Самостійна робота 3 з дисципліни **Data mining**

Студент - Сирота Сергій ТТП-42 (Serhii Syrota TTP-42)

Викладач - Криволап Андрій (PhD) (Kryvolap Andriy)

Task description:

```
Для одного з варіантів провести кластеризацію.
Використати методи: Ієрархічні (Single-link, Complete-link, Average-link),  k-середніх(k=3), k-медіан(k=3)
```

Task C

Single-link

Вхідні дані

__	X	Y
A	1	5
B	1	1
C	2	1
D	2	6
E	3	4
F	3	3
G	4	6
H	5	3
I	6	5
J	6	1

За дистанцію будемо брати Евклідову відстань. В цьому алгоритмі за відстань братимемо найменшу відстань між кластерами.

Рахуємо відстані

__	A	B	C	D	E	F	G	H	I	J
A	0	4	4.123	1.414	2.236	2.828	3.162	4.472	5	6.403
B		0	1	5.099	3.605	2.828	5.83	4.472	6.403	5
C			0	5	3.162	2.236	5.385	3.606	5.657	4
D				0	2.236	3.162	2	4.243	4.123	6.403
E					0	1	2.236	2.236	3.162	4.243
F						0	3.162	2	3.606	3.606

G							0	3.162	2.236	5.385
H								0	2.236	2.236
I									0	4
J										0

Матриця відстаней буде симетричною, тому значення під діагоналлю можемо не заповнювати.

Найменшу відстань мають точки B і C та E і F, тому зливаємо їх в кластер і знову знаходимо матрицю відстаней. Повторюємо цей алгоритм поки не залишиться один кластер.

__	A	B, C	D	E, F	G	H	I	J
A	0	4	1.414	2.236	3.162	4.472	5	6.403
B, C		0	5	2.236	5.385	3.606	5.657	4
D			0	2.236	2	4.243	4.123	6.403
E, F				0	2.236	2	3.162	3.606
G					0	3.162	2.236	5.385
H						0	2.236	2.236
I							0	4
J								0

Найменша відстань між A та D.

__	A, D	B, C	E, F	G	H	I	J
A, D	0	4	2.236	2	4.243	4.123	6.403
B, C		0	2.236	5.385	3.606	5.657	4
E, F			0	2.236	2	3.162	3.606
G				0	3.162	2.236	5.385
H					0	2.236	2.236
I						0	4
J							0

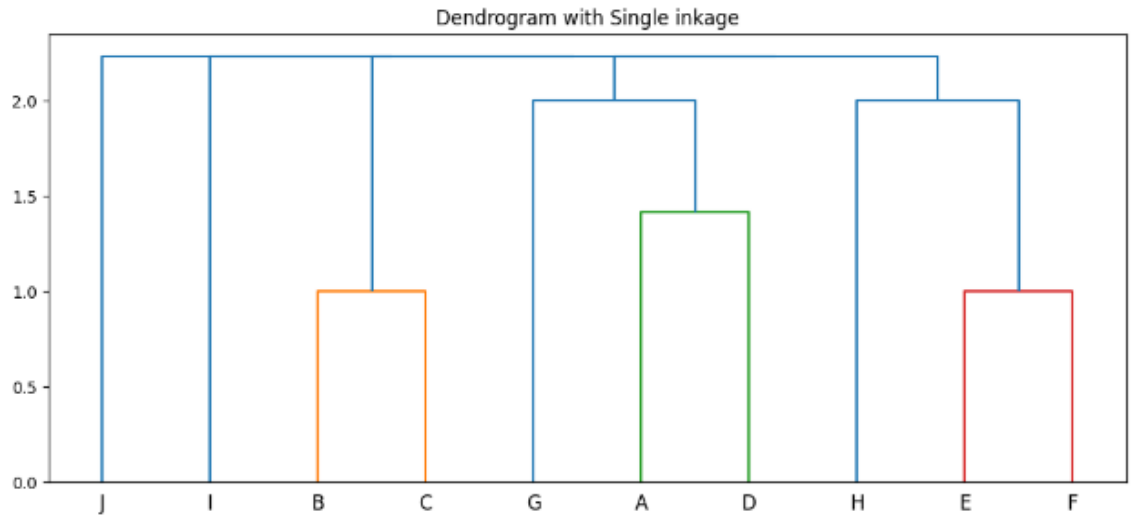
Найменша відстань між (A, D) та G, (E, F) та H

__	(A, D), G	B, C	(E, F), H	I	J
(A, D), G	0	4	2.236	2.236	5.385
B, C		0	2.236	5.657	4

(E, F), H			0	2.236	2.236
I				0	4
J					0

Найменша відстань між ((A, D), G) та ((E, F), H), ((A, D), G) та I, (B, C) та I, ((E, F), H) та I, ((E, F), H) та J

Бачимо, що тепер це все зливається в один кластер. Результат показано на наступній дендограмі:



Complete-link

Вхідні дані

__	X	Y
A	1	5
B	1	1
C	2	1
D	2	6
E	3	4
F	3	3
G	4	6
H	5	3
I	6	5
J	6	1

За дистанцію будемо брати Евклідову відстань. В цьому методі будемо брати найбільшу відстань між кластерами.

Рахуємо відстані

__	A	B	C	D	E	F	G	H	I	J
A	0	4	4.123	1.414	2.236	2.828	3.162	4.472	5	6.403
B		0	1	5.099	3.605	2.828	5.83	4.472	6.403	5
C			0	5	3.162	2.236	5.385	3.606	5.657	4
D				0	2.236	3.162	2	4.243	4.123	6.403
E					0	1	2.236	2.236	3.162	4.243
F						0	3.162	2	3.606	3.606
G							0	3.162	2.236	5.385
H								0	2.236	2.236
I									0	4
J										0

Найменші відстані мають B, C та E, F.

__	A	B, C	D	E, F	G	H	I	J
A	0	4.123	1.414	2.828	3.162	4.472	5	6.403
B, C		0	5	3.605	5.83	3.472	6.403	5
D			0	3.162	2	4.243	4.123	6.403
E, F				0	2.236	2	3.162	3.606
G					0	3.162	2.236	5.385
H						0	2.236	2.236
I							0	4
J								0

Найменша відстань між A та D.

__	A, D	B, C	E, F	G	H	I	J
A, D	0	5	3.162	3.162	4.472	5	6.403
B, C		0	3.605	5.83	3.472	6.403	5
E, F			0	2.236	2	3.162	3.606
G				0	3.162	2.236	5.385
H					0	2.236	2.236

I						0	4
J							0

Найменша відстань між (E, F) та H.

__	A, D	B, C	(E, F), H	G	I	J
A, D	0	5	4.472	3.162	5	6.403
B, C		0	3.605	5.83	6.403	5
(E, F), H			0	3.162	3.162	3.606
G				0	2.236	5.385
I					0	2.236
J						0

Найменша відстань між G, I, J

__	A, D	B, C	(E, F), H	G, I, J
A, D	0	5	4.472	6.403
B, C		0	3.605	6.403
(E, F), H			0	3.606
G, I, J				0

Найменша відстань між (B, C), ((E, F), H), (G, I, J)

__	A, D	(B, C), ((E, F), H), (G, I, J)
A, D	0	6.403
(B, C), ((E, F), H), (G, I, J)		0

Зливаємо 2 останні кластери в один та отримуємо ((A, D), ((B, C), ((E, F), H), (G, I, J)))

Average-link

Вхідні дані

__	X	Y
A	1	5
B	1	1
C	2	1
D	2	6

E	3	4
F	3	3
G	4	6
H	5	3
I	6	5
J	6	1

За дистанцію будемо брати Евклідову відстань. В цьому методі будемо брати середню арифметичну відстань між кластерами.

Рахуємо відстані

__	A	B	C	D	E	F	G	H	I	J
A	0	4	4.123	1.414	2.236	2.828	3.162	4.472	5	6.403
B		0	1	5.099	3.605	2.828	5.83	4.472	6.403	5
C			0	5	3.162	2.236	5.385	3.606	5.657	4
D				0	2.236	3.162	2	4.243	4.123	6.403
E					0	1	2.236	2.236	3.162	4.243
F						0	3.162	2	3.606	3.606
G							0	3.162	2.236	5.385
H								0	2.236	2.236
I									0	4
J										0

Найменші відстані мають B, C та E, F.

__	A	B, C	D	E, F	G	H	I	J
A	0	4.062	1.414	2.532	3.162	4.472	5	6.403
B, C		0	5.05	2.958	5.608	4.039	6.03	4.5
D			0	2.699	2	4.243	4.123	6.403
E, F				0	2.699	2.118	3.384	3.925
G					0	3.162	2.236	5.385
H						0	2.236	2.236
I							0	4

J								0

Найменші відстані мають A і D.

__	A, D	B, C	E, F	G	H	I	J
A, D	0	4.556	2.616	2.581	4.358	4.562	6.403
B, C		0	2.958	5.608	4.039	6.03	4.5
E, F			0	2.699	2.118	3.384	3.925
G				0	3.162	2.236	5.385
H					0	2.236	2.236
I						0	4
J							0

Найменші відстані мають (E, F) та H

__	A, D	B, C	(E, F), H	G	I	J
A, D	0	4.556	3.487	2.581	4.562	6.403
B, C		0	3.499	5.608	6.03	4.5
(E, F), H			0	2.931	2.81	3.081
G				0	2.236	5.385
I					0	4
J						0

Найменші відстані мають G та I

__	A, D	B, C	(E, F), H	G, I	J
A, D	0	4.556	3.487	3.572	6.403
B, C		0	3.499	5.819	4.5
(E, F), H			0	2.871	3.081
G, I				0	4.693
J					0

Найменші відстані мають (E, F), H та (G, I)

__	A, D	B, C	((E, F), H), (G, I)	J
A, D	0	4.556	3.53	6.403

B, C		0	4.659	4.5
((E, F), H), (G, I)			0	3.887
J				0

Найменші відстані мають (A, D) та (((E, F), H), (G, I))

__	**(A, D), (((E, F), H), (G, I))**	**B, C**	**J**
(A, D), (((E, F), H), (G, I))	0	4.608	5.145
B, C		0	4.5
J			0

Найменші відстані мають (B, C) та J. Після їхнього злиття залишається 2 кластери які ми зливаємо. Отримуємо результат:

(((A, D), (((E, F), H), (G, I))), ((B, C), J))

k-means

За умовою параметр k = 3.

Вхідні дані

__	**X**	**Y**
A	1	5
B	1	1
C	2	1
D	2	6
E	3	4
F	3	3
G	4	6
H	5	3
I	6	5
J	6	1

За дистанцію будемо брати Евклідову відстань. Алгоритм наступний:

1. Беремо початкові центри кожного кластеру, нехай ними будуть точки A, B, C.

2. Кожну точку присвоюємо до кластеру до центру якого вона є найближче.

3. Перераховуємо центри кластерів як середнє арифметичних всіх точок даного кластеру.

4. Виконуємо пункт 2 для нових центроїдів.

5. Якщо кластери не змінилися, алгоритм завершуємо.

Отже, на спочатку маємо кластери з центроїдами (A, B, C). В наступній таблиці показано 3 кластери та точки які належать до кожного з кластерів.

1 (Центр: (1, 5))	**A, D, E, G, I**
2 (Центр: (1, 1))	B
3 (Центр: (2, 1))	C, F, H, J

Будуємо табличку з новими центроїдами та точками, що належать до цих кластерів:

1 (Центр: (3.2, 5.2))	**A, D, E, G, I**
2 (Центр: (1, 1))	B, C
3 (Центр: (4, 2))	F, H, J

Будуємо табличку з новими центроїдами та точками, що належать до цих кластерів:

1 (Центр: (3.2, 5.2))	**A, D, E, G, I**
2 (Центр: (1.5, 1))	B, C
3 (Центр: (4.7, 2.3))	F, H, J

Кластери не змінилися, тому алгоритм завершено. Отримали наступні кластери:

1 – A, D, E, G, I

2 – B, C

3 – F, H, J

k-medians

За умовою параметр k = 3.

Вхідні дані

__	X	Y
A	1	5
B	1	1
C	2	1
D	2	6
E	3	4
F	3	3
G	4	6

H	5	3
I	6	5
J	6	1

За дистанцію будемо брати Евклідову відстань. Алгоритм наступний:

1. Беремо початкові центри кожного кластеру, нехай ними будуть точки A, B, C.

2. Кожну точку присвоюємо до кластеру до центру якого вона є найближче.

3. Перераховуємо центри кластерів як медіану всіх точок даного кластеру.

4. Виконуємо пункт 2 для нових центроїдів.

5. Якщо кластери не змінилися, алгоритм завершуємо.

Отже, на спочатку маємо кластери з центроїдами (A, B, C). В наступній таблиці показано 3 кластери та точки які належать до кожного з кластерів.

1 (Центр: (1, 5))	**A, D, E, G, I**
2 (Центр: (1, 1))	B
3 (Центр: (2, 1))	C, F, H, J

Будуємо табличку з новими центроїдами та точками, що належать до цих кластерів:

1 (Центр: (3, 5))	**A, D, E, G, I**
2 (Центр: (1, 1))	B, C
3 (Центр: (4, 2))	F, H, J

Будуємо табличку з новими центроїдами та точками, що належать до цих кластерів:

1 (Центр: (3, 5)) __	**A, D, E, G**
2 (Центр: (1.5, 1))	B, C
3 (Центр: (5, 3))	F, H, I, J

Будуємо табличку з новими центроїдами та точками, що належать до цих кластерів:

1 (Центр: (2.5, 5.5))	**A, D, E, G**
2 (Центр: (1.5, 1))	B, C
3 (Центр: (5.5, 3))	F, H, I, J

Кластери не змінилися, тому алгоритм завершено. Отримали наступні кластери:

1 – A, D, E, G

2 – B, C

3 – F, H, I, J

Як бачимо відповідь відрізняється від тієї, що вийшла в результаті алгоритму k-means