	X3	X4	X5
X1		3	5	5	4
X2			3	3	3
X3				5	3
X4					2

3-itemset (2 pts)	X1X2	X1X3	X1X4	X1X5	X2X3	X2X4	X2X5	X3X4	X3X5
X1X2		3	3	2				x	x
X1X3			5	2		x	x		
X1X4				2	x		x		x
X1X5					x	x		x	x
X2X3						3	2		
X2X4							2		x
X2X5								x	
X3X4									2

4-itemset (1 pt)	X1X2X3	X1X2X4	X1X3X4	X2X3X4
X1X2X3		3		
X1X2X4				
X1X3X4				
X2X3X4				

Recap

2-itemset	X1X2	X1X3	X1X4	X1X5	X2X3	X2X4	X2X5	X3X4	X3X5
3-itemset	X1X2X3	X1X2X4	X1X3X4	X2X3X4					
4-itemset	X1X2X3X4								

RULES	**(2 pts)**
X1X2X3X4	X1 → X2X3X4
	X2 → X1X3X4
	X3 → X1X2X4
	X4 → X1X2X3