+

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
  - $p_k$and q_k are, respectively, the *k*th attribute of data objects **P** and **q**.

T5-Q1

Distance calculation --- L_p norm

- Let r = 1:

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

Also known as L_1 norm distance measure:

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

- **Input:** $\boldsymbol{X = (1, 0, 5)}$ **;** $\boldsymbol{Y = (2, 4, 9)}$
  - $\boldsymbol{L_1(X, Y) = ?}$

T5-Q1

Distance calculation --- L_p norm

- L_1 norm distance measure:

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

- **Input:** $\boldsymbol{X} = (\mathbf{1}, \mathbf{0}, \mathbf{5})$ **;** $\boldsymbol{Y} = (\mathbf{2}, \mathbf{4}, \mathbf{9})$
  - $\boldsymbol{L_1}(\boldsymbol{X}, \boldsymbol{Y}) = ?$
    - **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:

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

T5-Q1

Distance calculation --- L_p norm

- Let r = 2:

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

Also known as L_2 norm distance measure:

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

- **Input:** $\boldsymbol{X = (1, 0, 5)\ ;\ Y = (2, 4, 9)}$
  - $\boldsymbol{L_2(X, Y) = ?}$

T5-Q1

Distance calculation --- L_p norm

- L_2 norm distance measure:

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

- **Input:** $\boldsymbol{X} = (\mathbf{1}, \mathbf{0}, \mathbf{5})$; $\boldsymbol{Y} = (\mathbf{2}, \mathbf{4}, \mathbf{9})$
  - $\boldsymbol{L_2(X,Y)} = ?$
    - **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:

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

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

T5-Q1

Distance calculation --- L_p norm

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

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

- **Input:** $\boldsymbol{X} = (\mathbf{1}, \mathbf{0}, \mathbf{5})$; $\boldsymbol{Y} = (\mathbf{2}, \mathbf{4}, \mathbf{9})$
  - $\boldsymbol{L}_\infty(\boldsymbol{X}, \boldsymbol{Y}) = ?$

T5-Q1

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:** $\boldsymbol{X} = (\mathbf{1}, \mathbf{0}, \mathbf{5})$ **;** $\boldsymbol{Y} = (\mathbf{2}, \mathbf{4}, \mathbf{9})$
  - $L_\infty(\boldsymbol{X}, \boldsymbol{Y}) = ?$
    - **n = 3**
    - $X_1 = 1$**,** $X_2 = 0$**,** $X_3 = 5$
    - $Y_1 = 2$**,** $Y_2 = 4$**,** $Y_3 = 9$

- L_∞ norm distance measure:

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

T5-Q1

Distance calculation --- L_p norm

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

T5-Q2

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

T5-Q2

Distance calculation --- edit distance

- **Input:** $X =$ "***university***" ; $Y =$ "***unverstiy***"
  - **Calculate their edit distance**

- "university" → "unversity" **(Delete 'i') (Unit Cost = 1)**
- "unversity" → "unversty" **(Delete 'i') (Unit Cost = 1)**
- "unversty" → "unverstiy" **(Insert 'i') (Unit Cost = 1)**

- **The edit distance is 3**

T5-Q2

Distance calculation --- edit distance

- Input: $X =$ "***university***" ; $Y =$ "***unverstiy***“
  - 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**

T5-Q3

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.

T5-Q3

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

T5-Q3

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)**
      - $\boldsymbol{min\{dist(x, y) : x \in X, y \in Y\}}$
        - X and Y are the cluster sets

    - Similarity of two clusters is based on the **two most similar (closest) points** in the different clusters

T5-Q3

min-merge clustering process:

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

Y

X	18	22	28	33	40	48
18						
22						
28						
33						
40						
48						

+

T5-Q3

min-merge clustering process:

Y

X

	18	22	28	33	40	48
18	0	4	10	15	22	30
22	4	0	6	11	18	26
28	10	6	0	5	12	20
33	15	11	5	0	7	15
40	22	18	12	7	0	8
48	30	26	20	15	8	0

- 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
  - n is the number of points in cluster X_k
  - m is the number of points in cluster Y_k

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

T5-Q3

min-merge clustering process

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

Y

X	18	22	28	33	40	48
18	0	4	10	15	22	30
22	4	0	6	11	18	26
28	10	6	0	5	12	20
33	15	11	5	0	7	15
40	22	18	12	7	0	8
48	30	26	20	15	8	0

New cluster

18,22

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

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

T5-Q3

min-merge clustering process

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

X (rows) / Y (columns)

	18	22	28	33	40	48
18	0	4	10	15	22	30
22	4	0	6	11	18	26
28	10	6	0	5	12	20
33	15	11	5	0	7	15
40	22	18	12	7	0	8
48	30	26	20	15	8	0

New cluster

18,22

Y

	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

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

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

T5-Q3

min-merge clustering process

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

X \ Y	18,22	28,33	40	48
18,22	0	**6**	18	26
28,33	**6**	0	7	15
40	18	7	0	8
48	26	15	8	0

New cluster

18,22, 28,33

X \ Y	18,22,28,33	40	48
18,22,28,33	0	**7**	15
40	**7**	0	8
48	15	8	0

New cluster

18,22,28,33,40

X \ Y	18,22,28,33,40	48
18,22,28,33,40	0	8
48	8	0

Complete group

{18, 22, 28, 33, 40, 48}

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

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

T5-Q3

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)**
      - $max\{dist(x, y) : x \in X, y \in Y\}$
        - 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

T5-Q3

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

X		18	22	28	33	40	48
	18						
	22						
	28						
	33						
	40						
	48						

T5-Q3

max-merge clustering process:

Y

X

	18	22	28	33	40	48
18	0	4	10	15	22	30
22	4	0	6	11	18	26
28	10	6	0	5	12	20
33	15	11	5	0	7	15
40	22	18	12	7	0	8
48	30	26	20	15	8	0

- 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
  - n is the number of points in cluster X_k
  - m is the number of points in cluster Y_k

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

T5-Q3

max-merge clustering process

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

Y

X

	18	22	28	33	40	48
18	0	4	10	15	22	30
22	4	0	6	11	18	26
28	10	6	0	5	12	20
33	15	11	5	0	7	15
40	22	18	12	7	0	8
48	30	26	20	15	8	0

New cluster

18,22

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

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

T5-Q3

max-merge clustering process

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

X \ Y	18	22	28	33	40	48
18	0	4	10	15	22	30
22	4	0	6	11	18	26
28	10	6	0	5	12	20
33	15	11	5	0	7	15
40	22	18	12	7	0	8
48	30	26	20	15	8	0

New cluster

18,22

X \ Y	18,22	28	33	40	48
18,22	0	10	15	22	30
28	10	0	**5**	12	20
33	15	**5**	0	7	15
40	22	12	7	0	8
48	30	20	15	8	0

New cluster

28,33

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

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

T5-Q3

max-merge clustering process

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

	18,22	28,33	40	48
18,22	0	15	22	30
28,33	15	0	12	20
40	22	12	0	8
48	30	20	8	0

New cluster: 40,48

	18,22	28,33	40,48
18,22	0	15	30
28,33	15	0	20
40,48	30	20	0

New cluster: 18,22,28,33

	18,22,28,33	40,48
18,22,28,33	0	30
40,48	30	0

Complete group: {18, 28, 22, 33, 40, 48}

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

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

Thanks for your attention