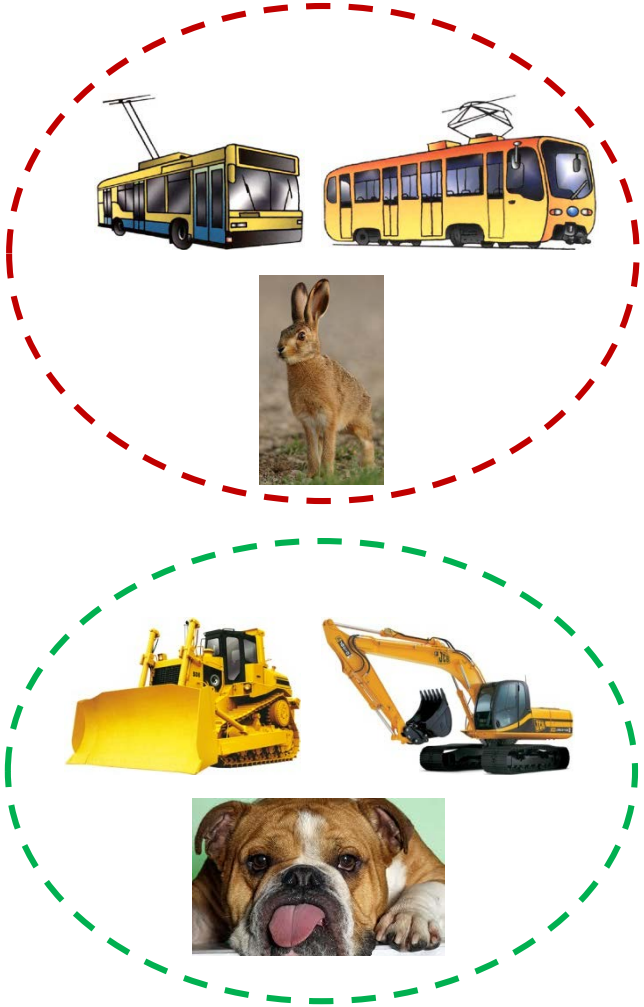
Задача кластеризации данных



Группа людей, действуя совместно, может свершить такое, о чем поодиночке они не могли бы и мечтать.

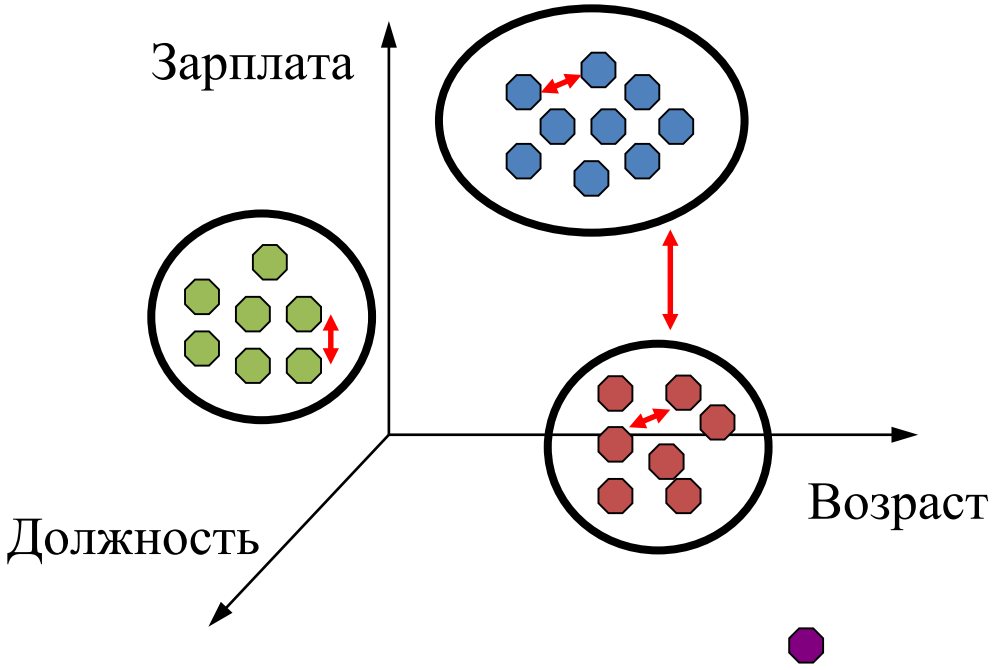
Франклин Рузвельт

Содержание

- **Основные концепции**
- **Разделительная кластеризация**
- Иерархическая кластеризация
- Меры качества кластеризации

Кластеризация

- Нахождение заранее неизвестных групп (*кластеров*) в множестве однотипных объектов, где объекты в одной группе существенно похожи, а объекты разных групп существенно отличны



Применение кластеризации

- Понимание данных
  - Биология: иерархии живых организмов
  - Землепользование: нахождение в базе наблюдений Земли сходных территорий
  - Маркетинг: таргетирование клиентов
  - Городское планирование: группы похожих зданий
  - Изучение землетрясений: кластеризация эпицентров
- Предобработка данных
  - Редукция данных: замена группы объектов их центроидом
  - Удаление выбросов: нахождение объектов, наиболее удаленных от всех кластеров
  - Восстановление пропущенных данных: использование координат центроидов
  - Нахождение ближайших соседей: поиск среди объектов того же кластера

Неоднозначность кластеризации: число групп

Сколько кластеров?

6

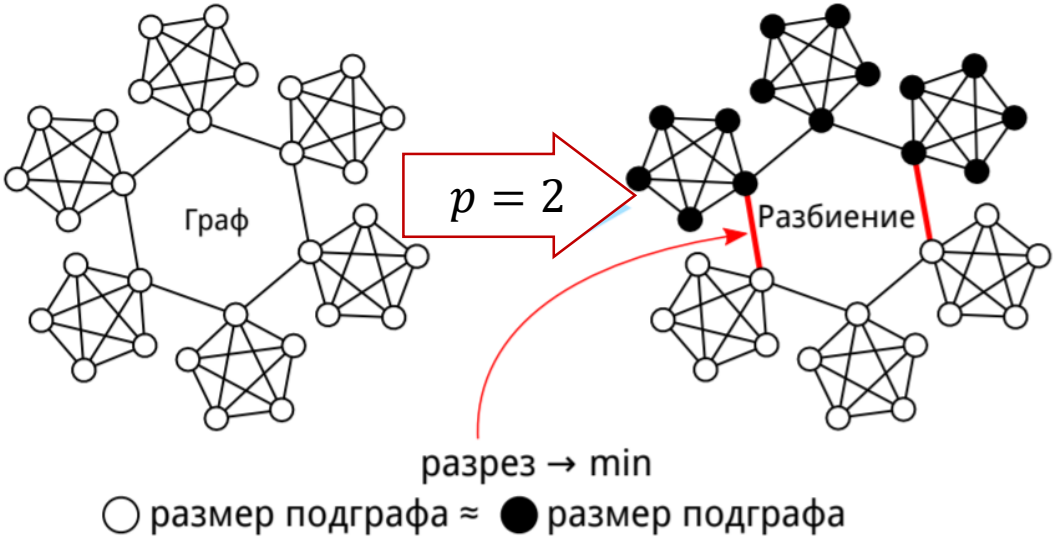
2

4

Неоднозначность кластеризации

- Выбор функции расстояния (или меры схожести)
- Размерность данных
- Типы атрибутов
- Зависимости между атрибутами
- Распределение данных
- Наличие шумов и выбросов в данных

Кластеризация как задача из другой предметной области



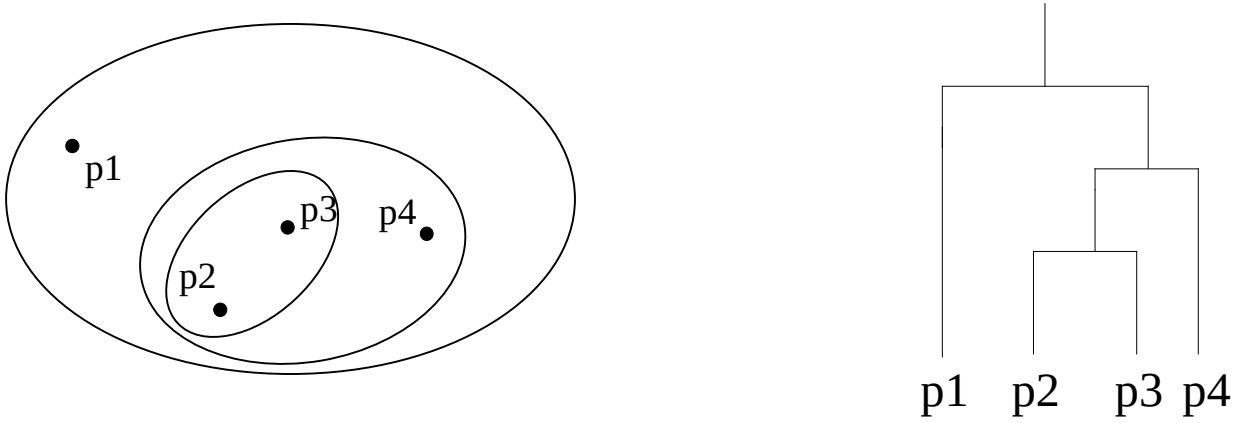
Граф $G(N, E, w)$

1. $N = \cup_{i=1}^{p} N_i,\ \forall i \neq j\ N_i \cap N_j = \emptyset,\ p > 1$
2. $w(N_i) \approx \frac{w(N)}{p}\ \ \forall i \in \{1, \dots, p\}$
3. $W_{cut} \to \min, W_{cut} = \sum_{e \in E_{cut}} w(e),$
   $E_{cut} = \{(u, v) \in E \mid u \in N_i, v \in N_j, 1 \leq i, j \leq p, i \neq j\}$

Базовые подходы: разделительная *vs.* иерархическая кластеризация

- Разделение объектов на непересекающиеся подмножества, каждый объект строго в одном подмножестве

- Иерархическое дерево пересекающихся подмножеств объектов

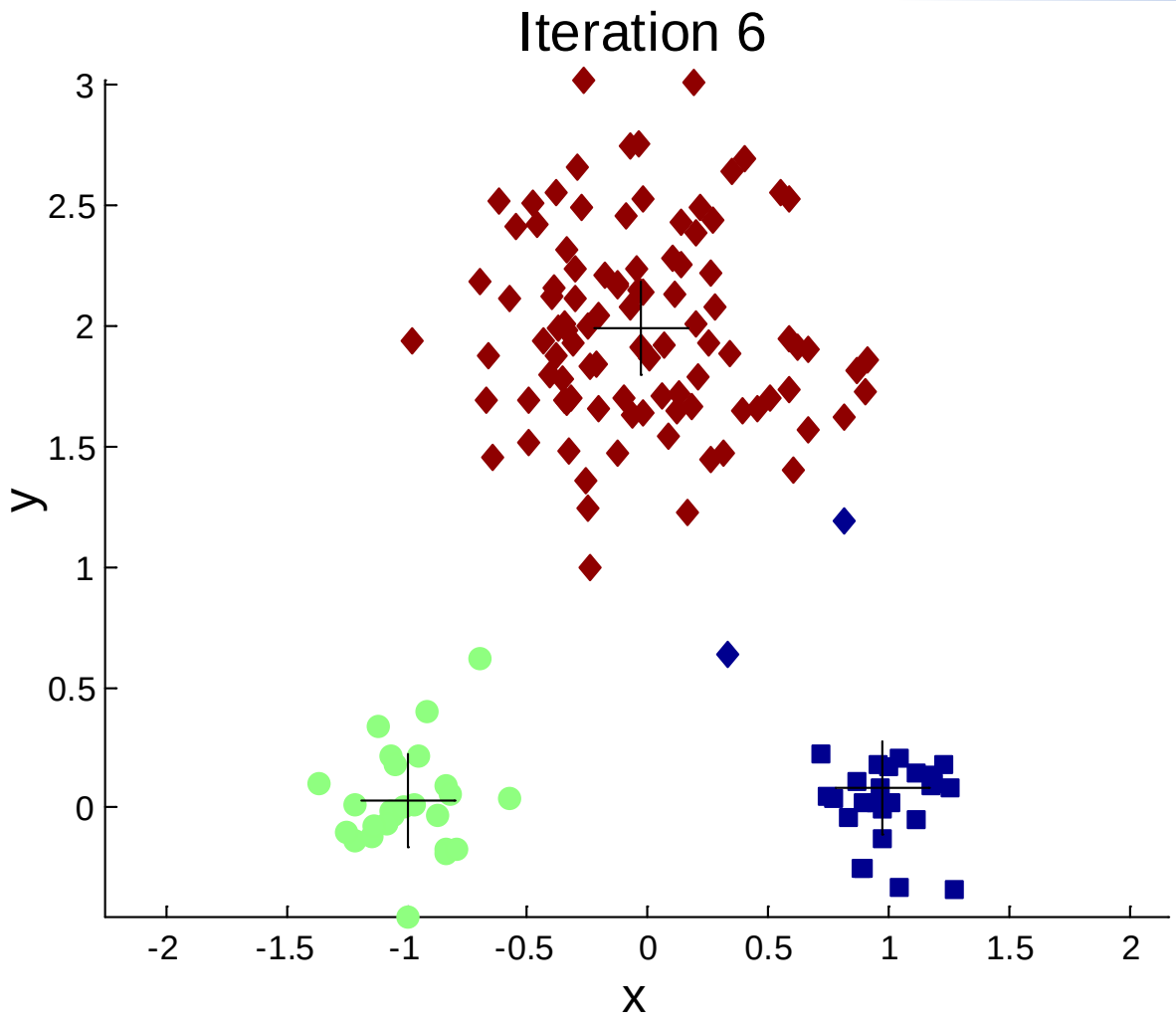


Разделительный алгоритм *k*-means (*k*-средних)

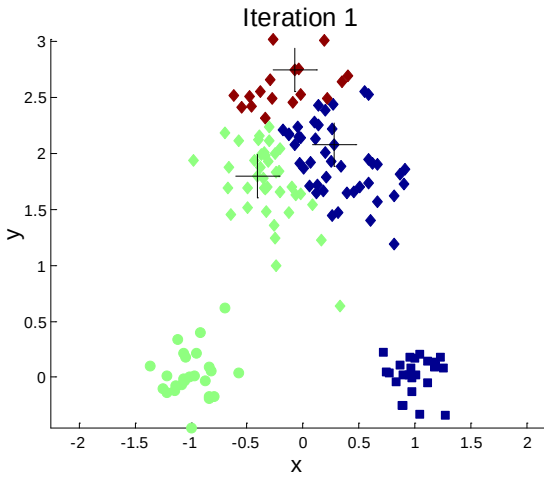
- Количество кластеров k – параметр алгоритма
- Кластер ассоциируется с его *центроидом* (центральной точкой)
- Объект принадлежит кластеру с ближайшим к нему центроидом

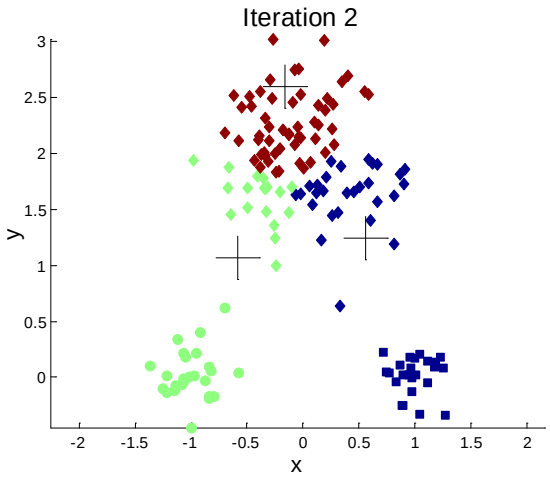
1: Select K points as the initial centroids.
2: **repeat**
3: Form K clusters by assigning all points to the closest centroid.
4: Recompute the centroid of each cluster.
5: **until** The centroids don't change

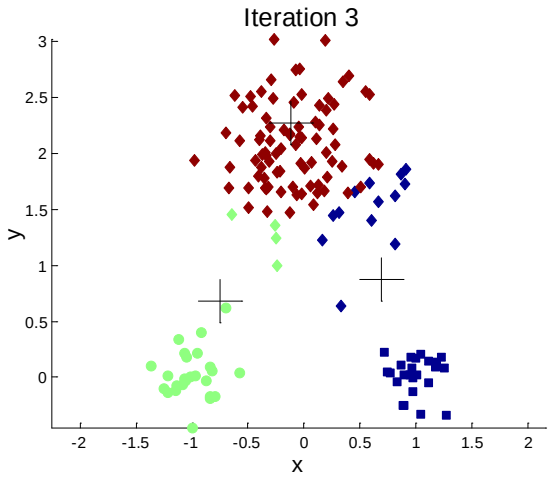
Пример работы *k*-means

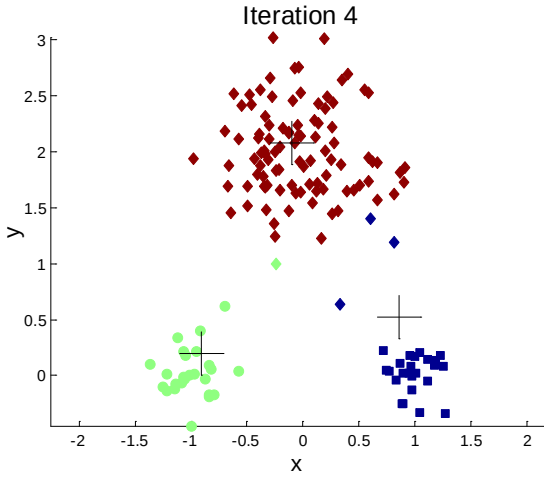


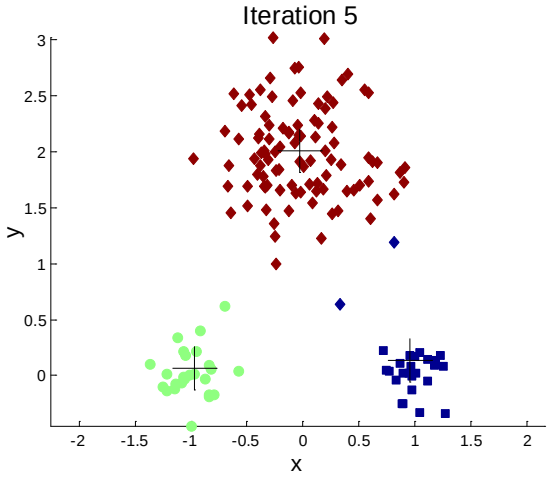
Пример работы *k*-means

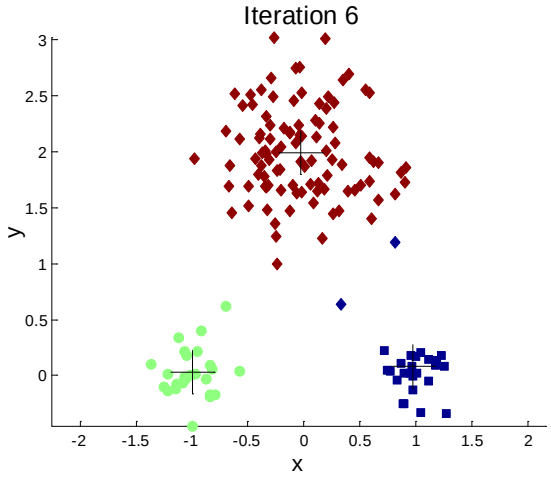












Детали алгоритма *k*-means

- Начальные центроиды берутся случайным образом. Результат кластеризации недетерминирован
- Центроид – обычно точка с усредненными координатами точек кластера
- Алгоритм сходится в начальных итерациях для общепринятых метрик
  - часто изменяется условие останова: у малого числа точек изменен кластер
  - может давать локальный минимум вместо глобального

Мера для выявления кластеров в k-means

- Сумма квадратов ошибок, Sum of Squared Errors

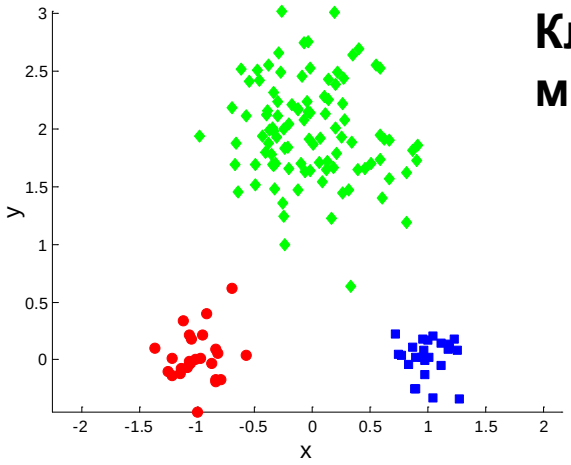
$$SSE = \sum_{i=1}^{k} \sum_{x \in C_i} dist^2(m_i, x)$$

  - x – точка кластера C_i $(1 < i < k)$
  - m_i – репрезентативная точка (центр) кластера
  - для точки ошибкой является расстояние до ближайшего кластера
- Из двух вариантов кластеризации выбирается имеющий меньшую суммарную ошибку
  - Увеличение k уменьшает ошибку. Хороший вариант кластеризации с меньшим k может иметь меньшее SSE, чем плохой вариант с б*о*льшим k

За и против k-means

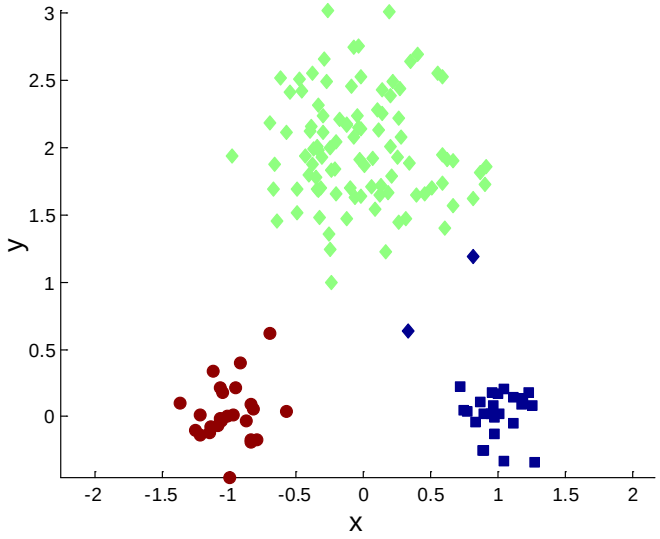
- Достоинства
  - Невысокая сложность: $O(n \cdot k \cdot d \cdot i)$, где n – мощность множества объектов, d – размерность объекта, i – число итераций (обычно $i \ll n$)
  - Сходится в начальных итерациях для общепринятых метрик
- Недостатки
  - Необходимость в задании параметра k
  - Недетерминированный результат
  - Неприменимость к категориальным данным (для них нужно использовать k-modes)
  - Неприменимость для кластеров невыпуклой формы
  - Чувствительность к размеру, плотности, шумам и выбросам в данных

Различные варианты кластеризации *k*-means

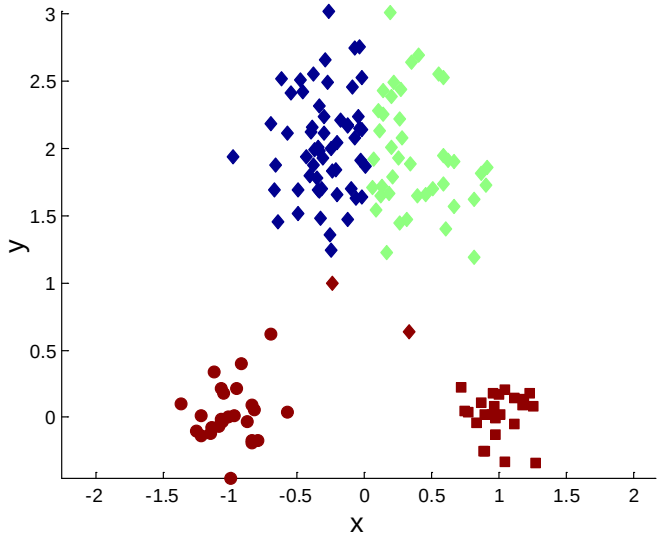


Кластеризуемое множество

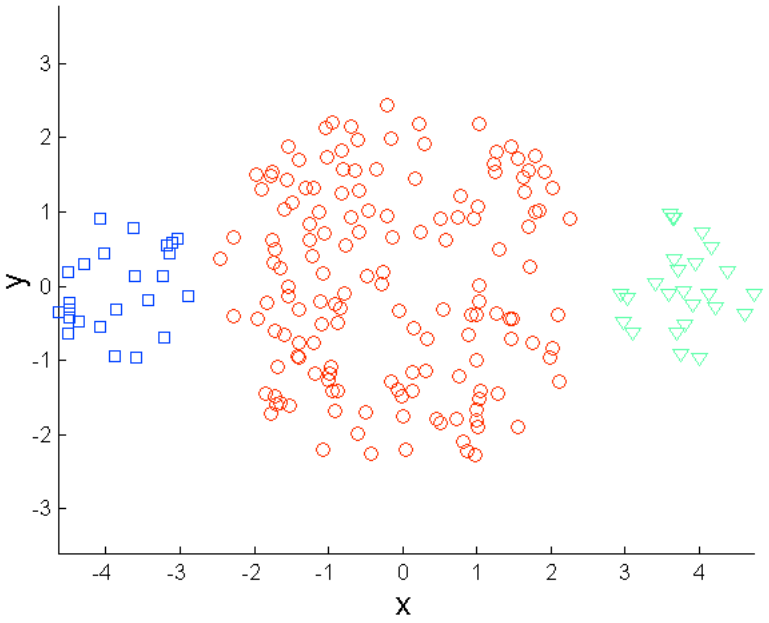
Оптимальная кластеризация



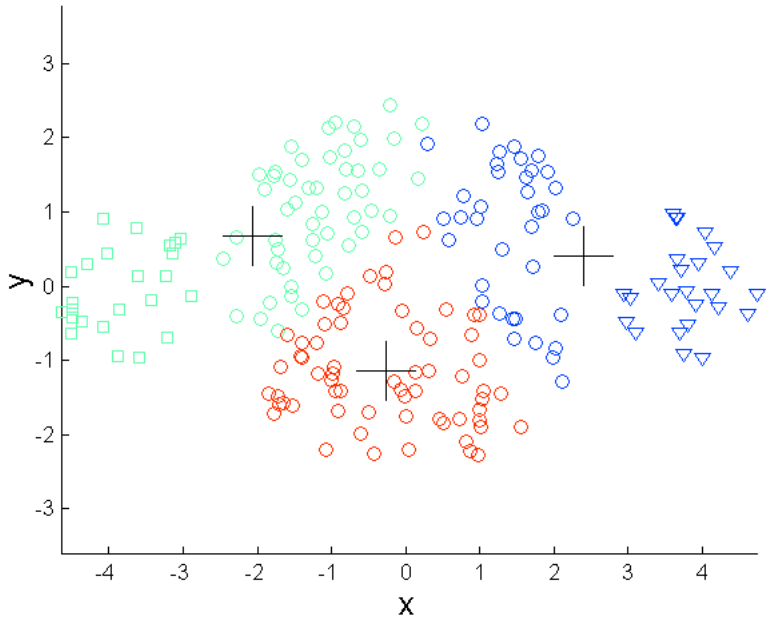
Суб-оптимальная кластеризация



Влияние размеров кластеров

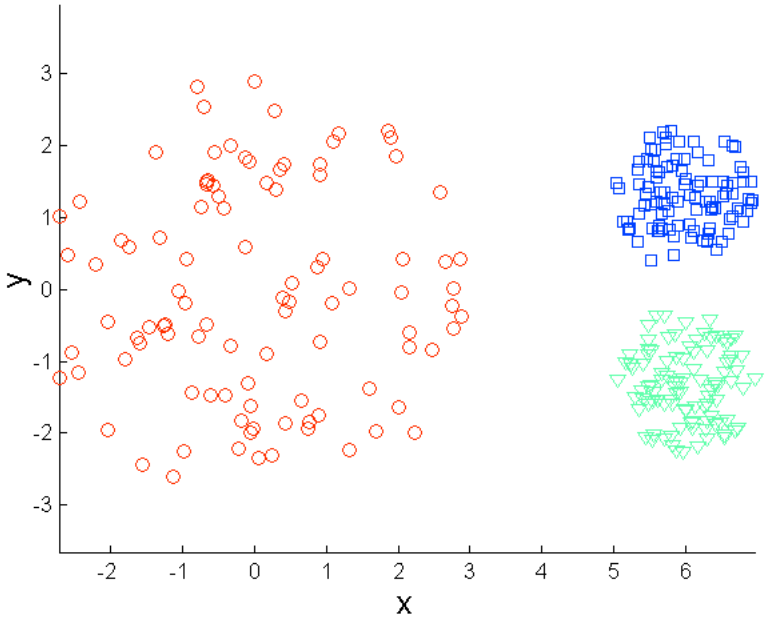


Кластеризуемое множество

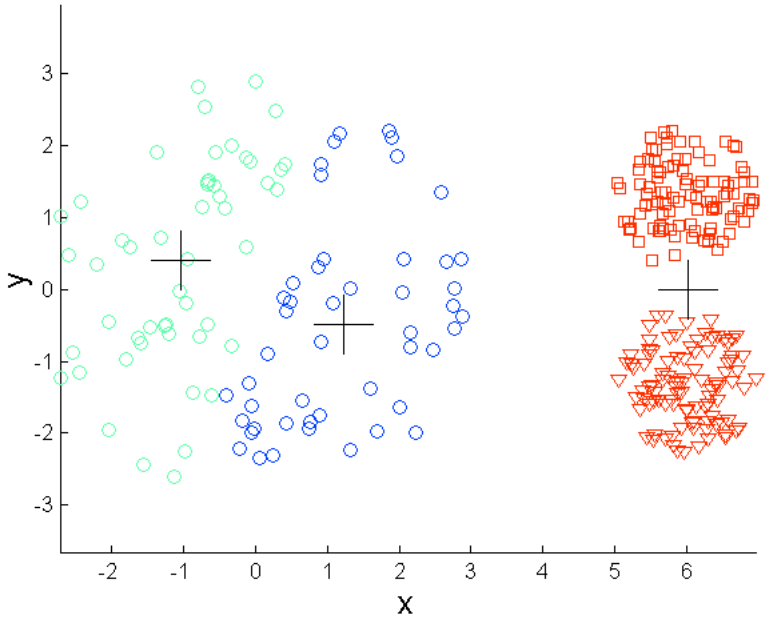


3-means

Влияние плотности кластеров

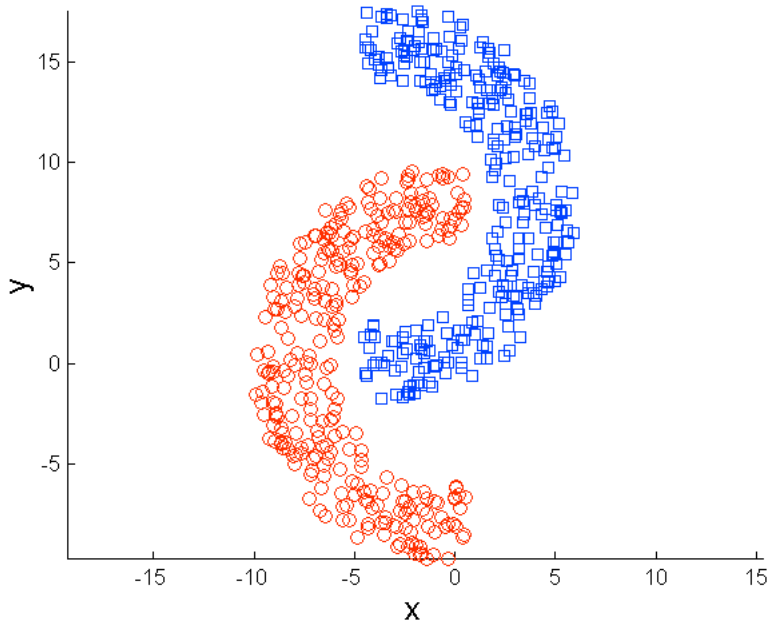


Кластеризуемое множество

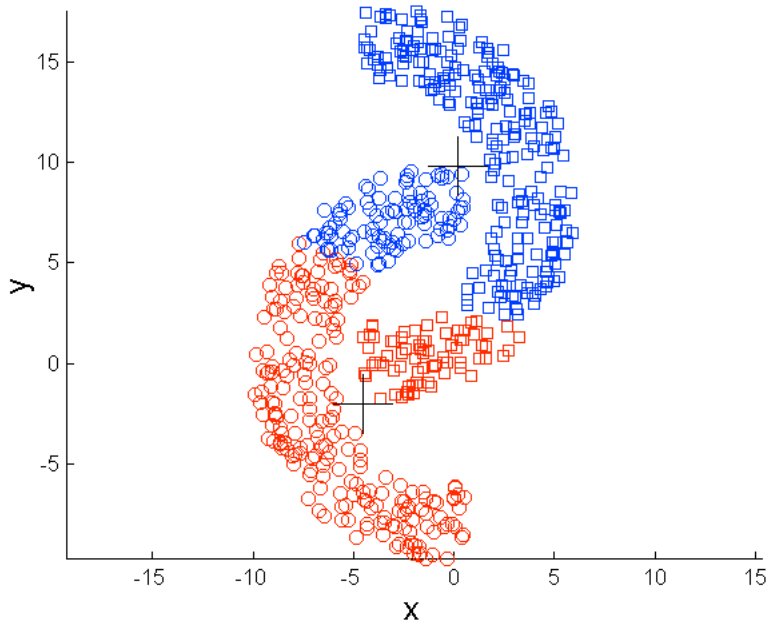


3-means

Влияние формы кластеров

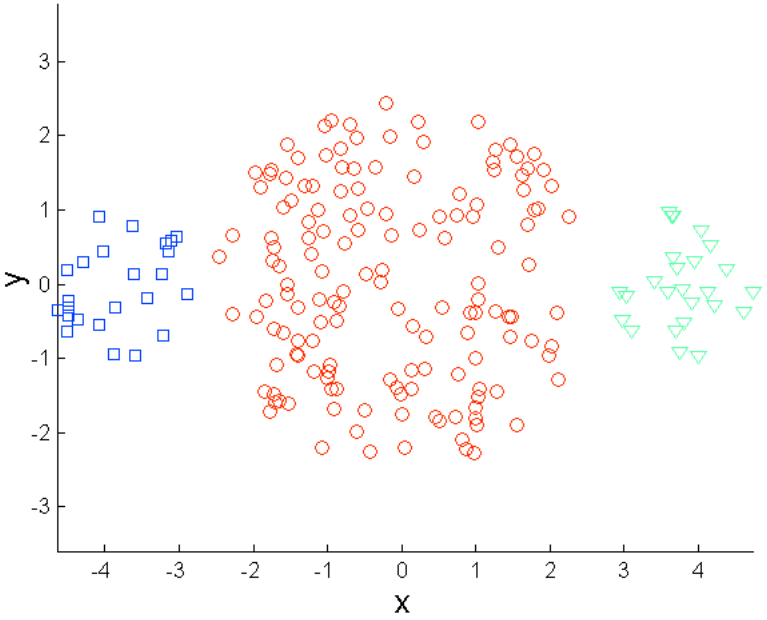


Кластеризуемое множество

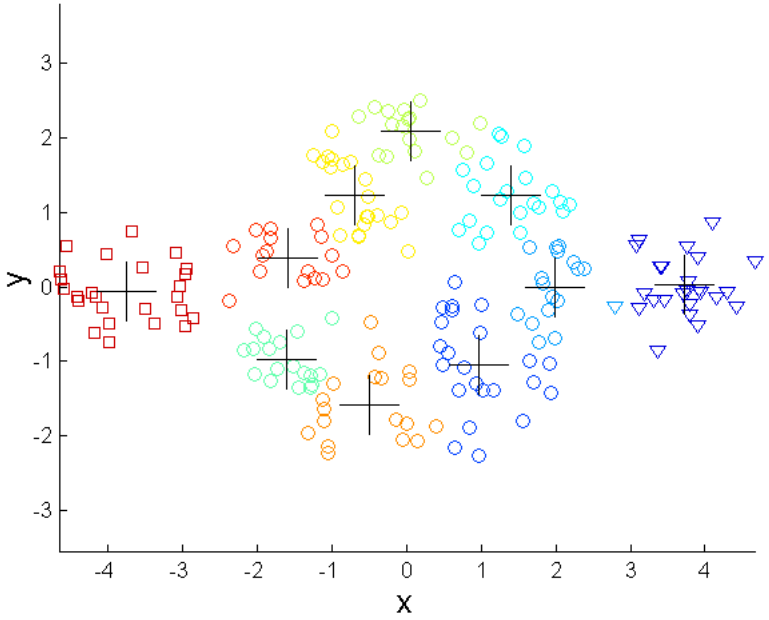


2-means

Увеличение количества кластеров



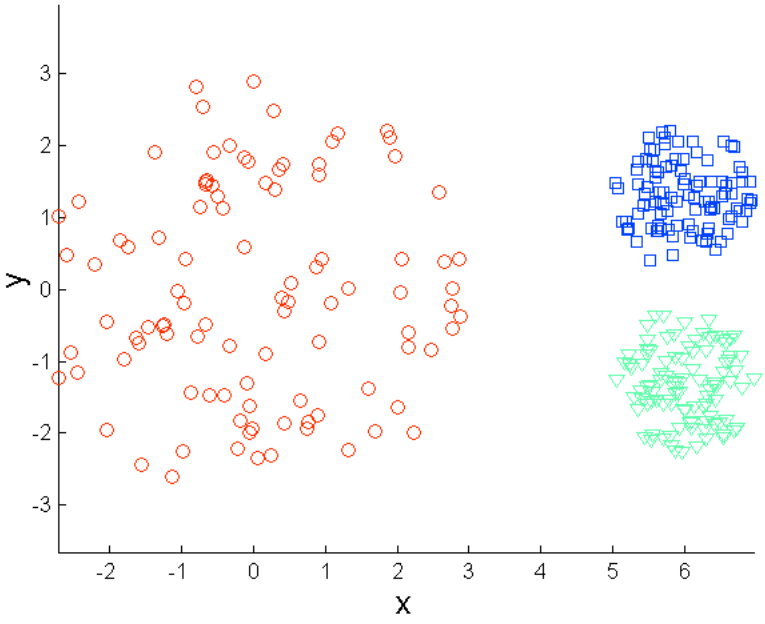
Кластеризуемое множество



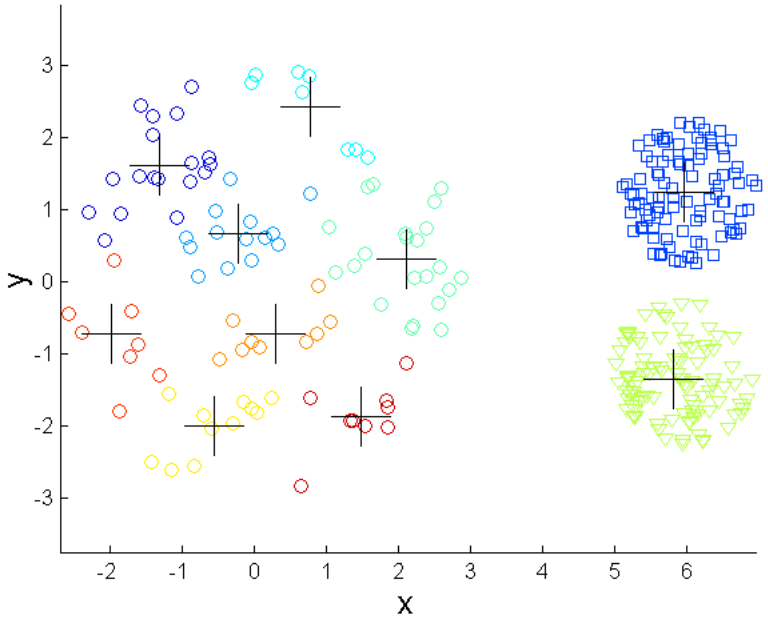
10-means

- Может помочь, но требует дальнейшего объединения некоторых полученных кластеров (например, с помощью иерархической кластеризации)

Увеличение количества кластеров



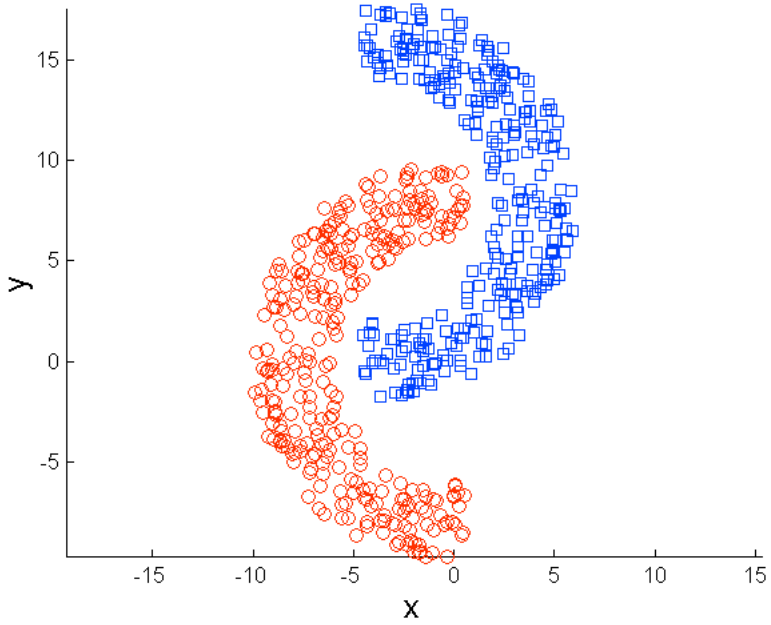
Кластеризуемое множество



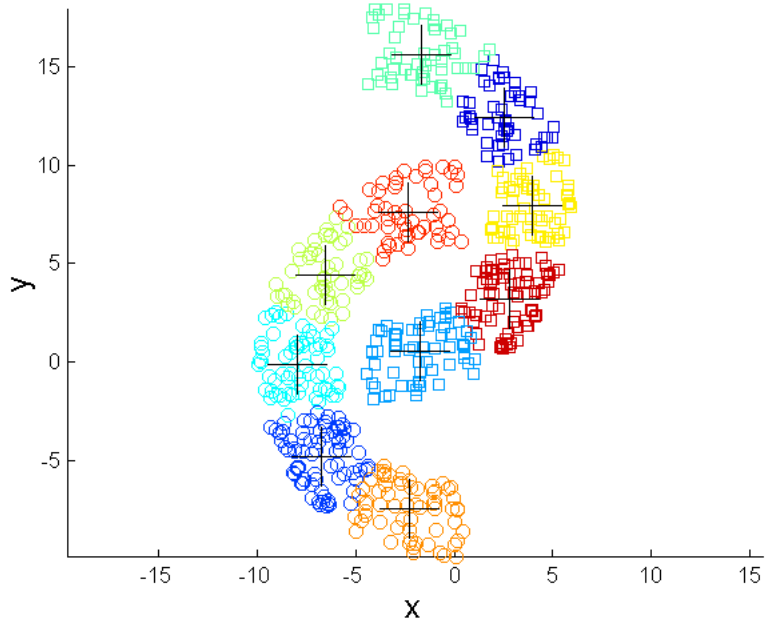
10-means

- Может помочь, но требует дальнейшего объединения некоторых полученных кластеров (например, с помощью иерархической кластеризации)

Увеличение количества кластеров



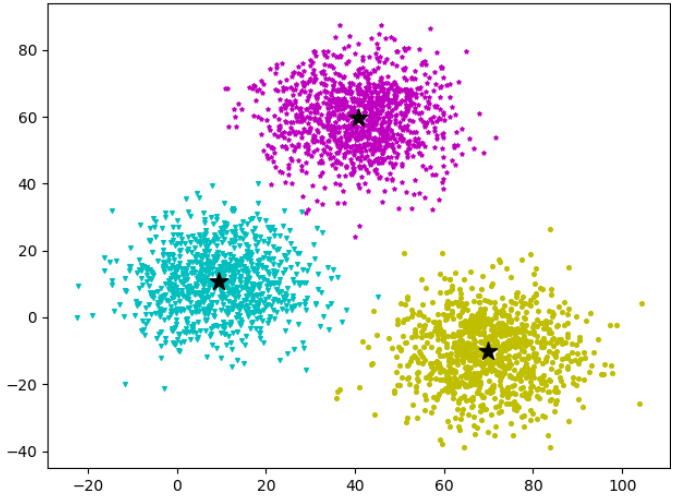
Кластеризуемое множество



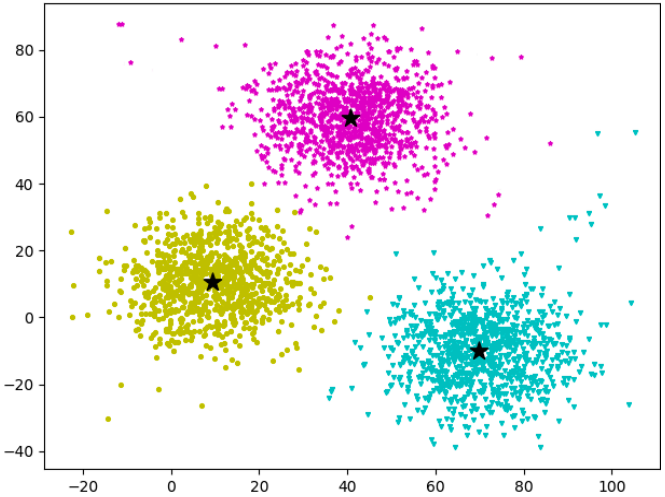
10-means

- Может помочь, но требует дальнейшего объединения некоторых полученных кластеров (например, с помощью иерархической кластеризации)

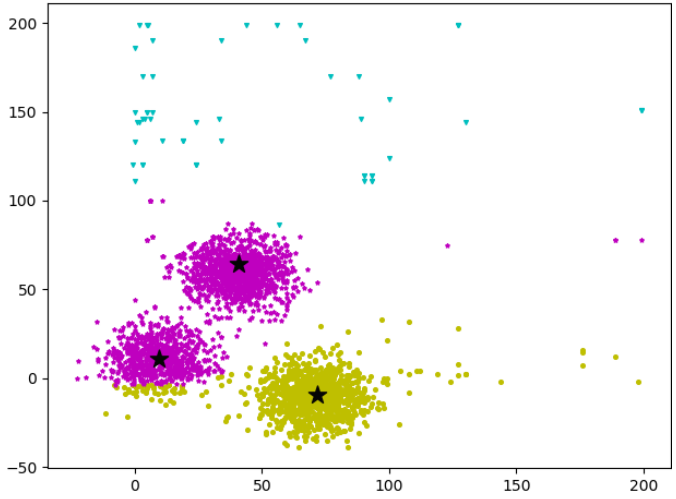
Влияние шумов и выбросов



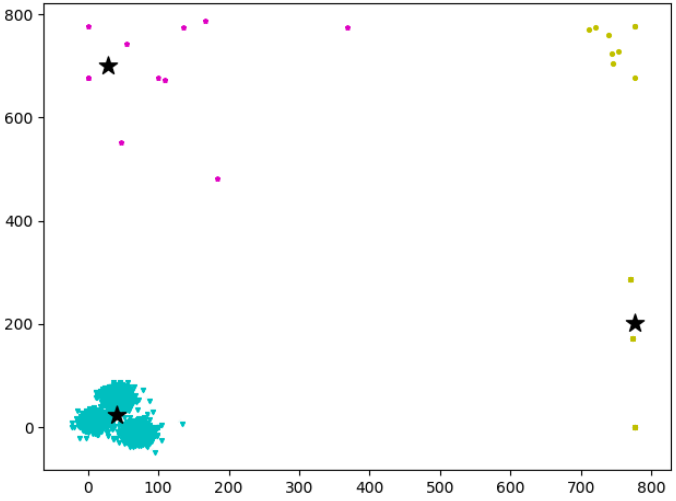
В данных
нет шума
S=0.88



В 3% данных
есть шум
S=0.62

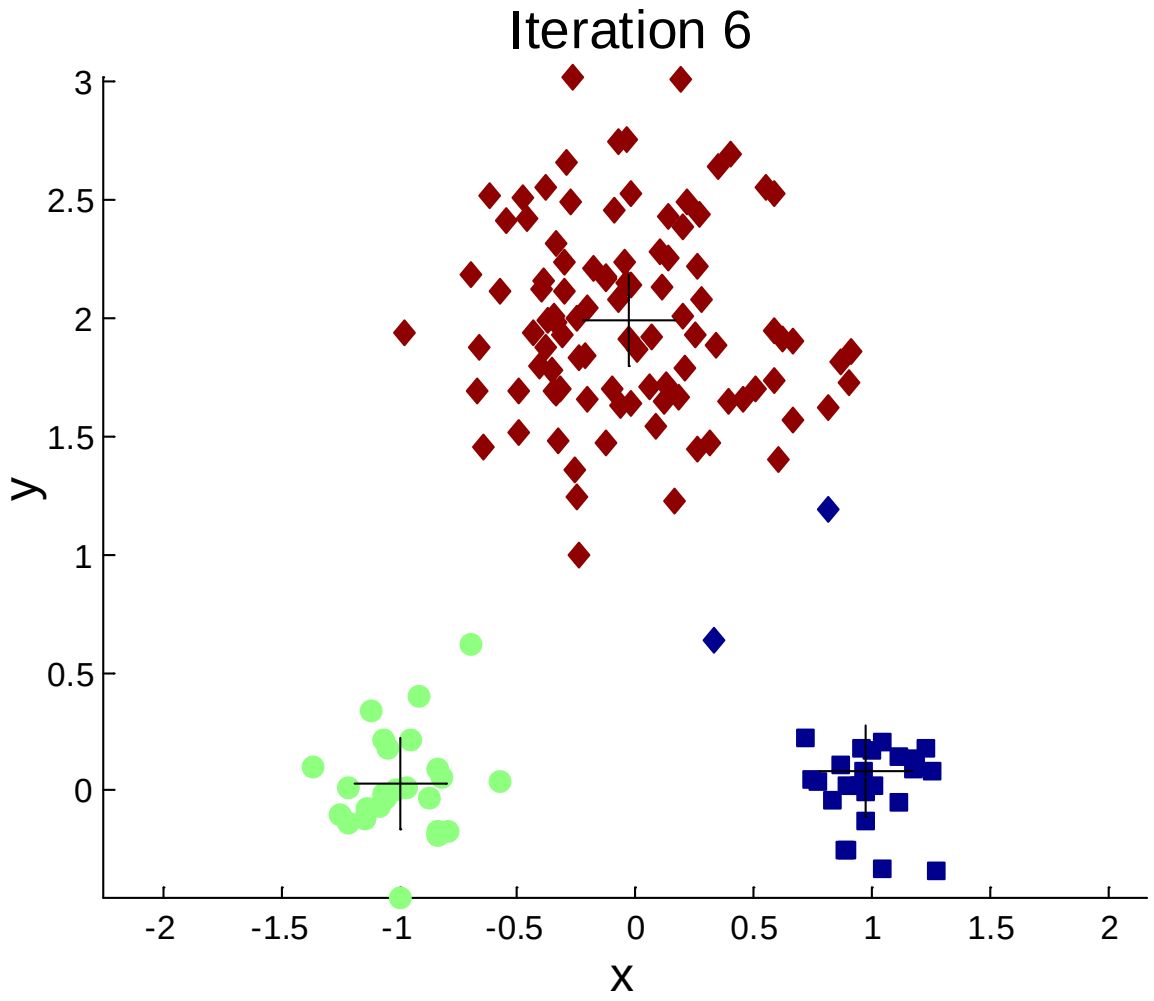


В 5% данных
есть шум
S=0.39

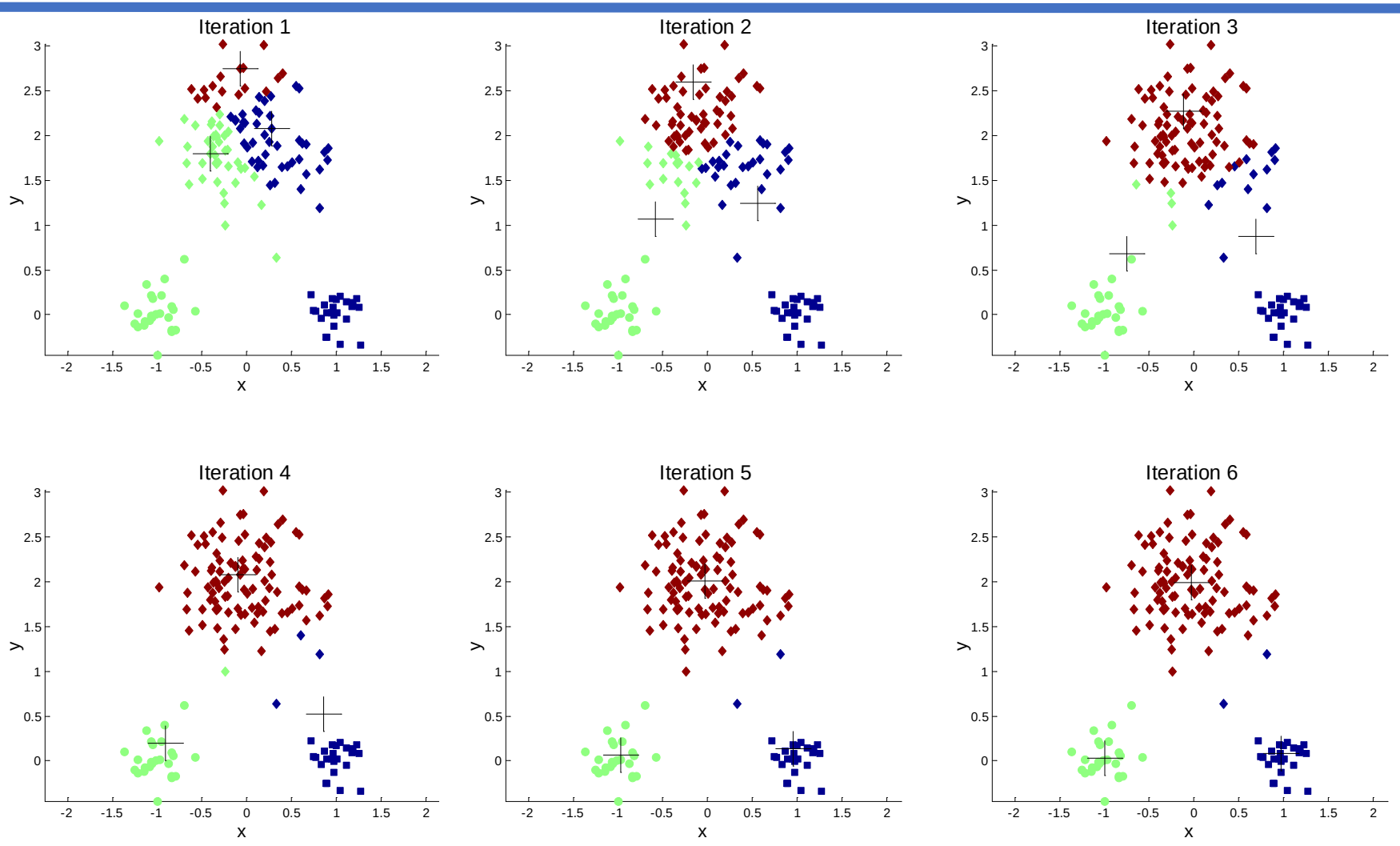


В 10% данных
есть шум
S=-0.05

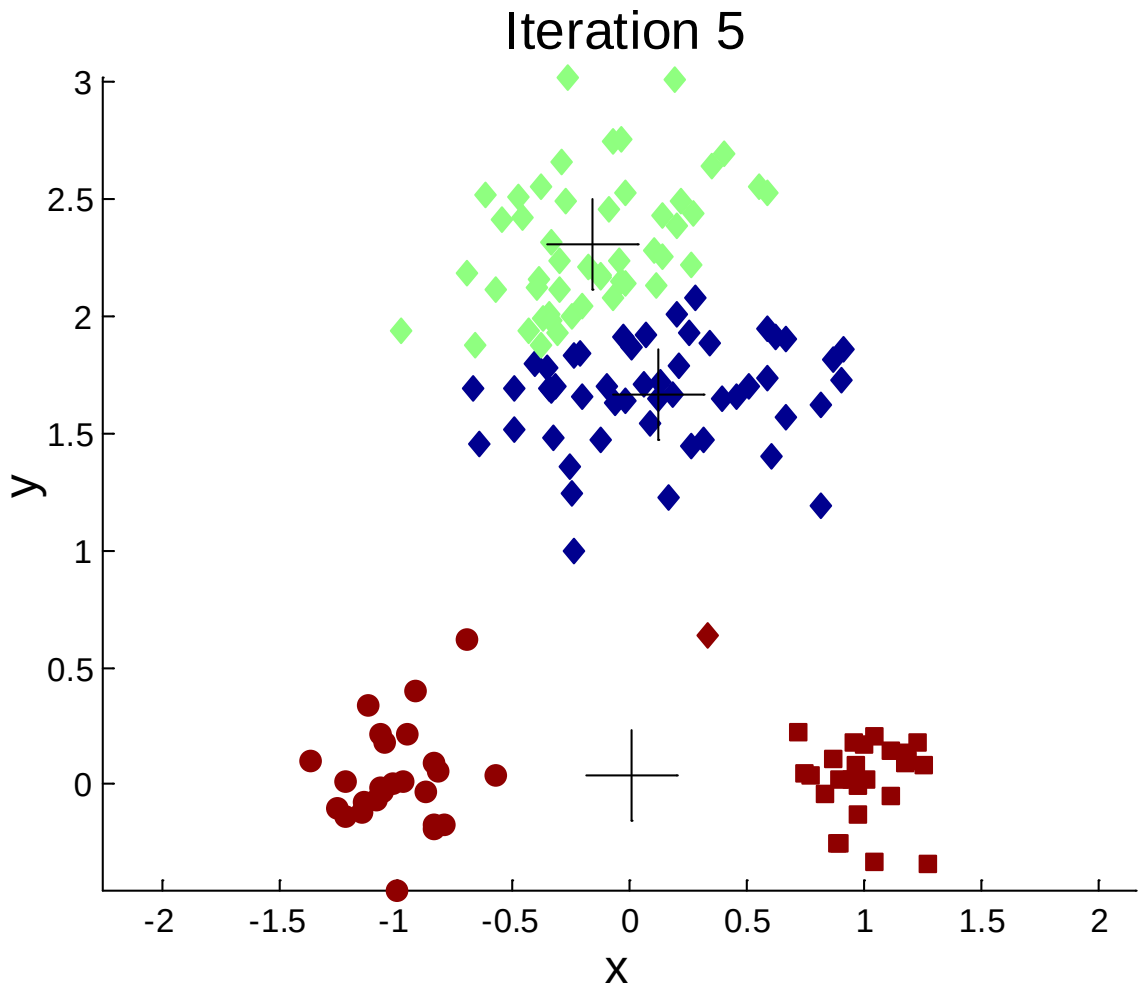
Влияние начального выбора центроидов (1)



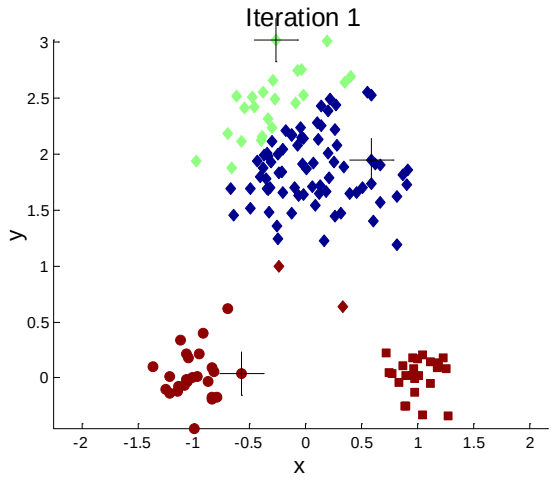
Влияние начального выбора центроидов (1)

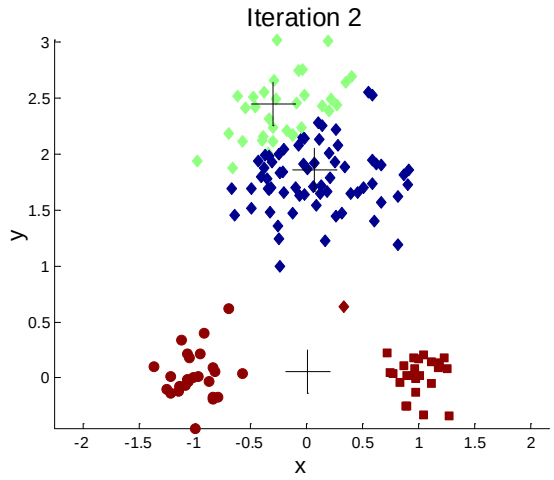


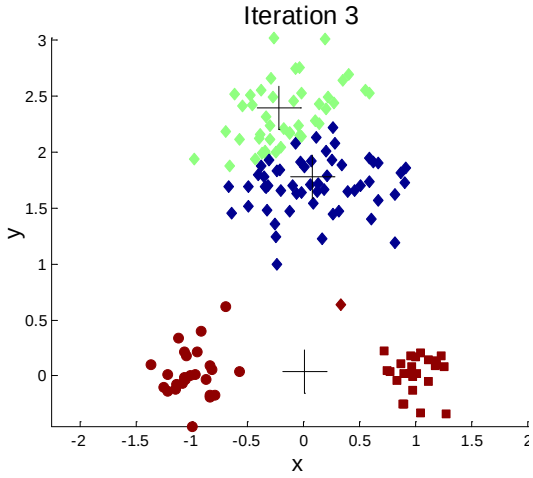
Влияние начального выбора центроидов (2)

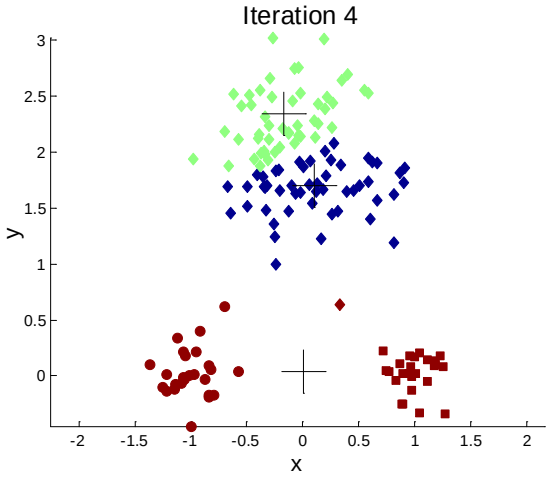


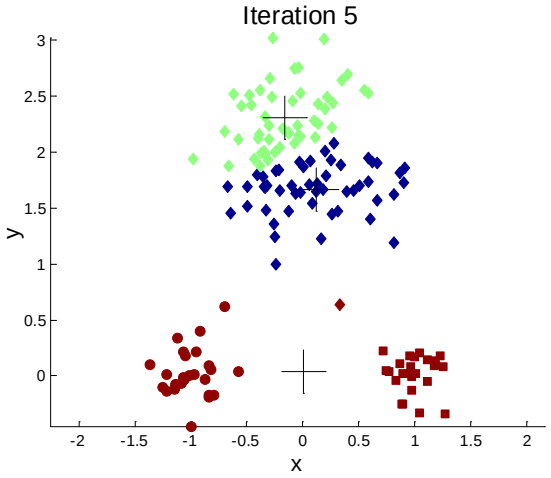
Влияние начального выбора центроидов (2)









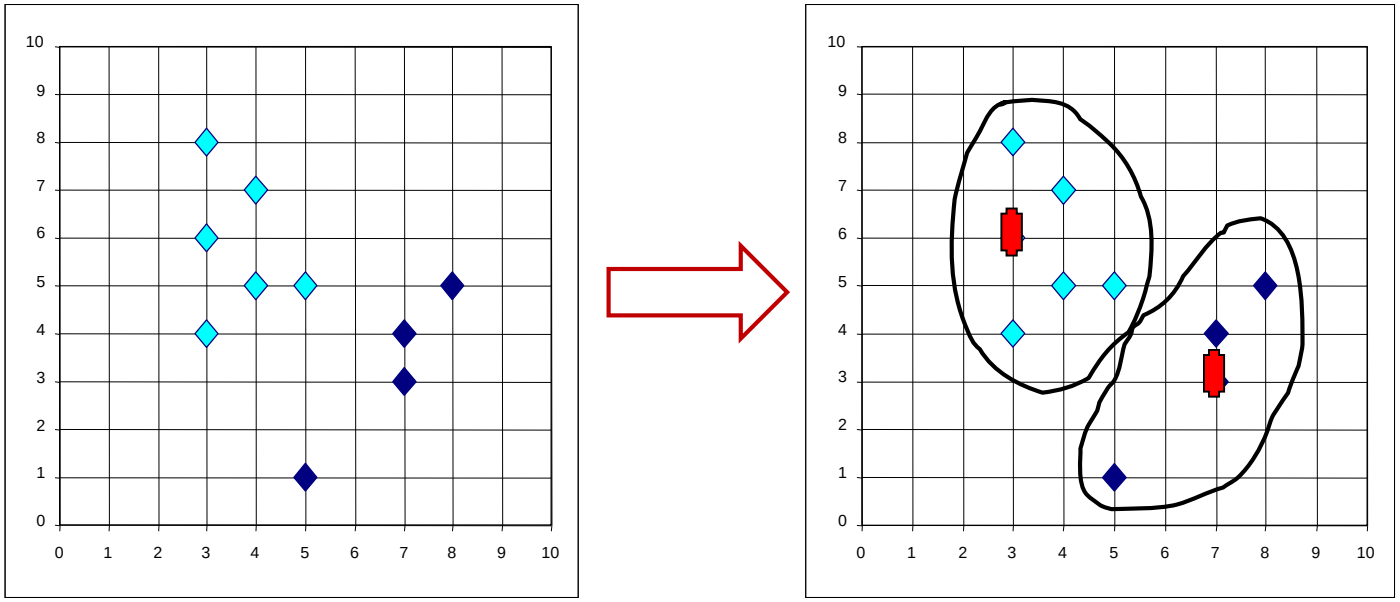


Подбор начальных центроидов

- Многократно запустить k-means, выбрать результат с минимальным значением SSE
  - Результат не обязан быть лучшим из возможных
- Выполнить иерархическую кластеризацию случайного подмножества исходных точек для k кластеров и взять центроиды этих кластеров
  - Работает для небольших подмножеств и значений k
- Взять центроид всех точек, затем $k - 1$ раз взять точку, наиболее удаленную от всех k выбранных до этого центроидов
  - В качестве центроида может быть взят выброс
  - Высокая трудоемкость
- Применить предыдущий подход для случайного подмножества точек

k-medoids (PAM, Partitioning Around Medoids)

- Медоид – репрезентативный объект кластера

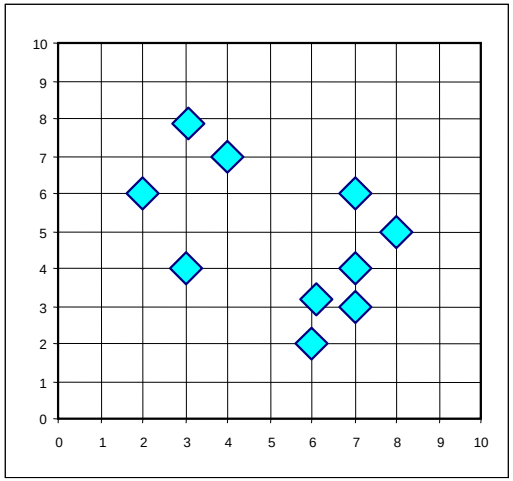


- Менее чувствителен к шумам и выбросам, чем *k*-means

k-medoids (PAM, Partitioning Around Medoids)

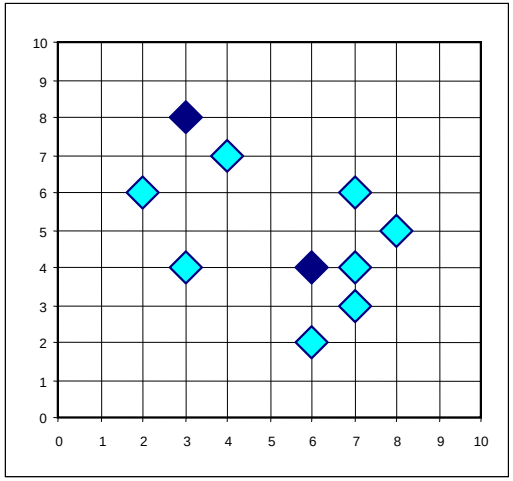
взять k случайных объектов в качестве медоидов $\breve{o}_1, \ldots, \breve{o}_k$
repeat
  назначить каждый объект кластеру с ближайшим медоидом
  случайным образом выбрать объект не-медоид o
  вычислить стоимость обмена местами o и $\breve{o}$
  $E = \sum_{i=1}^{k} \sum_{p \in \mathrm{C}_i} dist(p, o)^2 - \sum_{i=1}^{k} \sum_{p \in \mathrm{C}_i} dist(p, \breve{o})^2$
  if $E < 0$ **then**
    обменять местами объект o и медоид $\breve{o}$
until нет изменений

Пример работы *k*-medoids

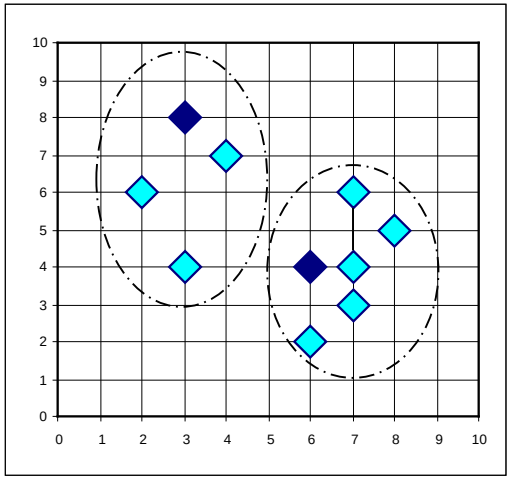


$k=2$

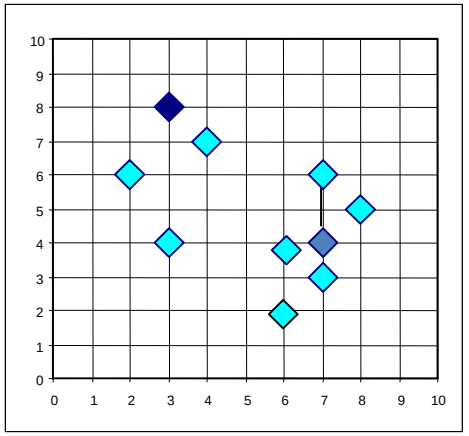
Выбрать k случайных объектов в качестве медоидов



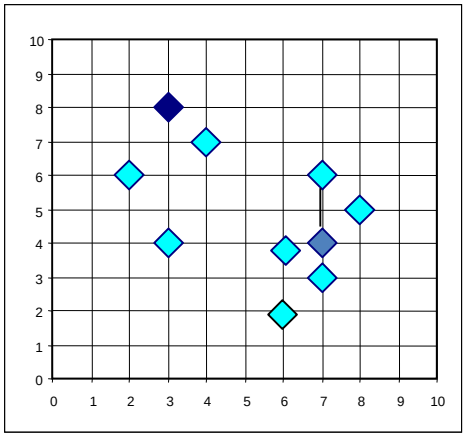
Присвоить оставшимся объектам кластеры по ближайшему медоиду



Выбрать случайный объект не-медоид, O_{random}



Вычислить стоимость обмена медоида и O_{random}



Если качество улучшилось, поменять медоид и O_{random}

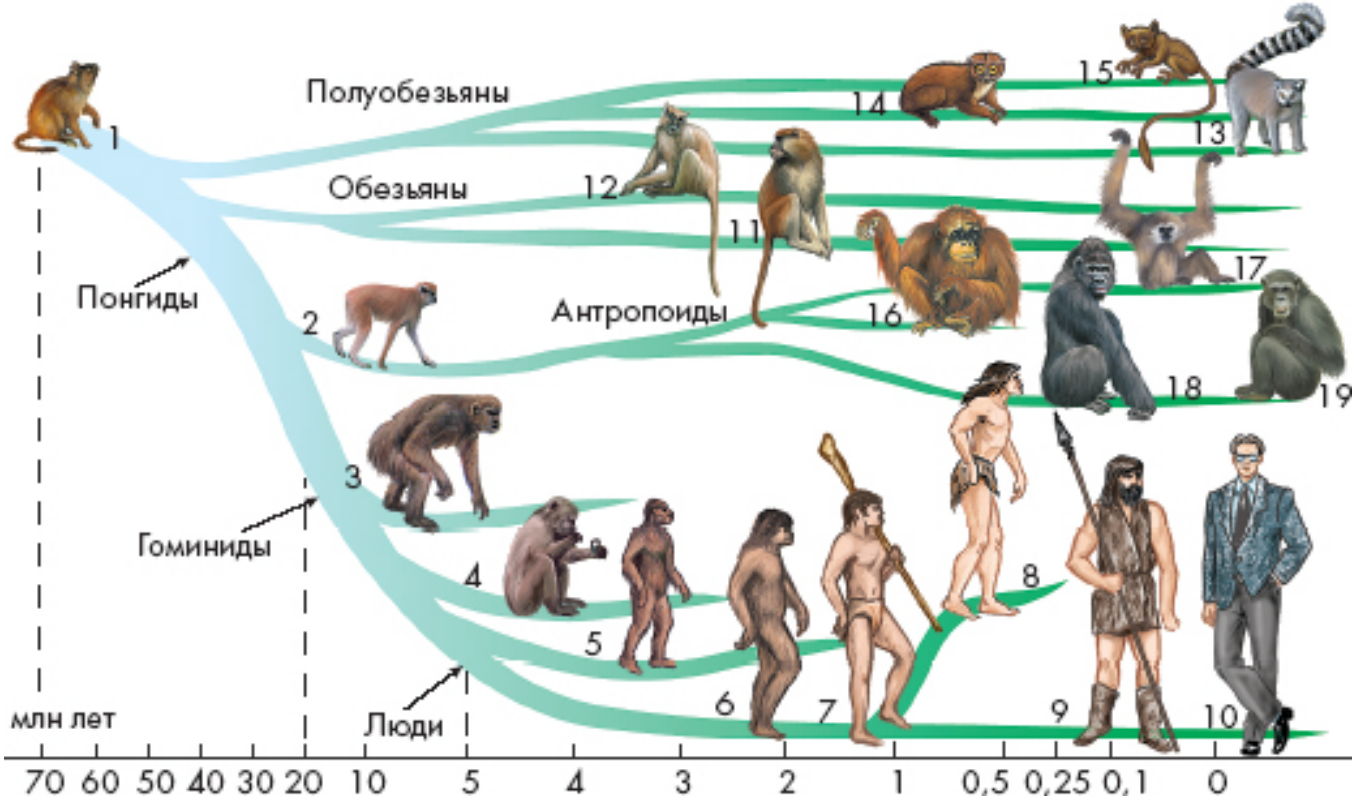
repeat

until нет изменений

Содержание

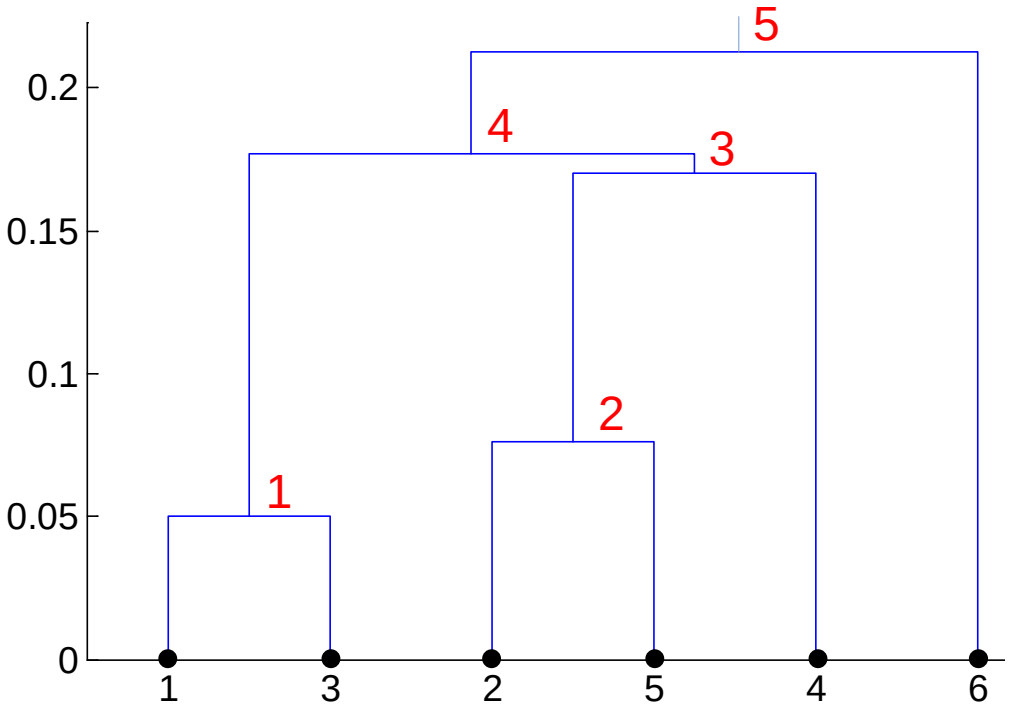
- Основные концепции
- Разделительная кластеризация
- **Иерархическая кластеризация**
- Меры качества кластеризации

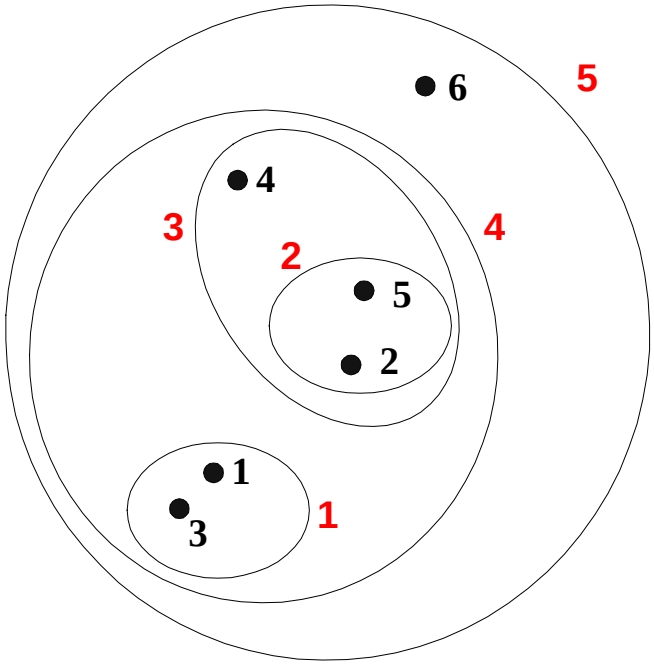
Иерархическая кластеризация



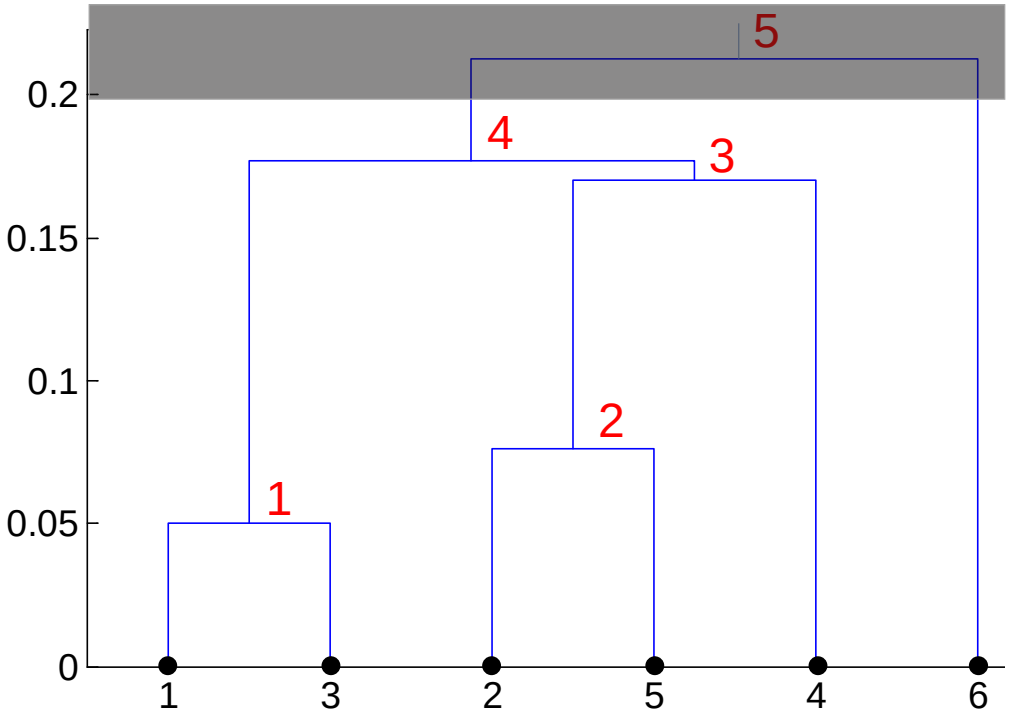
1 — плезиадацис (предок приматов); **2** — дрипитек африканский; **3** — рамапитек; **4** — австралопитек африканский; **5** — австралопитек; **6-7** — *homo erectus* (питекантроп, синантроп); **8** — неандерталец; **9** — *homo sapiens* (кроманьонец); **10** — современный человек; **11** — узконосые обезьяны; **12** — широконосые обезьяны; **13** — лемуры; **14** — лори; **15** — долгопяты; **16** — орангутаны; **17** — гиббоны; **18** — гориллы; **19** — шимпанзе

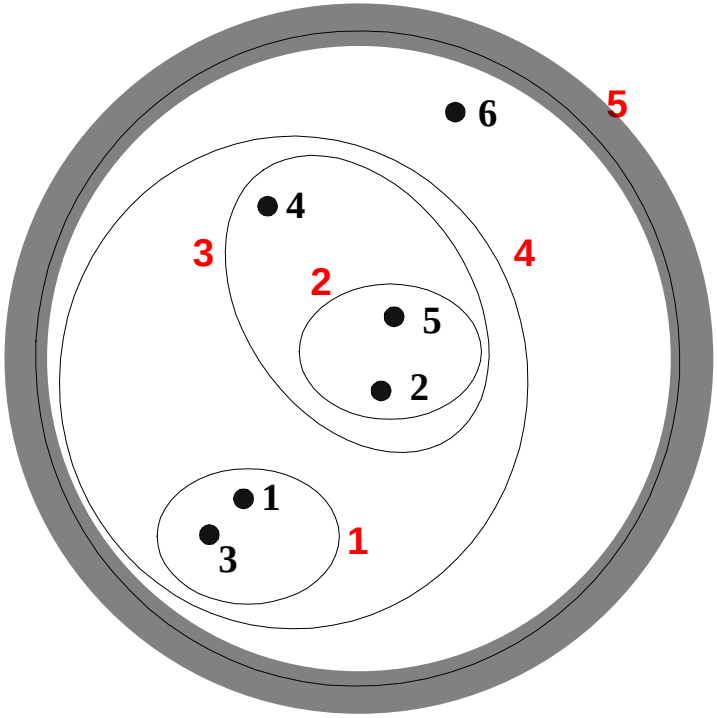
Дендрограммы



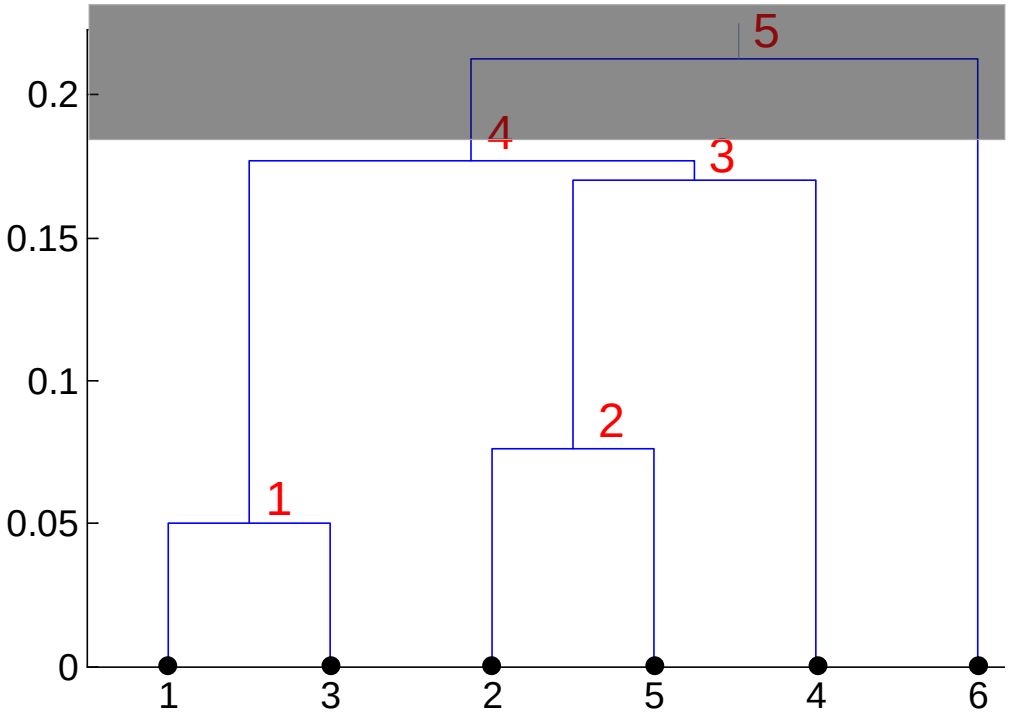


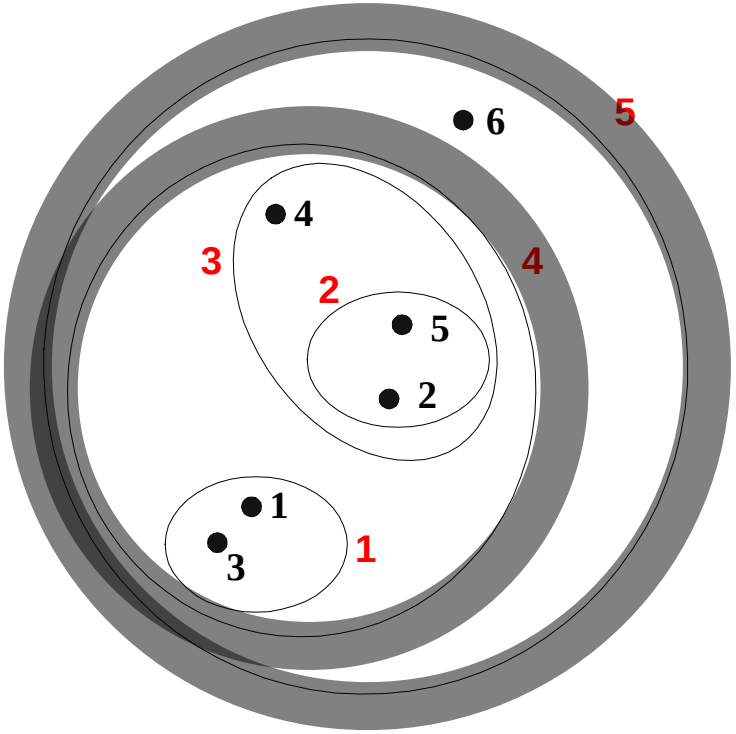
Дендрограммы



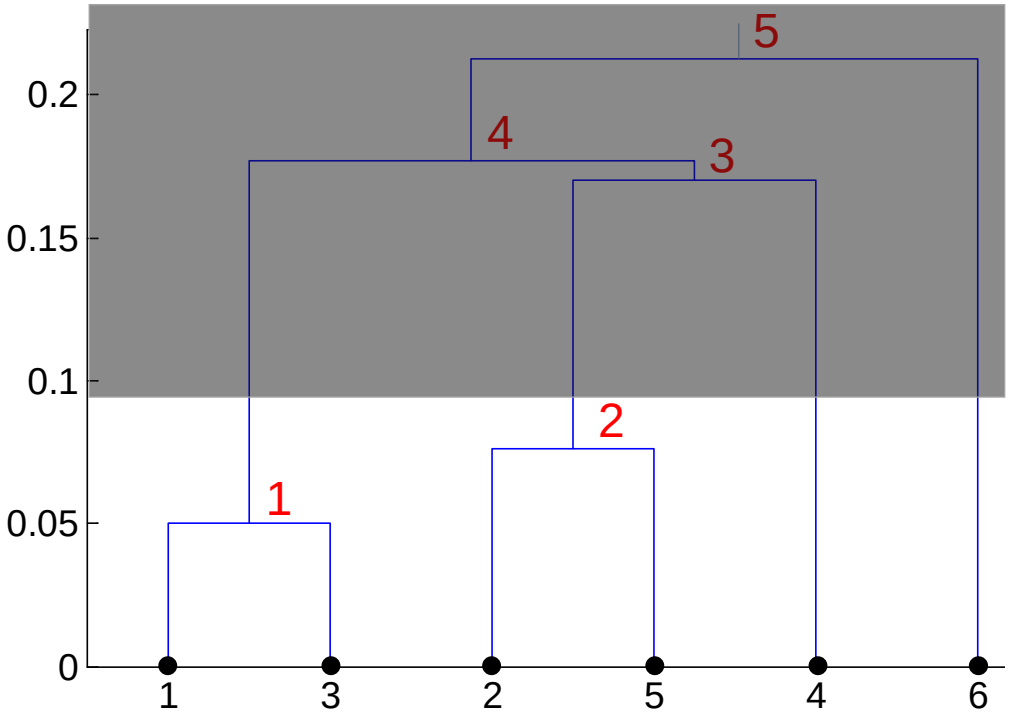


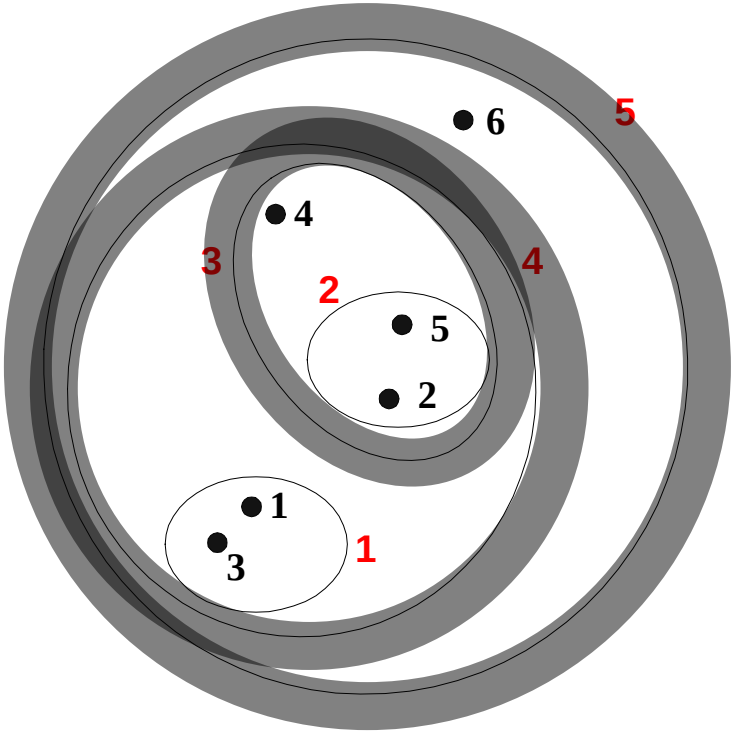
Дендрограммы



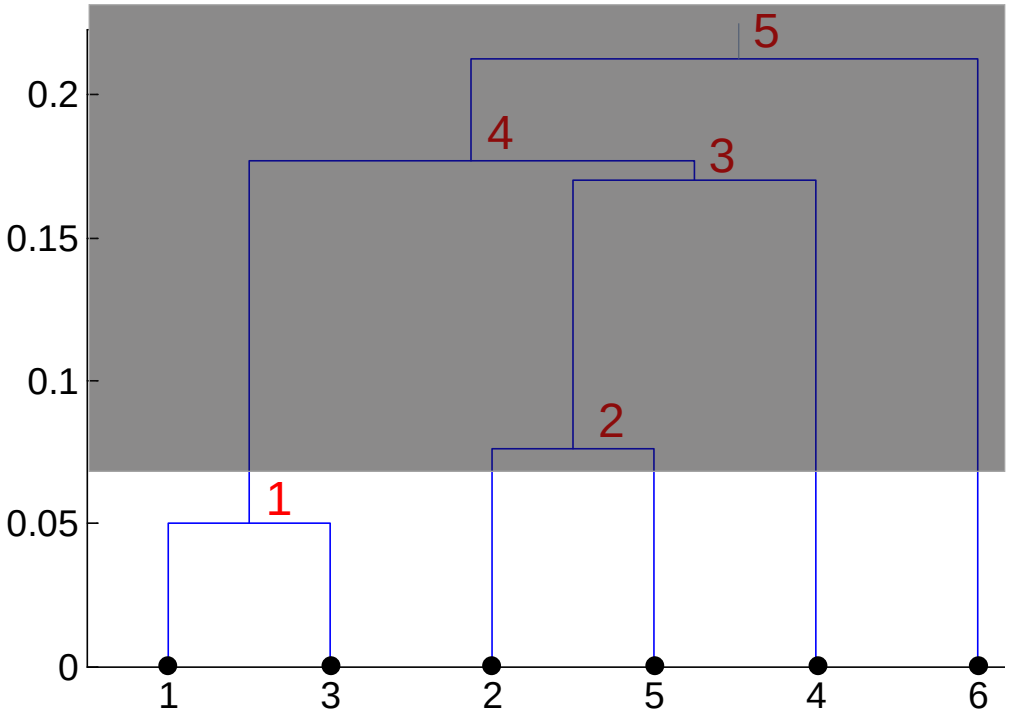


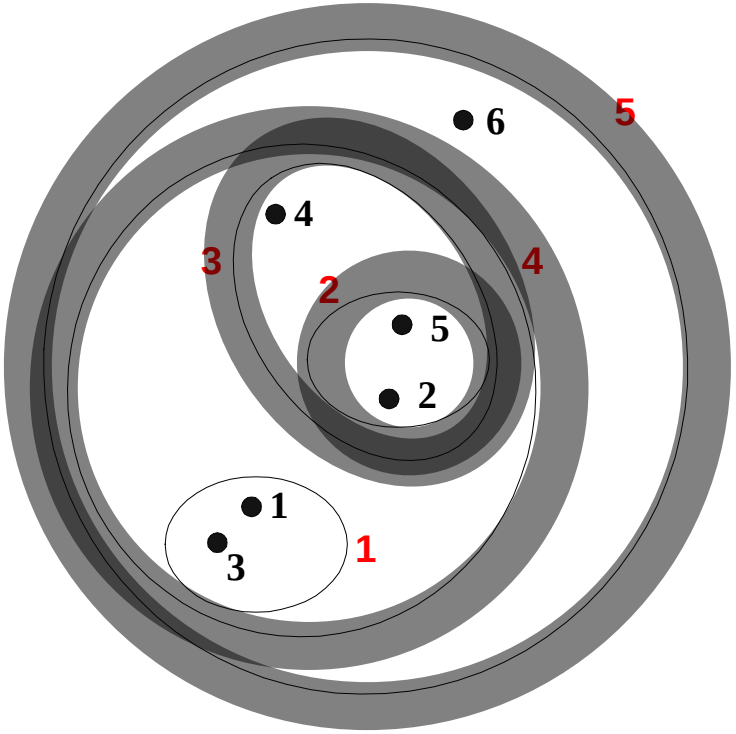
Дендрограммы



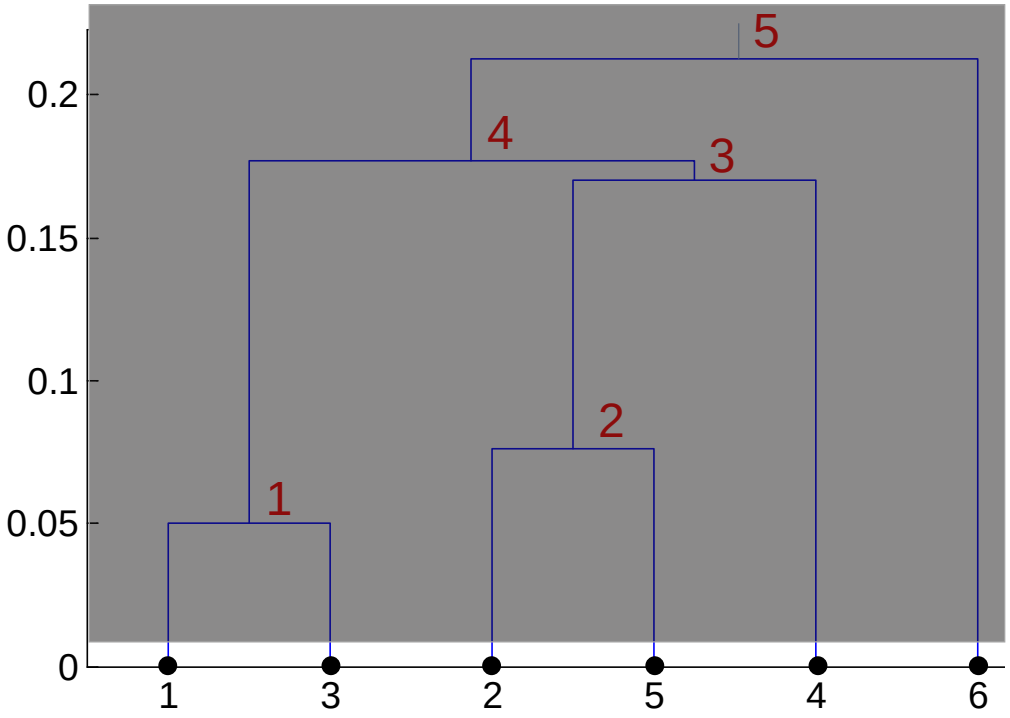


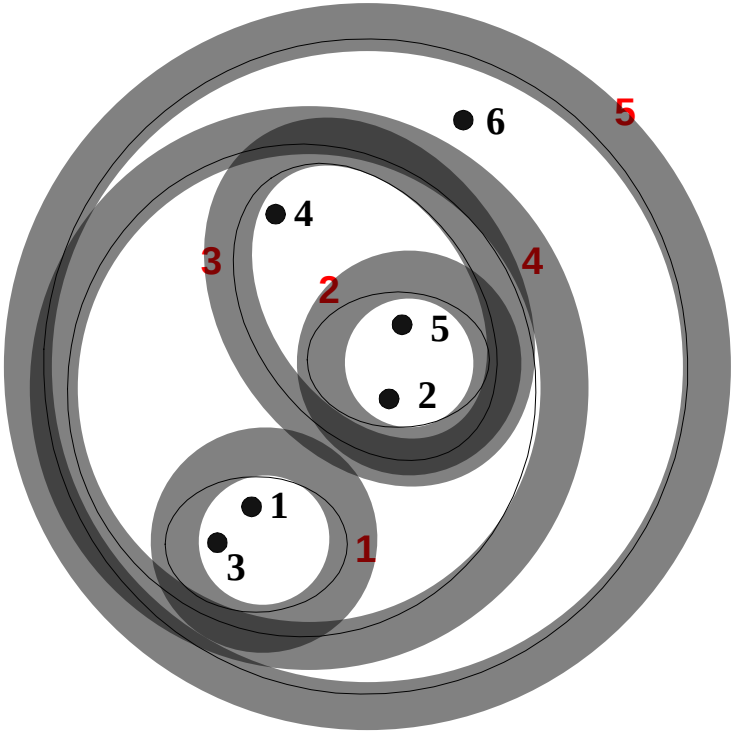
Дендрограммы



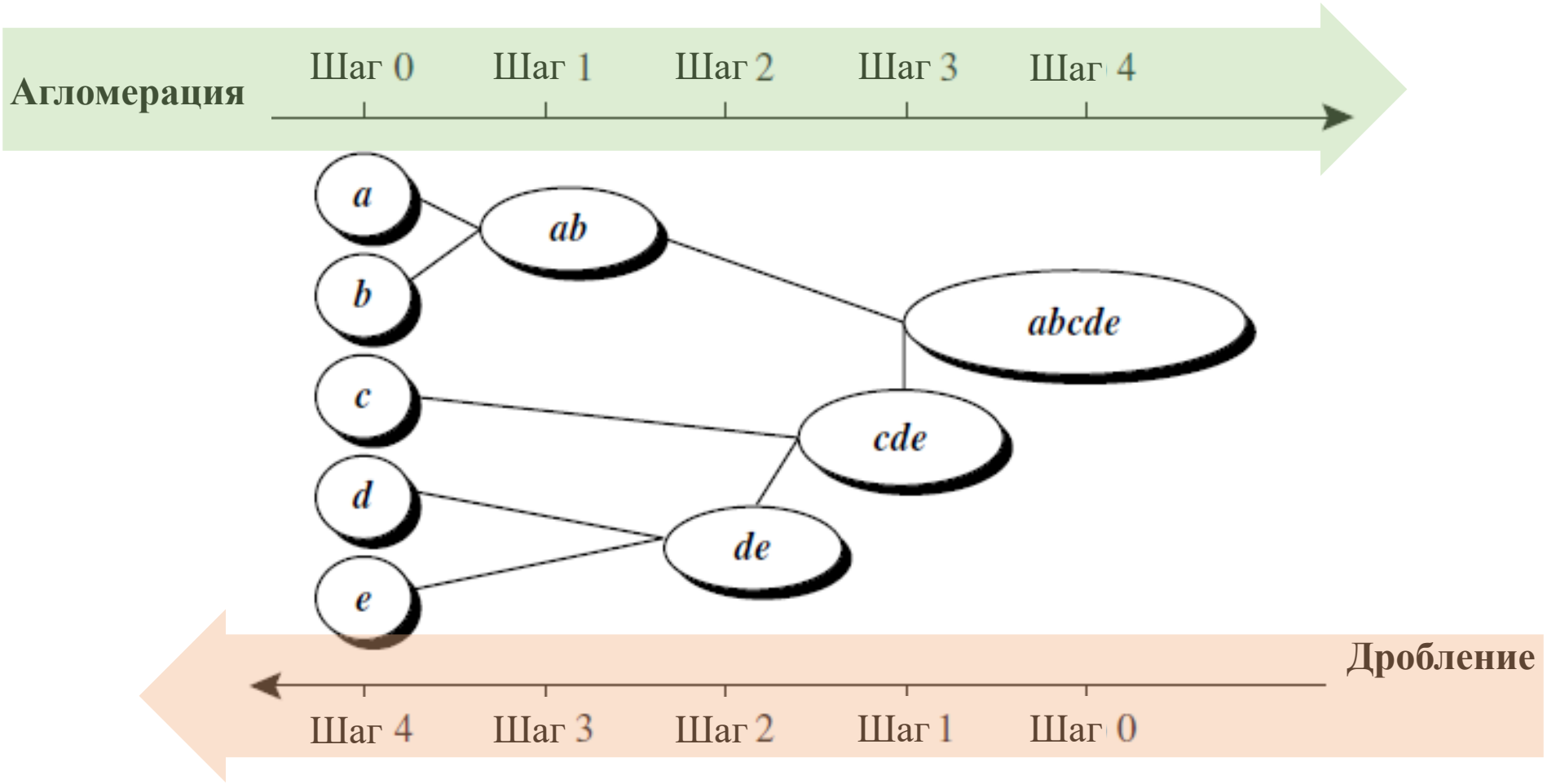


Дендрограммы





Агломерация *vs.* дробление

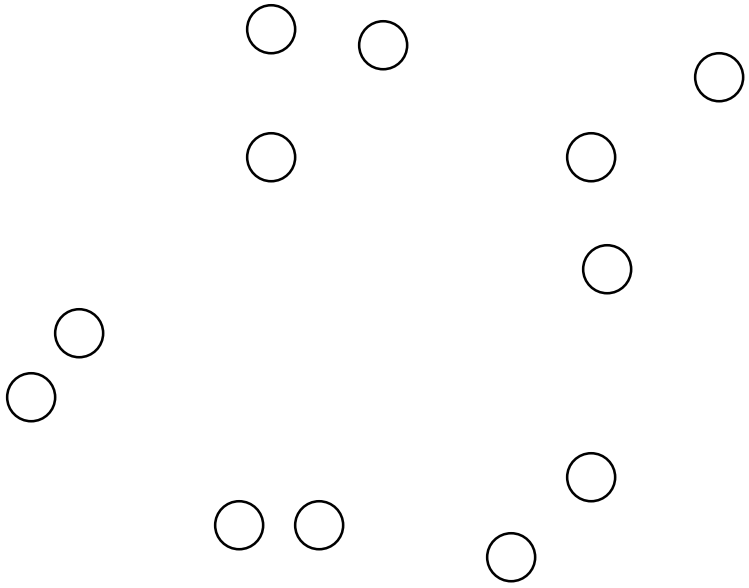


Агломеративная кластеризация

- Кластеризуемые объекты: $X = (x_1, x_2, \dots, x_n)$
- Кластеры: C

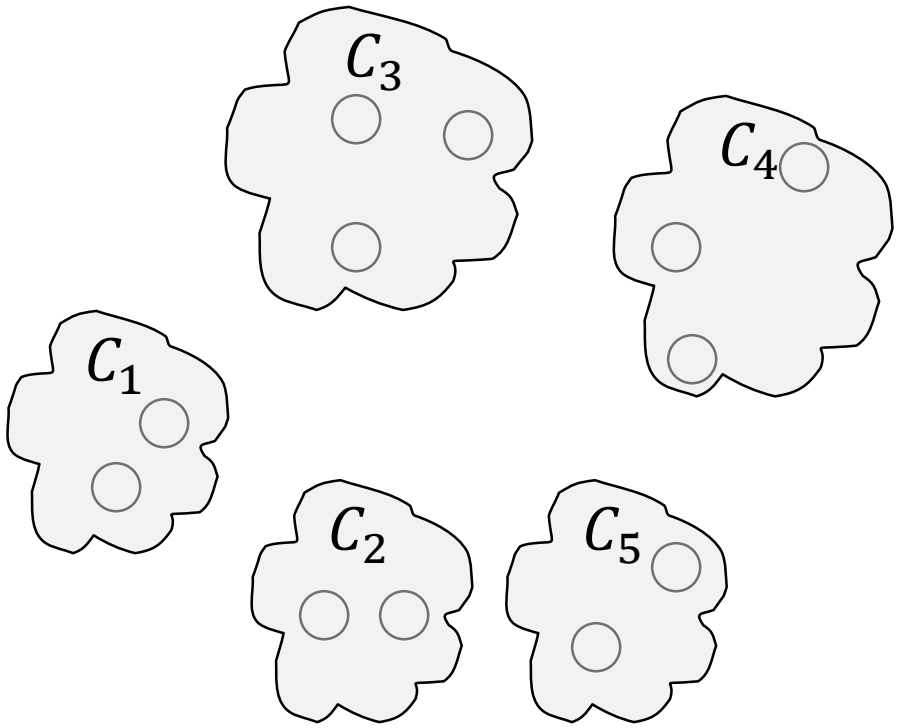
1. Вычислить матрицу расстояний $(d_{ij}) := dist(x_i, x_j)$
2. $\forall k\ (1 \le k \le n)\ c_k := x_k$
3. **repeat**
4. $c_{new} := c_i \cup c_j$, где $\{c_i, c_j\} = \arg\min_{c_p \neq c_q} dist(c_p, c_q)$
5. Обновить (d_{ij}): вычислить $dist(c_{new}, c_k)\ \forall k \neq i, j$
6. $C := C \cup c_{new}; C := C \setminus c_i; C := C \setminus c_j$
7. **until** $|C| = 1$

Агломеративная кластеризация (шаг 0)

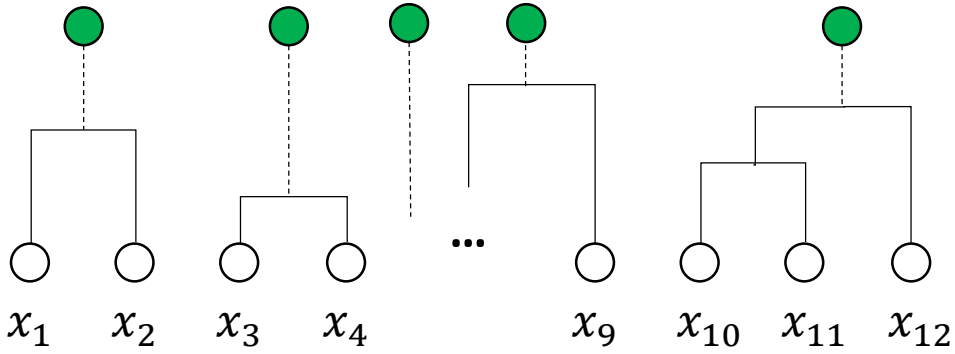


	x_1	x_2	x_3	x_4	x_5	...
x_1	0					
x_2		0				
x_3			0			
x_4				0		
x_5					0	
...						

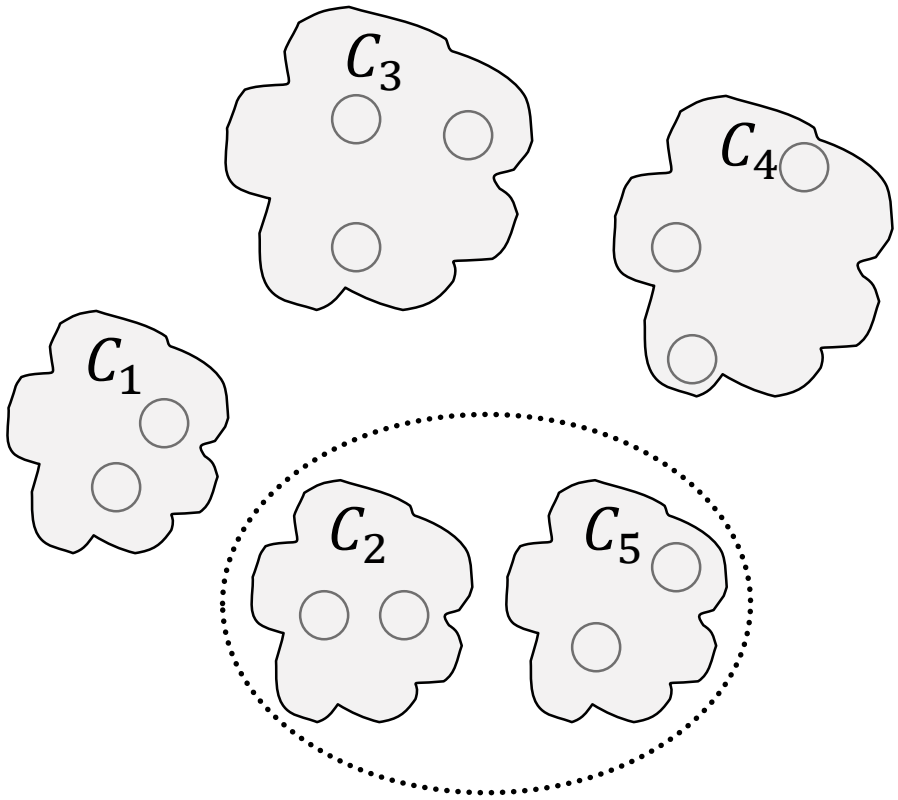
Агломеративная кластеризация (шаг k)



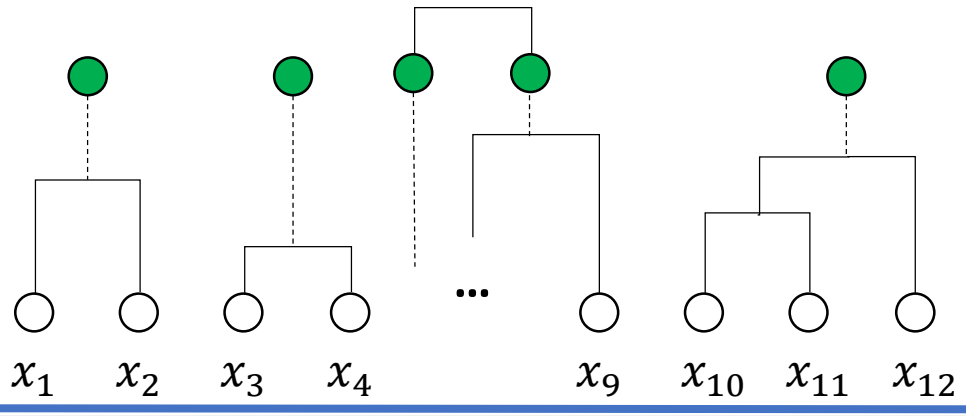
	c_1	c_2	c_3	c_4	c_5	…
c_1	0					
c_2		0				
c_3			0			
c_4				0		
c_5					0	
…						



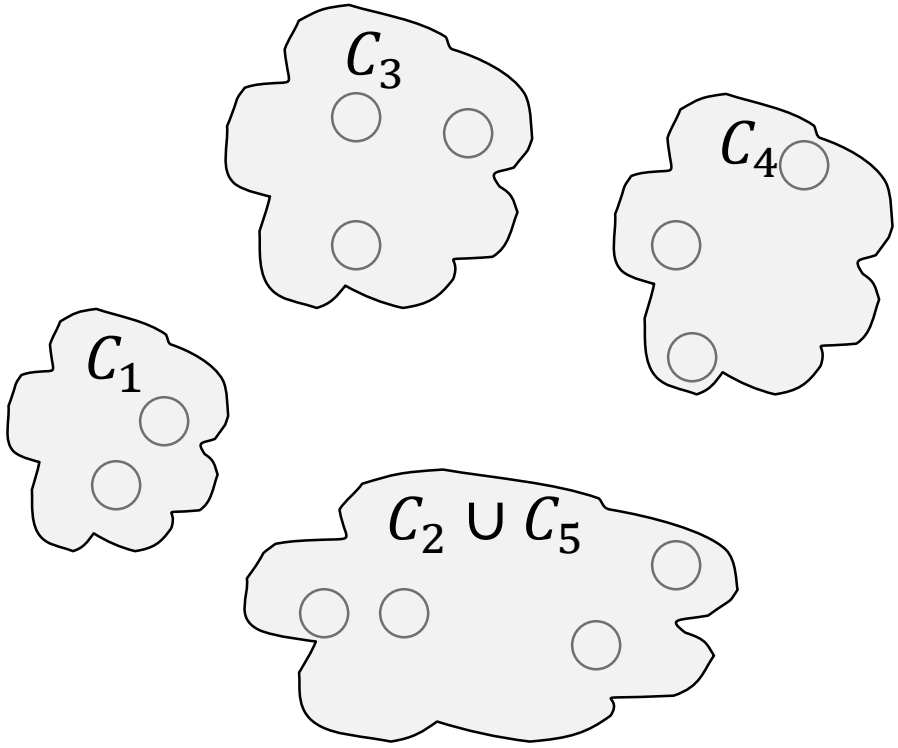
Агломеративная кластеризация (шаг $k+1$)



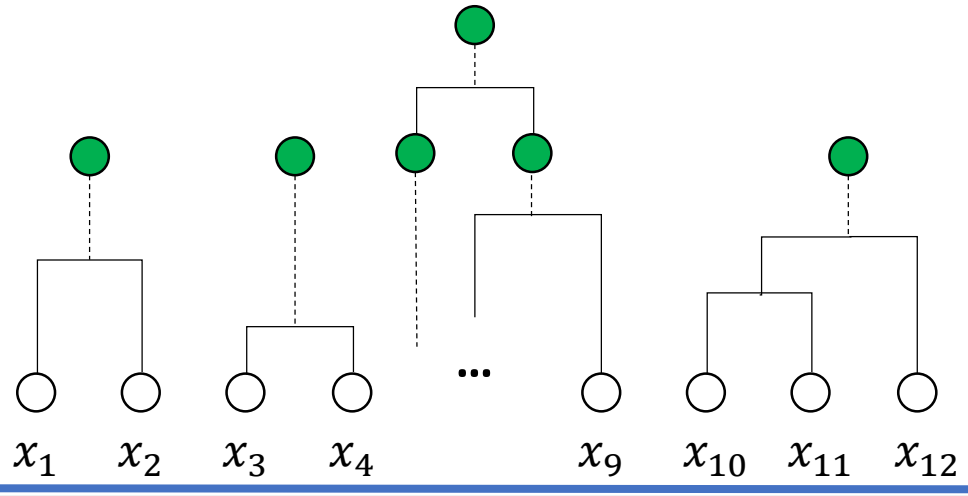
	c_1	c_2	c_3	c_4	c_5	…
c_1	0					
c_2		0				
c_3			0			
c_4				0		
c_5					0	
…						



Агломеративная кластеризация (шаг $k+1$)

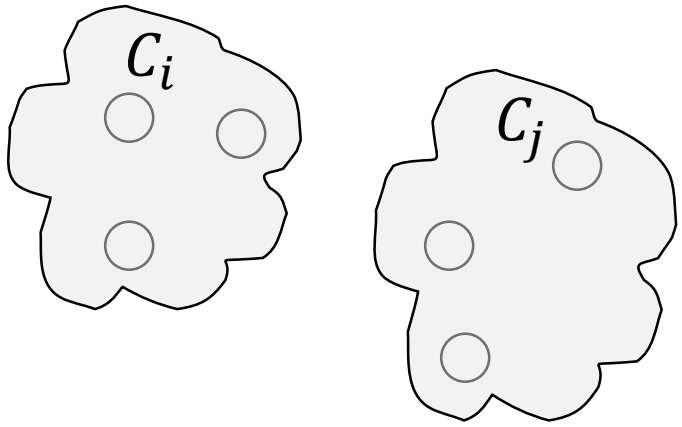


	c_1	$c_2 \cup c_5$	c_3	c_4	…
c_1	0				
$c_2 \cup c_5$		0			
c_3			0		
c_4				0	
…					

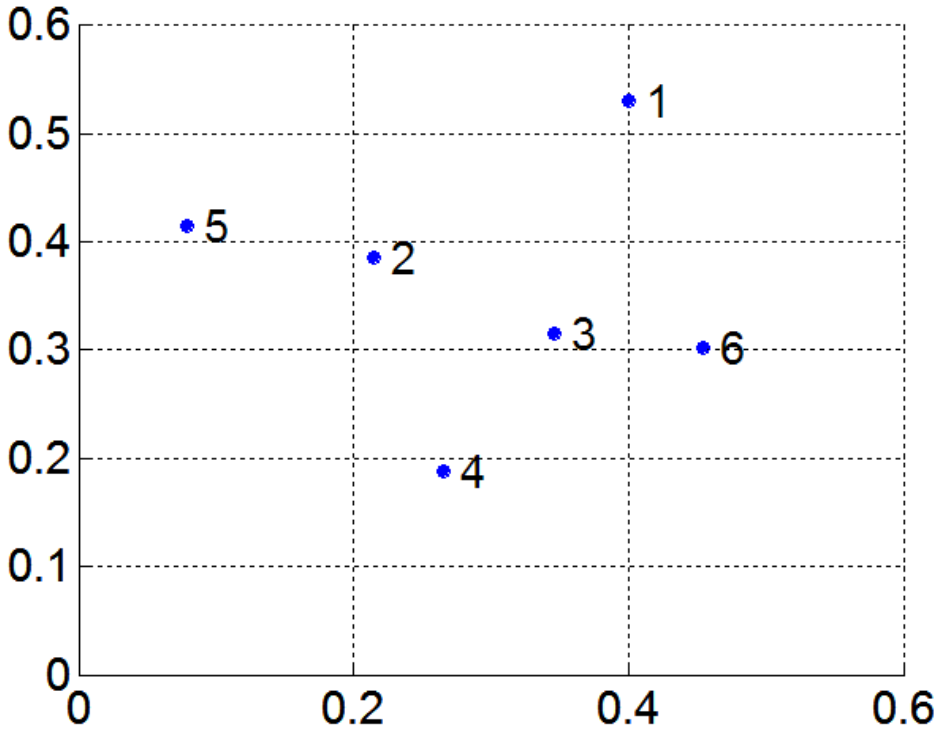


Меры схожести кластеров

- Min (Single linkage)
- Max (Complete linkage)
- Avg (Group average)
- Центроидная
- Медоидная
- Расстояние Уорда (Ward)



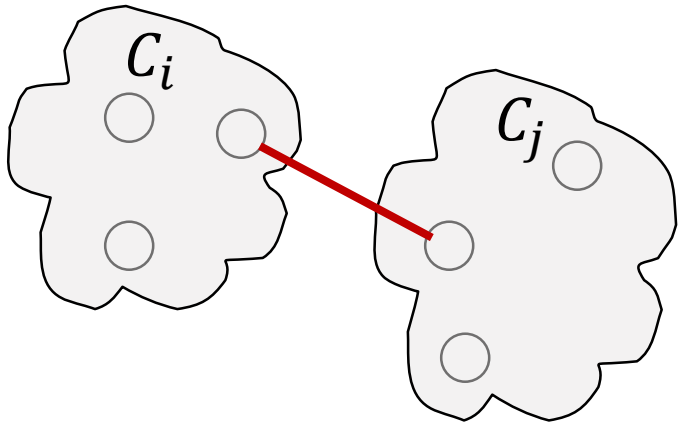
Min (Single linkage): пример



	p1	p2	p3	p4	p5	p6
p1	0.00	0.24	0.22	0.37	0.34	0.23
p2	0.24	0.00	0.15	0.20	0.14	0.25
p3	0.22	0.15	0.00	0.15	0.28	0.11
p4	0.37	0.20	0.15	0.00	0.29	0.22
p5	0.34	0.14	0.28	0.29	0.00	0.39
p6	0.23	0.25	0.11	0.22	0.39	0.00

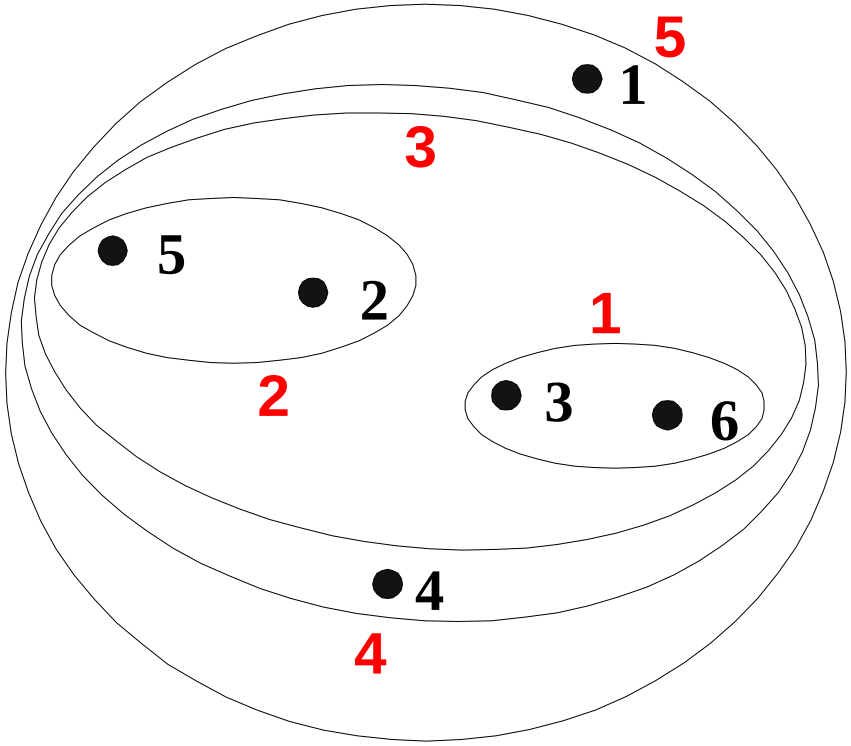
Меры схожести кластеров

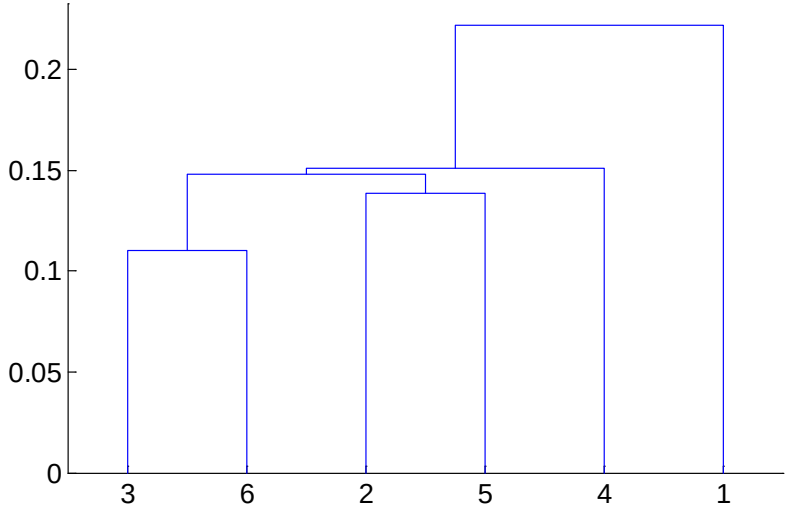
- **Min (Single linkage)**
- Max (Complete linkage)
- Avg (Group average)



Min (Single linkage): пример

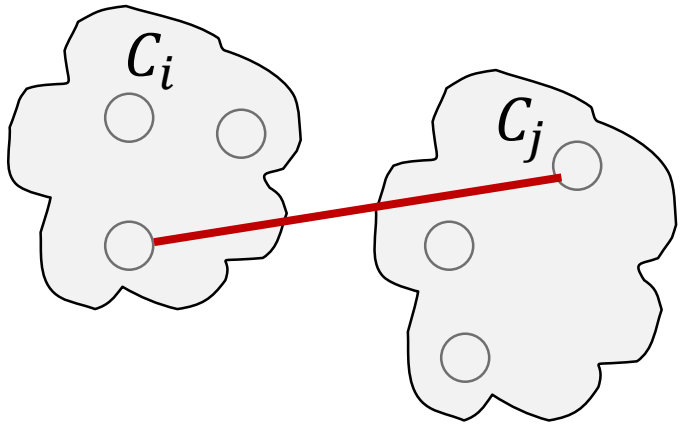
	p1	p2	p3	p4	p5	p6
p1	0.00	0.24	0.22	0.37	0.34	0.23
p2	0.24	0.00	0.15	0.20	0.14	0.25
p3	0.22	0.15	0.00	0.15	0.28	0.11
p4	0.37	0.20	0.15	0.00	0.29	0.22
p5	0.34	0.14	0.28	0.29	0.00	0.39
p6	0.23	0.25	0.11	0.22	0.39	0.00



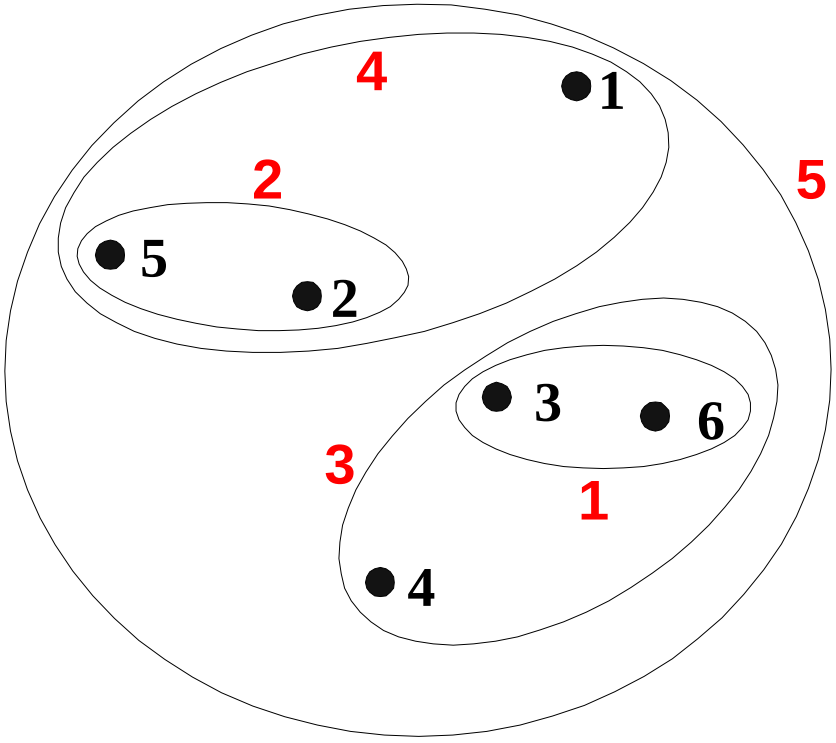


Меры схожести кластеров

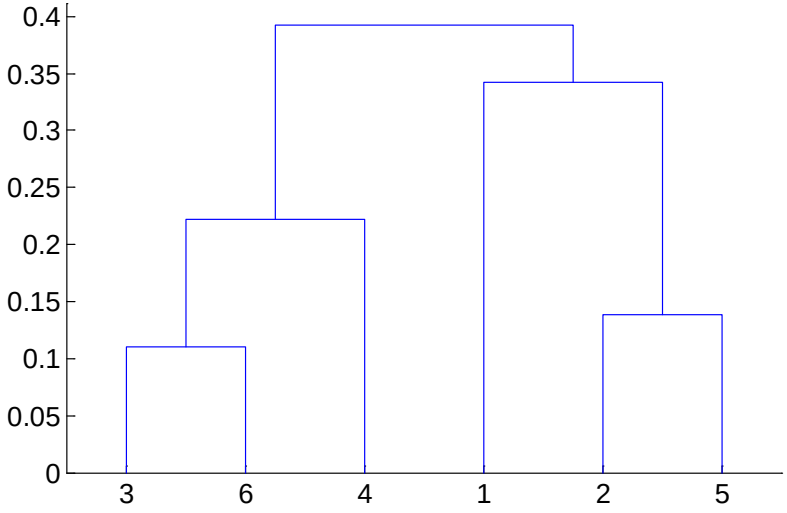
- Min (Single linkage)
- **Max (Complete linkage)**
- Avg (Group average)



Max (Complete linkage): пример

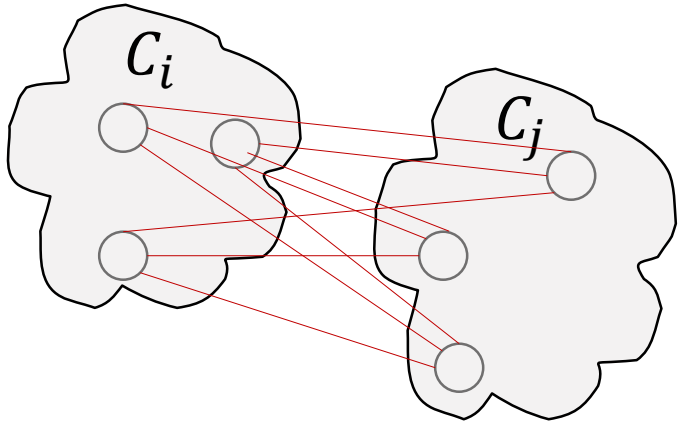


	p1	p2	p3	p4	p5	p6
p1	0.00	0.24	0.22	0.37	0.34	0.23
p2	0.24	0.00	0.15	0.20	0.14	0.25
p3	0.22	0.15	0.00	0.15	0.28	0.11
p4	0.37	0.20	0.15	0.00	0.29	0.22
p5	0.34	0.14	0.28	0.29	0.00	0.39
p6	0.23	0.25	0.11	0.22	0.39	0.00



Меры схожести кластеров

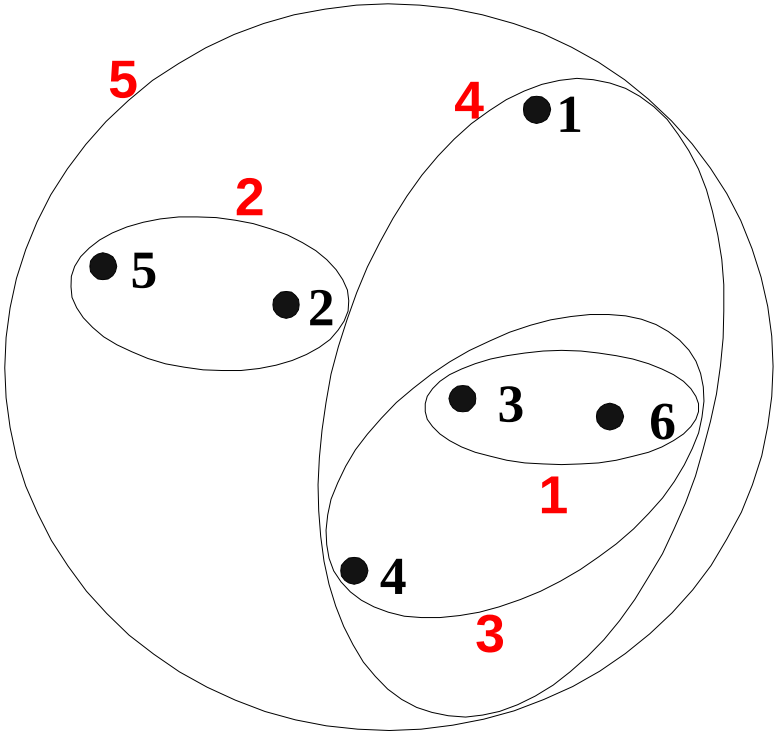
- Min (Single linkage)
- Max (Complete linkage)
- **Avg (Group average)**

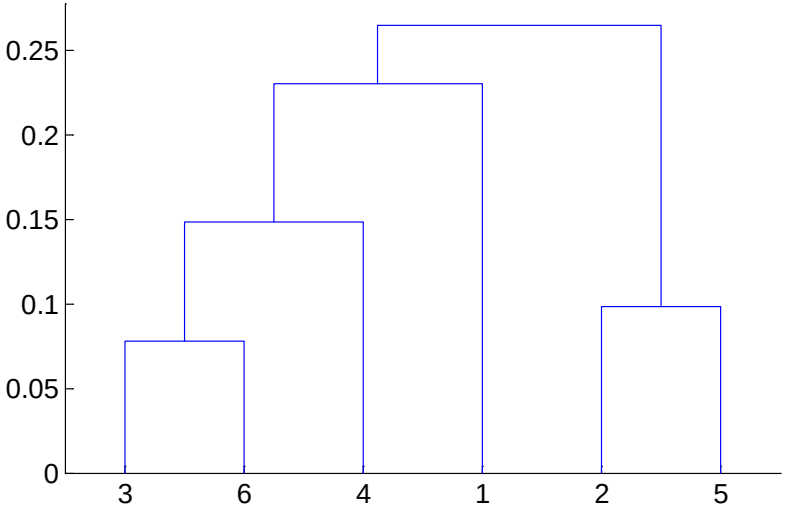


$$dist\left(C_i, C_j\right) = \frac{\sum_{x_i \in C_i, x_j \in C_j} dist(x_i, x_j)}{|C_i| \cdot |C_j|}$$

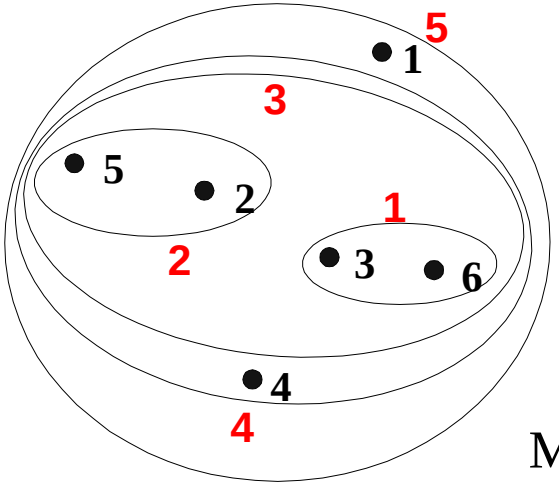
Avg (Group average): пример

	p1	p2	p3	p4	p5	p6
p1	0.00	0.24	0.22	0.37	0.34	0.23
p2	0.24	0.00	0.15	0.20	0.14	0.25
p3	0.22	0.15	0.00	0.15	0.28	0.11
p4	0.37	0.20	0.15	0.00	0.29	0.22
p5	0.34	0.14	0.28	0.29	0.00	0.39
p6	0.23	0.25	0.11	0.22	0.39	0.00

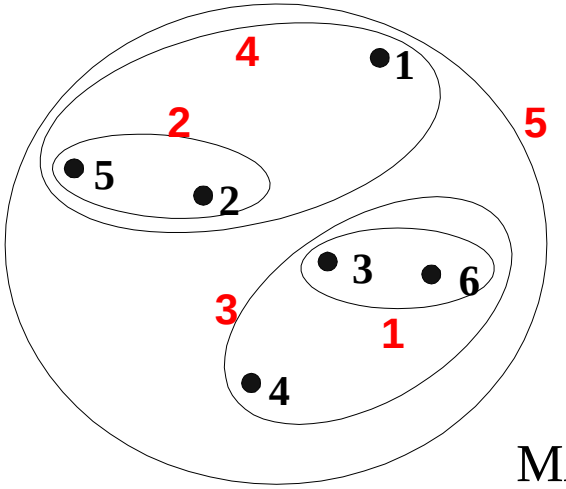




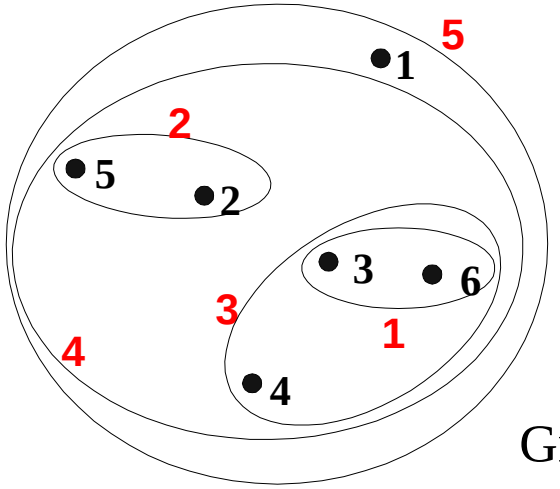
Сравнение мер



MIN

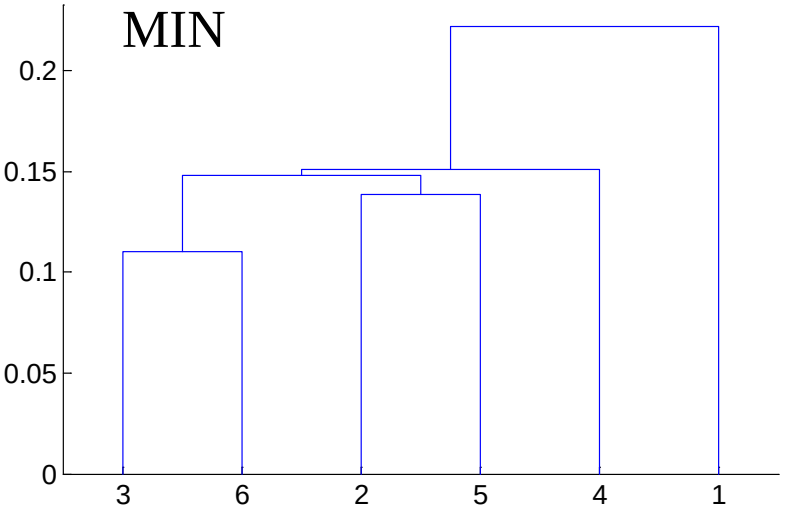


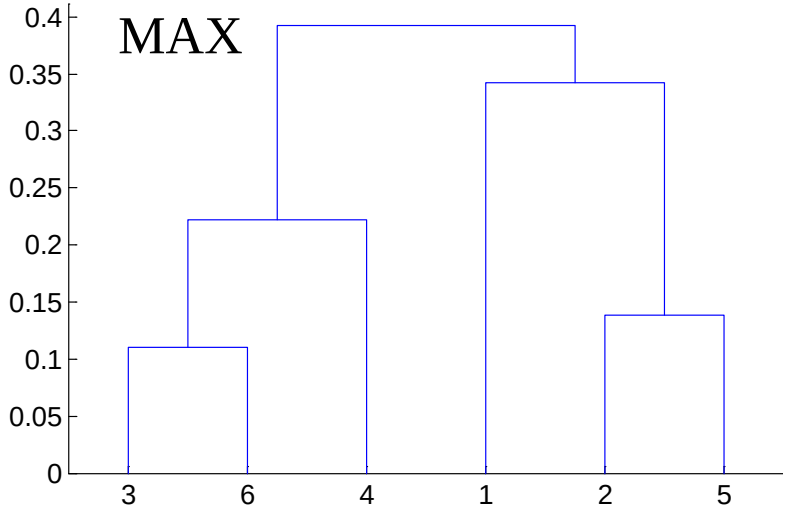
MAX

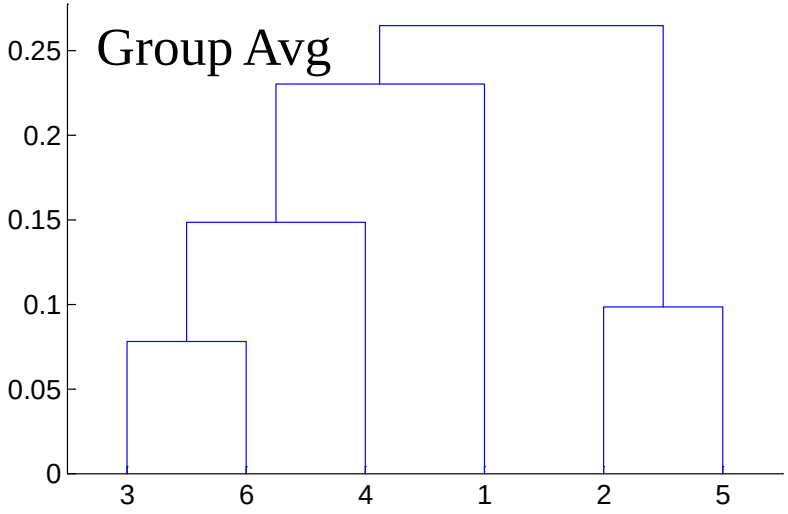


Group Average

Сравнение мер







	p1	p2	p3	p4	p5	p6
p1	0.00	0.24	0.22	0.37	0.34	0.23
p2	0.24	0.00	0.15	0.20	0.14	0.25
p3	0.22	0.15	0.00	0.15	0.28	0.11
p4	0.37	0.20	0.15	0.00	0.29	0.22
p5	0.34	0.14	0.28	0.29	0.00	0.39
p6	0.23	0.25	0.11	0.22	0.39	0.00

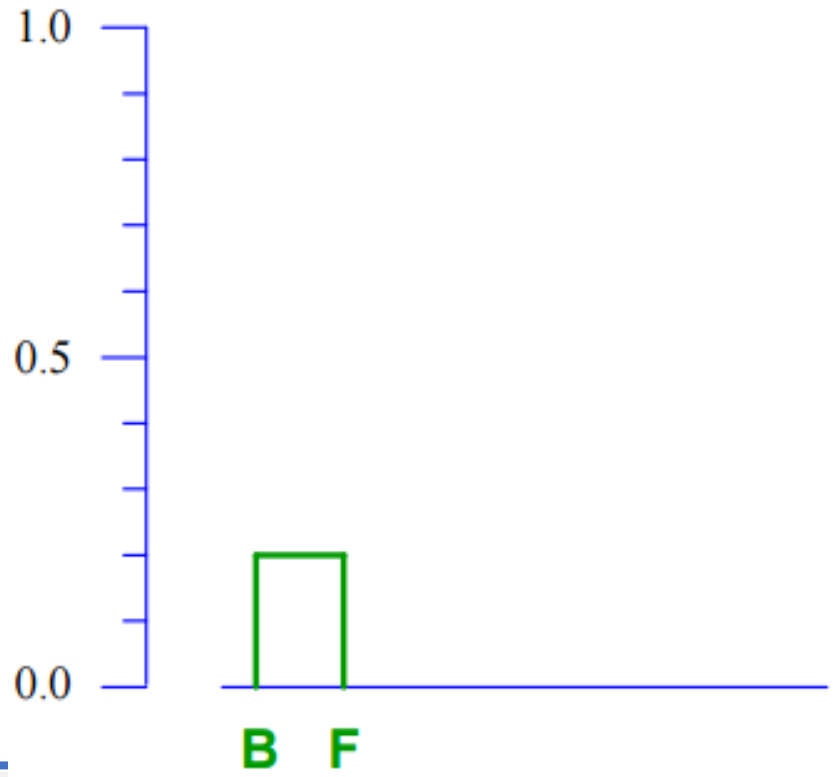
Иерархическая кластеризация: пример

samples	A	B	C	D	E	F	G
A	0	0.5000	0.4286	1.0000	0.2500	0.6250	0.3750
B	0.5000	0	0.7143	0.8333	0.6667	0.2000	0.7778
C	0.4286	0.7143	0	1.0000	0.4286	0.6667	0.3333
D	1.0000	0.8333	1.0000	0	1.0000	0.8000	0.8571
E	0.2500	0.6667	0.4286	1.0000	0	0.7778	0.3750
F	0.6250	0.2000	0.6667	0.8000	0.7778	0	0.7500
G	0.3750	0.7778	0.3333	0.8571	0.3750	0.7500	0

Матрица расстояний

Иерархическая кластеризация: пример (1)

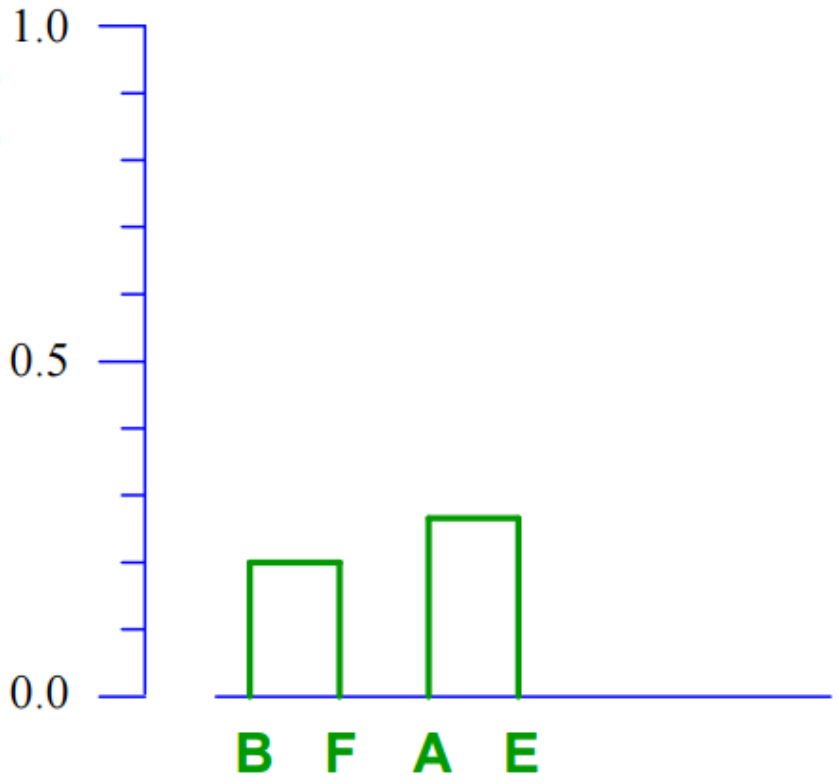
samples	A	B	C	D	E	F	G
A	0	0.5000	0.4286	1.0000	0.2500	0.6250	0.3750
B	0.5000	0	0.7143	0.8333	0.6667	0.2000	0.7778
C	0.4286	0.7143	0	1.0000	0.4286	0.6667	0.3333
D	1.0000	0.8333	1.0000	0	1.0000	0.8000	0.8571
E	0.2500	0.6667	0.4286	1.0000	0	0.7778	0.3750
F	0.6250	0.2000	0.6667	0.8000	0.7778	0	0.7500
G	0.3750	0.7778	0.3333	0.8571	0.3750	0.7500	0



- Min (Single linkage)

Иерархическая кластеризация: пример (2)

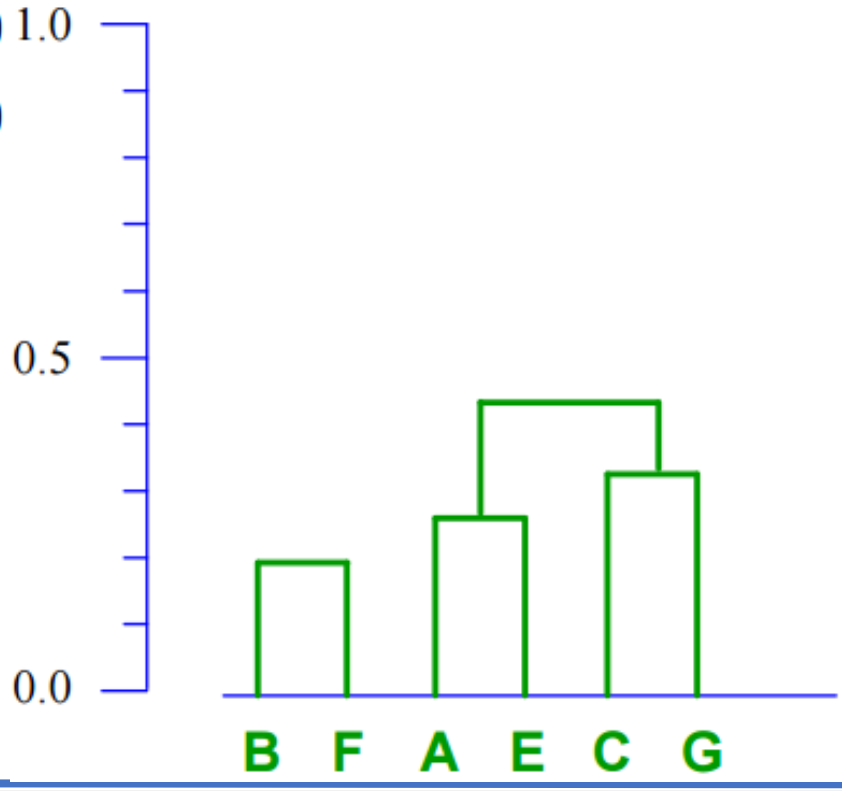
samples	A	(B,F)	C	D	E	G
A	0	0.6250	0.4286	1.0000	0.2500	0.3750
(B,F)	0.6250	0	0.7143	0.8333	0.7778	0.7778
C	0.4286	0.7143	0	1.0000	0.4286	0.3333
D	1.0000	0.8333	1.0000	0	1.0000	0.8571
E	0.2500	0.7778	0.4286	1.0000	0	0.3750
G	0.3750	0.7778	0.3333	0.8571	0.3750	0



- Min (Single linkage)

Иерархическая кластеризация: пример (3)

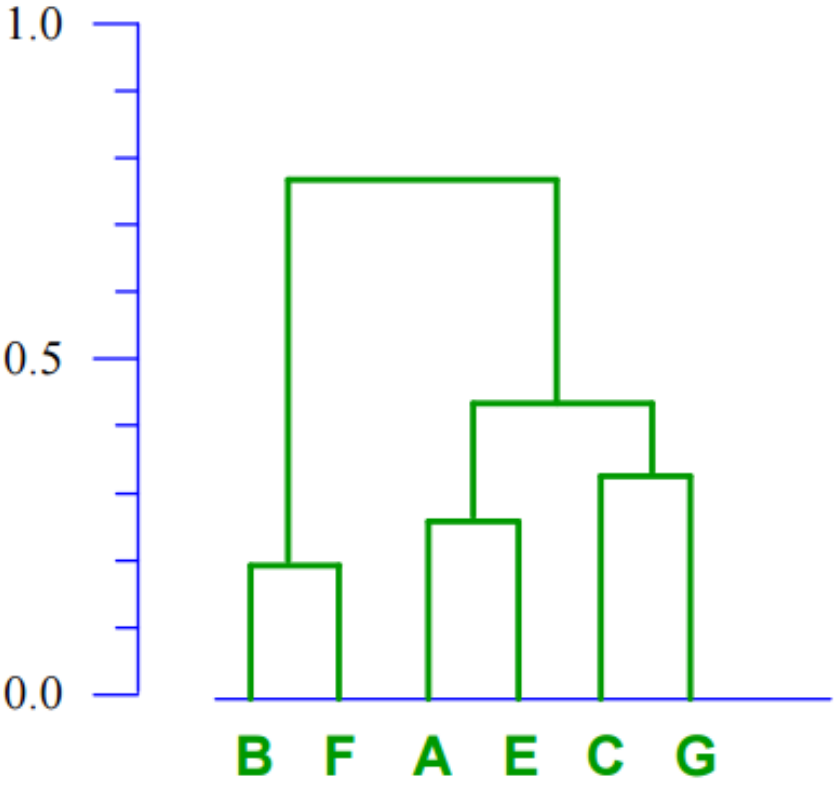
samples	(A,E)	(B,F)	(C,G)	D
(A,E)	0	0.7778	0.4286	1.0000
(B,F)	0.7778	0	0.7778	0.8333
(C,G)	0.4286	0.7778	0	1.0000
D	1.0000	0.8333	1.0000	0



- Min (Single linkage)

Иерархическая кластеризация: пример (4)

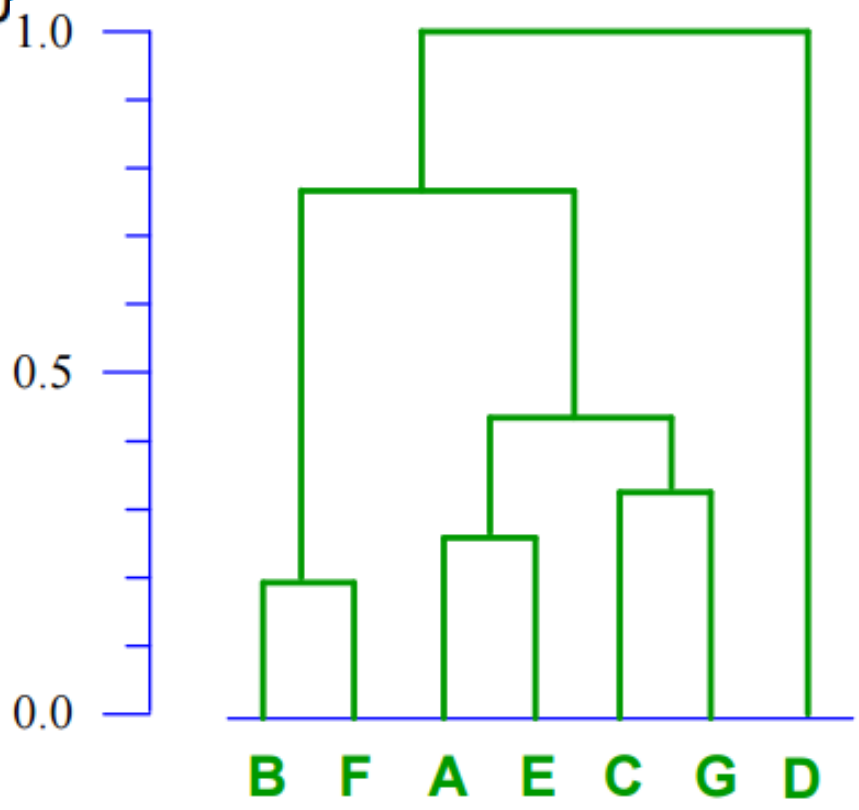
samples	(A,E,C,G)	(B,F)	D
(A,E,C,G)	0	0.7778	1.0000
(B,F)	0.7778	0	0.8333
D	1.0000	0.8333	0



- Min (Single linkage)

Иерархическая кластеризация: пример (5)

samples	(A,E,C,G,B,F)	D
(A,E,C,G,B,F)	0	1.0000
D	1.0000	0



- Min (Single linkage)

Иерархическая кластеризация: резюме

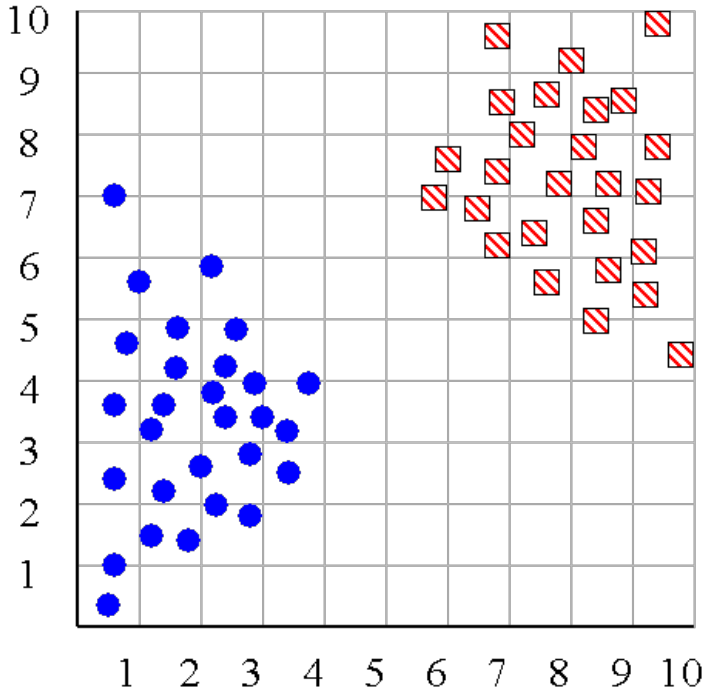
- Пространственная сложность: $O(n^2)$ (хранение матрицы расстояний)
- Временная сложность: как правило, $O(n^3)$
  - n шагов, на каждом из которых n^2 операций с матрицей расстояний
  - существуют приемы для сокращения сложности до $O(n^2 \log n)$
- Жадный алгоритм
- Отсутствует глобальная целевая функция, которая явно минимизируется
- Прочие сложности
  - Чувствительность к шумам и выбросам
  - Обработка кластеров различных мощностей и невыпуклых форм

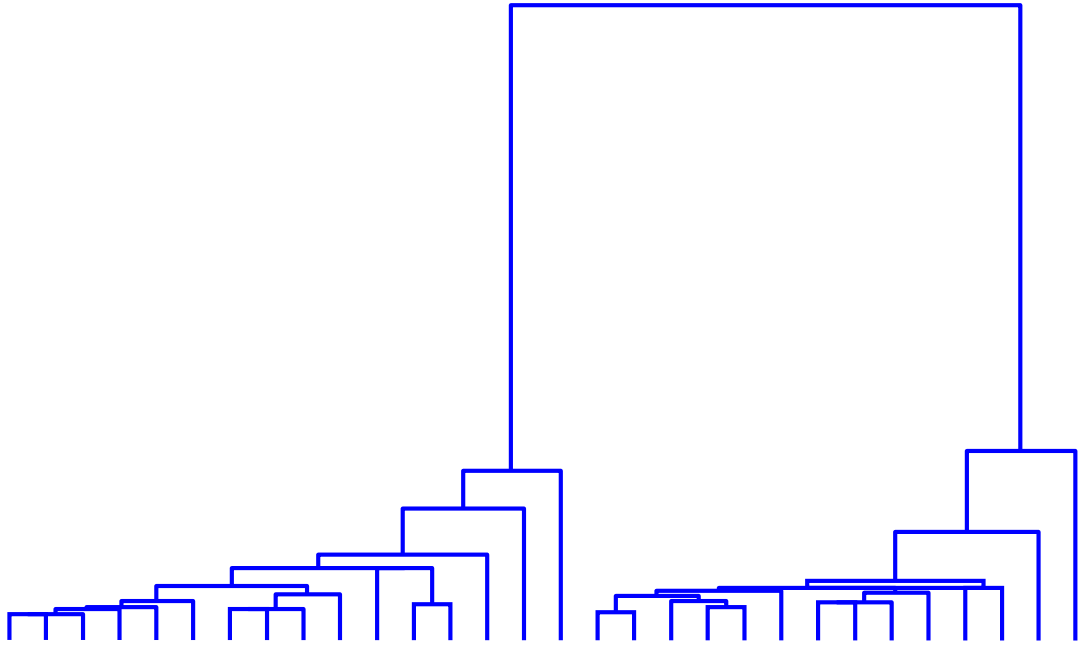
Содержание

- Основные концепции
- Разделительная кластеризация
- Иерархическая кластеризация
- **Меры качества кластеризации**

Оптимальное число кластеров: эмпирические методы

- $k = \sqrt{n/2}$
  - надеемся, что в каждом кластере $\sqrt{2n}$ объектов
- Визуализация иерархической кластеризации



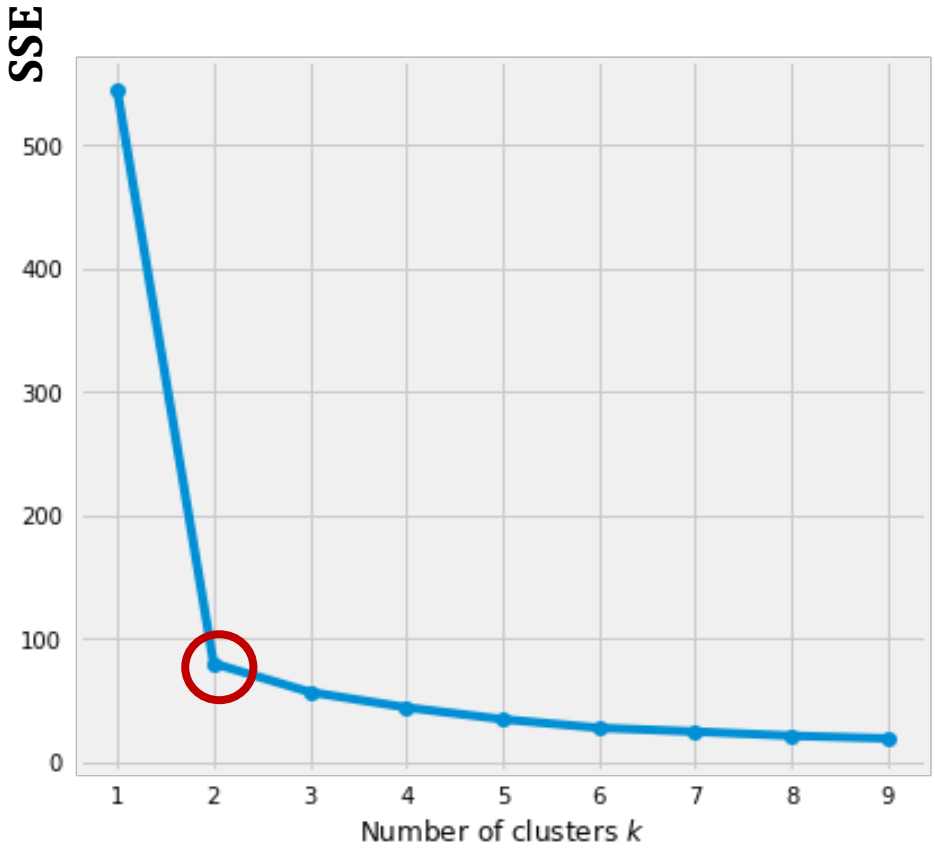


Оптимальное число кластеров: метод локтя

$$SSE(D) = \sum_{i=1}^{k} \sum_{j=1}^{|C_i|} dist^2(c_i, x_j)$$

$D = \{x_j\}$ – исходное множество

c_i – центроид кластера C_i

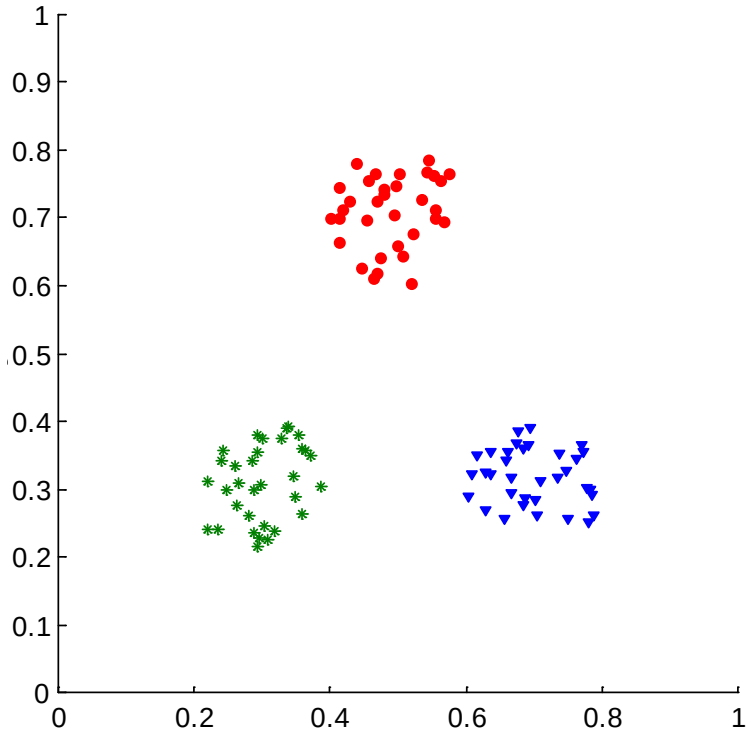


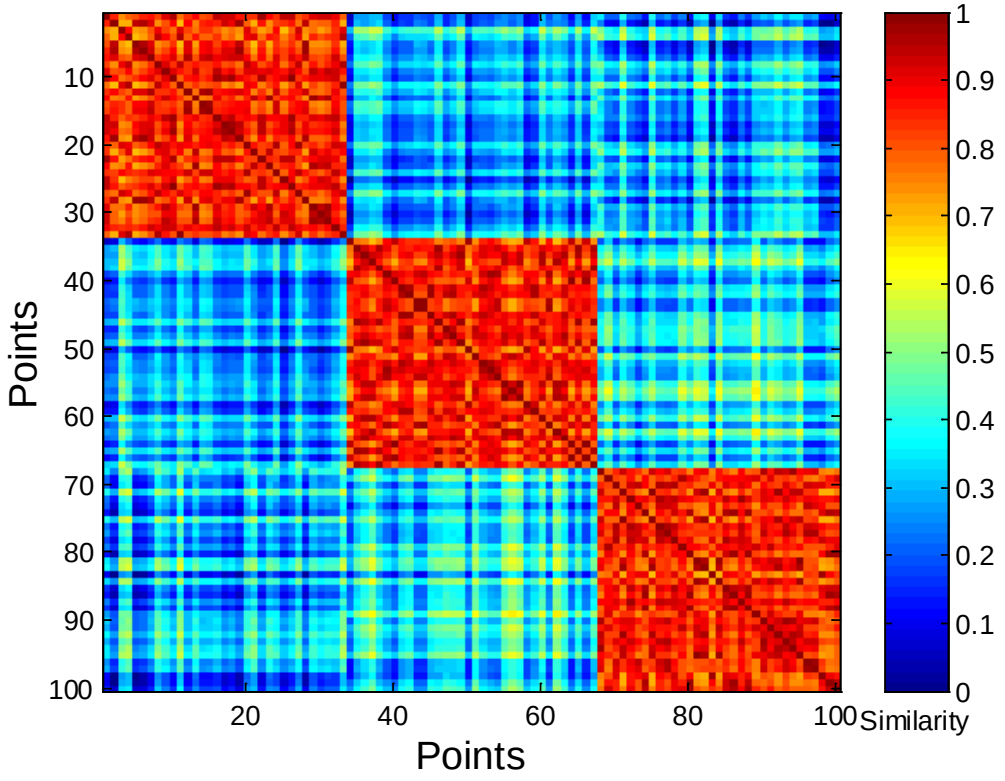
Кросс-валидация для оптимального числа кластеров

- Разбить исходное множество на m частей
- Для различных k
  - Выполнить кластеризацию $m - 1$ частей
  - Для m-й части вычислить $SSE = \Sigma dist^2(c, x)$, где x – объект этой части, c – ближайший к нему центроид
  - Повторить кластеризацию и вычисление SSE для всех частей
  - Вычислить среднее SSE
- Взять значение k, при котором значение SSE максимально

Визуальная оценка качества кластеризации по матрице расстояний (схожести)

- Упорядочить объекты кластеризованного множества по меткам кластеров и визуализировать матрицу расстояний (схожести)

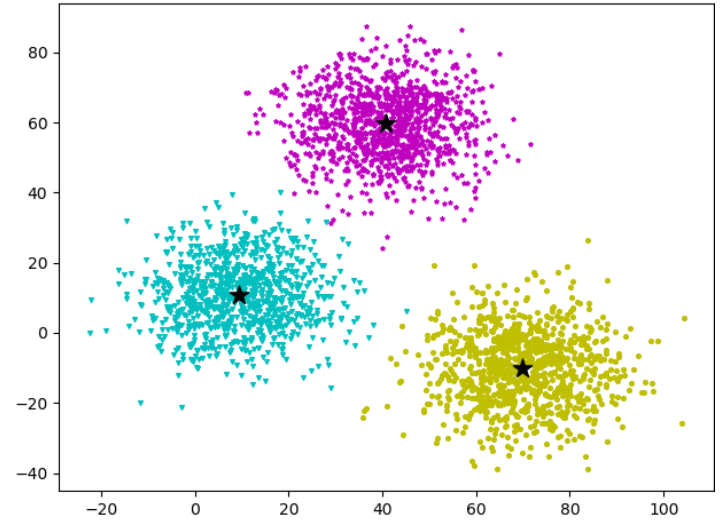




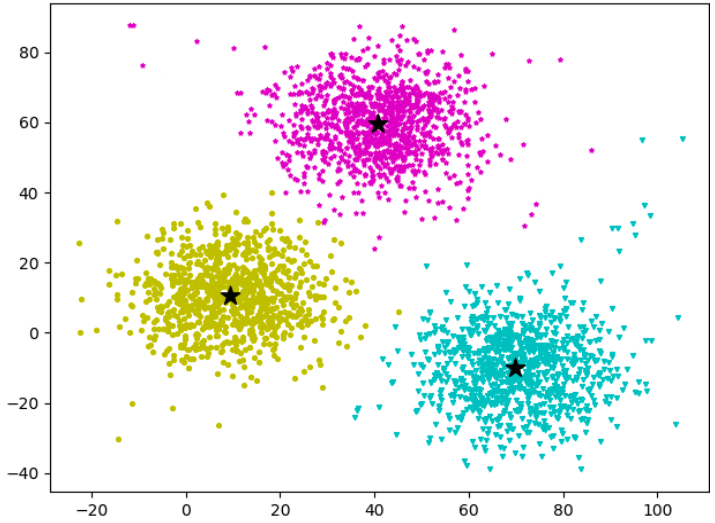
Силуэтный коэффициент

- ***Сцепление*** точки с другими точками того же кластера:
  среднее расстояние от данной точки до других точек кластера
  - $a(p) = \frac{1}{|C_i|-1} \sum_{q \in C_i \wedge q \neq p} dist(p,q)$
- ***Отдаленность*** точки от точек других кластеров:
  минимум среднего расстояния до точек других кластеров
  - $b(p) = \min\limits_{C_j \neq C_i} \frac{1}{|C_j|} \sum_{q \in C_j} dist(p,q)$
- Силуэтный коэффициент
  - для одной точки: $s(p) = \frac{b(p)-a(p)}{\max\{a(p),b(p)\}}$
  - для множества точек: $S(D) = \frac{1}{|D|} \sum_{p \in D} s(p)$
- $-1 \leq s, S \leq 1$
  - Чем ближе к 1, тем выше качество кластеризации
  - При отрицательном значении качество кластеризации низкое

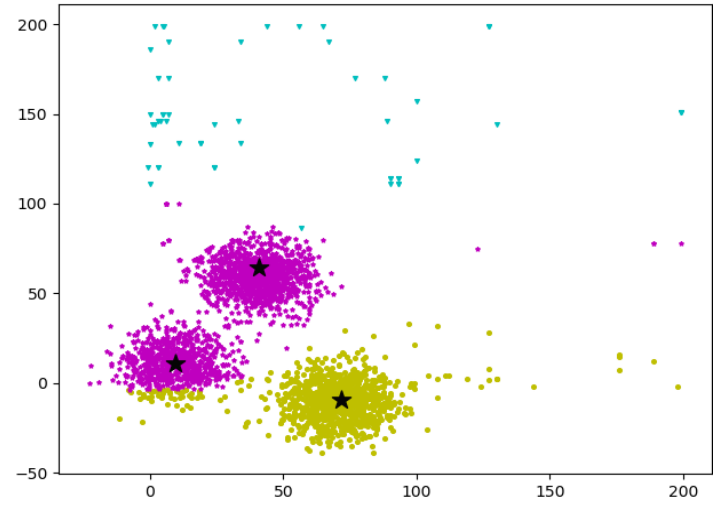
Пример: вычисление силуэтного коэффициента



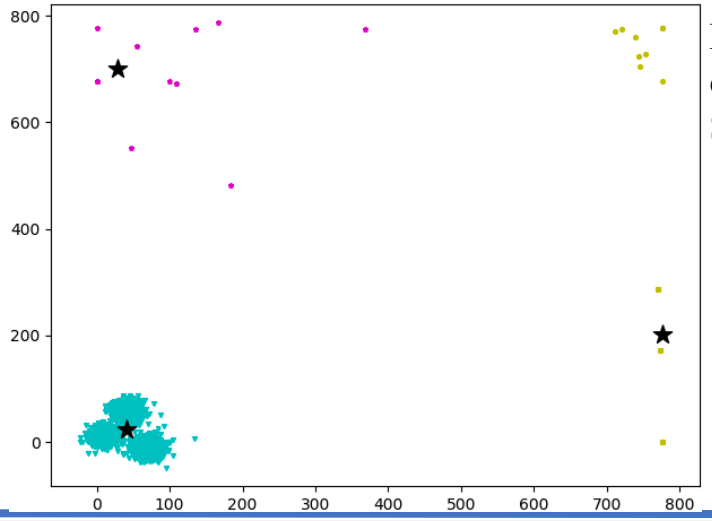
В данных
нет шума
S=0.88



В 3% данных
есть шум
S=0.62



В 5% данных
есть шум
S=0.39



В 10% данных
есть шум
S=-0.05

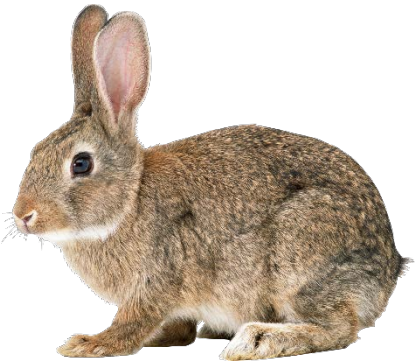
Содержание

- Основные концепции
- Разделительная кластеризация
- Иерархическая кластеризация
- Меры качества кластеризации
- **Нечеткая кластеризация**

Четкая *vs.* нечеткая кластеризация

A

B



C

D

E

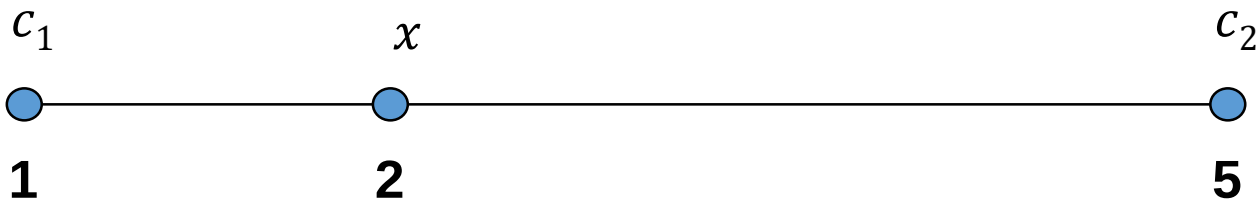
Objects	C_1	C_2
A	1	0
B	1	0
C	0	1
D	0	1
E	0	1

Objects	C_1	C_2
A	0.90	0.10
B	0.80	0.20
C	0.15	0.85
D	0.30	0.70
E	0.25	0.75

Четкая кластеризация

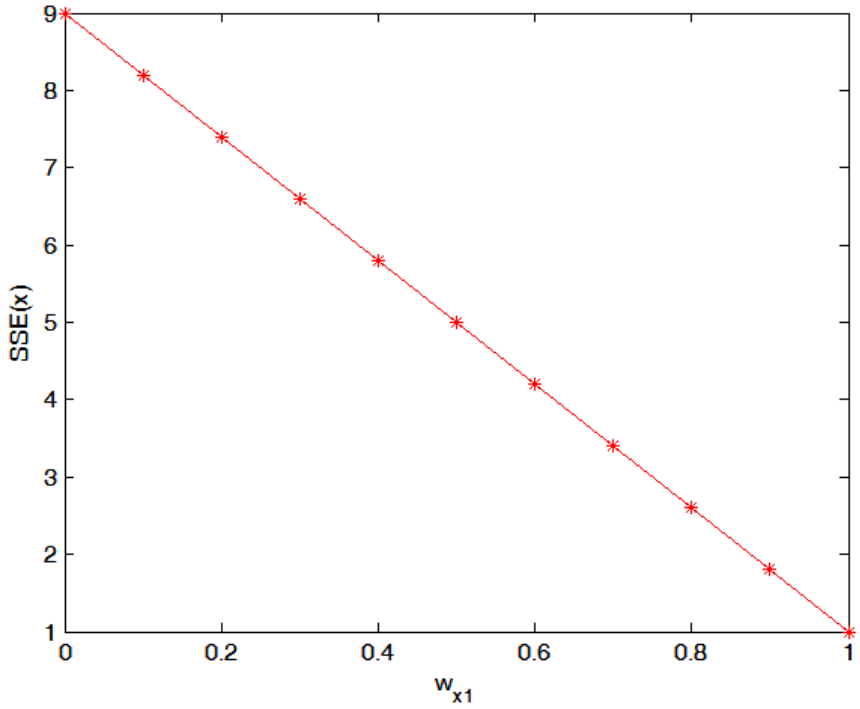
- $SSE = \sum_{j=1}^{k} \sum_{i=1}^{n} w_{ij}\, dist(x_i, c_j)^2$
- w_{ij} – вес факта $x_i \in c_j$, $\sum_{j=1}^{k} w_{ij} = 1$
- Минимизация SSE
  - Фиксировать c_j и найти w_{ij}
  - Фиксировать w_{ij} и вычислить c_j
- $w_{ij} \in \{0,1\}$

Четкая кластеризация



$$SSE(x) = w_{x1}(2-1)^2 + w_{x2}(5-2)^2 = w_{x1} + 9w_{x2}$$

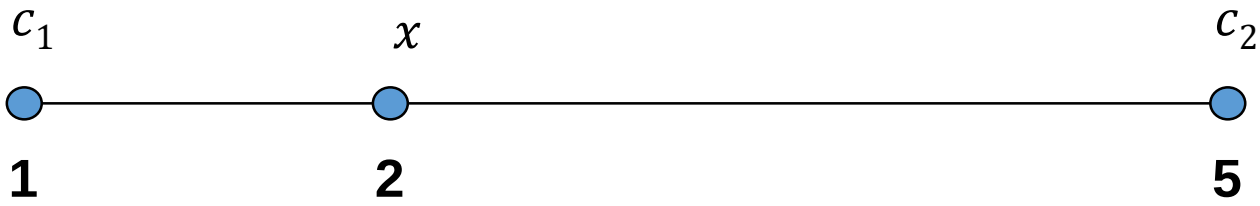
$SSE(x) \to min$ при
$w_{x1} = 1, w_{x2} = 0$



Нечеткая кластеризация

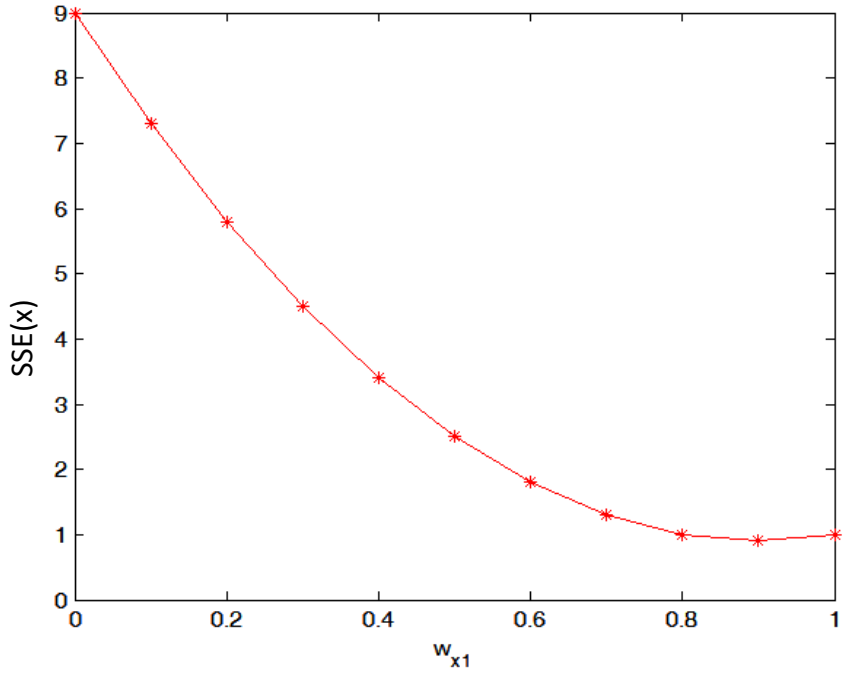
- $SSE = \sum_{j=1}^{k} \sum_{i=1}^{n} w_{ij}^{m} \, dist(x_i, c_j)^2$
- w_{ij} – вес факта $x_i \in c_j$, $\sum_{j=1}^{k} w_{ij} = 1$
- m $(m > 1)$ – «размытость» кластеров (обычно $m = 2$)
- Минимизация SSE
  - Фиксировать c_j и найти w_{ij}
  - Фиксировать w_{ij} и вычислить c_j
- $w_{ij} \in [0,1]$

Нечеткая кластеризация



$$SSE(x) = w_{x1}^2(2-1)^2 + w_{x2}^2(5-2)^2 = w_{x1}^2 + 9w_{x2}^2$$

$SSE(x) \to min$ при
$w_{x1} = 0.9, w_{x2} = 0.1$



Алгоритм Fuzzy *c*-Means

- k – количество кластеров
- $X = \{x_1, x_2, \ldots, x_n\}$ – множество d-мерных точек
- $C \in \mathbb{R}^{k \times d}$ – *матрица центроидов*
  - c_j – центр j-го кластера (d-мерный вектор)
- $U \in \mathbb{R}^{n \times k}$ – *матрица принадлежности*
  - $0 \le u_{ij} \le 1$ – степень принадлежности (расстояние) между точкой x_i и центроидом c_j
- Минимизируемая *целевая функция*
  - $J_{FCM}(X, k, m) = \sum_{j=1}^{k} \sum_{i=1}^{n} u_{ij}^{m} \, dist(x_i, c_j)^2$
  - $m > 1$ – размытость

x	$x_{i,1}$	…	$x_{i,d}$
1			
…			
n			

c	$c_{j,1}$	…	$c_{j,d}$
1			
…			
k			

u	1	…	k
1			
…			
n			

Алгоритм Fuzzy *c*-Means

Input: X, k, m, ε

Output: U

$s := 0, U^{(0)} := (u_{ij})$ {initialization}

repeat

{computation of new centroids' coordinates}

Compute $C^{(s)} := (c_j)$ using $\forall j, l \quad c_{jl} = \dfrac{\sum_{i=1}^{n} u_{ij}^m \cdot x_{il}}{\sum_{i=1}^{n} u_{ij}^m}$

where $u_{ij} \in U^{(s)}$

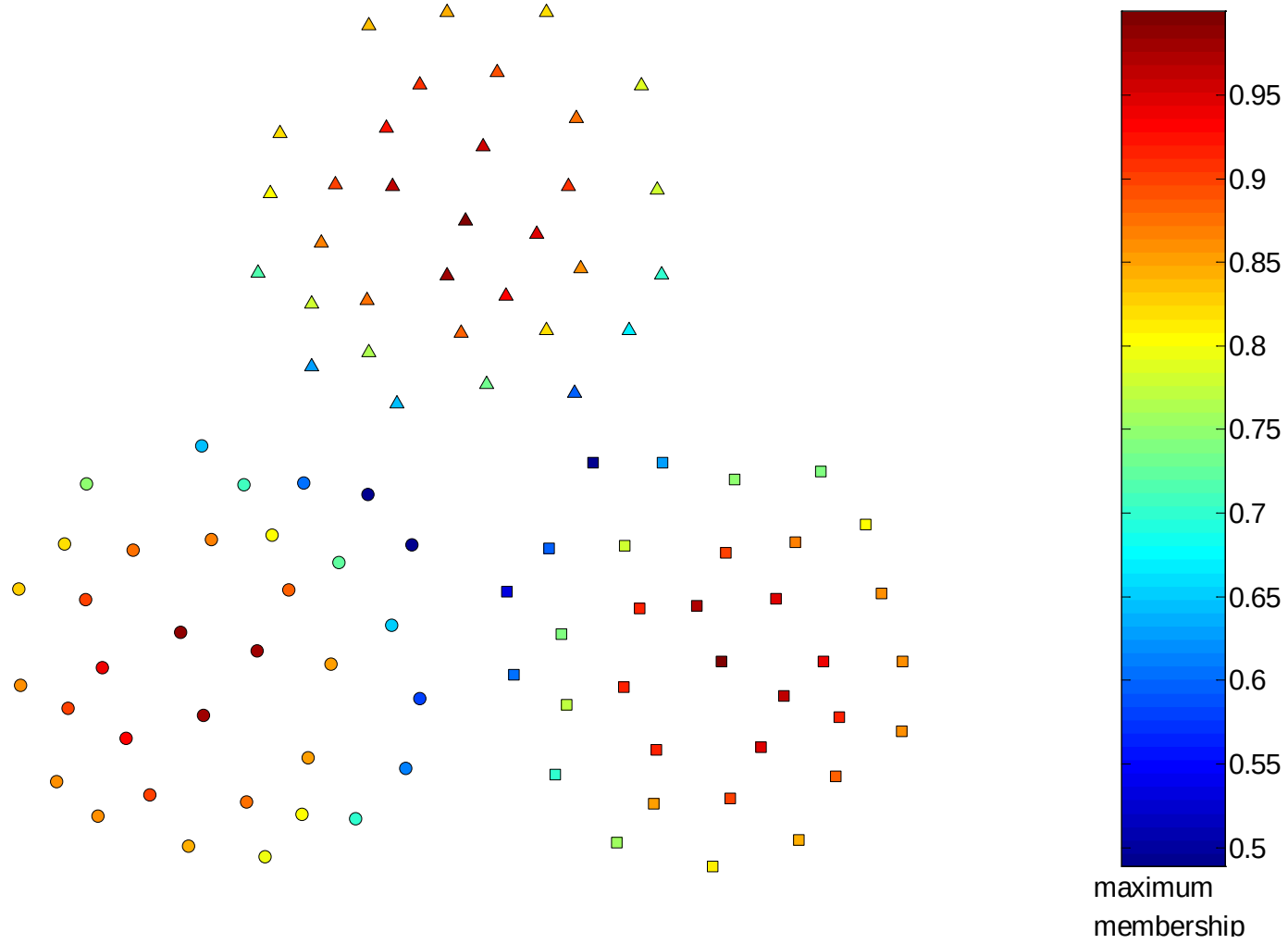
{update matrixes values}

Compute $U^{(s)}$ and $U^{(s+1)}$ using $u_{ij} = \sum_{t=1}^{k} \left(\dfrac{\rho(x_i, c_j)}{\rho(x_i, c_t)} \right)^{\frac{2}{1-m}}$

$s := s + 1$

until $\max_{ij}\{|u_{ij}^{(s)} - u_{ij}^{(s-1)}|\} \geqslant \varepsilon$

Алгоритм Fuzzy c-Means: пример работы

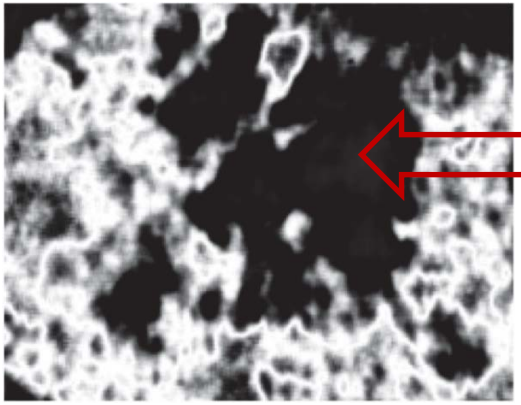


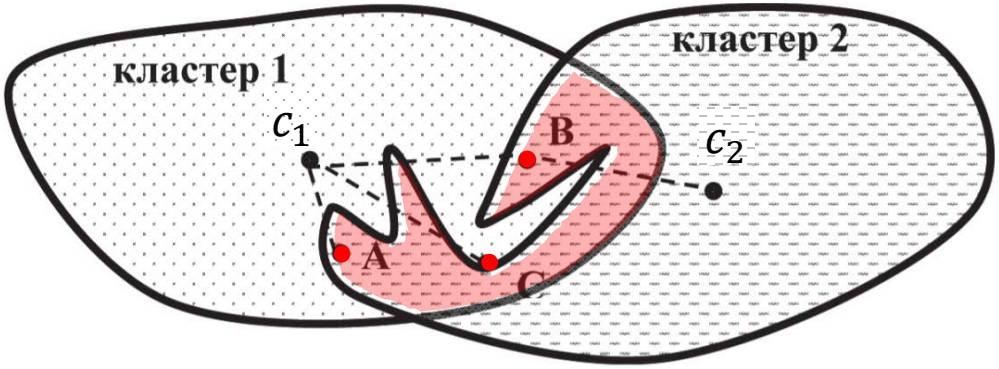
За и против Fuzzy *c*-Means

- Достоинства
  - Дает лучшие результаты для перекрывающихся кластеров и сравнительно лучшие результаты, чем k-means
- Недостатки
  - Количество кластеров необходимо задавать
  - Меньшее значение ε улучшает результаты, но ценой большего количества итераций

Применение нечеткой кластеризации

- Сегментация радиологических изображений





- Восстановление пропущенных данных
  - Нечеткая кластеризация точек без пропущенных координат
  - Создание прототипов: замена пропущенных координат соотв. координатами центроидов
  - Выбор прототипа с минимальным расстоянием до центроида

Литература

- Han J., Kamber M., Pei J. Data Mining: Concepts and Techniques. 3rd Edition. Morgan Kaufmann, 2011. 740 p. ISBN 978-0123814791
  - 10. Cluster Analysis: Basic Concepts and Methods; 10.1 Cluster Analysis; 10.2 Partitioning Methods, pp. 443-457
- Tan P.-N., Steinbach M., Karpatne A., Kumar V. Introduction to Data Mining. 2nd Edition. Pearson, 2019. 839 p. ISBN: 978-0-13-312890-1
  - 7. Cluster Analysis: Basic Concepts and Algorithms; 7.1 Overview; 7.2 K-means, pp. 525-553