КЛАСТЕРНЫЙ АНАЛИЗ

Задача 1

Дано:

Таблица 1

№ объекта (*i*)	1	2	3	4	5	6
$x_i(1)$	5	6	5	10	11	10
$x_i(2)$	10	12	13	9	9	7

а) *Провести классификацию, используя обычное евклидово расстояние и метод «ближайшего соседа»*

Согласно обычной евклидовой метрике расстояние между наблюдениями равно

$$d_{ij} = \sqrt{\sum_{k=1}^{m} (x_{ik} - x_{jk})^2}$$

По формуле находим расстояния между всеми шестью наблюдениями и строим матрицу расстояний

R1=

	1	2	3	4	5	6
1	0	2,236068	3	5,09902	6,082763	5,830952
2	2,236068	0	1,414214	5	5,830952	6,403124
3	3	1,414214	0	6,403124	7,211103	7,81025
4	5,09902	5	6,403124	0	1	2
5	6,082763	5,830952	7,211103	1	0	2,236068
6	5,830952	6,403124	7,81025	2	2,236068	0

Из матрицы расстояний следует, что объекты 4 и 5 наиболее близки $\rho_{4,5}=1$ и поэтому объединим их в один кластер. После объединения объектов имеем пять кластеров: S_1, S_2, S_3, $S_{(4,5)}$, S_6.

Расстояние между кластерами будем находить по принципу “ближайшего соседа”, воспользовавшись формулой пересчета. Так расстояние между кластером S_1 и кластером $S_{(4,5)}$ будет рассчитываться по формуле:

$$\rho_{1,(4,5)} = \rho(S_1, S_{(4,5)}) = \frac{1}{2}\rho_{1,4} + \frac{1}{2}\rho_{1,5} - \frac{1}{2}\left|\rho_{1,4} - \rho_{1,5}\right|$$

Проводя расчеты, получим новую матрицу расстояний.

R2=

	1	2	3	4,5	6
1	0	2,236068	3	5,09902	5,830952
2	2,236068	0	1,414214	5	6,403124
3	3	1,414214	0	6,403124	7,81025
4,5	5,09902	5	6,403124	0	2
6	5,830952	6,403124	7,81025	2	0

Объединим наблюдения 2 и 3, имеющие наименьшее расстояние $\rho_{2,3}$=1,41. После объединения имеем три кластера S_1, $S_{(2,3)}$, $S_{(4,5)}$ и S_6.

Вновь строим матрицу расстояний. Для этого необходимо рассчитать расстояние до кластера $S_{(2,3)}$. Воспользуемся матрицей расстояний R_2. Проведя аналогичные расчеты, получим матрицу расстояний

R3=

	1	2,3	4,5	6
1	0	2,236068	5,09902	5,830952
2,3	2,236068	0	5	6,403124
4,5	5,09902	5	0	2
6	5,830952	6,403124	2	0

Объединим кластеры $S_{(4,5)}$ и $S_{(6)}$, расстояние между которыми, согласно матрице $\mathbf{R}_3$, минимально $\rho_{(4,5),6}$=2. В результате этого получим три кластера: S_1, $S_{(2,3)}$ и $S_{(4,5,6)}$. Матрица расстояний будет иметь вид:

R4=

	1	2,3	4,5,6
1	0	2,236068	5,09902
2,3	2,236068	0	5
4,5,6	5,09902	5	0

Объединим наблюдения 1 и (2,3), имеющие наименьшее расстояние $\rho_{1,(2,3)}$=2,23. После объединения имеем три кластера $S_{(1,2,3)}$, $S_{(4,5,6)}$. Получим последнюю

матрицу расстояний R5.

R5=

	1,2,3	4,5,6
1,2,3	0	5
4,5,6	5	0

Из матрицы $\mathbf{R}_5$ следует, что на расстоянии $\rho_{(1,2,3),(4,5,6)}=5$ все шесть наблюдений объединяются в один кластер. На основании результатов кластерного анализа, можно сделать вывод, что наилучшим является разбиение шести объектов на два кластера: $S_{(1,2,3)}$ и $S_{(4,5,6)}$, когда пороговое расстояние находится в интервале $2{,}23<\rho_{пор}<5$.

б) *Провести классификацию, используя обычное евклидово расстояние и метод «дальнего соседа»*

Согласно обычной евклидовой метрике расстояние между наблюдениями равно

$$d_{ij}=\sqrt{\sum_{k=1}^{m}(x_{ik}-x_{jk})^2}$$

По формуле находим расстояния между всеми шестью наблюдениями и строим матрицу расстояний

R1=

	1	2	3	4	5	6
1	0	2,236068	3	5,09902	6,082763	5,830952
2	2,236068	0	1,414214	5	5,830952	6,403124
3	3	1,414214	0	6,403124	7,211103	7,81025
4	5,09902	5	6,403124	0	1	2
5	6,082763	5,830952	7,211103	1	0	2,236068
6	5,830952	6,403124	7,81025	2	2,236068	0

Из матрицы расстояний следует, что объекты 4 и 5 наиболее близки $\rho_{4,5}=1$ и

поэтому объединим их в один кластер. После объединения объектов имеем пять кластеров: S_1, S_2, S_3, $S_{(4,5)}$, S_6.

Расстояние между кластерами будем находить по принципу “дальнего соседа”, воспользовавшись формулой пересчета. Так расстояние между кластером S_1 и кластером $S_{(4,5)}$ будет рассчитываться по формуле:

$$d_{1,(4,5)} = d(S_1, S_{(4,5)}) = \frac{1}{2}d_{1,4} + \frac{1}{2}d_{1,5} + \frac{1}{2}|d_{1,4} - d_{1,5}|$$

Проводя расчеты, получим новую матрицу расстояний.

R2=

	1	2	3	4,5	6
1	0	2,236068	3	6,082763	5,830952
2	2,236068	0	1,414214	5,830952	6,403124
3	3	1,414214	0	7,211103	7,81025
4,5	6,082763	5,830952	7,211103	0	2,236068
6	5,830952	6,403124	7,81025	2,236068	0

Объединим наблюдения 2 и 3, имеющие наименьшее расстояние $\rho_{2,3}$=1,41. После объединения имеем три кластера S_1, $S_{(2,3)}$, $S_{(4,5)}$ и S_6.

Вновь строим матрицу расстояний. Для этого необходимо рассчитать расстояние до кластера $S_{(2,3)}$. Воспользуемся матрицей расстояний R_2. Проведя аналогичные расчеты, получим матрицу расстояний

R3=

	1	2,3	4,5	6
1	0	3	6,082763	5,830952
2,3	3	0	7,211103	6,403124
4,5	6,082763	7,211103	0	2,236068
6	5,830952	6,403124	2,236068	0

Объединим кластеры $S_{(4,5)}$ и $S_{(6)}$, расстояние между которыми, согласно матрице $\mathbf{R}_3$, минимально $\rho_{(4,5),6}$=2,23. В результате этого получим три кластера: S_1, $S_{(2,3)}$ и $S_{(4,5,6)}$. Матрица расстояний будет иметь вид:

R4=

	1	2,3	4,5,6
1	0	3	6,082763
2,3	3	0	7,211103

4,5,6	6,082763	7,211103	0

Объединим наблюдения 1 и (2,3), имеющие наименьшее расстояние $\rho_{1,(2,3)}=3$. После объединения имеем три кластера $S_{(1,2,3)}$, $S_{(4,5,6)}$. Получим последнюю матрицу расстояний R5.

R5=

	1,2,3	4,5,6
1,2,3	0	7,211103
4,5,6	7,211103	0

Из матрицы $\mathbf{R}_5$ следует, что на расстоянии $\rho_{(1,2,3),(4,5,6)}=7{,}21$ все шесть наблюдений объединяются в один кластер. На основании результатов кластерного анализа, можно сделать вывод, что наилучшим является разбиение шести объектов на два кластера: $S_{(1,2,3)}$ и $S_{(4,5,6)}$, когда пороговое расстояние находится в интервале $3<\rho_{пор}<7{,}21$.

в) *Провести классификацию, используя обычное евклидово расстояние и метод «средней связи»*

Согласно обычной евклидовой метрике расстояние между наблюдениями равно

$$d_{ij}=\sqrt{\sum_{k=1}^{m}(x_{ik}-x_{jk})^2}$$

По формуле находим расстояния между всеми шестью наблюдениями и строим матрицу расстояний

R1=

	1	2	3	4	5	6
1	0	2,236068	3	5,09902	6,082763	5,830952
2	2,236068	0	1,414214	5	5,830952	6,403124
3	3	1,414214	0	6,403124	7,211103	7,81025
4	5,09902	5	6,403124	0	1	2
5	6,082763	5,830952	7,211103	1	0	2,236068
6	5,830952	6,403124	7,81025	2	2,236068	0

Из матрицы расстояний следует, что объекты 4 и 5 наиболее близки $\rho_{4,5}=1$ и

поэтому объединим их в один кластер. После объединения объектов имеем пять кластеров: S_1, S_2, S_3, $S_{(4,5)}$, S_6.

Расстояние между кластерами будем находить по принципу “дальнего соседа”, воспользовавшись формулой пересчета. Так расстояние между кластером S_1 и кластером $S_{(4,5)}$ будет рассчитываться по формуле:

$$\rho_{(4,5),1} = \frac{1}{2}(\rho_{4,1} + \rho_{5,1})$$

Проводя расчеты, получим новую матрицу расстояний.

R2=

	1	2	3	4,5	6
1	0	2,236068	3	5,590891	5,830952
2	2,236068	0	1,414214	5,415476	6,403124
3	3	1,414214	0	6,807113	7,81025
4,5	5,590891	5,415476	6,807113	0	2,118034
6	5,830952	6,403124	7,81025	2,118034	0

Объединим наблюдения 2 и 3, имеющие наименьшее расстояние $\rho_{2,3}$=1,41. После объединения имеем три кластера S_1, $S_{(2,3)}$, $S_{(4,5)}$ и S_6.

Вновь строим матрицу расстояний. Для этого необходимо рассчитать расстояние до кластера $S_{(2,3)}$. Воспользуемся матрицей расстояний R_2. Проведя аналогичные расчеты, получим матрицу расстояний

R3=

	1	2,3	4,5	6
1	0	2,61803	5,590891	5,830952
2,3	2,61803	0	6,11129	7,10669
4,5	6,082763	6,11129	0	2,118034
6	5,830952	7,10669	2,118034	0

Объединим кластеры $S_{(4,5)}$ и $S_{(6)}$, расстояние между которыми, согласно матрице $\mathbf{R}_3$, минимально $\rho_{(4,5),6}$=2,11. В результате этого получим три кластера: S_1, $S_{(2,3)}$ и $S_{(4,5,6)}$. Матрица расстояний будет иметь вид:

R4=

	1	2,3	4,5,6
1	0	2,61803	5,71092
2,3	2,61803	0	6,60899

4,5,6	5,71092	6,60899	0

Объединим наблюдения 1 и (2,3), имеющие наименьшее расстояние $\rho_{1,(2,3)}$=2,61. После объединения имеем три кластера $S_{(1,2,3)}$, $S_{(4,5,6)}$. Получим последнюю матрицу расстояний R5.

R5=

	1,2,3	4,5,6
1,2,3	0	6,15996
4,5,6	6,15996	0

Из матрицы $\mathbf{R}_5$ следует, что на расстоянии $\rho_{(1,2,3),(4,5,6)}$= 6,16 все шесть наблюдений объединяются в один кластер. На основании результатов кластерного анализа, можно сделать вывод, что наилучшим является разбиение шести объектов на два кластера: $S_{(1,2,3)}$ и $S_{(4,5,6)}$, когда пороговое расстояние находится в интервале $2,61<\rho_{пор}<6,16$.

ВЫВОД: Таким образом, используя три алгоритма кластерного анализа, получился один вариант разбиения шести семей на две статистически однородные группы. Следовательно разбиении $S_{(1,2,3)}$ и $S_{(4,5,6)}$ является наиболее устойчивым.

Задача 2

Дано:

Таблица 2

№ хозяйства (*i*)	1	2	3	4
$x_i(1)$	1	7	1	9
$x_i(2)$	5	9	3	7

w1 = 0.1, w2 = 0.9

Классификация на основе "взвешенного евклидова расстояния" и принципа "ближайшего соседа".

Взвешенное евклидово расстояние между *i*-м и *l*-м наблюдениями определяется по формуле:

$$\rho_{\text{ВЕ}}(x_i, x_l) = \sqrt{\sum_{j=1}^{p} (x_i^{(j)} - x_l^{(j)})^2 \cdot \omega_j} .$$

Находим все остальные расстояния и строим матрицу расстояний:

1 2 3 4

R1=

1	0	4,242641	1,897367	3,162278
2	4,242641	0	6	2
3	1,897367	6	0	4,560702
4	3,162278	2	4,560702	0

Объединив S_1 и S_3, имеющих минимальное расстояние $\rho_{1,3}$=1,89 в кластер $S_{1,3}$, и применив принцип “ближайшего соседа” (формула та же, что и в задаче 1.б), получим матрицу расстояний:

R2=

	1,3	2	4
1,3	0	6	4,560702
2	6	0	2
4	4,560702	2	0

Образовав на расстоянии $\rho_{2,4}$=2 кластер $S_{2,4}$ вновь построим матрицу расстояний:

R3=

	1,3	2,4
1,3	0	6
2,4	6	0

Следовательно на расстояние $\rho_{(1,3),(2,4)}$=6 объединяются кластеры $S_{(1,3)}$ и $S_{(2,4)}$ и все четыре объекта образуют один кластер.

Результаты классификации представлены графически на рис.1

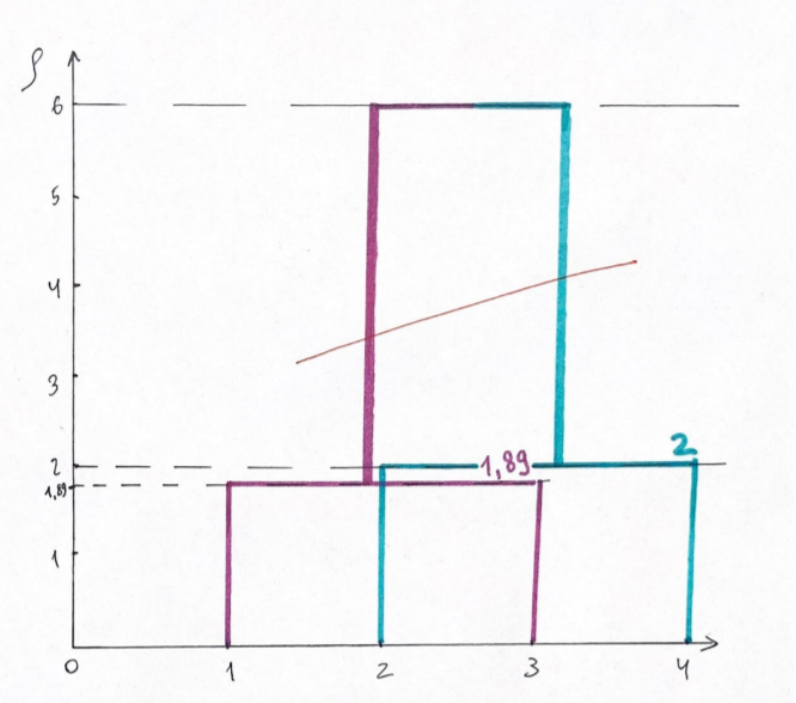

Рис.1 Дендрограмма (взвешенное евклидово расстояние, принцип ближайшего соседа)

Из матрицы $\mathbf{R}_3$ следует, что на расстоянии $\rho_{(1,3),(2,4,)}= 6$ все четыре наблюдения объединяются в один кластер. На основании результатов кластерного анализа, можно сделать вывод, что наилучшим является разбиение шести объектов на два кластера: $S_{(1,3)}$ и $S_{(2,4)}$, когда пороговое расстояние находится в интервале $2<\rho_{пор}<6$.