

Математический подход для построения простейших алгоритмов кластеризации.

Тверская Е. С.
e_tverskaya@bmstu.ru

Московский государственный технический университет им. Н. Э. Баумана
(национальный исследовательский университет)

Москва, 2025



ИУ-6
Компьютерные
системы и сети



Простейшие методы кластеризации

Метрические меры сходства.

Обобщенное («взвешенное») расстояние махаланобисского типа.

Определение

В общем виде данный тип расстояния можно записать следующим образом

$$d_0(X_i, X_j) = \sqrt{(X_i - X_j)^T \Lambda^T \Sigma^{-1} \Lambda (X_i - X_j)},$$

где

Σ – ковариационная матрица генеральной совокупности, из которой извлекаются наблюдения X_i .

Λ – некоторая симметрическая неотрицательно-определенная матрица «весовых» коэффициентов, которая чаще всего выбирается диагональной.



Простейшие методы кластеризации

Метрические меры сходства.

Примеры расстояния махаланобисского типа. Евклидово расстояние.

$$\begin{aligned}d_E(X_i, X_j) &= \sqrt{(X_i - X_j)^T (X_i - X_j)} = \\&= \sqrt{\left(x_i^{(1)} - x_j^{(1)}\right)^2 + \left(x_i^{(2)} - x_j^{(2)}\right)^2 + \dots + \left(x_i^{(p)} - x_j^{(p)}\right)^2}\end{aligned}$$

Когда использовать!

1. Наблюдения X извлекаются из генеральных совокупностей, описываемых многомерным нормальным законом с ковариационной матрицей вида $\sigma^2 I$, т. е. компоненты X взаимно независимы и имеют одну и ту же дисперсию.
2. Компоненты $x^{(1)}, x^{(2)}, \dots, x^{(p)}$ образа вектора наблюдений однородны и все они одинаково важны для отнесения объекта к тому или иному классу.
3. Признаковое пространство совпадает с геометрическим пространством, т. е. $p = 1, 2, 3$.



Простейшие методы кластеризации

Метрические меры сходства.

Примеры расстояния махаланобисского типа. Взвешенное Евклидово расстояние.

$$d_{\text{ВЕ}}(X_i, X_j) = \sqrt{\omega_1 (x_i^{(1)} - x_j^{(1)})^2 + \omega_2 (x_i^{(2)} - x_j^{(2)})^2 + \dots + \omega_p (x_i^{(p)} - x_j^{(p)})^2}$$

Когда использовать!

Данное расстояние используется, если каждой компоненте $x^{(k)}$ вектора наблюдений X можно присвоить некоторый «вес» ω_k , т. е. некоторую неотрицательную величину, которая характеризует степень важности данной компоненты с точки зрения решения. При этом полагают $0 \leq \omega_k \leq 1, k = 1, \dots, p$.



Простейшие методы кластеризации

Метрические меры сходства.

Примеры расстояния махаланобисского типа. Хеммингово расстояние.

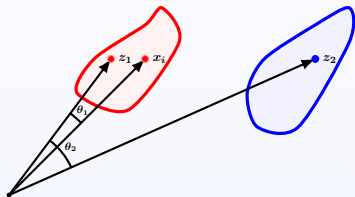
$$d_H(X_i, X_j) = \sum_{s=1}^p |x_i^{(s)} - x_j^{(s)}|$$

Когда использовать!

Данное расстояние используют как меру различия объектов, задаваемых дихотомическими признаками и равно числу несовпадений значений соответствующих признаков в *i*-м и *j*-м объектах.

Простейшие методы кластеризации

Неметрические меры сходства. Косинусное расстояние.



$$s(X_i, X_j) = \frac{(X_i)^T X_j}{\|X_i\| \|X_j\|}$$

$$s(x_i, z_1) = \cos(\theta_1) = \frac{(x_i)^T z_1}{\|x_i\| \|z_1\|}$$

$$s(x_i, z_2) = \cos(\theta_2) = \frac{(x_i)^T z_2}{\|x_i\| \|z_2\|}$$

Когда использовать!

Данной мерой **удобно** пользоваться, если кластеры стремятся расположиться вдоль координатных осей.

Недостатки данной меры: кластеры должны находиться на достаточном расстоянии друг от друга и от начала координат.

Простейшие методы кластеризации

Критерии кластеризации.

Эвристический подход: задание набора правил, которые обеспечивают использование выбранной меры сходства для отнесения образа к одному из кластеров. Например, для этой цели часто используют Евклидово расстояние, это связано с его естественной интерпретацией как меры близости объектов.

Оптимизация показателя качества: данный подход связан с разработкой процедуры, которая обеспечивает минимизацию или максимизацию выбранного показателя качества. Часто, для это цели используют сумму квадратов ошибки:

Простейшие методы кластеризации

Критерии кластеризации.

Рассмотрим подробнее предложенный вариант показателя качества

$$J = \sum_{j=1}^{N_c} \sum_{x \in S_j} \|x - m_j\|^2 \rightarrow \min, \quad \text{где}$$

N_c – число кластеров;

S_j – множество образов, относящихся к j -му кластеру;

$m_j = \frac{1}{N_j} \sum_{x \in S_j} x$ – вектор выборочных средних значений для множества

S_j ;

N_j – количество образов, входящих в множество S_j .

Смешанный подход: Часто встречаются алгоритмы основанные на совместном использовании эвристического подхода и подхода с использованием показателя качества.

Простейшие методы кластеризации

The simplest clustering algorithm.

Шаг 1

1. Задается множество образов $(x_1, x_2, \dots, x_N)$.
2. Произвольно выбирается величина $T > 0$ – пороговая величина.

Шаг 2

Произвольно выбирается центр 1-го кластера. Для определенности, пусть $z_1 = x_1$.

Шаг 3

Вычисляется расстояние $D_{21} = \rho(z_1, x_2) = \|x_2 - z_1\|$.
Если $D_{21} > T$, то $z_2 = x_2$, т.е. z_2 – центр нового кластера.
Если $D_{21} < T$, то $x_2 \in S_1$, где S_1 – множество образов, входящих в кластер с центром z_1 .

Шаг 4

Вычисляется расстояния $D_{31} = \rho(x_3, z_1) = \|x_3 - z_1\|$ и $D_{32} = \rho(x_3, z_2) = \|x_3 - z_2\|$.
Если $D_{31} > T$ и $D_{32} > T$, то $z_3 = x_3$, т.е. z_3 – центр нового кластера.
Иначе, образ x_3 зачиляется в тот класте, чей центр к нему ближе.

Простейшие методы кластеризации

The simplest clustering algorithm.

Аналогично, вычисляется расстояние от каждого нового образа до каждого известного центра кластера и сравнивается с пороговой величиной T . В результате, либо устанавливается новый центр кластера, либо, в противном случае, образ зачисляется в кластер с ближайшим к нему центром.

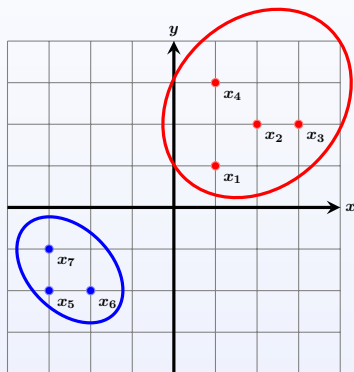
Недостатки алгоритма: алгоритм чувствителен к выбору исходной точки кластеризации и значению порога .

Достоинства алгоритма:

1. Алгоритм прост с вычислительной точки зрения.
2. Позволяет быстро получить приблизительные характеристики заданного набора данных.
3. Можно рассчитывать на получение достаточно хороших результатов, если данные имеют скученности и хорошо разделяются при соответствующем выборе исходной точки кластеризации и значения порога.

Простейшие методы кластеризации

The simplest clustering algorithm. Пример.



$$x_1 = (1, 1)$$

$$x_2 = (2, 2)$$

$$x_3 = (3, 2)$$

$$x_4 = (1, 3)$$

$$x_5 = (-3, -2)$$

$$x_6 = (-2, -2)$$

$$x_7 = (-3, -1)$$



Простейшие методы кластеризации

The simplest clustering algorithm. Пример.

Шаг 1

Выберем величину $T = 3$.

Шаг 2

Центром кластера будет $z_1 = x_1$.

Шаг 3,4

Вычислим

$$D_{21} = \sqrt{2} < T \Rightarrow x_2 \in S_1.$$

$$D_{31} = \sqrt{5} < T \Rightarrow x_3 \in S_1.$$

$$D_{41} = 2 < T \Rightarrow x_4 \in S_1.$$

$$D_{51} = 5 > T \Rightarrow$$

тогда $z_2 = x_5$ – центр нового кластера.

$$\text{Вычислим } D_{61} = 3\sqrt{2} > T \text{ и } D_{62} = 1 < T, \text{ тогда } x_6 \in S_2.$$

$$\text{Вычислим } D_{71} = 2\sqrt{5} > T \text{ и } D_{72} = 1 < T, \text{ тогда } x_7 \in S_2.$$

$$S_1 = \{x_1, x_2, x_3, x_4\} \text{ и } S_2 = \{x_5, x_6, x_7\}.$$



Простейшие методы кластеризации

Алгоритм *k - means*.

Шаг 1

1. Задается множество образов ($x_1, x_2, \dots, x_N$).
 2. Выбирается произвольным образом K ($K < N$) исходных центров кластеров: $z_1(1), \dots, z_K(1)$.
- Далее, не нарушая сути рассуждений, дальнейшую реализацию алгоритма можно рассматривать с шага m .

Шаг m

На данном шаге заданное множество образов распределяется по кластерам по следующему правилу:
 $x \in S_j(m)$, если $\|x - z_j\| < \|x - z_i\|$, $i = 1, \dots, K$, $i \neq j$,
где $S_j(m)$ – множество образов, входящих в кластер с центром $z_j(m)$.

Простейшие методы кластеризации

Алгоритм *k - means*.

Шаг $m + 1$

На данном шаге определяются новые центры кластеров $\mathbf{z}_j(k + 1)$, $j = 1, \dots, K$ таким образом, чтобы минимизировать показатель качества:

$$J_j = \sum_{\mathbf{x} \in S_j(m)} \|\mathbf{x} - \mathbf{z}_j(m + 1)\|^2, j = 1, \dots, K.$$

Необходимо выбрать центр $\mathbf{z}_j(m + 1)$, чтобы минимизировать показатель качества.

Центр $\mathbf{z}_j(m + 1)$, минимизирующий показатель качества является выборочным средним:

$$\mathbf{z}_j(m + 1) = \frac{1}{N_j} \sum_{\mathbf{x} \in S_j(m)} \mathbf{x}, j = 1, \dots, K, \text{ где}$$

N_j – число выборочных образов, входящих в $S_j(m)$.

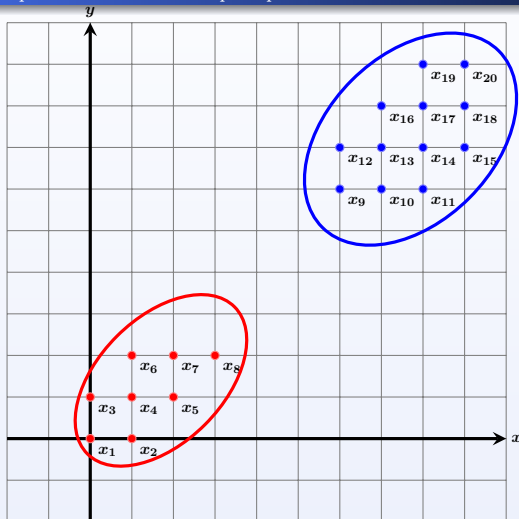
Шаг *end*

Равенство $\mathbf{z}_j(m) = \mathbf{z}_j(m + 1)$, $j = 1, \dots, K$ является условием сходимости алгоритма. При его достижении алгоритм заканчивает свою работу.

Иначе, алгоритм повторяется с шага m .

Простейшие методы кластеризации

Алгоритм ***k - means***. Пример.



$$x_1 = (0, 0)$$

$$x_2 = (1, 0)$$

$$x_3 = (0, 1)$$

$$x_4 = (1, 1)$$

$$x_5 = (2, 1)$$

$$x_6 = (1, 2)$$

$$x_7 = (2, 2)$$

$$x_8 = (3, 2)$$

$$x_9 = (6, 6)$$

$$x_{10} = (7, 6)$$

$$x_{11} = (8, 6)$$

$$x_{12} = (6, 7)$$

$$x_{13} = (7, 7)$$

$$x_{14} = (8, 7)$$

$$x_{15} = (9, 7)$$

$$x_{16} = (7, 8)$$

$$x_{17} = (8, 8)$$

$$x_{18} = (9, 8)$$

$$x_{19} = (8, 9)$$

$$x_{20} = (9, 9)$$



Простейшие методы кластеризации

Алгоритм *k - means*. Пример.

Шаг 1

Задается $K = 2$, $z_1(1) = x_1$ и $z_2(1) = x_2$.

Шаг 2

Оставшиеся образы распределяются по 2-м кластерам по правилу:
 $x \in S_1(1)$, если $\|x - z_1(1)\| < \|x - z_i(1)\|$, $i = 2$,
 $x \in S_2(1)$, если $\|x - z_2(1)\| < \|x - z_i(1)\|$, $i = 1$.

Итак, $S_1(1) = \{x_1, x_3\}$.
 $S_2(1) = \{x_2, x_4, x_5, x_6, \dots, x_{20}\}$.

Шаг 3

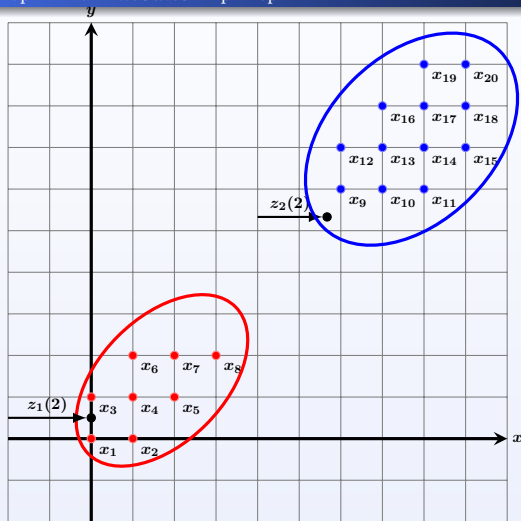
Коррекция центров кластера:

$$z_1(2) = \frac{1}{N_1} \sum_{x \in S_1(1)} x = \frac{1}{2} (x_1 + x_3) = \begin{pmatrix} 0, 0 \\ 0, 5 \end{pmatrix}.$$

$$z_2(2) = \frac{1}{N_2} \sum_{x \in S_2(1)} x = \frac{1}{18} (x_2 + x_4 + x_5 + \dots + x_{20}) = \begin{pmatrix} 5, 67 \\ 5, 33 \end{pmatrix}.$$

Простейшие методы кластеризации

Алгоритм *k - means*. Пример.



Шаг 4

Так как

$$z_i(1) \neq z_i(2), \quad i = 1, 2,$$

то возвращаемся к **шагу 2**.



Простейшие методы кластеризации

Алгоритм *k - means*. Пример.

Шаг 2

Образы распределяются по 2-м кластерам по правилу:

$x \in S_1(1)$, если $\|x - z_1(1)\| < \|x - z_i(1)\|$, $i = 2$,

$x \in S_2(1)$, если $\|x - z_2(1)\| < \|x - z_i(1)\|$, $i = 1$.

Итак, $S_1(2) = \{x_1, x_2, x_3, \dots, x_8\}$,

$S_2(2) = \{x_9, x_{10}, \dots, x_{20}\}$.

Шаг 3

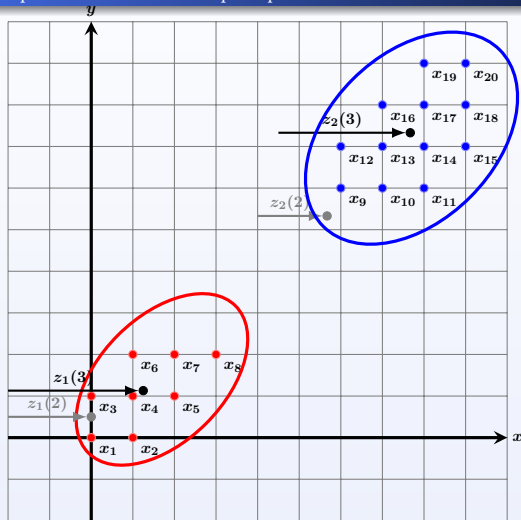
Коррекция центров кластера:

$$z_1(3) = \frac{1}{N_1} \sum_{x \in S_1(2)} x = \frac{1}{8} (x_1 + x_2 + \dots + x_8) = \begin{pmatrix} 1, 25 \\ 1, 13 \end{pmatrix}.$$

$$z_2(3) = \frac{1}{N_2} \sum_{x \in S_2(2)} x = \frac{1}{12} (x_9 + x_{10} + \dots + x_{20}) = \begin{pmatrix} 7, 67 \\ 7, 33 \end{pmatrix}.$$

Простейшие методы кластеризации

Алгоритм *k - means*. Пример.



Шаг 4

Так как,

$$z_i(2) \neq z_i(3), \quad i = 1, 2,$$

то возвращаемся к **шагу 2**.

Шаг 2

Образы разделяются по двум кластерам так же, как и на предыдущей итерации.

$$z_i(4) = z_i(3), \quad i = 1, 2,$$

$$z_1 = z_1(4) = \begin{pmatrix} 1, 25 \\ 1, 13 \end{pmatrix},$$

$$z_2 = z_2(4) = \begin{pmatrix} 7, 67 \\ 7, 33 \end{pmatrix}.$$

Простейшие методы кластеризации

Iterative Self-Organizing Data Analysis Techniques

ИСОМАД (Итеративный самоорганизующийся метод анализа данных).

Шаг 1

1. Задается множество образов ($\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_N$).
2. Задается множество исходных центров кластеров ($\mathbf{z}_1, \mathbf{z}_2, \dots, \mathbf{z}_{N_c}$).
3. K – необходимое число кластеров.
4. θ_N – параметр, с которым сравнивается количество выборочных образов, вошедших в кластер.
5. θ_s – параметр, характеризующий среднеквадратичное отклонение.
6. θ_c – параметр, характеризующий компактность.
7. L – максимальное количество пар центров кластеров, которые можно объединить.
8. I – допустимое число итераций.

Шаг 2

Заданные N образов распределяются по кластерам в соответствии с выбранными исходными центрами.

$\mathbf{x} \in S_j$, если $\|\mathbf{x} - \mathbf{z}_j\| < \|\mathbf{x} - \mathbf{z}_i\|$, $i = 1, \dots, N_c$, $i \neq j$,

где S_j – множество образов, входящих в кластер с центром $\mathbf{z}_j$.



Простейшие методы кластеризации

Iterative Self-Organizing Data Analysis Techniques

ИСОМАД (Итеративный самоорганизующийся метод анализа данных).

Шаг 3

Ликвидируются подмножества образов, в состав которых входит менее θ_N элементов. Т.е., если $N_j < \theta_N$ для некоторого j (N_j – число объектов, вошедших в подмножество S_j), то подмножество S_j исключается из рассмотрения и N_c уменьшается на 1, т.е.

$$N_c = N_c - 1.$$

Простейшие методы кластеризации

Iterative Self-Organizing Data Analysis Techniques

ИСОМАД (Итеративный самоорганизующийся метод анализа данных).

Шаг 3

Ликвидируются подмножества образов, в состав которых входит менее θ_N элементов. Т.е., если $N_j < \theta_N$ для некоторого j (N_j – число объектов, вошедших в подмножество S_j), то подмножество S_j исключается из рассмотрения и N_c уменьшается на 1, т.е.

$$N_c = N_c - 1.$$

Шаг 4

Центры кластеров корректируются посредством выборочного среднего, найденного по множеству S_j .

$$z_j = \frac{1}{N_j} \sum_{x \in S_j} x, \quad j = 1, \dots, N_c, \quad \text{где}$$

N_j – число выборочных образов, входящих в S_j .



Простейшие методы кластеризации

Iterative Self-Organizing Data Analysis Techniques

ИСОМАД (Итеративный самоорганизующийся метод анализа данных).

Шаг 5

Вычисляется среднее расстояние $\overline{D}_j$ между объектами, входящими в подмножество S_j , и соответствующими центрами кластера

$$\overline{D}_j = \frac{1}{N_j} \sum_{x \in S_j} \|x - z_j\|, \quad j = 1, \dots, N_c.$$



Простейшие методы кластеризации

Iterative Self-Organizing Data Analysis Techniques

ИСОМАД (Итеративный самоорганизующийся метод анализа данных).

Шаг 5

Вычисляется среднее расстояние $\overline{D}_j$ между объектами, входящими в подмножество S_j , и соответствующими центрами кластера

$$\overline{D}_j = \frac{1}{N_j} \sum_{x \in s_j} \|x - z_j\|, \quad j = 1, \dots, N_c.$$

Шаг 6

Вычисляется обобщенное среднее расстояние $\overline{D}$ между объектами, находящимися в отдельных кластерах и соответствующими центрами кластеров по формуле

$$\overline{D} = \frac{1}{N} \sum_{j=1}^{N_c} N_j \overline{D}_j.$$



Простейшие методы кластеризации

Iterative Self-Organizing Data Analysis Techniques

ИСОМАД (Итеративный самоорганизующийся метод анализа данных).

Шаг 7

(1) Если текущий цикл итераций – последний, то задается $\theta_c = 0$ и переход к **шагу 11**.

(2) Если выполняется условие $N_c \leq K/2$, то осуществляется переход на **шаг 8**.

(3) Если текущий цикл итераций имеет четный порядковый номер или выполняется условие $N_c \geq 2K$, то осуществляется переход на **шаг 11**.

Иначе, процесс итераций продолжается.

Простейшие методы кластеризации

Iterative Self-Organizing Data Analysis Techniques

ИСОМАД (Итеративный самоорганизующийся метод анализа данных).

Mar 7

(1) Если текущий цикл итераций – последний, то задается $\theta_c = \mathbf{0}$ и переход к *шагу 11*.

(2) Если выполняется условие $N_c \leq K/2$, то осуществляется переход на *шаг 8*.

(3) Если текущий цикл итераций имеет четный порядковый номер или выполняется условие $N_c \geq 2K$, то осуществляется переход на *шаг 11*.

Иначе, процесс итераций продолжается.

Mar 8

Для каждого подмножества S_j вычисляется вектор среднеквадратичного отклонения $\sigma_j = (\sigma_{1j}, \sigma_{2j}, \dots, \sigma_{nj})^T$

$$\sigma_{ij} = \sqrt{\frac{1}{N_j} \sum_{\mathbf{x} \in S_j} (x_{ik} - z_{ij})^2},$$

где n - размерность пространства образа;

x_{ik} - i -я компонента k -го образа в подмножестве S_j ;

z_{ij} - i -я компонента вектора $\mathbf{z}_j$, представляющего центр кластера S_j .

Каждая компонента вектора среднеквадратичного отклонения характеризует среднеквадратичное отклонение образа, входящего в подмножество S_j , по одной из главных осей.



Простейшие методы кластеризации

Iterative Self-Organizing Data Analysis Techniques

ИСОМАД (Итеративный самоорганизующийся метод анализа данных).

Шаг 9

В каждом векторе среднеквадратичного отклонения σ_j отыскивается максимальная компонента $\sigma_{j \max}$, $j = 1, \dots, N_c$.

Простейшие методы кластеризации

Iterative Self-Organizing Data Analysis Techniques

ИСОМАД (Итеративный самоорганизующийся метод анализа данных).

IIIar 9

В каждом векторе среднеквадратичного отклонения σ_j отыскивается максимальная компонента $\sigma_{j \max}$, $j = 1, \dots, N_c$.

Mar 10

Если для любого $\sigma_{j \max}, j = 1, \dots, N_c$ выполняется условие $\sigma_{j \max} > \theta_s$ и

$$(1) \boxed{\overline{D}_j > \overline{D} \text{ и } N_j > 2(\theta_N + 1)} \text{ или } (2) \boxed{N_c \leq K/2},$$

то кластер с центром $\mathbf{z}_j$ расщепляется на два новых кластера с центрами $\mathbf{z}_j^+$ и $\mathbf{z}_j^-$. При этом значение N_c увеличивается на 1. Для определения центра кластера $\mathbf{z}_j^+$ к компоненте вектора $\mathbf{z}_j$, соответствующей максимальной компоненте вектора $\boldsymbol{\sigma}_j$, прибавляется заданная величина γ_j . Центр кластера $\mathbf{z}_j^-$ определяется вычитанием этой же величины γ_j из той же самой компоненты вектора $\mathbf{z}_j$. Величину γ_j выбирают как

$$\gamma_j = k\sigma_{j \max}, \text{ где } 0 < k \leq 1.$$

При выборе γ_j следует руководствоваться тем, что она должна быть достаточно большой, чтобы расстояние от новых центров до произвольного образа было различимо, но и достаточно малой, чтобы общая структура кластеризации не нарушилась. Если расщепление произошло, то осуществляется переход на **шаг 2**. Если нет, то выполнение алгоритма продолжается.



Простейшие методы кластеризации

Iterative Self-Organizing Data Analysis Techniques

ИСОМАД (Итеративный самоорганизующийся метод анализа данных).

Шаг 11

Вычисляются расстояния D_{ij} между всеми парами центров кластеров:

$$D_{ij} = \|z_i - z_j\|, j = 1, \dots, N_c - 1, j = i + 1, \dots, N_c.$$



Простейшие методы кластеризации

Iterative Self-Organizing Data Analysis Techniques

ИСОМАД (Итеративный самоорганизующийся метод анализа данных).

Шаг 11

Вычисляются расстояния D_{ij} между всеми парами центров кластеров:

$$D_{ij} = \|z_i - z_j\|, j = 1, \dots, N_c - 1, j = i + 1, \dots, N_c.$$

Шаг 12

Расстояние D_{ij} сравнивают с параметром θ_c . Те L расстояний, которые оказались меньше θ_c , расставляются в порядке возрастания:

$$D_{i_1j_1} < D_{i_2j_2} < \dots < D_{i_Lj_L}.$$



Простейшие методы кластеризации

Iterative Self-Organizing Data Analysis Techniques

ИСОМАД (Итеративный самоорганизующийся метод анализа данных).

Шаг 11

Вычисляются расстояния D_{ij} между всеми парами центров кластеров:
 $D_{ij} = \|z_i - z_j\|, j = 1, \dots, N_c - 1, j = i + 1, \dots, N_c.$

Шаг 12

Расстояние D_{ij} сравнивают с параметром θ_c . Те L расстояний, которые оказались меньше θ_c , расставляются в порядке возрастания:
 $D_{i_1j_1} < D_{i_2j_2} < \dots < D_{i_Lj_L}.$

Шаг 13

Расстояние D_{ij} сравнивают с параметром θ_c . Те L расстояний, которые оказались меньше θ_c , расставляются в порядке возрастания:
 $D_{i_1j_1} < D_{i_2j_2} < \dots < D_{i_Lj_L}$ Данный шаг осуществляет процесс слияния кластеров. Кластеры с центрами z_{i_l} и z_{j_l} , объединяются (при условии, что процедура слияния в текущем цикле не применялась ни к тому, ни к другому кластеру). При этом центр нового кластера определяется по формуле

$$z_j^* = \frac{1}{N_{i_l} + N_{j_l}} [N_{i_l} z_{i_l} + N_{j_l} z_{j_l}].$$

Центры кластеров z_{i_l} и z_{j_l} ликвидируются и значение N_c уменьшается на 1.

Замечание: так как процедуру слияния к каждому центру можно применить только один раз, то ни при каких условиях не возможно получить L объединенных кластеров.



Простейшие методы кластеризации

Iterative Self-Organizing Data Analysis Techniques

ИСОМАД (Итеративный самоорганизующийся метод анализа данных). Пример.

Шаг 14



Если текущий цикл итерации — последний, то выполнение алгоритма прекращается. Иначе, следует возвратиться либо к **шагу 1** (если требуется корректировка параметров), либо к **шагу 2** (если параметры остаются неизменными).

Простейшие методы кластеризации

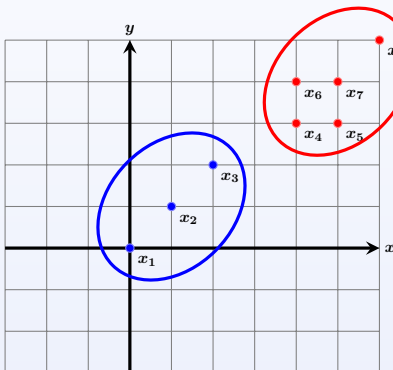
Iterative Self-Organizing Data Analysis Techniques

ИСОМАД (Итеративный самоорганизующийся метод анализа данных). Пример.

Шаг 14



Если текущий цикл итерации — последний, то выполнение алгоритма прекращается. Иначе, следует возвратиться либо к **шагу 1** (если требуется корректировка параметров), либо к **шагу 2** (если параметры остаются неизменными).



$$x_1 = (0, 0)$$

$$x_2 = (1, 1)$$

$$x_3 = (2, 2)$$

$$x_4 = (4, 3)$$

$$x_5 = (5, 3)$$

$$x_6 = (4, 4)$$

$$x_7 = (5, 4)$$

$$x_8 = (6, 5)$$

$$N = 8$$

$$n = 2$$

$$N_c = 1$$

$$z_1 = (0, 0)$$

$$K = 2$$

$$\theta_N = 1$$

$$\theta_s = 1$$

$$\theta_c = 4$$

$$L = 0$$

$$I = 4$$



Простейшие методы кластеризации

Iterative Self-Organizing Data Analysis Techniques

ИСОМАД (Итеративный самоорганизующийся метод анализа данных). Пример.

Шаг 2

Так как задан только один центр кластера, то

$$S_1 = \{x_1, x_2, x_3, x_4, \dots, x_8\} \quad \text{и} \quad N = 8.$$



Простейшие методы кластеризации

Iterative Self-Organizing Data Analysis Techniques

ИСОМАД (Итеративный самоорганизующийся метод анализа данных). Пример.

Шаг 2

Так как задан только один центр кластера, то

$$S_1 = \{x_1, x_2, x_3, x_4, \dots, x_8\} \quad \text{и} \quad N = 8.$$

Шаг 3

Так как $N_1 > \theta_N$, то ни одно подмножество не ликвидируется



Простейшие методы кластеризации

Iterative Self-Organizing Data Analysis Techniques

ИСОМАД (Итеративный самоорганизующийся метод анализа данных). Пример.

Шаг 2

Так как задан только один центр кластера, то

$$S_1 = \{x_1, x_2, x_3, x_4, \dots, x_8\} \quad \text{и} \quad N = 8.$$

Шаг 3

Так как $N_1 > \theta_N$, то ни одно подмножество не ликвидируется

Шаг 4

Корректируется положение центра кластера:

$$z_1 = \frac{1}{N_1} \sum_{x \in S_1} x = \begin{pmatrix} 3,38 \\ 2,75 \end{pmatrix}.$$

Простейшие методы кластеризации

Iterative Self-Organizing Data Analysis Techniques

ИСОМАД (Итеративный самоорганизующийся метод анализа данных). Пример.

Шаг 2

Так как задан только один центр кластера, то

$$S_1 = \{x_1, x_2, x_3, x_4, \dots, x_8\} \quad \text{и} \quad N = 8.$$

Шаг 3

Так как $N_1 > \theta_N$, то ни одно подмножество не ликвидируется

Шаг 4

Корректируется положение центра кластера:

$$z_1 = \frac{1}{N_1} \sum_{x \in S_1} x = \begin{pmatrix} 3,38 \\ 2,75 \end{pmatrix}.$$

Шаг 5

Вычисляется расстояние $\overline{D}_j$:

$$\overline{D}_1 = \frac{1}{N_1} \sum_{x \in S_j} \|x - z_1\| = 2,26.$$



Простейшие методы кластеризации

Iterative Self-Organizing Data Analysis Techniques

ИСОМАД (Итеративный самоорганизующийся метод анализа данных). Пример.

Шаг 6

Вычисляется расстояние $\overline{D}$:

$$\overline{D} = \overline{D}_1 = 2,26.$$



Простейшие методы кластеризации

Iterative Self-Organizing Data Analysis Techniques

ИСОМАД (Итеративный самоорганизующийся метод анализа данных). Пример.

Шаг 6

Вычисляется расстояние $\overline{D}$:

$$\overline{D} = \overline{D}_1 = 2,26.$$

Шаг 7

Так как данный цикл итераций не последний и $N_c = K/2$, то осуществляется переход к **шагу 8**.



Простейшие методы кластеризации

Iterative Self-Organizing Data Analysis Techniques

ИСОМАД (Итеративный самоорганизующийся метод анализа данных). Пример.

Шаг 6

Вычисляется расстояние $\overline{D}$:

$$\overline{D} = \overline{D}_1 = 2,26.$$

Шаг 7

Так как данный цикл итераций не последний и $N_c = K/2$, то осуществляется переход к **шагу 8**.

Шаг 8

Вычисляется вектор среднеквадратичного отклонения для заданного S_1 :

$$\sigma_1 = \begin{pmatrix} 1,99 \\ 1,56 \end{pmatrix}.$$



Простейшие методы кластеризации

Iterative Self-Organizing Data Analysis Techniques

ИСОМАД (Итеративный самоорганизующийся метод анализа данных). Пример.

Шаг 6

Вычисляется расстояние $\overline{D}$:

$$\overline{D} = \overline{D}_1 = 2,26.$$

Шаг 7

Так как данный цикл итераций не последний и $N_c = K/2$, то осуществляется переход к **шагу 8**.

Шаг 8

Вычисляется вектор среднеквадратичного отклонения для заданного S_1 :

$$\sigma_1 = \begin{pmatrix} 1,99 \\ 1,56 \end{pmatrix}.$$

Шаг 9

Максимальная компонента вектора σ_1 равна 1,99, следовательно $\sigma_{1\max} = 1,99$.

Простейшие методы кластеризации

Iterative Self-Organizing Data Analysis Techniques

ИСОМАД (Итеративный самоорганизующийся метод анализа данных). Пример.

Шаг 6

Вычисляется расстояние $\bar{D}$:

$$\bar{D} = \bar{D}_1 = 2, 26.$$

Шаг 7

Так как данный цикл итераций не последний и $N_c = K/2$, то осуществляется переход к **шагу 8**.

Шаг 8

Вычисляется вектор среднеквадратичного отклонения для заданного S_1 :

$$\sigma_1 = \begin{pmatrix} 1, 99 \\ 1, 56 \end{pmatrix}.$$

Шаг 9

Максимальная компонента вектора σ_1 равна 1, 99, следовательно $\sigma_{1 \max} = 1, 99$.

Шаг 10

Так как $\sigma_{1 \max} > \theta_s$ и $N_c = K/2$, то кластер с центром z_1 расщепляется на два новых кластера. Выбираем $\gamma_j = 0, 5\sigma_{1 \max} \approx 1, 0$. Тогда

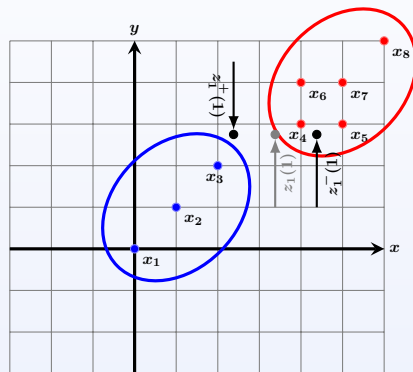
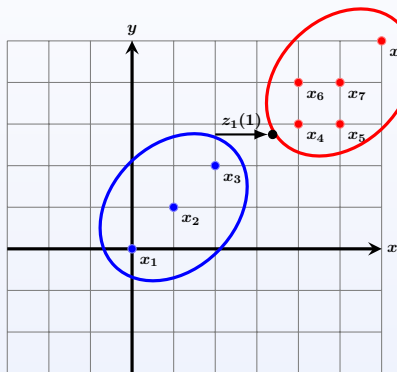
$$z_1 = z_1^+ = \begin{pmatrix} 4, 38 \\ 2, 75 \end{pmatrix}, \quad z_2 = z_1^- = \begin{pmatrix} 2, 38 \\ 2, 75 \end{pmatrix}.$$

Значение N_c увеличивается на 1. Переход к **шагу 2**.

Простейшие методы кластеризации

Iterative Self-Organizing Data Analysis Techniques

ИСОМАД (Итеративный самоорганизующийся метод анализа данных). Пример.





Простейшие методы кластеризации

Iterative Self-Organizing Data Analysis Techniques

ИСОМАД (Итеративный самоорганизующийся метод анализа данных). Пример.

Шаг 2

$$S_1 = \{x_4, x_5, x_6, x_7, \dots, x_8\} \quad \text{и} \quad N_1 = 5,$$

$$S_2 = \{x_1, x_2, x_3\} \quad \text{и} \quad N_2 = 3.$$



Простейшие методы кластеризации

Iterative Self-Organizing Data Analysis Techniques

ИСОМАД (Итеративный самоорганизующийся метод анализа данных). Пример.

Шаг 2

$$S_1 = \{x_4, x_5, x_6, x_7, \dots, x_8\} \quad \text{и} \quad N_1 = 5,$$

$$S_2 = \{x_1, x_2, x_3\} \quad \text{и} \quad N_2 = 3.$$

Шаг 3

Так как $N_1 > \theta_N$ и $N_2 > \theta_N$, то ни одно подмножество не ликвидируется.



Простейшие методы кластеризации

Iterative Self-Organizing Data Analysis Techniques

ИСОМАД (Итеративный самоорганизующийся метод анализа данных). Пример.

Шаг 2

$$S_1 = \{x_4, x_5, x_6, x_7, \dots, x_8\} \quad \text{и} \quad N_1 = 5,$$

$$S_2 = \{x_1, x_2, x_3\} \quad \text{и} \quad N_2 = 3.$$

Шаг 3

Так как $N_1 > \theta_N$ и $N_2 > \theta_N$, то ни одно подмножество не ликвидируется.

Шаг 4

Корректируются положения центров кластеров:

$$z_1 = \frac{1}{N_1} \sum_{x \in S_1} x = \begin{pmatrix} 4,80 \\ 3,80 \end{pmatrix}, \quad z_2 = \frac{1}{N_2} \sum_{x \in S_2} x = \begin{pmatrix} 1,00 \\ 1,00 \end{pmatrix}.$$



Простейшие методы кластеризации

Iterative Self-Organizing Data Analysis Techniques

ИСОМАД (Итеративный самоорганизующийся метод анализа данных). Пример.

Шаг 2

$$S_1 = \{x_4, x_5, x_6, x_7, \dots, x_8\} \text{ и } N_1 = 5,$$

$$S_2 = \{x_1, x_2, x_3\} \text{ и } N_2 = 3.$$

Шаг 3

Так как $N_1 > \theta_N$ и $N_2 > \theta_N$, то ни одно подмножество не ликвидируется.

Шаг 4

Корректируются положения центров кластеров:

$$z_1 = \frac{1}{N_1} \sum_{x \in S_1} x = \begin{pmatrix} 4,80 \\ 3,80 \end{pmatrix}, \quad z_2 = \frac{1}{N_2} \sum_{x \in S_2} x = \begin{pmatrix} 1,00 \\ 1,00 \end{pmatrix}.$$

Шаг 5

Вычисляется расстояние $\overline{D}_j, j = 1, 2$:

$$\overline{D}_1 = \frac{1}{N_1} \sum_{x \in S_1} \|x - z_1\| = 0,80, \quad \overline{D}_2 = \frac{1}{N_2} \sum_{x \in S_2} \|x - z_2\| = 0,94.$$



Простейшие методы кластеризации

Iterative Self-Organizing Data Analysis Techniques

ИСОМАД (Итеративный самоорганизующийся метод анализа данных). Пример.

Шаг 6

Вычисляется расстояние $\overline{D}$:

$$\overline{D} = \frac{1}{N} \sum_{j=1}^{N_c} N_j \overline{D}_j = \frac{1}{8} \sum_{j=1}^2 N_j \overline{D}_j = 0,85.$$



Простейшие методы кластеризации

Iterative Self-Organizing Data Analysis Techniques

ИСОМАД (Итеративный самоорганизующийся метод анализа данных). Пример.

Шаг 6

Вычисляется расстояние $\overline{D}$:

$$\overline{D} = \frac{1}{N} \sum_{j=1}^{N_c} N_j \overline{D}_j = \frac{1}{8} \sum_{j=1}^2 N_j \overline{D}_j = 0,85.$$

Шаг 7

Так как данная итерация имеет четный порядок (выполняется условие **(3) шага 7**), то осуществляется переход к **шагу 11**.

Простейшие методы кластеризации

Iterative Self-Organizing Data Analysis Techniques

ИСОМАД (Итеративный самоорганизующийся метод анализа данных). Пример.

Шаг 6

Вычисляется расстояние $\overline{D}$:

$$\overline{D} = \frac{1}{N} \sum_{j=1}^{N_c} N_j \overline{D}_j = \frac{1}{8} \sum_{j=1}^2 N_j \overline{D}_j = 0,85.$$

Шаг 7

Так как данная итерация имеет четный порядок (выполняется условие **(3) шага 7**), то осуществляется переход к **шагу 11**.

Шаг 11

Вычислим расстояние между парами центров кластеров:

$$D_{12} = \|z_1 - z_2\| = 4,72.$$

Простейшие методы кластеризации

Iterative Self-Organizing Data Analysis Techniques

ИСОМАД (Итеративный самоорганизующийся метод анализа данных). Пример.

Шаг 6

Вычисляется расстояние $\bar{D}$:

$$\bar{D} = \frac{1}{N} \sum_{j=1}^{N_c} N_j \bar{D}_j = \frac{1}{8} \sum_{j=1}^2 N_j \bar{D}_j = 0,85.$$

Шаг 7

Так как данная итерация имеет четный порядок (выполняется условие **(3) шага 7)**, то осуществляется переход к **шагу 11**.

Шаг 11

Вычислим расстояние между парами центров кластеров:

$$D_{12} = \|z_1 - z_2\| = 4,72.$$

Шаг 12

Величина $D_{12} > \theta_c$.



Простейшие методы кластеризации

Iterative Self-Organizing Data Analysis Techniques

ИСОМАД (Итеративный самоорганизующийся метод анализа данных). Пример.

Шаг 13

Результаты **шага 12** показывают, что объединение кластеров не произойдет.

Простейшие методы кластеризации

Iterative Self-Organizing Data Analysis Techniques

ИСОМАД (Итеративный самоорганизующийся метод анализа данных). Пример.

Шаг 13

Результаты **шага 12** показывают, что объединение кластеров не произойдет.

Шаг 14

Данный цикл итераций не последний, следовательно нужно принять решение: вносить или не вносить изменения в параметры процесса кластеризации на **шаге 1**.

Выводы: 1. Число выделенных кластеров соответствует заданному
2. Расстояние между кластерами больше среднего разброса (характеризуется среднеквадратичным отклонением)
3. Каждый кластер содержит существенную часть образов
Делаем вывод, что переходим к **шагу 2**.

Простейшие методы кластеризации

Iterative Self-Organizing Data Analysis Techniques

ИСОМАД (Итеративный самоорганизующийся метод анализа данных). Пример.

Шаг 13

Результаты **шага 12** показывают, что объединение кластеров не произойдет.

Шаг 14

Данный цикл итераций не последний, следовательно нужно принять решение: вносить или не вносить изменения в параметры процесса кластеризации на **шаге 1**.

Выводы: 1. Число выделенных кластеров соответствует заданному
2. Расстояние между кластерами больше среднего разброса (характеризуется среднеквадратичным отклонением)
3. Каждый кластер содержит существенную часть образов
Делаем вывод, что переходим к **шагу 2**.

Шаг 2-6

Те же результаты, что и в предыдущей итерации.

Простейшие методы кластеризации

Iterative Self-Organizing Data Analysis Techniques

ИСОМАД (Итеративный самоорганизующийся метод анализа данных). Пример.

Шаг 13

Результаты **шага 12** показывают, что объединение кластеров не произойдет.

Шаг 14

Данный цикл итераций не последний, следовательно нужно принять решение: вносить или не вносить изменения в параметры процесса кластеризации на **шаге 1**.

Выводы: 1. Число выделенных кластеров соответствует заданному
2. Расстояние между кластерами больше среднего разброса (характеризуется среднеквадратичным отклонением)
3. Каждый кластер содержит существенную часть образов
Делаем вывод, что переходим к **шагу 2**.

Шаг 2-6

Те же результаты, что и в предыдущей итерации.

Шаг 7

Ни одно из условий (1)-(3) не выполнено. Следовательно, переходим к **шагу 8**.



Простейшие методы кластеризации

Iterative Self-Organizing Data Analysis Techniques

ИСОМАД (Итеративный самоорганизующийся метод анализа данных). Пример.

Шаг 8

Для множеств S_1 и S_2

$$\sigma_1 = \begin{pmatrix} 0,75 \\ 0,75 \end{pmatrix}, \quad \sigma_2 = \begin{pmatrix} 0,82 \\ 0,82 \end{pmatrix}.$$



Простейшие методы кластеризации

Iterative Self-Organizing Data Analysis Techniques

ИСОМАД (Итеративный самоорганизующийся метод анализа данных). Пример.

Шаг 8

Для множеств S_1 и S_2

$$\sigma_1 = \begin{pmatrix} 0,75 \\ 0,75 \end{pmatrix}, \quad \sigma_2 = \begin{pmatrix} 0,82 \\ 0,82 \end{pmatrix}.$$

Шаг 9

Максимальные компоненты векторов σ_1 и σ_2 :

$$\sigma_{1 \max} = 0,75 \quad \text{и} \quad \sigma_{2 \max} = 0,82.$$



Простейшие методы кластеризации

Iterative Self-Organizing Data Analysis Techniques

ИСОМАД (Итеративный самоорганизующийся метод анализа данных). Пример.

Шаг 8

Для множеств S_1 и S_2

$$\sigma_1 = \begin{pmatrix} 0,75 \\ 0,75 \end{pmatrix}, \quad \sigma_2 = \begin{pmatrix} 0,82 \\ 0,82 \end{pmatrix}.$$

Шаг 9

Максимальные компоненты векторов σ_1 и σ_2 :

$$\sigma_{1 \max} = 0,75 \quad \text{и} \quad \sigma_{2 \max} = 0,82.$$

Шаг 10

Условие расщепление кластеров не выполняется. Переходим к **шагу 11**.

Простейшие методы кластеризации

Iterative Self-Organizing Data Analysis Techniques

ИСОМАД (Итеративный самоорганизующийся метод анализа данных). Пример.

Шаг 8

Для множеств S_1 и S_2

$$\sigma_1 = \begin{pmatrix} 0,75 \\ 0,75 \end{pmatrix}, \quad \sigma_2 = \begin{pmatrix} 0,82 \\ 0,82 \end{pmatrix}.$$

Шаг 9

Максимальные компоненты векторов σ_1 и σ_2 :

$$\sigma_{1 \max} = 0,75 \quad \text{и} \quad \sigma_{2 \max} = 0,82.$$

Шаг 10

Условие расщепления кластеров не выполняется. Переходим к шагу 11.

Шаг 11

Вычислим расстояние между парами центров кластеров:

$$D_{12} = \|z_1 - z_2\| = 4,72.$$



Простейшие методы кластеризации

Iterative Self-Organizing Data Analysis Techniques

ИСОМАД (Итеративный самоорганизующийся метод анализа данных). Пример.

Шаг 12-13

Результаты идентичные предыдущему циклу.



Простейшие методы кластеризации

Iterative Self-Organizing Data Analysis Techniques

ИСОМАД (Итеративный самоорганизующийся метод анализа данных). Пример.

Шаг 12-13

Результаты идентичные предыдущему циклу.

Шаг 14

Данный цикл интераций не последний, следовательно нужно принять решение: вносить или не вносить изменения в параметры процесса кластеризации на **шаге 1**.

Новые результаты были получены только для векторов среднеквадратичного отклонения. Следовательно, переходим к **шагу 2**.



Простейшие методы кластеризации

Iterative Self-Organizing Data Analysis Techniques

ИСОМАД (Итеративный самоорганизующийся метод анализа данных). Пример.

Шаг 12-13

Результаты идентичные предыдущему циклу.

Шаг 14

Данный цикл интераций не последний, следовательно нужно принять решение: вносить или не вносить изменения в параметры процесса кластеризации на **шаге 1**.

Новые результаты были получены только для векторов среднеквадратичного отклонения. Следовательно, переходим к **шагу 2**.

Шаг 2-6

Те же результаты, что и в предыдущей итерации.



Простейшие методы кластеризации

Iterative Self-Organizing Data Analysis Techniques

ИСОМАД (Итеративный самоорганизующийся метод анализа данных). Пример.

Шаг 12-13

Результаты идентичные предыдущему циклу.

Шаг 14

Данный цикл итераций не последний, следовательно нужно принять решение: вносить или не вносить изменения в параметры процесса кластеризации на **шаге 1**.

Новые результаты были получены только для векторов среднеквадратичного отклонения. Следовательно, переходим к **шагу 2**.

Шаг 2-6

Те же результаты, что и в предыдущей итерации.

Шаг 7

Данный цикл итераций – последний, задаем $\theta_c = 0$ и переходим к **шагу 11**.

Простейшие методы кластеризации

Iterative Self-Organizing Data Analysis Techniques

ИСОМАД (Итеративный самоорганизующийся метод анализа данных). Пример.

Шаг 12-13

Результаты идентичные предыдущему циклу.

Шаг 14

Данный цикл итераций не последний, следовательно нужно принять решение: вносить или не вносить изменения в параметры процесса кластеризации на **шаге 1**.

Новые результаты были получены только для векторов среднеквадратичного отклонения. Следовательно, переходим к **шагу 2**.

Шаг 2-6

Те же результаты, что и в предыдущей итерации.

Шаг 7

Данный цикл итераций – последний, задаем $\theta_c = 0$ и переходим к **шагу 11**.

Шаг 11

Вычислим расстояние между парами центров кластеров: $D_{12} = ||z_1 - z_2|| = 4,72$.

Простейшие методы кластеризации

Iterative Self-Organizing Data Analysis Techniques

ИСОМАД (Итеративный самоорганизующийся метод анализа данных). Пример.

Шаг 12-13

Результаты идентичные предыдущему циклу.

Шаг 14

Данный цикл итераций не последний, следовательно нужно принять решение: вносить или не вносить изменения в параметры процесса кластеризации на **шаге 1**.

Новые результаты были получены только для векторов среднеквадратичного отклонения. Следовательно, переходим к **шагу 2**.

Шаг 2-6

Те же результаты, что и в предыдущей итерации.

Шаг 7

Данный цикл итераций – последний, задаем $\theta_c = 0$ и переходим к **шагу 11**.

Шаг 11

Вычислим расстояние между парами центров кластеров: $D_{12} = ||z_1 - z_2|| = 4,72$.

Шаг 12

Результаты идентичны предыдущему циклу.

Простейшие методы кластеризации

Iterative Self-Organizing Data Analysis Techniques

ИСОМАД (Итеративный самоорганизующийся метод анализа данных). Пример.

Шаг 13

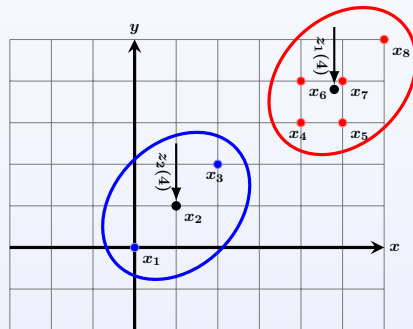
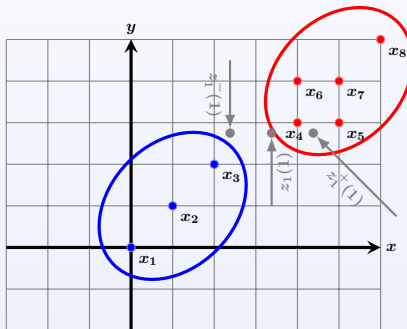


Шаг 14



Результаты **шага 12** показывают, что объединение кластеров не произойдет.

Данный цикл итераций последний. Алгоритм заканчивает свою работу.





Простейшие методы кластеризации

Полезные ссылки по простейшим методам кластеризации (центроидным)

- <https://www.youtube.com/watch?v=BCiZc0n6COY&list=PLruBu5BI5n4aFpG32iMbdWoRVAA-Vcso6&index=1>
- <https://inference.org.uk/mackay/Book.html>
- <https://link.springer.com/book/10.1007/978-0-387-84858-7>
- <https://scikit-learn.org/stable/index.html>