ОП «Политология», 2022-23
Введение в МСА
Иерархический кластерный анализ (разбор задачи 1)
Д. В. Сальникова, А. А. Тамбовцева

Задача 1. Перед вами небольшая таблица, содержащая данные по четырём респондентам (x – число килограммов, набранных за новогодние праздники, y – среднее число часов, проведённых на свежем воздухе):

id	x	y
1	0	4
2	7	0
3	1	2
4	4	2

(a) Постройте диаграмму рассеивания для представленного набора данных.
(b) Реализуйте иерархический кластерный анализ по следующему алгоритму.
   i. Вычислите средние значения $\bar{x}$ и $\bar{y}$, стандартные отклонения s_x и s_y.
   ii. Выполните нормировку предложенных данных, вычитая из каждого значения среднее значение по столбцу и деля результат на стандартное отклонение столбца[1].
   iii. Используя в качестве метрики манхэттенское расстояние, запишите матрицу расстояний для преобразованного набора данных.
   iv. Используя в качестве метода агрегирования метод дальнего соседа, реализуйте иерархический кластерный анализ и постройте дендрограмму.

Решение.
Реализуйте иерархический кластерный анализ по следующему алгоритму.
i. Вычислите средние значения $\bar{x}$ и $\bar{y}$, стандартные отклонения s_x и s_y.

$$\bar{x} = 3$$

$$\bar{y} = 2$$

$$s_x = \sqrt{\frac{(0-3)^2 + (7-3)^2 + (1-3)^2 + (4-3)^2}{3}} = \sqrt{10} \approx 3$$

$$s_y = \sqrt{\frac{(4-2)^2 + (0-2)^2 + (2-2)^2 + (2-2)^2}{3}} = \sqrt{2.67} \approx 2$$

ii. Выполните нормировку предложенных данных, вычитая из каждого значения среднее значение по столбцу и деля результат на стандартное отклонение столбца.

[1] для удобства стандартное отклонение можно округлить до целых

id	x′	y′
1	−1	1
2	1.3	−1
3	−0.7	0
4	0.3	0

Примеры нормировки:

$$x'_1 = \frac{0-3}{3} = -1$$

$$x'_2 = \frac{7-3}{3} \approx 1.3$$

$$y'_1 = \frac{4-2}{2} = 1$$

$$y'_3 = \frac{2-2}{2} = 0$$

iii. Используя в качестве метрики манхэттенское расстояние, запишите матрицу расстояний для преобразованного набора данных.

$$D = \begin{bmatrix} & \mathbf{1} & \mathbf{2} & \mathbf{3} & \mathbf{4} \\ \mathbf{1} & 0 & 4.3 & 1.3 & 2.3 \\ \mathbf{2} & 4.3 & 0 & 3 & 2 \\ \mathbf{3} & 1.3 & 3 & 0 & 1 \\ \mathbf{4} & 2.3 & 2 & 1 & 0 \end{bmatrix}$$

Примеры вычисления расстояний:

$$d(1,2) = |-1-1.3| + |1-(-1)| = 4.3$$

$$d(1,3) = |-1-(-0.7)| + |1-0| = 1.3$$

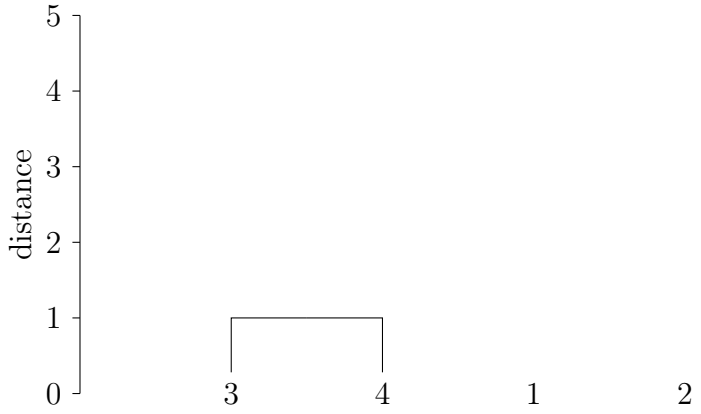
$$d(2,3) = |1.3-(-0.7)| + |-1-0| = 3$$

iv. Используя в качестве метода агрегирования метод дальнего соседа, реализуйте иерархический кластерный анализ и постройте дендрограмму.

Шаг 1. На первом шаге у нас 4 кластера: 1, 2, 3, 4.

Шаг 2. Теперь объединим в кластер те точки, которые ближе всего друг к другу. Это точки 3 и 4, расстояние между ними 1. Соединим эти точки, а на вертикальной оси отметим расстояние 1.[2]

[2]Заранее узнать, как удобнее располагать точки, невозможно, здесь я сразу предлагаю такой порядок, чтобы на итоговой дендрограмме одни «ветки» не перечеркивали другие при объединении в группы. Можно было нарисовать как есть, а потом переставить посимпатичнее.



По итогу этого шага получаем уже 3 кластера: $3 + 4$, 1, 2.

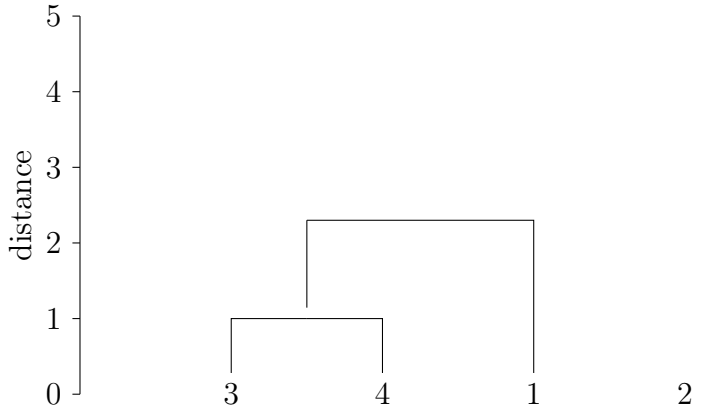
Шаг 3. Объединяем точки в более крупные кластеры. Для этого необходимо определить попарные расстояния между точками 1 и 2, а также расстояние между кластером $3 + 4$ и точкой 1, между кластером $3 + 4$ и точкой 2. Итак, с учетом метода дальнего соседа:

$$d(1, 2) = 4.3$$

$$d(3 + 4, 1) = \max\{d(3, 1), d(4, 1)\} = \max\{1.3, 2.3\} = 2.3$$

$$d(3 + 4, 2) = \max\{d(3, 2), d(4, 2)\} = \max\{3, 2\} = 3$$

Объединяем ближайшие группы, то есть те, расстояние между которыми является самым маленьким. Самое маленькое расстояние из полученных – расстояние между кластером 3+4 и точкой 1. Значит, присоединяем ее к этому кластеру и фиксируем расстояние 2.3.



По итогу этого шага получаем уже 2 кластера: $3 + 4 + 1$, 2.

Шаг 4. Завершаем кластеризацию – формируем один большой кластер, надо

только выяснить, на каком расстоянии происходит объединение в группы. Вычисляем:

$$d(3+4+1,2) = \max\{d(3,2), d(4,2), d(1,2)\} = \max\{3, 2, 4.3\} = 4.3$$

Итоговая дендрограмма: