حل تمرین فصل ۸ داده‌کاوی

مصطفی سبیلو (۴۰۱۰۸۲۵۴)

مهدی محمدی (۴۰۱۱۲۰۹۴)

۱) با فرض داشتن دو نمونه به صورت (۲۲, ۱, ۴۲, ۱۰) و (۲۰, ۰, ۳۶, ۸) موارد زیر را انجام دهید.

الف) محاسبه فاصله اقلیدسی

O1 = (10, 42, 1, 22)

O2 = (8, 36, 0, 20)

$$Distance(O_1, O_2) = \sqrt{\sum_{k=1}^{m}(x_{1k} - x_{2k})^2} = \sqrt{(10-8)^2 + (42-36)^2 + (1-0)^2 + (22-20)^2}$$
$$= \sqrt{4 + 36 + 1 + 4} = \sqrt{45} \approx 6.7082$$

ب) محاسبه فاصله منهتن

$$Distance(O_1, O_2) = \sum_{k=1}^{m}|x_{1k} - x_{2k}| = |10-8| + |42-36| + |1-0| + |22-20| = 2 + 6 + 1 + 2$$
$$= 11$$

پ) محاسبه فاصله مینکوفسکی با مقدار p = 3

$$Distance(O_1, O_2) = \sqrt[3]{\sum_{k=1}^{m}(|x_{1k} - x_{2k}|)^3}$$
$$= \sqrt[3]{(|10-8|)^3 + (|42-36|)^3 + (|1-0|)^3 + (|22-20|)^3} = \sqrt[3]{8 + 216 + 1 + 8}$$
$$= \sqrt[3]{233} \approx 6.1534$$

ت) محاسبه معیار چبیشف

$$Distance(O_1, O_2) = Max_{k=1,2,3,4}|x_{1k} - x_{2k}| = Max(|10-8|, |42-36|, |1-0|, |22-20|) = 6$$

۲) با فرض داشتن داده‌های زیر برای هشت نقطه در قالب سه خوشه A و B و C و A1 و B1 و C1 به‌عنوان مراکز خوشه‌ها با به کارگیری معیار فاصله اقلیدسی مراحل اجرای الگوریتم k-means را روی این داده‌ها تکمیل نمایید.

A1(2, 10), A2(2, 5), A3(8, 4), B1(5, 8), B2(7, 5), B3(6, 4), C1(1, 2), C2(4, 9)

فاصله هر نمونه از مرکزهای اولیه A1 و B1 و C1 به دست می‌آوریم. هر نمونه به هر مرکزی نزدیک‌تر بود، به آن خوشه تعلق می‌گیرد.

نمونه	فاصله تا A1	فاصله تا B1	فاصله تا C1	کلاس
A1	0	3.6055	8.0622	A
A2	5	4.2426	3.1622	C
A3	8.4852	5	7.2801	B
B1	3.6055	0	7.2111	B
B2	7.0710	3.6055	6.7082	B
B3	7.2111	4.1231	5.3851	B
C1	8.0622	7.2111	0	C
C2	2.2360	1.4142	7.6157	B

میانگین نمونه‌ها را در هر خوشه محاسبه می‌کنیم و به عنوان مراکز جدید خوشه‌ها در نظر می‌گیریم.

$$Centroid_A = Mean_A = \left(\frac{2}{1}, \frac{10}{1}\right) = (2,10)$$

$$Centroid_B = Mean_B = \left(\frac{8+5+7+6+4}{5}, \frac{4+8+5+4+9}{5}\right) = (6,6)$$

$$Centroid_C = Mean_C = \left(\frac{2+1}{2}, \frac{5+2}{2}\right) = (1.5,3.5)$$

فاصله همه نمونه‌ها را با مراکز جدید محاسبه می‌کنیم و هر نمونه به هر مرکزی نزدیک‌تر بود، به آن خوشه تعلق می‌گیرد.

نمونه	فاصله تا مرکز خوشه A	فاصله تا مرکز خوشه B	فاصله تا مرکز خوشه C	کلاس
A1	0	5.6568	6.5192	A
A2	5	4.1231	1.5811	C
A3	8.4852	2.8284	6.5192	B
B1	3.60555	2.2360	5.7008	B
B2	7.0710	1.4142	5.7008	B
B3	7.2111	2	4.5276	B
C1	8.0622	6.4031	1.5811	C
C2	2.2360	3.6055	6.0415	A

میانگین نمونه‌ها را در هر خوشه محاسبه می‌کنیم و به عنوان مراکز جدید خوشه‌ها در نظر می‌گیریم.

$$Centroid_A = Mean_A = \left(\frac{2+4}{2}, \frac{10+9}{2}\right) = (3,9.5)$$

$$Centroid_B = Mean_B = \left(\frac{8+5+7+6}{4}, \frac{4+8+5+4}{4}\right) = (6.5,5.25)$$

$$Centroid_C = Mean_C = \left(\frac{2+1}{2}, \frac{5+2}{2}\right) = (1.5,3.5)$$

فاصله همه نمونه‌ها را با مراکز جدید محاسبه می‌کنیم و هر نمونه به هر مرکزی نزدیک‌تر بود، به آن خوشه تعلق می‌گیرد.

نمونه	فاصله تا مرکز خوشه A	فاصله تا مرکز خوشه B	فاصله تا مرکز خوشه C	کلاس
A1	1.1180	6.5431	6.5192	A
A2	4.6097	4.5069	1.5811	C
A3	7.4330	1.9525	6.5192	B
B1	2.5	3.1324	5.7008	A
B2	6.0207	0.5590	5.7008	B
B3	6.2649	1.3462	4.5276	B
C1	7.7620	6.3884	1.5811	C
C2	1.1180	4.5069	6.0415	A

میانگین نمونه‌ها را در هر خوشه محاسبه می‌کنیم و به عنوان مراکز جدید خوشه‌ها در نظر می‌گیریم.

$$Centroid_A = Mean_A = \left(\frac{2+5+4}{3}, \frac{10+8+9}{3}\right) = (3.6666,9)$$

$$Centroid_B = Mean_B = \left(\frac{8+7+6}{3}, \frac{4+5+4}{3}\right) = (7,4.3333)$$

$$Centroid_C = Mean_C = \left(\frac{2+1}{2}, \frac{5+2}{2}\right) = (1.5,3.5)$$

فاصله همه نمونه‌ها را با مراکز جدید محاسبه می‌کنیم و هر نمونه به هر مرکزی نزدیک‌تر بود، به آن خوشه تعلق می‌گیرد.

نمونه	فاصله تا مرکز خوشه A	فاصله تا مرکز خوشه B	فاصله تا مرکز خوشه C	کلاس
A1	1.9435	7.5572	6.5192	A
A2	4.3333	5.0442	1.5811	C
A3	6.6165	1.0540	6.5192	B
B1	1.6667	4.1766	5.7008	A
B2	5.2068	0.6666	5.7008	B
B3	5.5176	1.0540	4.5276	B
C1	7.4907	6.4377	1.5811	C
C2	0.3334	5.5478	6.0415	A

با توجه به اینکه تغییری در خوشه‌ها اتفاق نیفتاد، بنابراین در پایان نمونه‌ها به صورت زیر خوشه‌بندی می‌شوند:

A: {A1, B1, C2}

B: {A3, B2, B3}

C: {A2, C1}

۳) پس از انجام اصلاحات زیر در جدول ۸-۶ اسلاید ۴۱، الگوریتم مزبور را مجددا اجرا و کلیه مراحل را بیان نمایید.

الف) تغییر مقادیر O4 به X=2 و Y=2

ب) افزودن O6 با مقادیر X=3 و Y=2

Attribute	O1	O2	O3	O4	O5	O6
X	1	5	6	2	4	3
Y	2	2	1	2	1	2

شرط الگوریتم این است که در هر مرحله فقط یک خوشه با خوشه دیگر ادغام شود.

در مرحله اول هر نمونه را به عنوان یک خوشه در نظر می‌گیریم.

ماتریس تشابه براساس فاصله اقلیدسی

	O1	O2	O3	O4	O5	O6
O1	0					
O2	4	0				
O3	5.099	1.414	0			
O4	1	3	4.123	0		
O5	3.162	1.414	2	2.236	0	
O6	2	2	3.162	1	1.414	0

دو خوشه O1 و O4 همچنین دو خوشه O4 و O6 کمترین مقدار را در ماتریس شباهت دارند، با توجه به شرط الگوریتم به صورت تصادفی دو خوشه O1 و O4 را برای ادغام انتخاب می‌کنیم.

	{O1, O4}	O2	O3	O5	O6
{O1, O4}	0				
O2	3	0			
O3	4.123	1.414	0		
O5	2.236	1.414	2	0	
O6	1	2	3.162	1.414	0

دو خوشه {O1, O4} و O6 کمترین مقدار را در ماتریس شباهت دارند، بنابراین آن‌ها را باهم ادغام انتخاب می‌کنیم.

	{O1, O4, O6}	O2	O3	O5
{O1, O4, O6}	0			
O2	2	0		
O3	3.162	1.414	0	
O5	1.414	1.414	2	0

خوشه‌های {O1, O4, O6} و O5، خوشه‌های O2 و O3 و خوشه‌های O2 و O5 کمترین مقدار را در ماتریس شباهت دارند، با توجه به شرط الگوریتم به صورت تصادفی دو خوشه O2 و O3 را برای ادغام انتخاب می‌کنیم.

	{O1, O4, O6}	{O2, O3}	O5
{O1, O4, O6}	0		
{O2, O3}	2	0	
O5	1.414	1.414	0

خوشه‌های {O1, O4, O6} و O5 و خوشه‌های {O2, O3} و O5 کمترین مقدار را در ماتریس شباهت دارند، با توجه به شرط الگوریتم به صورت تصادفی دو خوشه {O2, O3} و O5 را برای ادغام انتخاب می‌کنیم.

	{O1, O4, O6}	{O2, O3, O5}
{O1, O4, O6}	0	
{O2, O3, O5}	1.414	0

در مرحله آخر دو خوشه باقی مانده را نیز باهم ادغام می‌کنیم.

۴) الگوریتم DIANA را روی داده‌های سوال شماره ۳ بالا انجام دهید.

Attribute	O1	O2	O3	O4	O5	O6
X	1	5	6	2	4	3
Y	2	2	1	2	1	2

همه نمونه‌ها را در خوشه C قرار می‌دهیم.

دو خوشه A و B که هر خالی هستند، ایجاد می‌کنیم.

تمام نمونه‌ها را از خوشه C به خوشه A منتقل می‌کنیم.

یک نمونه از خوشه A انتخاب کرده و فاصله آن نمونه تا نمونه‌های دیگر را محاسبه می‌کنیم و باهم جمع می‌کنیم و بر تعداد یکی کمتر از تعداد نمونه‌ها تقسیم می‌کنیم. نمونه‌ای که عدد به دست آمده برای آن از نمونه‌های دیگر بزرگ‌تر بود را به خوشه B منتقل می‌کنیم.

	O1	O2	O3	O4	O5	O6	مجموع
O1	0	4	5.099	1	3.162	2	15.2612
O2	4	0	1.414	3	1.414	2	11.8284
O3	5.099	1.414	0	4.123	2	3.162	15.7986
O4	1	3	4.123	0	2.236	1	11.3591
O5	3.162	1.414	2	2.236	0	1.414	10.2267
O6	2	2	3.162	1	1.414	0	9.5764

$$D(O_i, A - \{O_i\}) = \frac{1}{|A| - 1} \times \sum_{O_j \in A, O_j \neq O_i} d(O_i, O_j)$$

$D(O_1, A - \{O_1\}) = 3.0522$

$D(O_2, A - \{O_2\}) = 2.3656$

$D(O_3, A - \{O_3\}) = 3.1597$

$D(O_4, A - \{O_4\}) = 2.2718$

$D(O_5, A - \{O_5\}) = 2.0453$

$D(O_6, A - \{O_6\}) = 1.9152$

چون مقدار محاسبه شده برای نمونه O3 از بقیه نمونه‌ها بیشتر است، بنابراین نمونه O3 را به خوشه B منتقل می‌کنیم.