THE CHINESE UNIVERSITY OF HONG KONG
Department of Mathematics
Exercises on Agglomerative Hierarchical Clustering

1 Introduction to Agglomerative Hierarchical Clustering

Hierarchical clustering relies on pairwise distance matrices as the core clustering criterion. Characteristics:

- Uses a full pairwise distance matrix to guide cluster merging decisions.
- Does not require pre-specified number of clusters k as input; clustering terminates either when all points form one single cluster or a custom distance threshold is reached.
- Requires formal definition of inter-cluster distance $d(C_i, C_j)$ to quantify separation between two disjoint clusters C_i and C_j.

2 Single Linkage Hierarchical Clustering

Definition and Core Concept

Single linkage clustering (also known as nearest-neighbor agglomerative clustering) is a bottom-up (agglomerative) hierarchical clustering algorithm. The inter-cluster distance is defined as the *minimum Euclidean distance* between any point belonging to the first cluster and any point belonging to the second cluster. This minimum-distance merging rule often produces elongated, chain-shaped clusters because merging only depends on the single closest pair of points across two groups.

Formal definition of inter-cluster single linkage distance for clusters C_i, C_j:

$$d(C_i, C_j) = \min \{d(x, y) \mid x \in C_i,\ y \in C_j\}$$

where the Euclidean distance between two 2-dimensional points $x = (x_1, x_2)$ and $y = (y_1, y_2)$ is:

$$d(x, y) = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2}$$

Formal Algorithm Steps

The iterative single linkage clustering procedure follows five fixed stages:

1. **Initialization**: Every raw data point forms its own singleton cluster.
2. **Compute Full Inter-Cluster Distances**: Calculate all pairwise distances between current clusters using the single linkage minimum-distance rule.
3. **Merge Closest Pair**: Locate the pair of clusters with the smallest inter-cluster distance and combine them into one new merged cluster.
4. **Refresh Distance Matrix**: Recalculate distances between the newly merged cluster and all unchanged remaining clusters, applying the single linkage rule.

5. **Iterate Termination**: Repeat Steps 2–4 repeatedly until all data points belong to one global cluster, or stop early if a target number of clusters / distance cutoff threshold is satisfied.

Numerical Example Calculation

We demonstrate single linkage clustering on a dataset of six 2D sample points. All Euclidean distance values are rounded to two decimal places for consistent precision.

Data Point Label	x-Coordinate	y-Coordinate
A	1	3
B	2	4
C	5	10
D	2	8
E	8	5
F	11	12

Table 1: 2D Raw Dataset for Single Linkage Clustering

Step 1: Initial State — All Points as Singleton Clusters

First compute every pairwise Euclidean distance with full algebraic expansion for complete transparency:

1. $d(A,B) = \sqrt{(1-2)^2 + (3-4)^2} = \sqrt{(-1)^2 + (-1)^2} = \sqrt{1+1} = \sqrt{2} \approx 1.41$
2. $d(A,C) = \sqrt{(1-5)^2 + (3-10)^2} = \sqrt{(-4)^2 + (-7)^2} = \sqrt{16+49} = \sqrt{65} \approx 8.06$
3. $d(A,D) = \sqrt{(1-2)^2 + (3-8)^2} = \sqrt{(-1)^2 + (-5)^2} = \sqrt{1+25} = \sqrt{26} \approx 5.10$
4. $d(A,E) = \sqrt{(1-8)^2 + (3-5)^2} = \sqrt{(-7)^2 + (-2)^2} = \sqrt{49+4} = \sqrt{53} \approx 7.28$
5. $d(A,F) = \sqrt{(1-11)^2 + (3-12)^2} = \sqrt{(-10)^2 + (-9)^2} = \sqrt{100+81} = \sqrt{181} \approx 13.45$
6. $d(B,C) = \sqrt{(2-5)^2 + (4-10)^2} = \sqrt{(-3)^2 + (-6)^2} = \sqrt{9+36} = \sqrt{45} \approx 6.71$
7. $d(B,D) = \sqrt{(2-2)^2 + (4-8)^2} = \sqrt{0^2 + (-4)^2} = \sqrt{0+16} = 4.00$
8. $d(B,E) = \sqrt{(2-8)^2 + (4-5)^2} = \sqrt{(-6)^2 + (-1)^2} = \sqrt{36+1} = \sqrt{37} \approx 6.08$
9. $d(B,F) = \sqrt{(2-11)^2 + (4-12)^2} = \sqrt{(-9)^2 + (-8)^2} = \sqrt{81+64} = \sqrt{145} \approx 12.04$
10. $d(C,D) = \sqrt{(5-2)^2 + (10-8)^2} = \sqrt{(3)^2 + (2)^2} = \sqrt{9+4} = \sqrt{13} \approx 3.61$
11. $d(C,E) = \sqrt{(5-8)^2 + (10-5)^2} = \sqrt{(-3)^2 + (5)^2} = \sqrt{9+25} = \sqrt{34} \approx 5.83$
12. $d(C,F) = \sqrt{(5-11)^2 + (10-12)^2} = \sqrt{(-6)^2 + (-2)^2} = \sqrt{36+4} = \sqrt{40} \approx 6.32$
13. $d(D,E) = \sqrt{(2-8)^2 + (8-5)^2} = \sqrt{(-6)^2 + (3)^2} = \sqrt{36+9} = \sqrt{45} \approx 6.71$
14. $d(D,F) = \sqrt{(2-11)^2 + (8-12)^2} = \sqrt{(-9)^2 + (-4)^2} = \sqrt{81+16} = \sqrt{97} \approx 9.85$
15. $d(E,F) = \sqrt{(8-11)^2 + (5-12)^2} = \sqrt{(-3)^2 + (-7)^2} = \sqrt{9+49} = \sqrt{58} \approx 7.62$

Symmetric initial pairwise distance matrix (minimum value highlighted in red):

Cluster	$A(1,3)$	$B(2,4)$	$C(5,10)$	$D(2,8)$	$E(8,5)$	$F(11,12)$
$A(1,3)$	0.00	1.41	8.06	5.10	7.28	13.45
$B(2,4)$	1.41	0.00	6.71	4.00	6.08	12.04
$C(5,10)$	8.06	6.71	0.00	3.61	5.83	6.32
$D(2,8)$	5.10	4.00	3.61	0.00	6.71	9.85
$E(8,5)$	7.28	6.08	5.83	6.71	0.00	7.62
$F(11,12)$	13.45	12.04	6.32	9.85	7.62	0.00

Table 2: Initial Euclidean Distance Matrix (6 Singleton Clusters)

Step 2: First Merge Operation

The global minimum distance from Table 2 is $d(A,B) \approx 1.41$. We merge singletons A and B to form a new composite cluster AB.

Step 3: Rebuild Distance Matrix (Current Clusters: $AB,\ C,\ D,\ E,\ F$)

Single linkage rule for distance between merged cluster AB and external cluster X:

$$d(AB,\ X) = \min\left\{d(A,X),\ d(B,X)\right\}$$

Compute all required inter-cluster distances step-by-step:

$$\begin{aligned}
d(AB,C) &= \min\left(d(A,C)=8.06,\ d(B,C)=6.71\right) = 6.71\\
d(AB,D) &= \min\left(d(A,D)=5.10,\ d(B,D)=4.00\right) = 4.00\\
d(AB,E) &= \min\left(d(A,E)=7.28,\ d(B,E)=6.08\right) = 6.08\\
d(AB,F) &= \min\left(d(A,F)=13.45,\ d(B,F)=12.04\right) = 12.04
\end{aligned}$$

Distances between unchanged clusters C, D, E, F remain identical to the initial matrix. Updated distance matrix:

Cluster	AB	$C(5,10)$	$D(2,8)$	$E(8,5)$	$F(11,12)$
AB	0.00	6.71	4.00	6.08	12.04
$C(5,10)$	6.71	0.00	3.61	5.83	6.32
$D(2,8)$	4.00	3.61	0.00	6.71	9.85
$E(8,5)$	6.08	5.83	6.71	0.00	7.62
$F(11,12)$	12.04	6.32	9.85	7.62	0.00

Table 3: Distance Matrix After Merging $A + B \to AB$

Step 4: Second Merge Operation

The new minimum distance in Table 3 is $d(C,D) \approx 3.61$. Merge singletons C, D into composite cluster CD.

Step 5: Rebuild Distance Matrix (Current Clusters: $AB,\ CD,\ E,\ F$)

Single linkage rule for merged cluster CD:

$$d(CD,\ X) = \min\left\{d(C,X),\ d(D,X)\right\}$$

Calculations:

$$d(CD, AB) = \min\big(d(C, AB) = 6.71,\ d(D, AB) = 4.00\big) = 4.00$$
$$d(CD, E) = \min\big(d(C, E) = 5.83,\ d(D, E) = 6.71\big) = 5.83$$
$$d(CD, F) = \min\big(d(C, F) = 6.32,\ d(D, F) = 9.85\big) = 6.32$$

Updated matrix (new minimum distance highlighted red):

Cluster	AB	CD	$E(8,5)$	$F(11,12)$
AB	0.00	4.00	6.08	12.04
CD	4.00	0.00	5.83	6.32
$E(8,5)$	6.08	5.83	0.00	7.62
$F(11,12)$	12.04	6.32	7.62	0.00

Table 4: Distance Matrix After Merging $C + D \to CD$

Step 6: Third Merge Operation

Minimum distance in Table 4 is $d(AB, CD) \approx 4.00$. Merge clusters AB and CD into large composite cluster $ABCD$.

Step 7: Rebuild Distance Matrix (Current Clusters: $ABCD,\ E,\ F$)

Single linkage rule for merged cluster $ABCD$:

$$d(ABCD,\ X) = \min\big\{d(AB, X),\ d(CD, X)\big\}$$

Calculations:

$$d(ABCD, E) = \min\big(d(AB, E) = 6.08,\ d(CD, E) = 5.83\big) = 5.83$$
$$d(ABCD, F) = \min\big(d(AB, F) = 12.04,\ d(CD, F) = 6.32\big) = 6.32$$

Updated matrix:

Cluster	$ABCD$	$E(8,5)$	$F(11,12)$
$ABCD$	0.00	5.83	6.32
$E(8,5)$	5.83	0.00	7.62
$F(11,12)$	6.32	7.62	0.00

Table 5: Distance Matrix After Merging $AB + CD \to ABCD$

Step 8: Fourth Merge Operation

Minimum distance in Table 5 is $d(ABCD, E) \approx 5.83$. Merge clusters $ABCD$ and E into composite cluster $ABCDE$.

Step 9: Rebuild Distance Matrix (Current Clusters: $ABCDE,\ F$)

Single linkage rule for merged cluster $ABCDE$:

$$d(ABCDE,\ F) = \min\big\{d(ABCD, F),\ d(E, F)\big\} = \min(6.32,\ 7.62) = 6.32$$

Final two-cluster distance matrix:

Cluster	$ABCDE$	$F(11, 12)$
$ABCDE$	0.00	6.32
$F(11, 12)$	6.32	0.00

Table 6: Distance Matrix After Merging $ABCD + E \to ABCDE$

Step 10: Final Global Merge Operation

Only two disjoint clusters remain: $ABCDE$ and F, with inter-cluster distance equal to 6.32. Merge the two clusters to form one universal cluster $ABCDEF$ containing all six original data points. The clustering process terminates.

Merge Sequence Summary (Single Linkage Hierarchy)

Ordered list of all merge events with corresponding cut distance threshold:

1. Merge singleton cluster A and singleton cluster B at distance threshold 1.41
2. Merge singleton cluster C and singleton cluster D at distance threshold 3.61
3. Merge composite cluster AB and composite cluster CD at distance threshold 4.00
4. Merge composite cluster $ABCD$ and singleton cluster E at distance threshold 5.83
5. Merge composite cluster $ABCDE$ and singleton cluster F at distance threshold 6.32