Data mining

Dr. Ohidujjaman Tuhin

2.

Area	Perimeter	Compactness
15.26	18.84	0.871
14.88	14.57	0.8811
14.29	14.09	0.905
13.84	13.94	0.8955
16.14	14.99	0.9034
14.38	14.21	0.8951

Use k-mean standard algorithm to determine three clusters from the above data set.

Soln:

Let Area = x
Perimeter = y
Compactness = z

Assume, $A \equiv (x, y, z) \equiv (15.26, 14.84, 0.871)$

$B \equiv (14.88, 14.57, 0.8811)$

$C \equiv (14.29, 14.09, 0.905)$

$D \equiv (13.84, 13.94, 0.8955)$

$E \equiv (16.14, 14.99, 0.9034)$

$F \equiv (14.38, 14.21, 0.8951)$

Let, centroid selection

class-I: $B \equiv (14.88, 14.57, 0.8811)$

class-II: $D \equiv (13.84, 13.94, 0.8955)$

Class-III: $F \equiv (14.38, 14.21, 0.8951)$

Iteration I:

Data	class-I: B (14.88, 14.57, 0.8811)	class-II: D (13.84, 13.94, .8955)	class-III: F (14.38, 14.21, .895)	class
A (15.26, 14.84, 0.871)	0.466	1.68	1.08	I
B (14.88, 14.57, 0.8811)	0	1.216	0.616	I
C (14.29, 14.09, 0.905)	0.76	0.47	0.15	III
D (13.84, 13.94, .8955)	1.216	0	0.60	II
E (16.14, 14.99, .9034)	1.32	2.52	1.92	I
F (14.38, 14.21, 0.8951)	0.616	0.60	0	III

∴ class-I: A B E

(15.26, 14.84, .871), (14.88, 14.57, .8811), (16.14, 14.99, 0.9034)

$$\therefore \text{Centroid} = \left(\frac{15.26+14.88+16.14}{3}, \frac{14.84+14.57+14.99}{3}, \frac{0.871+0.8811+0.9034}{3} \right)$$

$$= (15.43, 14.8, 0.885)$$

class II: D ≡ (13.84, 13.94, 0.8955)

class III: C F

(14.29, 14.09, 0.905), (14.38, 14.21, 0.8951)

$$\therefore \text{Centroid} = \left(\frac{14.29 + 14.38}{2}, \frac{14.09 + 14.21}{2}, \frac{0.905 + 0.8951}{2} \right)$$

$$= (14.335, 14.15, 0.90)$$

Iteration II:

	Class I: (15.43, 14.8, 0.885)	Class II: (13.84, 13.94, 0.8955)	Class III: (14.335, 14.15, 0.90)	Class
A (15.26, 14.84, 0.871)	0.175	1.68	1.15	I
B (14.88, 14.57, 0.8811)	0.596	1.216	0.68	I
C (14.29, 14.09, 0.905)	1.34	0.47	0.075	III
D (13.84, 13.94, 0.8955)	1.80	0	0.53	II
E (16.14, 14.99, 0.9034)	0.73	2.52	1.99	I
F (14.38, 14.21, 0.8951)	1.20	0.60	0.075	III

$\therefore$ Class I: A, B, E ; centroid = (15.43, 14.8, 0.885)

Class II: D ; centroid = (13.84, 13.94, 0.8955)

Class III: C, F ; centroid = (14.335, 14.15, 0.9)

$\therefore$ Cluster I: A (15.26, 14.84, 0.871), B (14.88, 14.57, 0.8811), E (16.14, 14.99, 0.9034)

Cluster II: D (13.84, 13.94, 0.8955)

Cluster III: C (14.29, 14.09, 0.905), F (14.38, 14.21, 0.8951)

3.

Student	Score	Height
Julie	11	5.5
John	11	6
Ryan	13	6.2
Bob	18	4.8
Prince	15	5.8
Mathew	10	6.1

Apply Single linkage and complete linkage clustering algorithm to form hierarchical clustering from the above data set and also draw dendogram.

Sol^n: Let Score = x, Height = y

$\therefore$ Julie $\equiv (x, y) = (11, 5.5)$

$\therefore$ Jhon $\equiv (11, 6)$ Bob $\equiv (18, 4.8)$

Ryan $\equiv (13, 6.2)$ Prince $\equiv (15, 5.8)$

Mathew $\equiv (10, 6.1)$,

P.T.O

Input distance matrix:

student	Julie (11, 5.5)	Jhon (11, 6)	Ryan (13, 6.2)	Bob (18, 4.8)	Prince (15, 5.8)	Mathew (10, 6.1)
Julie (11, 5.5)	0	0.5	2.1	7.03	4.01	1.17
Jhon (11, 6)		0	2.01	7.1	4.0	1.00
Ryan (13, 6.2)			0	5.1	2.04	3.00
Bob (18, 4.8)				0	3.16	8.1
Prince (15, 5.8)					0	5.01
Mathew (10, 6.1)						0

∴ Julie and Jhon are combind.

Iteration 1:

student	Julie/Jhon (11, 5.5)/(11, 6)	Ryan (13, 6.2)	Bob (18, 4.8)	Prince (15, 5.8)	Mathew (10, 6.1)
Julie/Jhon (11, 5.5)/(11, 6)	0	2.01	7.03	4.0	1.0
Ryan (13, 6.2)		0	5.1	2.04	3.0
Bob (18, 4.8)			0	3.16	8.1
Prince (15, 5.8)				0	5.01
Mathew					0

$distance_1 = \min(Julie/Jhon, Ryan)$

$= \min(2.1, 2.01)$

$= 2.01$

$d_2 = \min(Julie/Jhon, Bob)$

$= \min(7.03, 7.1)$

$= 7.03$

$d_3 = \min(Julie/Jhon, Prince)$

$= \min(4.01, 4.0) = 4.0$

$d_4 = \min(\text{Julie/Jhon}, \text{Mathew})$

$= \min(1.67, 1.0)$

$= 1.0$

∴ Julie/Jhon and Mathew are combind.

Iteration II:

Student	Julie/Jhon/Mathew	Ryan	Bob	Prince
Julie/Jhon/Math	0	2.01	7.03	4.0
Ryan		0	5.1	2.04
Bob			0	3.16
Prince				0

$d_1 = \min(\text{Julie/Jhon/Mathew}, \text{Ryan})$

$= \min(2.01, 3.00)$

$= 2.01$

$d_2 = \min($

$d_2 = \min(\text{Julie/Jhon/Mathew}, \text{Bob})$

$= \min(7.03, 8.1)$

$= 7.03$

$d_3 = \min(\text{Julie/Jhon/Mathew}, \text{Prince})$

$= \min(4.0, 5.01) = 4.0$

∴ Julie/Jhon/Mathew and Ryan are combind.

Iteration III :

student	Julie/Jhon/Mathew/Ryan	Bob	Prince
Julie/Jhon/Mathew/Ryan	0	5.1	2.04
Bob		0	3.16
Prince			0

$d_1 = \min$ (Julie/Jhon/Mathew/Ryan , Bob)

$= \min (7.03, 5.1)$

$= 5.1$

$d_2 = \min$ (Julie/Jhon/Mathew/Ryan , Prince)

$= \min (4.0, 2.04)$

$= 2.04$

∴ Julie/Jhon/Mathew/Ryan and Prince are combind.

Iteration IV :

student	Julie/Jhon/Mat/R/P	Bob
Julie/Jhon/Mat/R/P	0	3.16
Bob		0

$d_1 = \min$ (Julie/Jhon/Math/Ray/Pri , Bob)

$= \min (5.1, 3.16)$

∴ Ju|Jh|Math|Ry|P and Bob are combind.

Iteration V:

student	Ju|Jh|Ma|R|P|B
Ju|Jh|Mat|R|P|B	0

∴ The Dendogram of the above solution we have as follows:—

Data set: Julie Jhon Ryan Bob prince Mathew

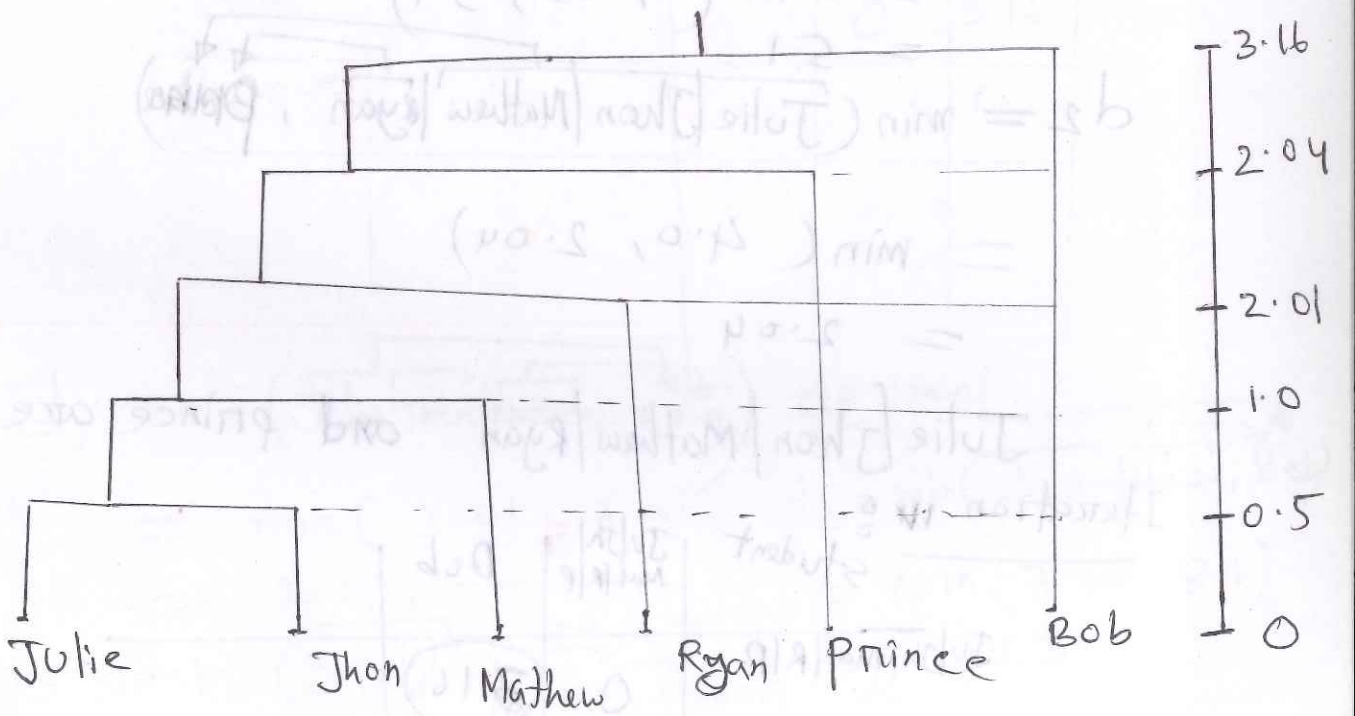

figure: Dendogram

Note: Complete linkage is similar to single linkage, except choosing value of two data. That means, need to select maximum data value rather than minimum
e.g. max(A|B, C) = max( )