香港中文大學
數學系
凝聚層次聚類練習

1 凝聚層次聚類簡介

層次聚類依賴成對距離矩陣作為核心聚類準則。特點：

- 使用完整的成對距離矩陣來指導聚類合併決策。
- 不需要預先指定聚類數量k 作為輸入；當所有點形成單一聚類或達到自訂的距離閾值時，聚類即告終止。
- 需要正式定義簇間距離$d(C_i, C_j)$，以量化兩個不相交聚類C_i 和C_j 之間的分離程度。

2 單一連結層次聚類

定義與核心概念

單一連結聚類（又稱最近鄰凝聚聚類）是一種自下而上（凝聚式）的層次聚類演算法。簇間距離被定義為屬於第一個聚類的任何點與屬於第二個聚類的任何點之間的最小歐幾里得距離。這種最小距離合併規則通常會產生細長、鏈狀的聚類，因為合併僅取決於跨越兩個群組的單對最接近的點。

聚類C_i, C_j 之間的簇間單一連結距離的正式定義：

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

其中兩個二維點$x = (x_1, x_2)$ 和$y = (y_1, y_2)$ 之間的歐幾里得距離為：

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

正式演算法步驟

迭代的單一連結聚類程序遵循五個固定階段：

1. **初始化**：每個原始數據點各自形成一個單例聚類。
2. **計算完整的簇間距離**：使用單一連結最小距離規則計算目前聚類之間的所有成對距離。
3. **合併最接近的配對**：找出簇間距離最小的一對聚類，並將它們組合成一個新的合併聚類。
4. **更新距離矩陣**：應用單一連結規則，重新計算新合併的聚類與所有未改變的剩餘聚類之間的距離。
5. **迭代至終止**：重複步驟2–4，直到所有數據點都屬於一個全局聚類，或者如果達到了目標聚類數量/ 距離截斷閾值，則提早停止。

數值例子計算

我們在一個包含六個二維樣本點的數據集上示範單一連結聚類。所有歐幾里得距離值均四捨五入至小數點後兩位，以保持精度一致。

數據點標籤	x 座標	y 座標
A	1	3
B	2	4
C	5	10
D	2	8
E	8	5
F	11	12

Table 1: 單一連結聚類的二維原始數據集

步驟1：初始狀態— 所有點作為單例聚類

首先計算所有成對的歐幾里得距離，並提供完整的代數展開以確保計算過程完全透明：

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. $d(E,F) = \sqrt{(8-11)^2 + (5-12)^2} = \sqrt{(-3)^2 + (-7)^2} = \sqrt{9+49} = \sqrt{58} \approx 7.62$

對稱的初始成對距離矩陣（最小值以紅色標示）：

步驟2：第一次合併操作

表2 中的全局最小距離為$d(A,B) \approx 1.41$。我們將單例聚類A 和B 合併，形成一個新的複合聚類AB。

聚類	$A(1,3)$	$B(2,4)$	$C(5,10)$	$D(2,8)$	$E(8,5)$	$F(11,12)$
$A(1,3)$	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: 初始歐幾里得距離矩陣（6 個單例聚類）

步驟3：重建距離矩陣（目前聚類：$AB,\ C,\ D,\ E,\ F$）

合併聚類AB 與外部聚類X 之間距離的單一連結規則：

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

逐步計算所有所需的簇間距離：

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

未改變的聚類C, D, E, F 之間的距離與初始矩陣保持一致。更新後的距離矩陣：

聚類	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: 合併$A + B \to AB$ 後的距離矩陣

步驟4：第二次合併操作

表3 中的新最小距離為$d(C,D) \approx 3.61$。將單例聚類C, D 合併成複合聚類CD。

步驟5：重建距離矩陣（目前聚類：$AB,\ CD,\ E,\ F$）

合併聚類CD 的單一連結規則：

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

計算：

$$\begin{aligned}
d(CD,AB) &= \min\left(d(C,AB)=6.71,\ d(D,AB)=4.00\right) = 4.00\\
d(CD,E) &= \min\left(d(C,E)=5.83,\ d(D,E)=6.71\right) = 5.83\\
d(CD,F) &= \min\left(d(C,F)=6.32,\ d(D,F)=9.85\right) = 6.32
\end{aligned}$$

更新後的矩陣（新的最小距離以紅色標示）：

聚類	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: 合併$C + D \to CD$ 後的距離矩陣

步驟6：第三次合併操作

表4 中的最小距離為$d(AB, CD) \approx 4.00$。將聚類AB 和CD 合併成大型複合聚類$ABCD$。

步驟7：重建距離矩陣（目前聚類：$ABCD,\ E,\ F$）

合併聚類$ABCD$ 的單一連結規則：

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

計算：

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

更新後的矩陣：

聚類	$ABCD$	$E(8,5)$	$F(11,12)$
$ABCD$	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: 合併$AB + CD \to ABCD$ 後的距離矩陣

步驟8：第四次合併操作

表5 中的最小距離為$d(ABCD, E) \approx 5.83$。將聚類$ABCD$ 和E 合併成複合聚類$ABCDE$。

步驟9：重建距離矩陣（目前聚類：$ABCDE,\ F$）

合併聚類$ABCDE$ 的單一連結規則：

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

最終的雙聚類距離矩陣：

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

Table 6: 合併$ABCD + E \to ABCDE$ 後的距離矩陣

步驟10：最終的全局合併操作

僅剩下兩個不相交的聚類：$ABCDE$ 和F，其簇間距離等於6.32。合併這兩個聚類以形成一個包含所有六個原始數據點的全局聚類$ABCDEF$。聚類過程終止。

合併順序摘要（單一連結層次結構）

所有合併事件及其對應截斷距離閾值的排序列表：

1. 在距離閾值1.41 處，合併單例聚類A 與單例聚類B
2. 在距離閾值3.61 處，合併單例聚類C 與單例聚類D
3. 在距離閾值4.00 處，合併複合聚類AB 與複合聚類CD
4. 在距離閾值5.83 處，合併複合聚類$ABCD$ 與單例聚類E
5. 在距離閾值6.32 處，合併複合聚類$ABCDE$ 與單例聚類F