CSCI 347: Introduction to Data Mining

Hierarchical Clustering Example

AgglomerativeClustering(D, k):

1.
$$\mathscr{C} \leftarrow \{C_i = \{x_i\} \mid x_i \in D\}$$

2.
$$\Delta = \{ \delta(x_i, x_j) : x_i, x_j \in D \}$$

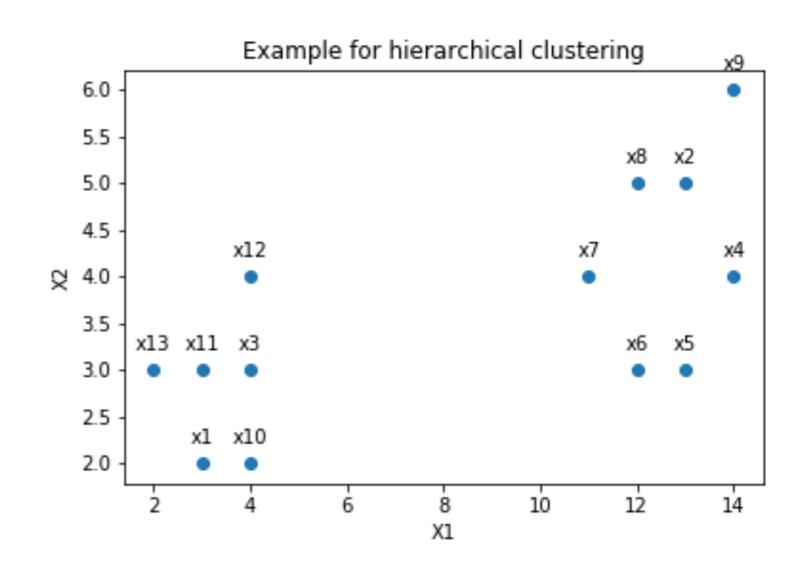
- 3. repeat:
 - 4. Find the closest pair of clusters $C_i, C_i \in \mathcal{C}$

5.
$$C_{i,j} = C_i \cup C_j$$

6.
$$\mathscr{C} \leftarrow (\mathscr{C} \setminus \{C_i, C_j\}) \cup \{C_{i,j}\}$$

- 7. Update distance matrix Δ to reflect new clustering
- 4. Until $|\mathscr{C}| = k$

		X_1	X_2
	x_1	3	2
	x_2	13	5
	x_3	4	3
D =	\mathcal{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 (D, k):

1.
$$\mathscr{C} \leftarrow \{C_i = \{x_i\} \mid x_i \in D\}$$

2.
$$\Delta = \{ \delta(x_i, x_j) : x_i, x_j \in D \}$$

- 3. repeat:
 - 4. Find the closest pair of clusters $C_i, C_i \in \mathcal{C}$

$$5. C_{i,j} = C_i \cup C_j$$

6.
$$\mathscr{C} \leftarrow (\mathscr{C} \setminus \{C_i, C_j\}) \cup \{C_{i,j}\}$$

- 7. Update distance matrix Δ to reflect new clustering
- 4. Until $|\mathscr{C}| = k$

	X_1	X_2
x_1	3	2
x_2	13	5
x_3	4	3
\mathcal{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

	X 1	X2	X3	X4	X5	X ₆	X7	X8	X 9	X ₁₀	X ₁₁	X ₁₂	X ₁₃
x_1	0												
X2		0											
X3			0										
X4				0									
X5					0								
X ₆						0							
X7							0						
X8								0					
X 9									0				
X ₁₀										0			
X ₁₁											0		
X ₁₂												0	
X ₁₃													0

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

		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

		X 1	X2	X3	X4	X5	X ₆	X7	X8	X 9	X ₁₀	X ₁₁	X ₁₂	X ₁₃
	X 1	0												
	X ₂		0											
	X3			0										
	X4				_	-		f L:		:!_				
$\Delta = 0$	X5			6.0		EX	ample	for ni	erarcn	ical clu	usterin	g	x9	
_	X ₆			5.5 - 5.0 -							1	x8 x2		
	X7			4.5 -		x12					A		x4	
	X8		_ X	4.0 - 3.5 -	13						•	cc	•	
	X 9			3.0 -	x13 x	• •					1	x6 x5		
	X ₁₀			2.5 -		1 x20								
	X ₁₁				2	4		6	8 X1	10		12	14	
	X ₁₂												0	
	X ₁₃													0

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

		X_1	X_2
	x_1	3	2
	x_2	13	5
	x_3	4	3
=	\mathcal{X}_4	14	4
	x_5	13	3
	x_6	12	3
	x_7	11	4
	x_8	12	5
	x_9	14	6
	<i>x</i> ₁₀	4	2
	x_{11}	3	3
	<i>x</i> ₁₂	4	4
	<i>x</i> ₁₃	2	3

		x ₁	X2	X3	X4	X5	X6	X7	X8	X 9	X ₁₀	X ₁₁	X ₁₂	X ₁₃
	\mathbf{x}_1	0												
	X 2	10.44	0											
	X3			0										
	X4				_	Evar	nnlo f	or bior	archic	al clust	oring			
\ = L	X5			6.0 -		Exai	iipie it	n mei	archic	ai Ciusi	tering		9	
	X 6			5.5 - 5.0 -							x8	x2		
	X 7			4.5 -		x12					KI .		κ 4	
	X8		_ X	35 -	13 ×11	x3			/		х6	x5	•	
	X 9			3.0 - 2.5 -	• •	•	/				•	•		
	X ₁₀			2.0 -	•	×20			,					
	X ₁₁				2	4	6		8 X1	10	12]	4	
	X ₁₂												0	
	X ₁₃													0

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

		X_1	X_2
	x_1	3	2
	x_2	13	5
	x_3	4	3
=	\mathcal{X}_4	14	4
	x_5	13	3
		12	3
	x_7	11	4
	x_8	12	5
	x_9	14	6
	<i>x</i> ₁₀	4	2
	x_{11}	3	3
	<i>x</i> ₁₂	4	4
	x_{13}	2	3

		x ₁	X2	X3	X ₄	X5	X6	X7	X8	X 9	X ₁₀	X ₁₁	X ₁₂	X ₁₃
	X 1	0												
	X2	10.44	0											
	X 3	1.41		0										
	X4				_	F		b.:		-1 -14				
$\Delta = $	X5		(5.0 -		Exar	npie ro	or nier	arcnic	al clust	ering	;	x9	
	X ₆			5.5 -							x8	x2		
	X7		4	4.5 -		x12				2	x7	2	x4	
	X8		— ₂ 4	3.5 -		•					•		•	
	X 9			3.0 -	(13 x11	, x3					x6	x5		
	X ₁₀			2.5 -	×1	×10								
	X ₁₁				2	4	6		8 X1	10	12]	14	
	X ₁₂												0	
	X ₁₃													0

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

		X 1	X2	X3	X4	X5	X ₆	X7	X8	X 9	X ₁₀	X ₁₁	X ₁₂	X ₁₃
	X 1	0												
	X2	10.44	0											
	X3	1.41	9.21	0										
	X4	11.18	1.41	10.05	0									
: [X5	10.05	2	9	1.41	0								
	X6	9.06	2.24	8	2.24	1	0							
	X7	8.25	2.24	7.07	3	2.24	1.41	0						
	X8	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	
	X13	1.41	11.18	2	12.04	11	10	9.06	10.20	12.37	2.24	1	2.24	0

Agglomerative Clustering (D, k):

1.
$$\mathscr{C} \leftarrow \{C_i = \{x_i\} \mid x_i \in D\}$$

2.
$$\Delta = \{ \delta(x_i, x_j) : x_i, x_j \in D \}$$

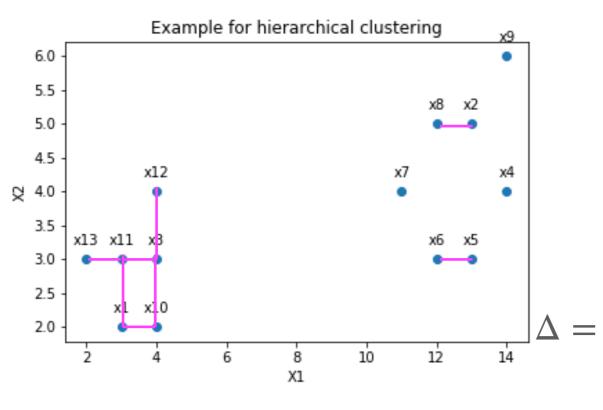
- 3. repeat:
 - 4. Find the closest pair of clusters $C_i, C_i \in \mathscr{C}$

$$5. C_{i,j} = C_i \cup C_j$$

6.
$$\mathscr{C} \leftarrow (\mathscr{C} \setminus \{C_i, C_j\}) \cup \{C_{i,j}\}$$

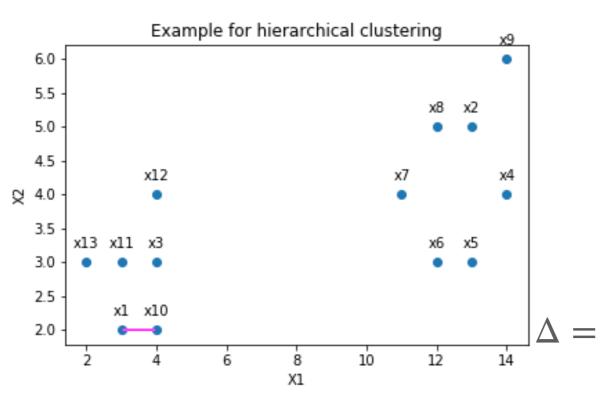
- 7. Update distance matrix Δ to reflect new clustering
- 4. Until $|\mathscr{C}| = k$

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



	X 1	X2	X3	X4	X5	X6	X7	X8	X 9	X ₁₀	X ₁₁	X ₁₂	X ₁₃
X 1	0												
X2	10.44	0											
X3	1.41	9.21	0										
X4	11.18	1.41	10.05	0									
X5	10.05	2	9	1.41	0								
X ₆	9.06	2.24	8	2.24	1	0							
X7	8.25	2.24	7.07	3	2.24	1.41	0						
X8	9.49	1	8.25	2.23	2.24	2	1.41	0					
X9	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

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



	X 1	X2	X3	X4	X5	X ₆	X7	X8	X9	X ₁₀	X ₁₁	X ₁₂	X ₁₃
X 1	0												
X2	10.44	0											
X3	1.41	9.21	0										
X4	11.18	1.41	10.05	0									
X5	10.05	2	9	1.41	0								
X ₆	9.06	2.24	8	2.24	1	0							
X7	8.25	2.24	7.07	3	2.24	1.41	0						
X8	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(D, k):

1.
$$\mathscr{C} \leftarrow \{C_i = \{x_i\} \mid x_i \in D\}$$

2.
$$\Delta = \{ \delta(x_i, x_j) : x_i, x_j \in D \}$$

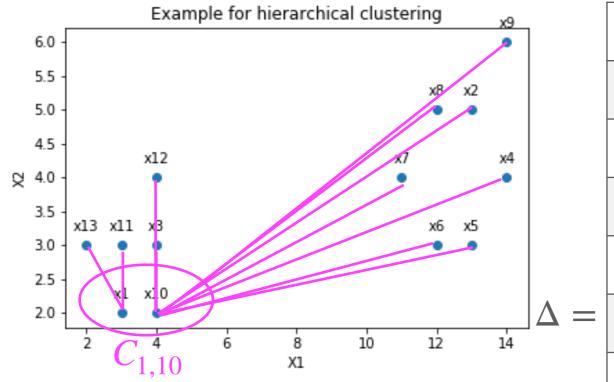
- 3. repeat:
 - 4. Find the closest pair of clusters $C_i, C_i \in \mathcal{C}$

$$5. C_{i,j} = C_i \cup C_j$$

6.
$$\mathscr{C} \leftarrow (\mathscr{C} \setminus \{C_i, C_j\}) \cup \{C_{i,j}\}$$

- 7. Update distance matrix Δ to reflect new clustering
- 8. Until $|\mathscr{C}| = k$

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

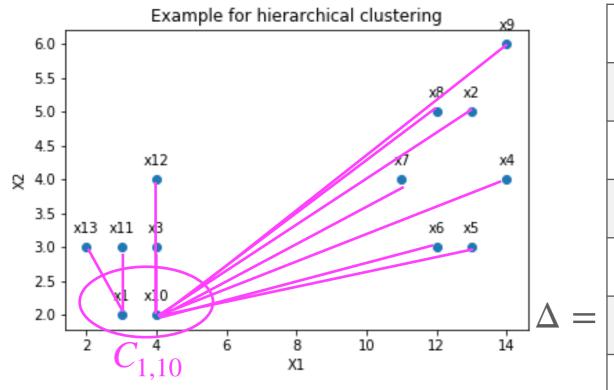


$$C_{i,j} = C_i \cup C_j$$

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

Update Δ

	{x ₁ ,x ₁₀ }	X2	X3	X4	X5	X ₆	X7	X8	X 9	X ₁₁	X ₁	X ₁₃
$\{x_{1,}x_{10}\}$	0											
X ₂	9.49	0										
X3	1	9.21	0									
X4	10.20	1.41	10.05	0								
X5	9.06	2	9	1.41	0							
X ₆	3.61	2.24	8	2.24	1	0						
X7	7.28	2.24	7.07	3	2.24	1.41	0					
X8	8.54	1	8.25	2.23	2.24	2	1.41	0				
X9	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



	$\{x_{1,}x_{10}\}$	X2	X3	X4	X5	X6	X7	X8	X9	X ₁₁	\mathbf{x}_1	X ₁₃
$\{x_{1},x_{10}\}$	0											
X2	9.49	0										
X3	1	9.21	0									
X4	10.20	1.41	10.05	0								
X5	9.06	2	9	1.41	0							
X ₆	3.61	2.24	8	2.24	1	0						
X7	7.28	2.24	7.07	3	2.24	1.41	0					
X8	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 (D, k):

1.
$$\mathscr{C} \leftarrow \{C_i = \{x_i\} \mid x_i \in D\}$$

2.
$$\Delta = \{ \delta(x_i, x_j) : x_i, x_j \in D \}$$

3. repeat:

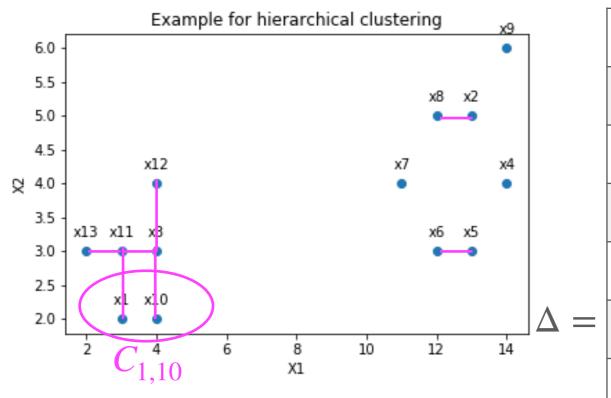
4. Find the closest pair of clusters $C_i, C_j \in \mathscr{C}$

5.
$$C_{i,j} = C_i \cup C_j$$

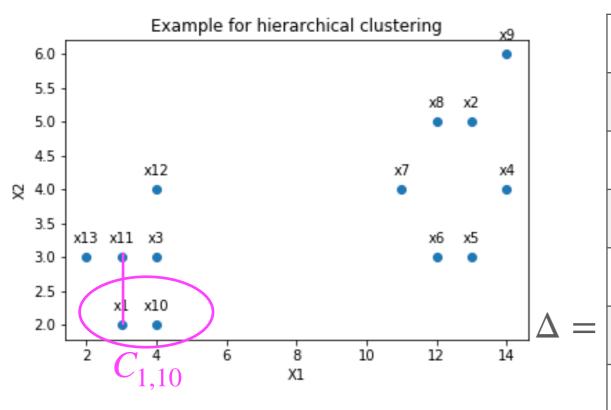
6.
$$\mathscr{C} \leftarrow (\mathscr{C} \setminus \{C_i, C_j\}) \cup \{C_{i,j}\}$$

7. Update distance matrix Δ to reflect new clustering

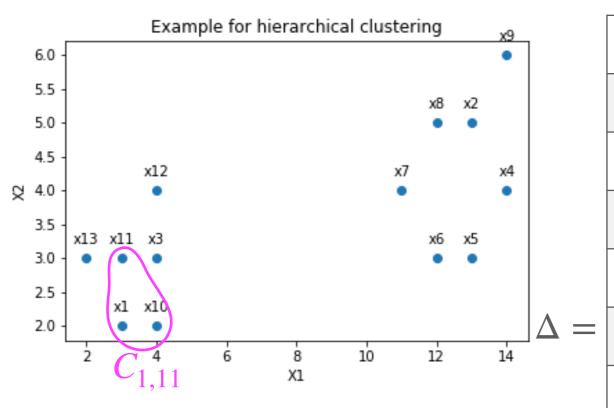
8. Until
$$|\mathscr{C}| = k$$



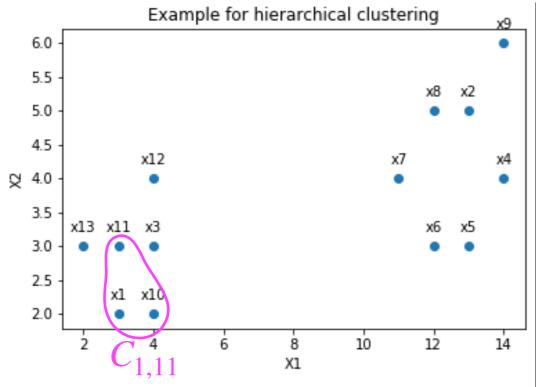
	$\{x_{1,}x_{10}\}$	X2	X3	X4	X5	X ₆	X ₇	X8	X 9	X ₁₁	X 1	X ₁₃
$\{x_{1},x_{10}\}$	0											
X2	9.49	0										
X3	1	9.21	0									
X4	10.20	1.41	10.05	0								
X5	9.06	2	9	1.41	0							
X ₆	3.61	2.24	8	2.24	1	0						
X ₇	7.28	2.24	7.07	3	2.24	1.41	0					
X8	8.54	1	8.25	2.23	2.24	2	1.41	0				
X 9	10.77	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



	$\{x_{1,}x_{10}\}$	X2	X3	X4	X5	X ₆	X ₇	X8	X 9	X ₁₁	\mathbf{x}_1	X ₁₃
$\{x_{1,}x_{10}\}$	0											
X2	9.49	0										
X3	1	9.21	0									
X4	10.20	1.41	10.05	0								
X5	9.06	2	9	1.41	0							
X ₆	3.61	2.24	8	2.24	1	0						
X7	7.28	2.24	7.07	3	2.24	1.41	0					
X8	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



	$\{x_{1,}x_{10}\}$	X2	X3	X4	X5	X ₆	X7	X8	X 9	X ₁₁	X 1	X ₁₃
$\{x_{1},x_{10}\}$	0											
X2	9.49	0										
X3	1	9.21	0									
X4	10.20	1.41	10.05	0								
X5	9.06	2	9	1.41	0							
X ₆	3.61	2.24	8	2.24	1	0						
X7	7.28	2.24	7.07	3	2.24	1.41	0					
X8	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	
X13	1.41	11.18	2	12.04	11	10	9.06	10.20	12.37	1	2.24	0



Δ	=
---	---

	$\{x_{1,}x_{10,}x_{11}\}$	X2	X3	X4	X5	X ₆	X7	X8	X 9	X ₁₂	X ₁₃
$\{x_{1,}x_{10,}x_{11}\}$	0										
X ₂	9.49	0									
X3	1	9.21	0								
X4	10.20	1.41	10.05	0							
X5	9.06	2	9	1.41	0						
X ₆	3.61	2.24	8	2.24	1	0					
X7	7.28	2.24	7.07	3	2.24	1.41	0				
X8	8.54	1	8.25	2.23	2.24	2	1.41	0			
X9	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

- ightharpoonup L₁ norm, L₂ norm, etc...
- ➤ Single Linkage
- Complete Linkage
- ➤ Group Average
- ➤ Mean Distance
- ➤ Minimum Variance/Ward's Method