CSCI 347: Introduction to Data Mining

Hierarchical Clustering Example

AGGLOMERATIVE CLUSTERING ALGORITHM

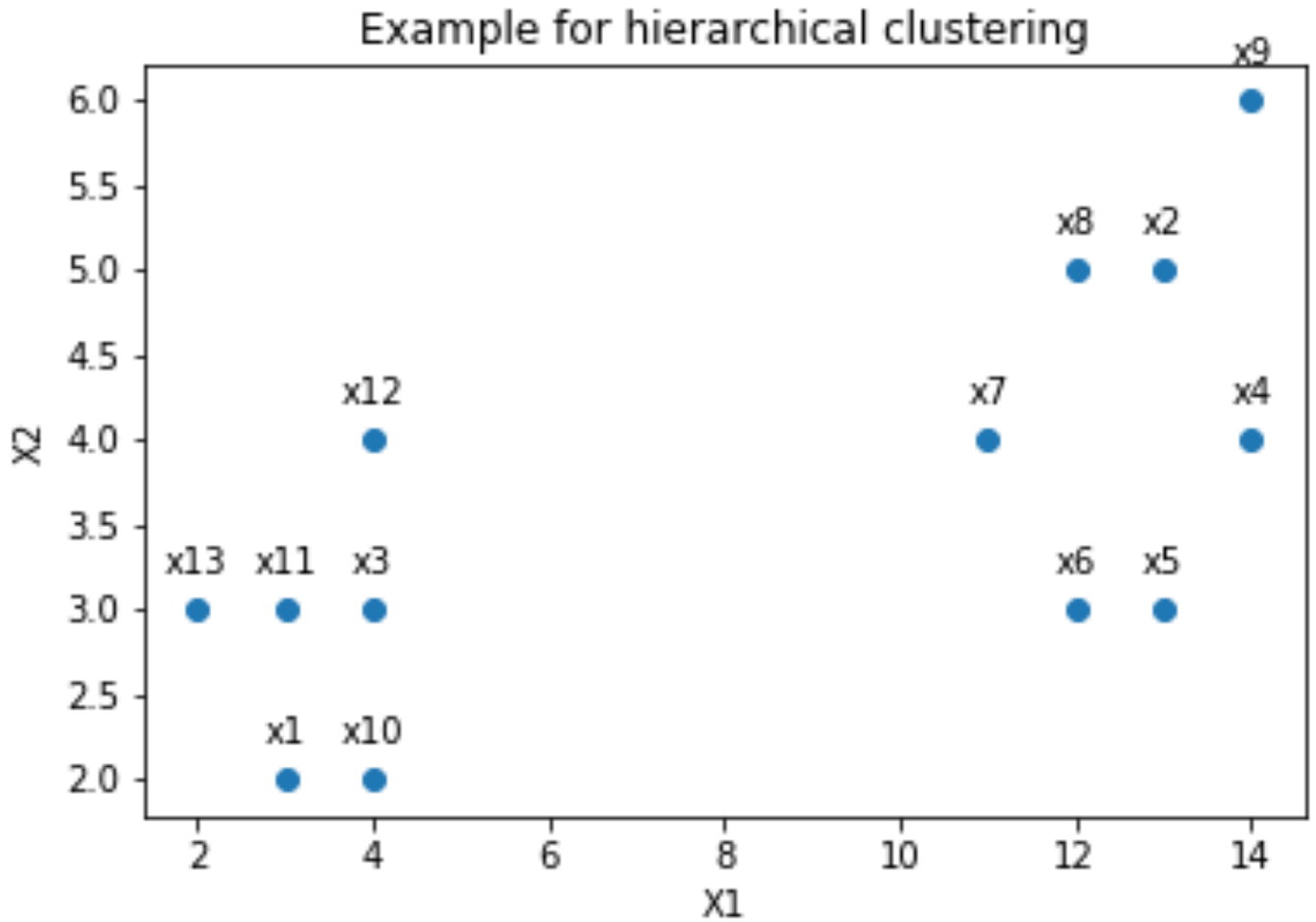
AgglomerativeClustering(D, k) :

1. $\mathscr{C} \leftarrow \{C_i = \{x_i\} \,|\, x_i \in D\}$
2. $\Delta = \{\delta(x_i, x_j) : x_i, x_j \in D\}$
3. repeat:
    4. Find the closest pair of clusters $C_i, C_j \in \mathscr{C}$
    5. $C_{i,j} = C_i \cup C_j$
    6. $\mathscr{C} \leftarrow (\mathscr{C} \backslash \{C_i, C_j\}) \cup \{C_{i,j}\}$
    7. Update distance matrix Δ to reflect new clustering
4. Until $|\mathscr{C}| = k$

EXAMPLE

$D =$

	X_1	X_2
x_1	3	2
x_2	13	5
x_3	4	3
x_4	14	4
x_5	13	3
x_6	12	3
x_7	11	4
x_8	12	5
x_9	14	6
x_{10}	4	2
x_{11}	3	3
x_{12}	4	4
x_{13}	2	3



AGGLOMERATIVE CLUSTERING ALGORITHM

AgglomerativeClustering(D, k) :

1. $\mathscr{C} \leftarrow \{C_i = \{x_i\} \,|\, x_i \in D\}$
2. $\Delta = \{\delta(x_i, x_j) : x_i, x_j \in D\}$
3. repeat:
    4. Find the closest pair of clusters $C_i, C_j \in \mathscr{C}$
    5. $C_{i,j} = C_i \cup C_j$
    6. $\mathscr{C} \leftarrow (\mathscr{C} \backslash \{C_i, C_j\}) \cup \{C_{i,j}\}$
    7. Update distance matrix Δ to reflect new clustering
4. Until $|\mathscr{C}| = k$

EXAMPLE

$D =$

	X_1	X_2
x_1	3	2
x_2	13	5
x_3	4	3
x_4	14	4
x_5	13	3
x_6	12	3
x_7	11	4
x_8	12	5
x_9	14	6
x_{10}	4	2
x_{11}	3	3
x_{12}	4	4
x_{13}	2	3

$\Delta =$

	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}	x_{11}	x_{12}	x_{13}
x_1	0												
x_2		0											
x_3			0										
x_4				0									
x_5					0								
x_6						0							
x_7							0						
x_8								0					
x_9									0				
x_{10}										0			
x_{11}											0		
x_{12}												0	
x_{13}													0

EXAMPLE

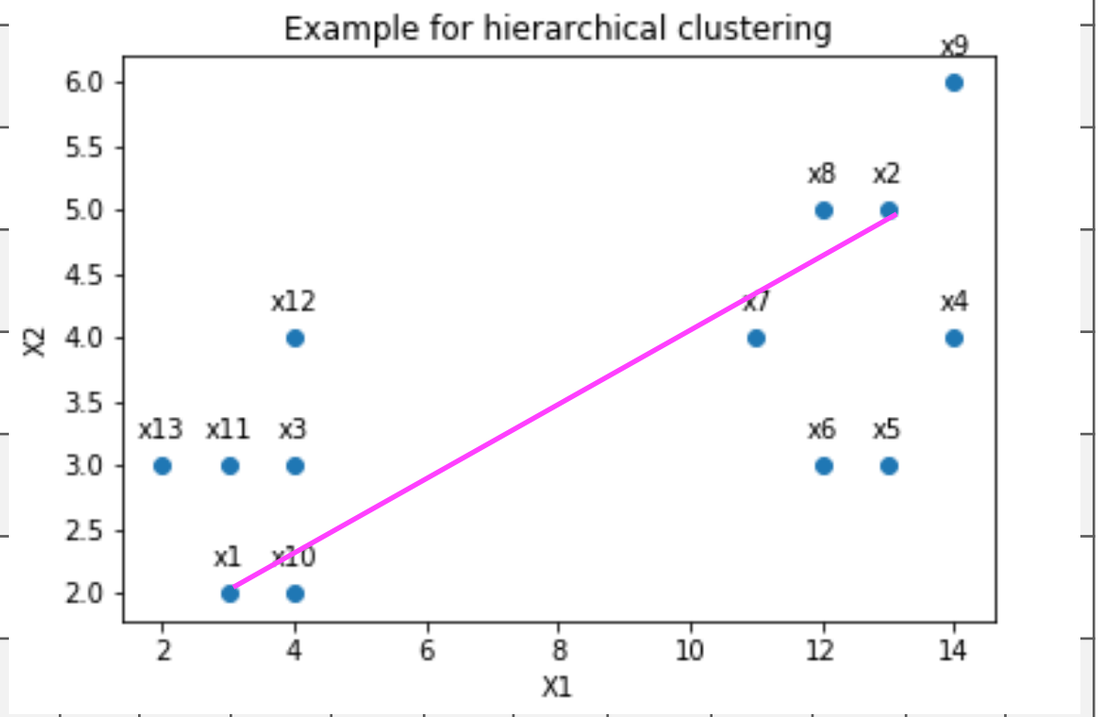
$$\delta(x_1, x_2) = \sqrt{(3-13)^2 + (2-5)^2} = \sqrt{100+9} = \sqrt{109} = 10.44$$

$D =$

	X_1	X_2
x_1	3	2
x_2	13	5
x_3	4	3
x_4	14	4
x_5	13	3
x_6	12	3
x_7	11	4
x_8	12	5
x_9	14	6
x_{10}	4	2
x_{11}	3	3
x_{12}	4	4
x_{13}	2	3

$\Delta =$

	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}	x_{11}	x_{12}	x_{13}
x_1	0												
x_2		0											
x_3			0										
x_4													
x_5													
x_6													
x_7													
x_8													
x_9													
x_{10}													
x_{11}													
x_{12}												0	
x_{13}													0



EXAMPLE

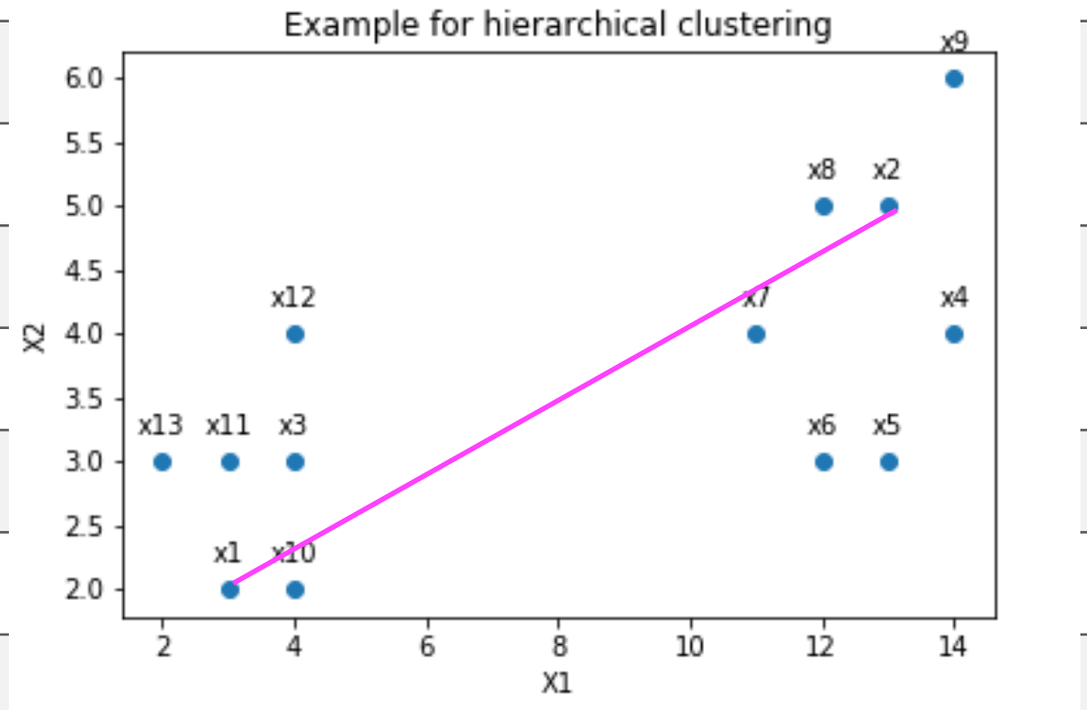
$$\delta(x_1, x_2) = \sqrt{(3-13)^2 + (2-5)^2} = \sqrt{100+9} = \sqrt{109} = 10.44$$

$D =$

	X_1	X_2
x_1	3	2
x_2	13	5
x_3	4	3
x_4	14	4
x_5	13	3
x_6	12	3
x_7	11	4
x_8	12	5
x_9	14	6
x_{10}	4	2
x_{11}	3	3
x_{12}	4	4
x_{13}	2	3

$\Delta =$

	x1	x2	x3	x4	x5	x6	x7	x8	x9	x10	x11	x12	x13
x1	0												
x2	10.44	0											
x3			0										
x4													
x5													
x6													
x7													
x8													
x9													
x10													
x11													
x12												0	
x13													0



EXAMPLE

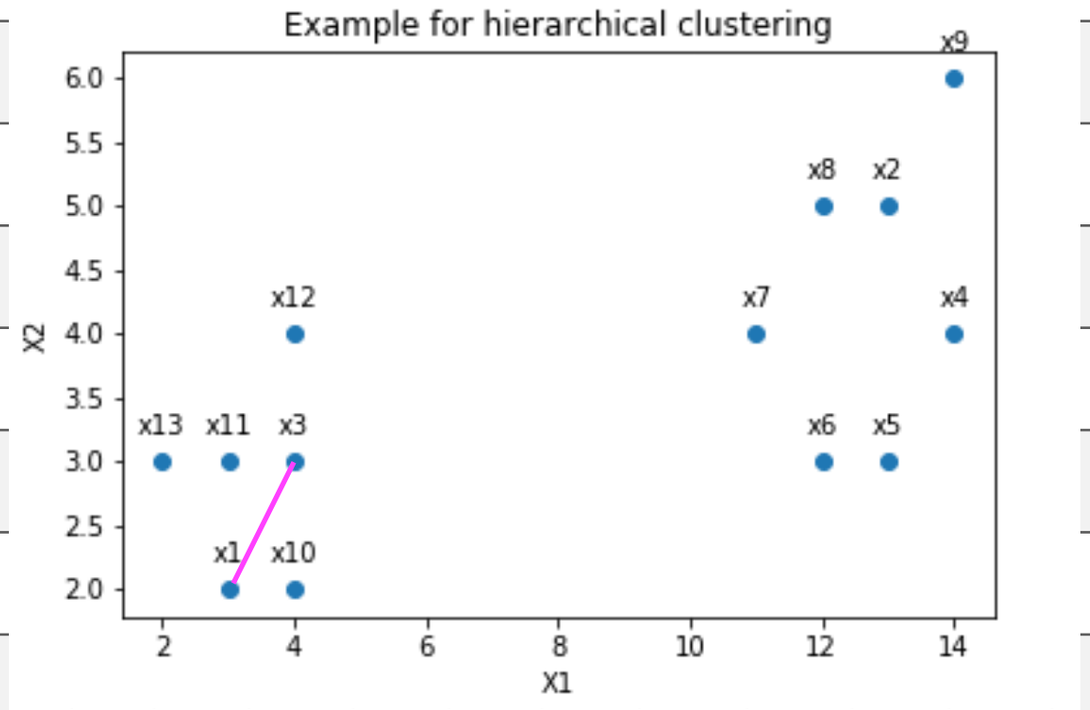
$$\delta(x_1, x_3) = \sqrt{(3-4)^2 + (2-3)^2} = \sqrt{1+1} = \sqrt{2} = 1.41$$

$D =$

	X_1	X_2
x_1	3	2
x_2	13	5
x_3	4	3
x_4	14	4
x_5	13	3
x_6	12	3
x_7	11	4
x_8	12	5
x_9	14	6
x_{10}	4	2
x_{11}	3	3
x_{12}	4	4
x_{13}	2	3

$\Delta =$

	x1	x2	x3	x4	x5	x6	x7	x8	x9	x10	x11	x12	x13
x1	0												
x2	10.44	0											
x3	1.41		0										
x4													
x5													
x6													
x7													
x8													
x9													
x10													
x11													
x12												0	
x13													0



EXAMPLE

$D =$

	X_1	X_2
x_1	3	2
x_2	13	5
x_3	4	3
x_4	14	4
x_5	13	3
x_6	12	3
x_7	11	4
x_8	12	5
x_9	14	6
x_{10}	4	2
x_{11}	3	3
x_{12}	4	4
x_{13}	2	3

$\Delta =$

	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}	x_{11}	x_{12}	x_{13}
x_1	0												
x_2	10.44	0											
x_3	1.41	9.21	0										
x_4	11.18	1.41	10.05	0									
x_5	10.05	2	9	1.41	0								
x_6	9.06	2.24	8	2.24	1	0							
x_7	8.25	2.24	7.07	3	2.24	1.41	0						
x_8	9.49	1	8.25	2.23	2.24	2	1.41	0					
x_9	11.70	1.41	10.44	2	3.16	3.61	3.61	2.24	0				
x_{10}	1	9.49	1	10.20	9.06	8.06	7.28	8.54	10.77	0			
x_{11}	1	10.20	1	11.05	10	9	8.06	9.22	11.40	1.41	0		
x_{12}	2.24	9.06	1	10	9.06	8.06	7	8.06	10.20	2	1.41	0	
x_{13}	1.41	11.18	2	12.04	11	10	9.06	10.20	12.37	2.24	1	2.24	0

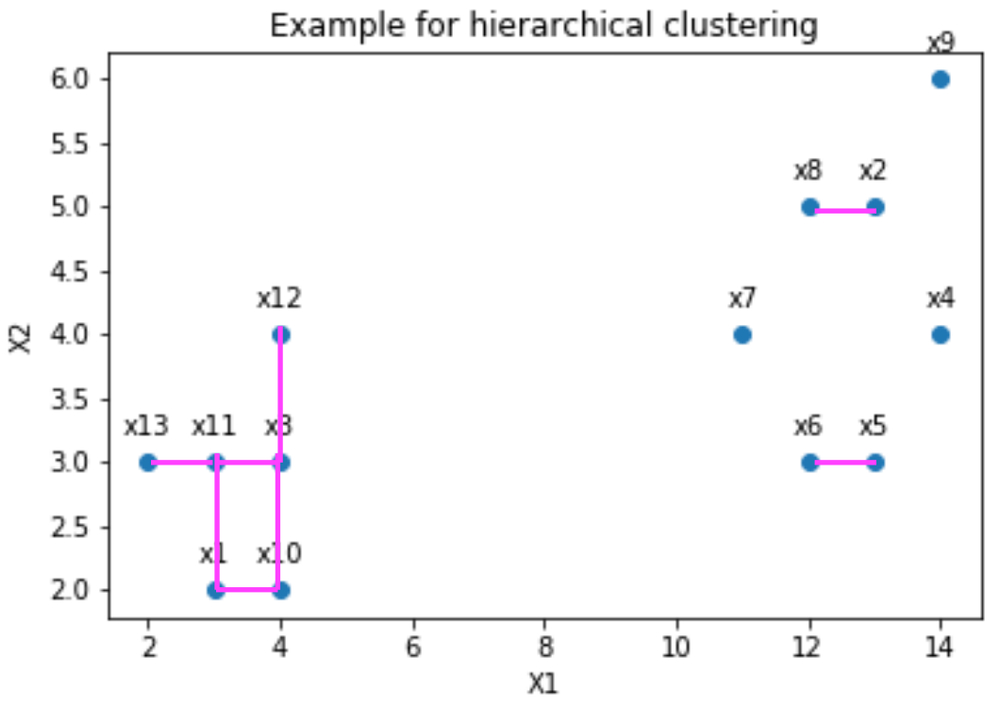
AGGLOMERATIVE CLUSTERING ALGORITHM

AgglomerativeClustering(D, k) :

1. $\mathscr{C} \leftarrow \{C_i = \{x_i\} \,|\, x_i \in D\}$
2. $\Delta = \{\delta(x_i, x_j) : x_i, x_j \in D\}$
3. repeat:
    4. Find the closest pair of clusters $C_i, C_j \in \mathscr{C}$
    5. $C_{i,j} = C_i \cup C_j$
    6. $\mathscr{C} \leftarrow (\mathscr{C} \backslash \{C_i, C_j\}) \cup \{C_{i,j}\}$
    7. Update distance matrix Δ to reflect new clustering
4. Until $|\mathscr{C}| = k$

EXAMPLE

4. Find the closest pair of clusters $C_i, C_j \in \mathscr{C}$

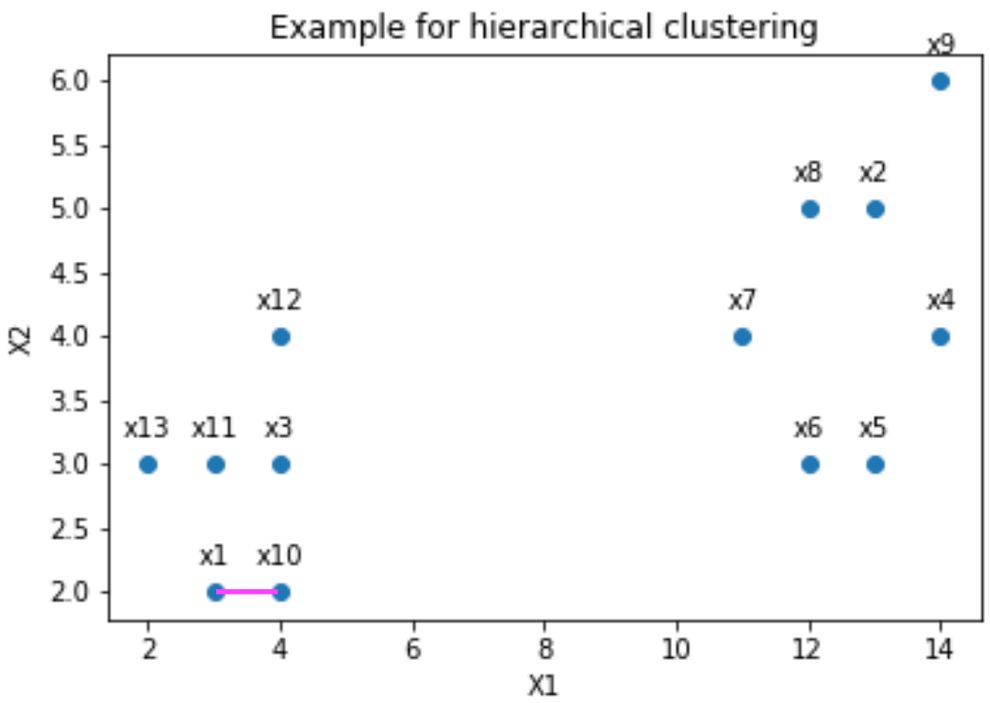


$\Delta =$

	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}	x_{11}	x_{12}	x_{13}
x_1	0												
x_2	10.44	0											
x_3	1.41	9.21	0										
x_4	11.18	1.41	10.05	0									
x_5	10.05	2	9	1.41	0								
x_6	9.06	2.24	8	2.24	1	0							
x_7	8.25	2.24	7.07	3	2.24	1.41	0						
x_8	9.49	1	8.25	2.23	2.24	2	1.41	0					
x_9	11.70	1.41	10.44	2	3.16	3.61	3.61	2.24	0				
x_{10}	1	9.49	1	10.20	9.06	8.06	7.28	8.54	10.77	0			
x_{11}	1	10.20	1	11.05	10	9	8.06	9.22	11.40	1.41	0		
x_{12}	2.24	9.06	1	10	9.06	8.06	7	8.06	10.20	2	1.41	0	
x_{13}	1.41	11.18	2	12.04	11	10	9.06	10.20	12.37	2.24	1	2.24	0

EXAMPLE

4. Find the closest pair of clusters $C_i, C_j \in \mathscr{C}$



$\Delta =$

	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}	x_{11}	x_{12}	x_{13}
x_1	0												
x_2	10.44	0											
x_3	1.41	9.21	0										
x_4	11.18	1.41	10.05	0									
x_5	10.05	2	9	1.41	0								
x_6	9.06	2.24	8	2.24	1	0							
x_7	8.25	2.24	7.07	3	2.24	1.41	0						
x_8	9.49	1	8.25	2.23	2.24	2	1.41	0					
x_9	11.70	1.41	10.44	2	3.16	3.61	3.61	2.24	0				
x_{10}	1	9.49	1	10.20	9.06	8.06	7.28	8.54	10.77	0			
x_{11}	1	10.20	1	11.05	10	9	8.06	9.22	11.40	1.41	0		
x_{12}	2.24	9.06	1	10	9.06	8.06	7	8.06	10.20	2	1.41	0	
x_{13}	1.41	11.18	2	12.04	11	10	9.06	10.20	12.37	2.24	1	2.24	0

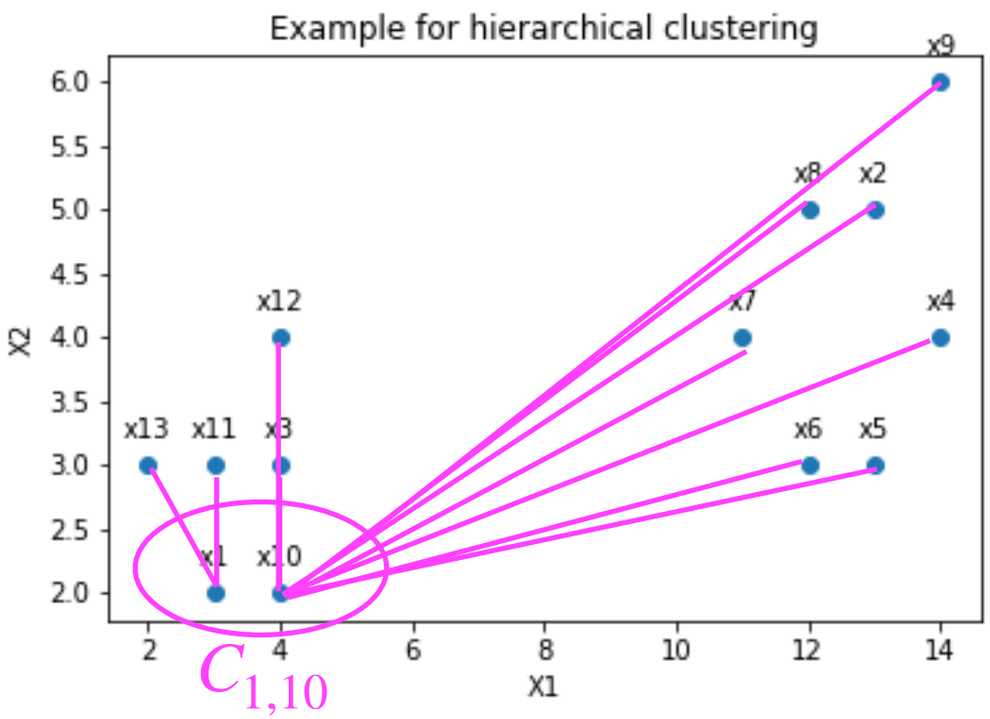
AGGLOMERATIVE CLUSTERING ALGORITHM

AgglomerativeClustering(D, k) :

1. $\mathscr{C} \leftarrow \{C_i = \{x_i\} \,|\, x_i \in D\}$
2. $\Delta = \{\delta(x_i, x_j) : x_i, x_j \in D\}$
3. repeat:
    4. Find the closest pair of clusters $C_i, C_j \in \mathscr{C}$
    5. $C_{i,j} = C_i \cup C_j$
    6. $\mathscr{C} \leftarrow (\mathscr{C} \backslash \{C_i, C_j\}) \cup \{C_{i,j}\}$
    7. Update distance matrix Δ to reflect new clustering
8. Until $|\mathscr{C}| = k$

EXAMPLE

Single linkage: $\delta(C_i, C_j) = \min\{\delta(x, y) \mid x \in C_i, y \in C_j\}$



$C_{1,10}$

$\Delta =$

	$\{x_1,x_{10}\}$	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{11}	x_1	x_{13}
$\{x_1,x_{10}\}$	0											
x_2	9.49	0										
x_3	1	9.21	0									
x_4	10.20	1.41	10.05	0								
x_5	9.06	2	9	1.41	0							
x_6	3.61	2.24	8	2.24	1	0						
x_7	7.28	2.24	7.07	3	2.24	1.41	0					
x_8	8.54	1	8.25	2.23	2.24	2	1.41	0				
x_9	10.77	1.41	10.44	2	3.16	3.61	3.61	2.24	0			
x_{11}	1	10.20	1	11.05	10	9	8.06	9.22	11.40	0		
x_{12}	2	9.06	1	10	9.06	8.06	7	8.06	10.20	1.41	0	
x_{13}	1.41	11.18	2	12.04	11	10	9.06	10.20	12.37	1	2.24	0

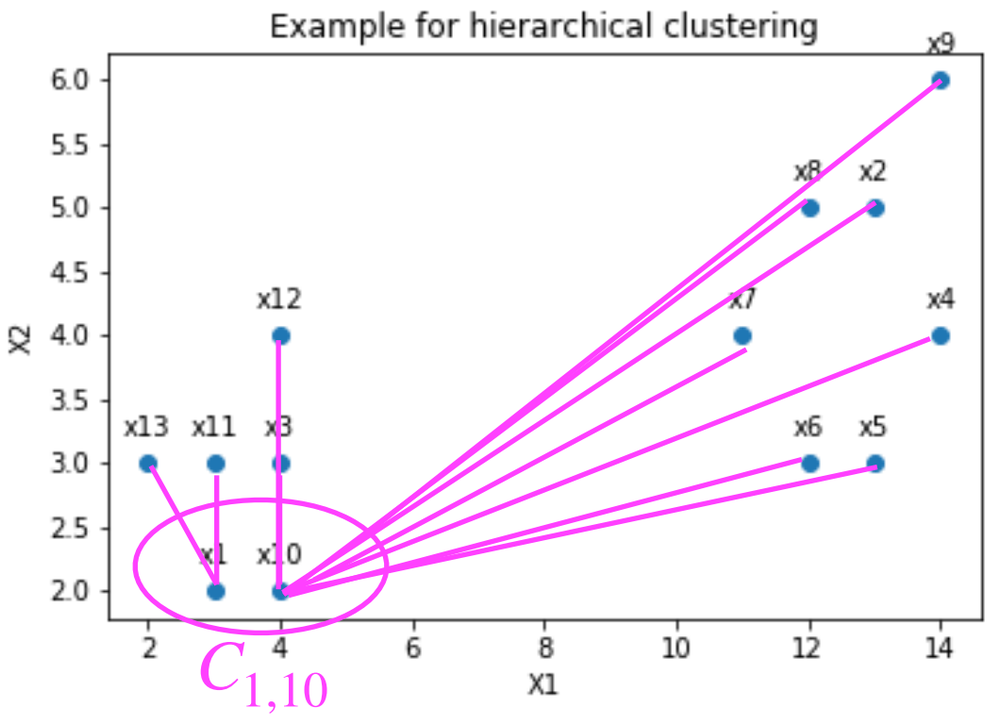
$$C_{i,j} = C_i \cup C_j$$

$$\mathscr{C} \leftarrow (\mathscr{C} \backslash \{C_i, C_j\}) \cup \{C_{i,j}\}$$

Update Δ

EXAMPLE

Single linkage: $\delta(C_i, C_j) = \min\{\delta(x, y) \mid x \in C_i, y \in C_j\}$



$C_{1,10}$

$\Delta =$

	$\{x_1,x_{10}\}$	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{11}	x_1	x_{13}
$\{x_1,x_{10}\}$	0											
x_2	9.49	0										
x_3	1	9.21	0									
x_4	10.20	1.41	10.05	0								
x_5	9.06	2	9	1.41	0							
x_6	3.61	2.24	8	2.24	1	0						
x_7	7.28	2.24	7.07	3	2.24	1.41	0					
x_8	8.54	1	8.25	2.23	2.24	2	1.41	0				
x_9	10.77	1.41	10.44	2	3.16	3.61	3.61	2.24	0			
x_{11}	1	10.20	1	11.05	10	9	8.06	9.22	11.40	0		
x_{12}	2	9.06	1	10	9.06	8.06	7	8.06	10.20	1.41	0	
x_{13}	1.41	11.18	2	12.04	11	10	9.06	10.20	12.37	1	2.24	0

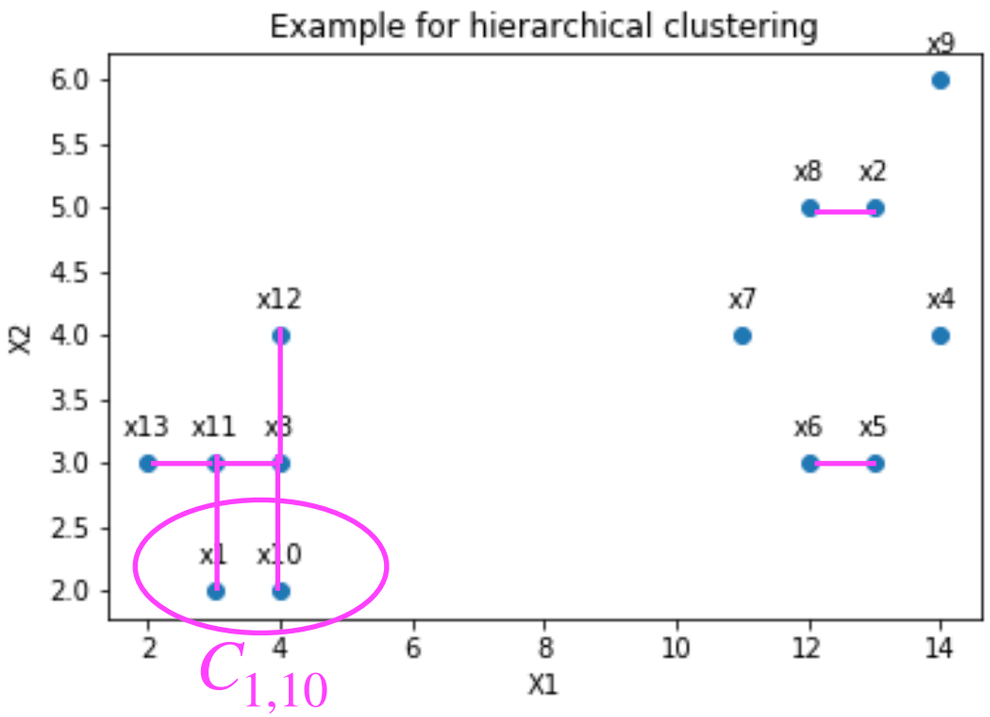
AGGLOMERATIVE CLUSTERING ALGORITHM

AgglomerativeClustering(D, k) :

1. $\mathscr{C} \leftarrow \{C_i = \{x_i\} \,|\, x_i \in D\}$
2. $\Delta = \{\delta(x_i, x_j) : x_i, x_j \in D\}$
3. repeat:
    4. Find the closest pair of clusters $C_i, C_j \in \mathscr{C}$
    5. $C_{i,j} = C_i \cup C_j$
    6. $\mathscr{C} \leftarrow (\mathscr{C} \backslash \{C_i, C_j\}) \cup \{C_{i,j}\}$
    7. Update distance matrix Δ to reflect new clustering
8. Until $|\mathscr{C}| = k$

EXAMPLE

Single linkage: $\delta(C_i, C_j) = \min\{\delta(x, y) \mid x \in C_i, y \in C_j\}$



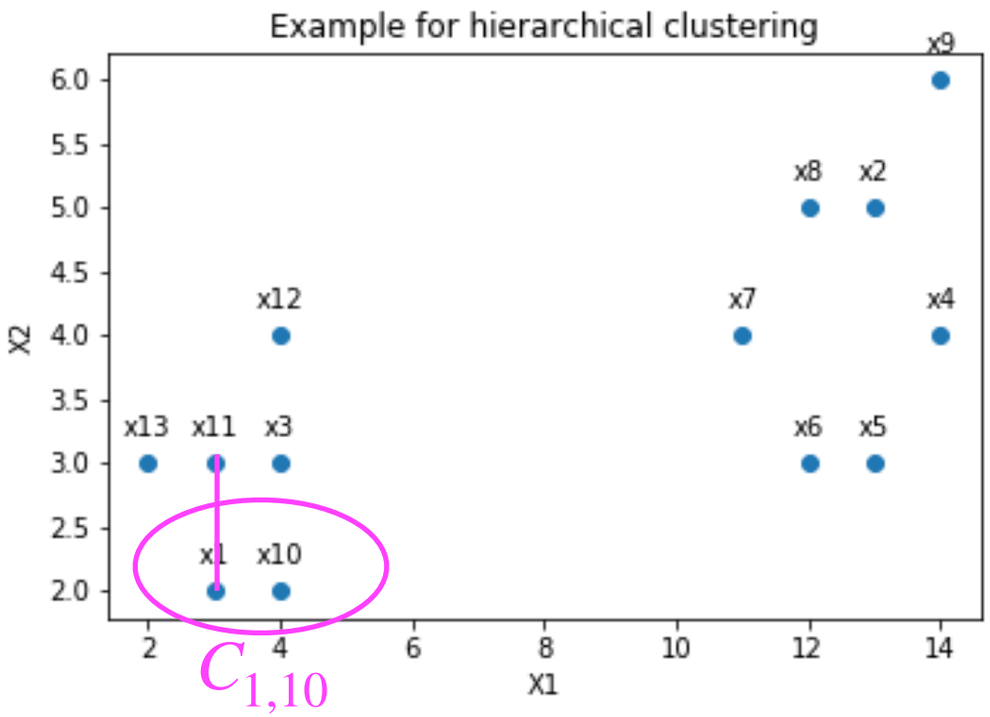
$C_{1,10}$

$\Delta =$

	$\{x_1,x_{10}\}$	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{11}	x_1	x_{13}
$\{x_1,x_{10}\}$	0											
x_2	9.49	0										
x_3	1	9.21	0									
x_4	10.20	1.41	10.05	0								
x_5	9.06	2	9	1.41	0							
x_6	3.61	2.24	8	2.24	1	0						
x_7	7.28	2.24	7.07	3	2.24	1.41	0					
x_8	8.54	1	8.25	2.23	2.24	2	1.41	0				
x_9	10.77	1.41	10.44	2	3.16	3.61	3.61	2.24	0			
x_{11}	1	10.20	1	11.05	10	9	8.06	9.22	11.40	0		
x_{12}	2	9.06	1	10	9.06	8.06	7	8.06	10.20	1.41	0	
x_{13}	1.41	11.18	2	12.04	11	10	9.06	10.20	12.37	1	2.24	0

EXAMPLE

Single linkage: $\delta(C_i, C_j) = \min\{\delta(x, y) \mid x \in C_i, y \in C_j\}$

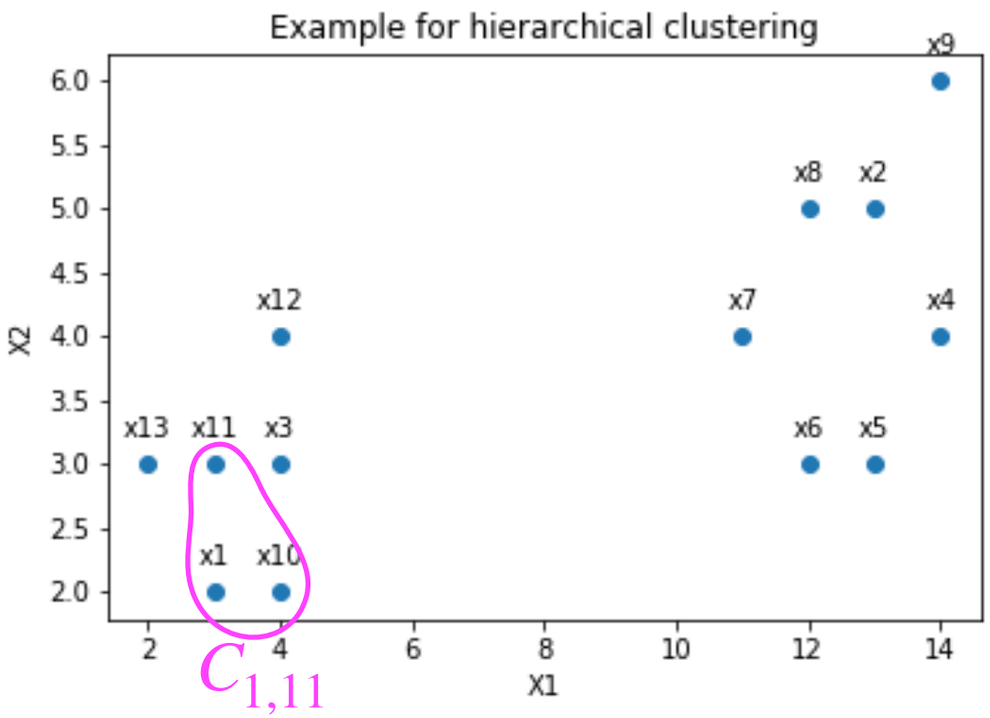


$\Delta =$

	$\{x_1,x_{10}\}$	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{11}	x_1	x_{13}
$\{x_1,x_{10}\}$	0											
x_2	9.49	0										
x_3	1	9.21	0									
x_4	10.20	1.41	10.05	0								
x_5	9.06	2	9	1.41	0							
x_6	3.61	2.24	8	2.24	1	0						
x_7	7.28	2.24	7.07	3	2.24	1.41	0					
x_8	8.54	1	8.25	2.23	2.24	2	1.41	0				
x_9	10.77	1.41	10.44	2	3.16	3.61	3.61	2.24	0			
x_{11}	1	10.20	1	11.05	10	9	8.06	9.22	11.40	0		
x_{12}	2	9.06	1	10	9.06	8.06	7	8.06	10.20	1.41	0	
x_{13}	1.41	11.18	2	12.04	11	10	9.06	10.20	12.37	1	2.24	0

EXAMPLE

Single linkage: $\delta(C_i, C_j) = \min\{\delta(x, y) \mid x \in C_i, y \in C_j\}$

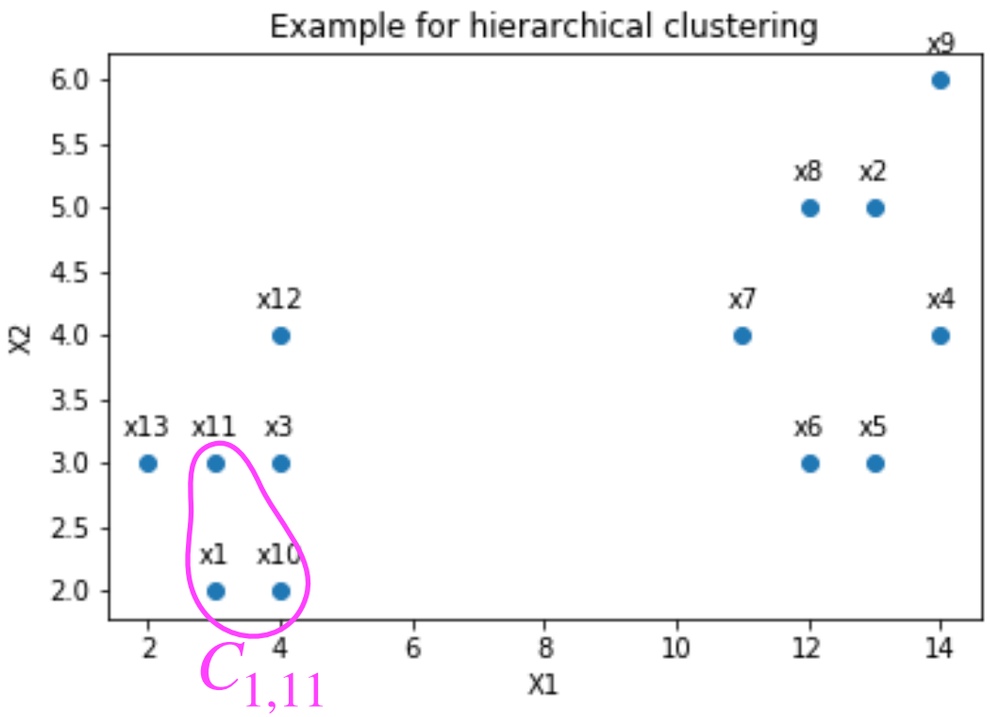


$\Delta =$

	$\{x_1,x_{10}\}$	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{11}	x_1	x_{13}
$\{x_1,x_{10}\}$	0											
x_2	9.49	0										
x_3	1	9.21	0									
x_4	10.20	1.41	10.05	0								
x_5	9.06	2	9	1.41	0							
x_6	3.61	2.24	8	2.24	1	0						
x_7	7.28	2.24	7.07	3	2.24	1.41	0					
x_8	8.54	1	8.25	2.23	2.24	2	1.41	0				
x_9	10.77	1.41	10.44	2	3.16	3.61	3.61	2.24	0			
x_{11}	1	10.20	1	11.05	10	9	8.06	9.22	11.40	0		
x_{12}	2	9.06	1	10	9.06	8.06	7	8.06	10.20	1.41	0	
x_{13}	1.41	11.18	2	12.04	11	10	9.06	10.20	12.37	1	2.24	0

EXAMPLE

Single linkage: $\delta(C_i, C_j) = \min\{\delta(x, y) \mid x \in C_i, y \in C_j\}$



$C_{1,11}$

$\Delta =$

	$\{x_1,x_{10},x_{11}\}$	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{12}	x_{13}
$\{x_1,x_{10},x_{11}\}$	0										
x_2	9.49	0									
x_3	1	9.21	0								
x_4	10.20	1.41	10.05	0							
x_5	9.06	2	9	1.41	0						
x_6	3.61	2.24	8	2.24	1	0					
x_7	7.28	2.24	7.07	3	2.24	1.41	0				
x_8	8.54	1	8.25	2.23	2.24	2	1.41	0			
x_9	10.77	1.41	10.44	2	3.16	3.61	3.61	2.24	0		
x_{12}	1.41	9.06	1	10	9.06	8.06	7	8.06	10.20	0	
x_{13}	1	11.18	2	12.04	11	10	9.06	10.20	12.37	2.24	0

DIFFERENT DISTANCE MEASURES WILL AFFECT RESULTS

- L_1 norm, L_2 norm, etc…
- Single Linkage
- Complete Linkage
- Group Average
- Mean Distance
- Minimum Variance/Ward's Method