

INFS 4203 / 7203 Data Mining Tutorial 5: Clustering

Seun Aremu o.aremu@uq.edu.au

+ Clustering

Key concepts

- Similarity and distance
 - L_1 , Euclidean distance (L_2 norm), L_{∞} norm
 - Edit distance
- Clustering
 - Agglomerative hierarchical clustering:



+ Clustering

Similarity and distance

- How to evaluate similarity between observations:
 - Distance-based (e.g. L_p norm)
 - Edit distance



+

Clustering

Distance calculation --- L_p norm

■ Minkowski Distance: generalization of Euclidian distance

$$Distance = \left(\sum_{k=1}^{n} |p_k - q_k|^r\right)^{\frac{1}{r}}$$

- r is a parameter
- n is the number of attributes
- $\blacksquare p_k$ and q_k are, respectively, the kth attribute of data objects **P** and **Q**.



Distance calculation --- L_p norm

■ Let r = 1:

Distance =
$$(\sum_{k=1}^{n} |p_k - q_k|^1)^{\frac{1}{1}}$$

Also known as L_1 norm distance measure:

$$L_1(X,Y) = \sum_{k=1}^n |X_k - Y_k|$$

- Input: X = (1, 0, 5); Y = (2, 4, 9)
 - $L_1(X,Y) = ?$



Distance calculation --- L_p norm

■ L_1 norm distance measure:

$$L_1(X,Y) = \sum_{k=1}^{n} |X_k - Y_k|$$

- Input: X = (1, 0, 5); Y = (2, 4, 9)
 - $L_1(X,Y) = ?$
 - n = 3
 - $X_1 = 1$, $X_2 = 0$, $X_3 = 5$
 - $Y_1 = 2$, $Y_2 = 4$, $Y_3 = 9$
- L_1 norm distance measure:

$$L_1(X,Y) = |X_1 - Y_1| + |X_2 - Y_2| + |X_3 - Y_3| = |\mathbf{1} - \mathbf{2}| + |\mathbf{0} - \mathbf{4}| + |\mathbf{5} - \mathbf{9}| = \mathbf{9}$$



Distance calculation --- L_p norm

■ Let r = 2:

Distance =
$$(\sum_{k=1}^{n} |p_k - q_k|^2)^{\frac{1}{2}}$$

Also known as L_2 norm distance measure:

$$L_2(X,Y) = \sqrt{\sum_{k=1}^{n} (X_k - Y_k)^2}$$

- Input: X = (1, 0, 5); Y = (2, 4, 9)
 - $L_2(X,Y) = ?$

+

T5-Q1

Distance calculation --- L_p norm

■ L_2 norm distance measure:

$$L_2(X,Y) = \sqrt{\sum_{k=1}^{n} (X_k - Y_k)^2}$$

- Input: X = (1, 0, 5); Y = (2, 4, 9)
 - $L_2(X,Y) = ?$
 - n = 3
 - $X_1 = 1, X_2 = 0, X_3 = 5$
 - $Y_1 = 2$, $Y_2 = 4$, $Y_3 = 9$
- L_2 norm distance measure:

$$L_2(X,Y) = \sqrt{(X_1 - Y_1)^2 + (X_2 - Y_2)^2 + (X_3 - Y_3)^2} = \sqrt{(1 - 2)^2 + (0 - 4)^2 + (5 - 9)^2}$$

$$L_2(X,Y) = \sqrt{33} = 5.74$$



Distance calculation --- L_p norm

■ L_{∞} norm or L_{max} norm distance measure:

$$L_{\infty}(X,Y) = \max_{k=1,\dots,n} |X_k - Y_k|$$

- Input: X = (1, 0, 5); Y = (2, 4, 9)
 - $L_{\infty}(X,Y) = ?$



Distance calculation --- L_p norm

■ L_{∞} norm or L_{max} norm distance measure:

$$L_{\infty}(X,Y) = \max_{k=1,\dots,n} |X_k - Y_k|$$

- Input: X = (1, 0, 5); Y = (2, 4, 9)
 - $L_{\infty}(X,Y) = ?$
 - n = 3
 - $X_1 = 1, X_2 = 0, X_3 = 5$
 - $Y_1 = 2$, $Y_2 = 4$, $Y_3 = 9$
- L_{∞} norm distance measure:

$$L_{\infty}(X,Y) = \max(|X_1 - Y_1|, |X_2 - Y_2|, |X_3 - Y_3|) = \max(|1 - 2|, |0 - 4|, |5 - 9|) = 4$$





Distance calculation --- L_p norm

- Input: X = (1, 0, 5); Y = (2, 4, 9)
 - $L_1(X,Y) = 9$
 - $L_2(X,Y) = 5.74$
 - $L_{\infty}(X,Y)=4$

Distance calculation --- edit distance

- Edit distance:
 - Minimum number of edit operations to change string X to string Y
- Edit operations:
 - Insert a symbol
 - **Delete** a symbol
 - Substitute a symbol
- Cost of edit operations:
 - **■** Insert = 1
 - **■ Delete** = 1
 - Substitute = 1



Distance calculation --- edit distance

- Input: X = "university"; Y = "unversity"
 - Calculate their edit distance
- "unversity" "unversty" (Delete 'i') (Unit Cost = 1)
- "unversty" ——— "unverstiy" (Insert 'i') (Unit Cost = 1)

■ The edit distance is 3



Distance calculation --- edit distance

- Input: X = "university"; Y = "unversity"
 - Calculate their edit distance
- "university" ——— "unversity" (Delete 'i') (Unit Cost = 1)
- "unversity" ——— "unverstty" (Substitute 'i' for 't') (Unit Cost = 1)
- "unverstty" ——— "unverstiy" (Substitute 't' for 'i') (Unit Cost = 1)

■ The edit distance is 3



Hierarchical clustering method

- Given a set of ages, {18, 22, 28, 33, 40, 48}
 - Use **Agglomerative Hierarchical Clustering** algorithm to group them step by step.
 - Use min to merge two closest clusters and update Proximity Matrix correspondingly.
 - Use max to merge two closest clusters and update Proximity Matrix correspondingly.



Hierarchical clustering

- Agglomerative:
 - Start with singleton clusters, continuously merge two clusters at a time to build a bottom-up hierarchy of clusters
 - Single link (nearest neighbor --- min)
 - Complete link (diameter --- max)
- Algorithm
 - 1. Compute the proximity matrix for each point
 - Let each data point be a cluster
 - 2. Merge the two closest clusters
 - 3. Update the proximity matrix
 - 4. Repeat steps 2 and 3 until only a single cluster remains



Hierarchical clustering

- Agglomerative:
 - Start with singleton clusters, continuously merge two clusters at a time to build a bottom-up hierarchy of clusters
 - Proximity measure by Single link (nearest neighbor --- min)
 - - *X* and *Y* are the cluster sets
 - Similarity of two clusters is based on the **two most similar (closest) points** in the different clusters



min-merge clustering process:

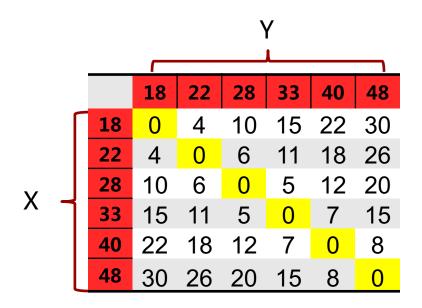
- Algorithm
 - 1. Compute the proximity matrix for each point in **X** and **Y**
 - Let each data point be a cluster

				•	Υ		
		18	22	28	33	40	48
	18						
	22						
X -	28						
^	33						
	40						
	48						



⁺ T5-Q3

min-merge clustering process:



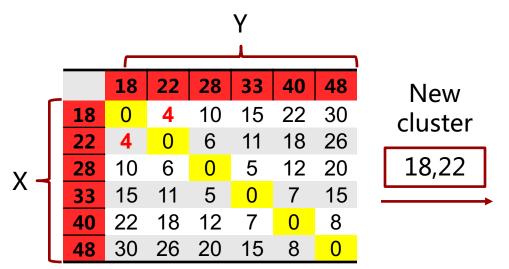
- Find the minimum distance between points in cluster $X_{k,n}$ and cluster $Y_{k,m}$
 - k is the number of clusters in the cluster set
 - \blacksquare n is the number of points in cluster X_k
 - \blacksquare m is the number of points in cluster Y_k

$$D(X_k, Y_k) = \min_{\substack{i=1,\dots,n\\j=1,\dots,m}} \left\{ dist(x_i, y_j) \right\}$$



min-merge clustering process

- Algorithm
 - 2. Merge the two closest clusters
 - 3. Update the proximity matrix

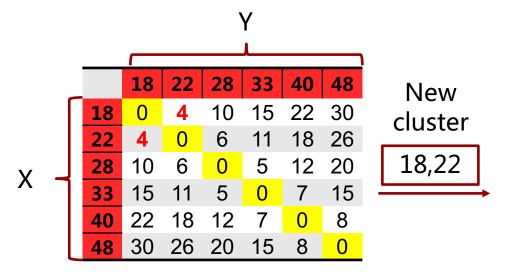


$$D(X_k, Y_k) = \min_{\substack{i=1,\dots,n\\j=1,\dots,m}} \left\{ dist(x_i, y_j) \right\}$$

New Cluster =
$$\min_{1,...,k} \{D(X_k, Y_k)\}$$

min-merge clustering process

- Algorithm
 - 2. Merge the two closest clusters
 - 3. Update the proximity matrix



			Υ		
					$\overline{}$
	18,22	28	33	40	48
18,22	0	6	11	18	26
28	6	0	5	12	20
33	11	5	0	7	15
40	18	12	7	0	8
48	26	20	15	8	0

New cluster 28,33

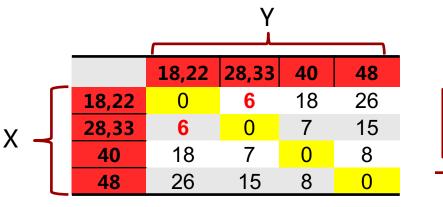
$$D(X_k, Y_k) = \min_{\substack{i=1,\dots,n\\j=1,\dots,m}} \left\{ dist(x_i, y_j) \right\}$$

New Cluster =
$$\min_{1,...,k} \{D(X_k, Y_k)\}$$



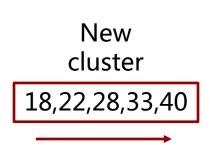
min-merge clustering process

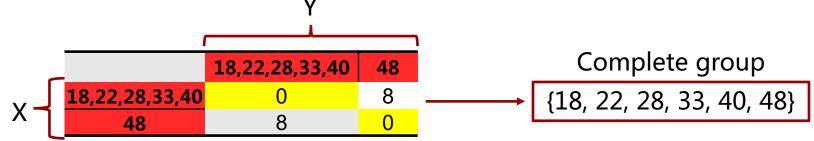
- Algorithm
 - Repeat steps 2 and 3 until only a single cluster remains





	Υ				
		<u> </u>			
	18,22,28,33	40	48		
18,22,28,33	0	7	15		
40	7	0	8		
48	15	8	0		





$$D(X_k, Y_k) = \min_{\substack{i=1,\dots,n\\j=1,\dots,m}} \left\{ dist(x_i, y_j) \right\}$$

New Cluster =
$$\min_{1,...,k} \{D(X_k, Y_k)\}$$

Hierarchical clustering

- Agglomerative:
 - Start with singleton clusters, continuously merge two clusters at a time to build a bottom-up hierarchy of clusters
 - Proximity measure by Complete link (diameter --- max)
 - - X and Y are the cluster sets
 - Similarity of two clusters is based on the **two least similar (most distant)** points in the different clusters



max-merge clustering process:

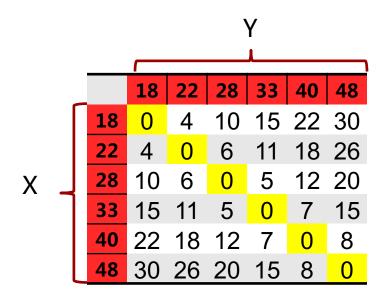
- Algorithm
 - 1. Compute the proximity matrix for each point in **X** and **Y**
 - Let each data point be a cluster

				`	Y		
		18	22	28	33	40	48
	18						
	22						
Χ -	28						
^	33						
	40						
	48	·	·				



⁺ T5-Q3

max-merge clustering process:



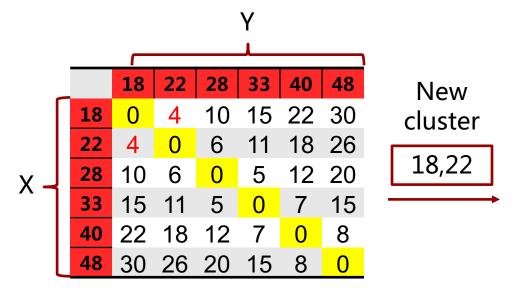
- Find the max distance between points in cluster $X_{k,n}$ and cluster $Y_{k,m}$
 - k is the number of clusters in the cluster set
 - \blacksquare *n* is the number of points in cluster X_k
 - \blacksquare m is the number of points in cluster Y_k

$$D(X_k, Y_k) = \max_{\substack{i=1,\dots,n\\j=1,\dots,m}} \{dist(x_i, y_j)\}$$



max-merge clustering process

- Algorithm
 - 2. Merge the two closest clusters
 - 3. Update the proximity matrix



$$D(X_k, Y_k) = \max_{\substack{i=1,\dots,n\\j=1,\dots,m}} \{dist(x_i, y_j)\}$$

New Cluster =
$$\min_{1,...,k} \{D(X_k, Y_k)\}$$

max-merge clustering process

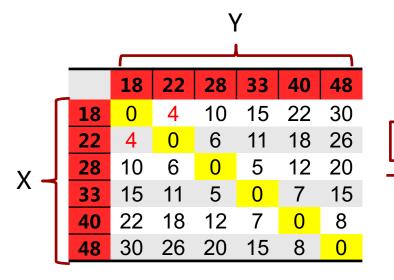
- Algorithm
 - 2. Merge the two closest clusters

New

cluster

18,22

■ 3. Update the proximity matrix



			Y	
	18,22	28	33	4
18,22	0	10	15	2
28	10	0	5	1
33	15	5	0	•
40	22	12	7	(
48	30	20	15	,

New cluster
28,33
\rightarrow

30

20

15

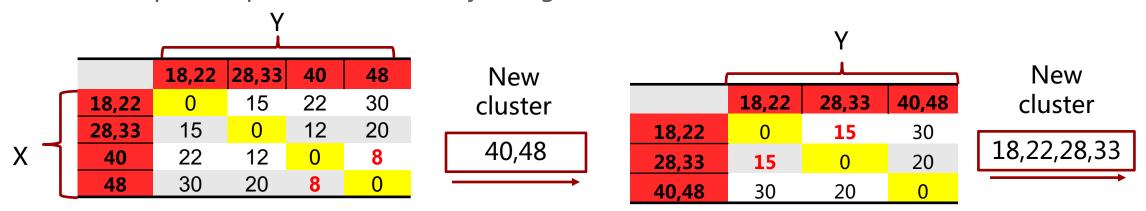
$$D(X_k, Y_k) = \max_{\substack{i=1,\dots,n\\j=1,\dots,m}} \left\{ dist(x_i, y_j) \right\}$$

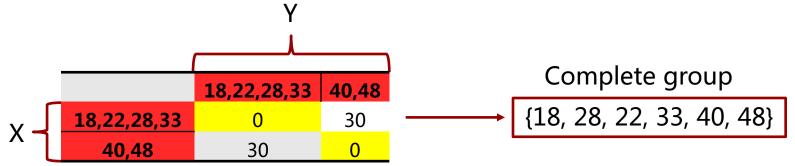
New Cluster =
$$\min_{1,...,k} \{D(X_k, Y_k)\}$$



max-merge clustering process

- Algorithm
 - Repeat steps 2 and 3 until only a single cluster remains





$$D(X_k, Y_k) = \max_{\substack{i=1,\dots,n\\j=1,\dots,m}} \left\{ dist(x_i, y_j) \right\}$$

New Cluster =
$$\min_{1,...,k} \{D(X_k, Y_k)\}$$



Thanks for your attention

