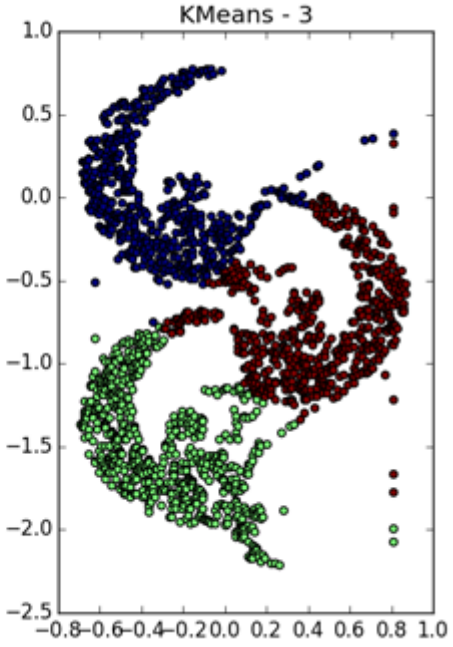
Задача кластеризации данных



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

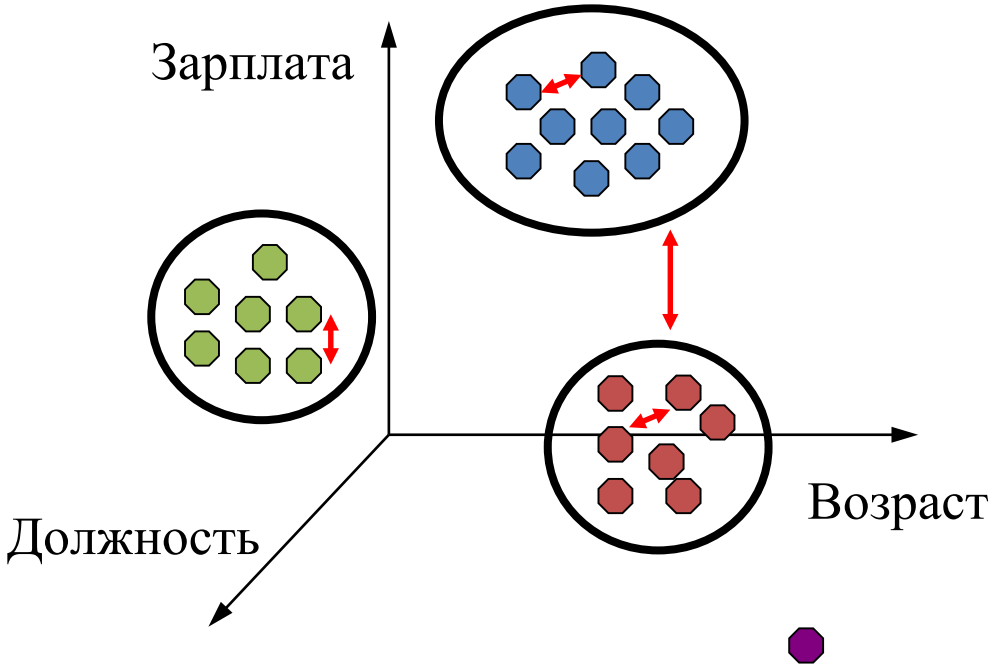
Франклин Рузвельт

Содержание

- **Основные концепции**
- **Разделительная кластеризация**
- Иерархическая кластеризация
- Плотностная кластеризация
- Нечеткая 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

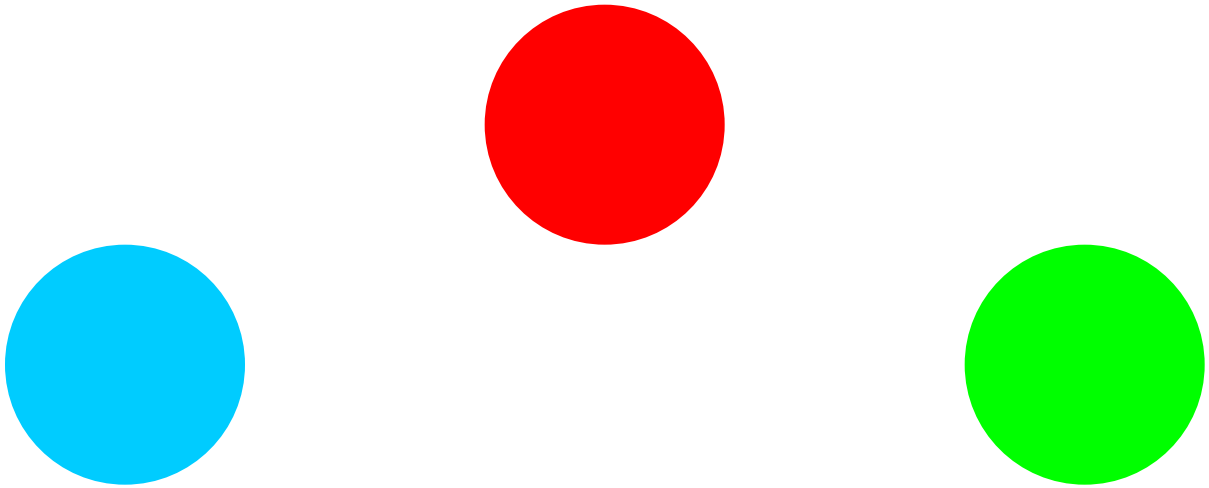
- Выбор функции расстояния (или меры схожести)
- Размерность данных
- Типы атрибутов
- Зависимости между атрибутами
- Распределение данных
- Наличие шумов и выбросов в данных

Виды кластеризации

- Эксклюзивная *vs.* инклюзивная
  - Объект принадлежит одному или нескольким кластерам
- Нечеткая *vs.* четкая
  - Объект принадлежит каждому кластеру с некоторым весом (вероятностью), сумма весов равна 1
- Частичная *vs.* полная
  - Выполняется кластеризация части либо всего исходного множества объектов
- Гетерогенная *vs.* гомогенная
  - Допускается ли неоднородность в размерах, плотности и форме кластеров

Хорошо отделимые кластеры

- Любой объект в кластере ближе (более похож) к любому другому объекту в кластере, чем к любому объекту вне данного кластера



3 хорошо отделимых кластера

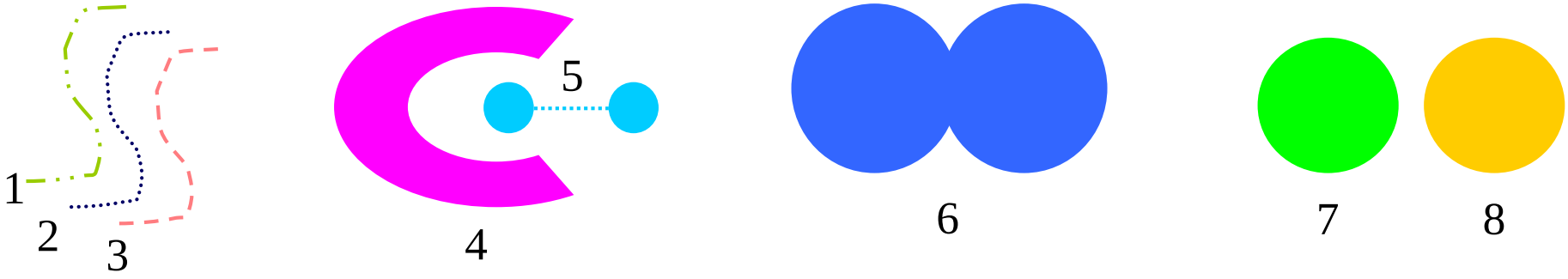
Центро-ориентированные кластеры

- Объект в кластере ближе (более похож) к «центру» кластера, чем к центру любого другого кластера
- Центр кластера
  - центроид: усреднение координат всех объектов в кластере
  - медоид: наиболее «репрезентативный» объект кластера

4 центро-ориентированных кластера

Смежно-ориентированные кластеры

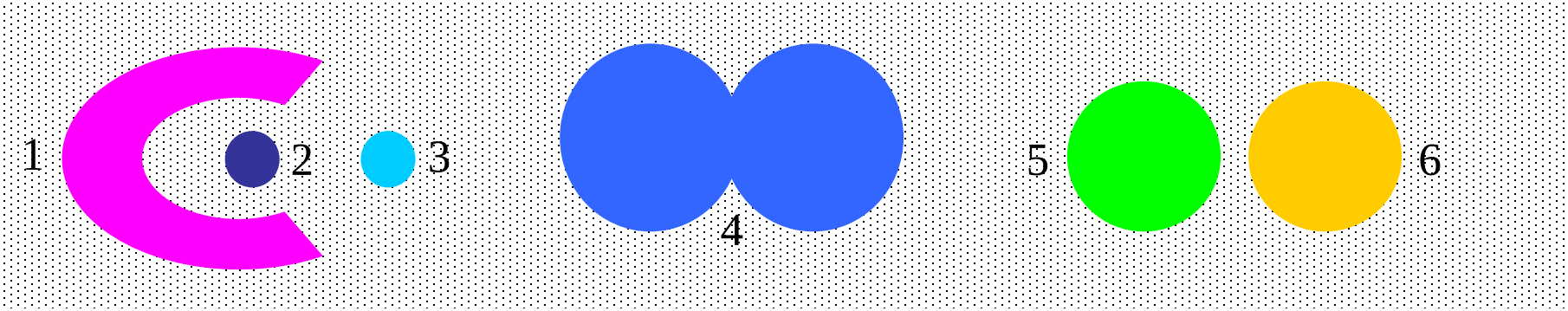
- Объект в кластере ближе (или более похож) к одному или нескольким другим объектам в кластере, чем к любому объекту вне кластера



8 смежно-ориентированных кластеров

Плотностно-ориентированные кластеры

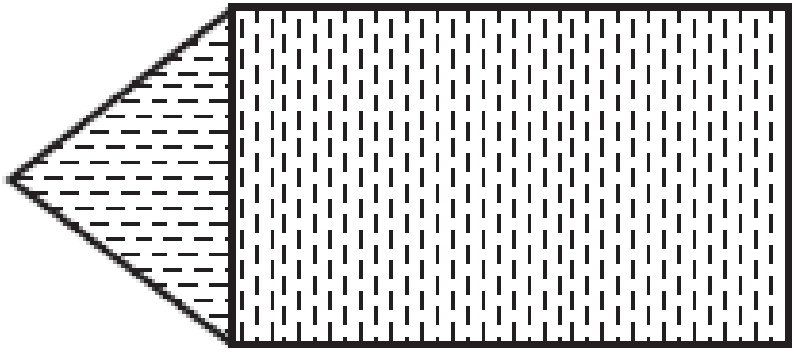
- Плотное скопление объектов, которое отделено скоплениями с низкой плотностью объектов от других скоплений с высокой плотностью
  - нерегулярные или переплетенные кластеры
  - шумы и выбросы



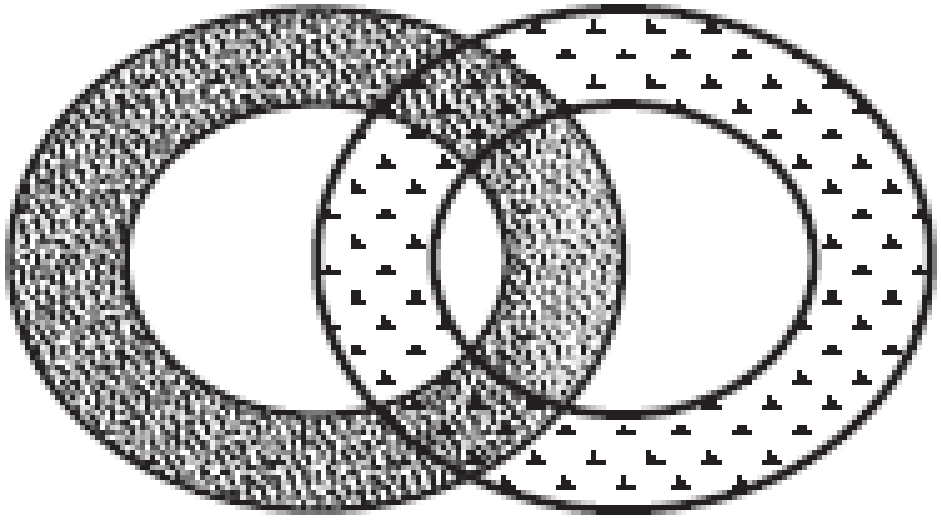
6 плотностно-ориентированных кластеров

Концептуальные кластеры

- Объекты кластера имеют одно и то же свойство или выражают одну и ту же концепцию



2 концептуальных кластера



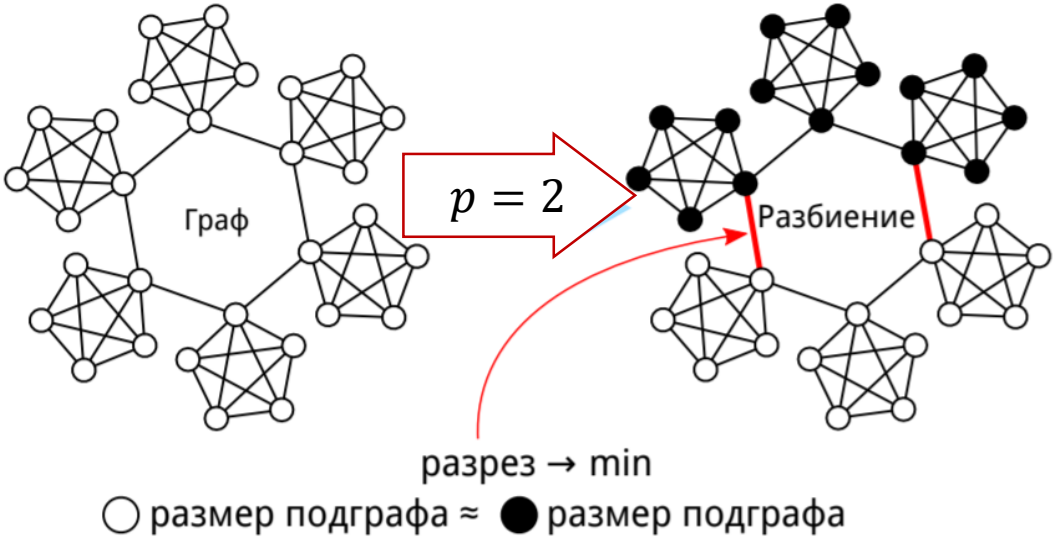
2 концептуальных кластера

(точки в областях пересечения принадлежат обоим кластерам)

Кластеры на базе целевой функции

- Объекты кластеров минимизируют или максимизируют целевую функцию
- Перечисление всех возможных способов разделения объектов на кластеры, оценка «качества» каждого набора кластеров заданной целевой функцией (NP-трудная задача)
- Глобальные или локальные цели
  - Иерархическая кластеризация: обычно локальные цели
  - Разделительная кластеризация: обычно глобальные цели
- Использование глобальной целевой функции: подогнать данные к параметризованной модели
  - Параметры модели определяются по данным
  - Модели смешивания предполагают, что данные представляют собой «смесь» ряда статистических распределений

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. $W_{cut} \to \min, W_{cut} = \sum_{e \in E_{cut}} w(e),$
   $E_{cut} = \{(u, v) \in E \mid u \in N_i, v \in N_j, 1 \leq i, j \leq p, i \neq j\}$

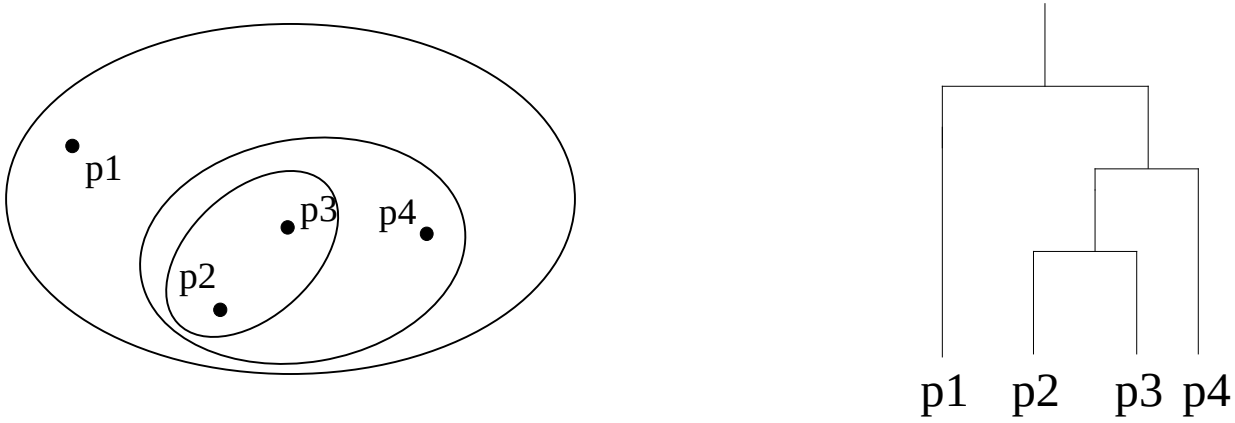
Особенности, требования и вызовы кластеризации

- Масштабируемость
  - Быстродействие и возможность эффективного распараллеливания
- Умение работать с разными типами атрибутов
  - Числовые, булевы, категориальные, порядковые и их сочетание
- Кластеризация с ограничениями
  - Пользователь может вводить данные об ограничениях
  - Использование знаний о предметной области для определения входных параметров
- Интерпретируемость и удобство использования
- Прочее
  - Обнаружение кластеров произвольной формы
  - Устойчивость к шумам
  - Инкрементальная кластеризация, независимость от порядка ввода
  - Большая размерность данных

Базовые подходы: разделительная *vs.* иерархическая кластеризация

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

- Иерархическое дерево пересекающихся подмножеств объектов



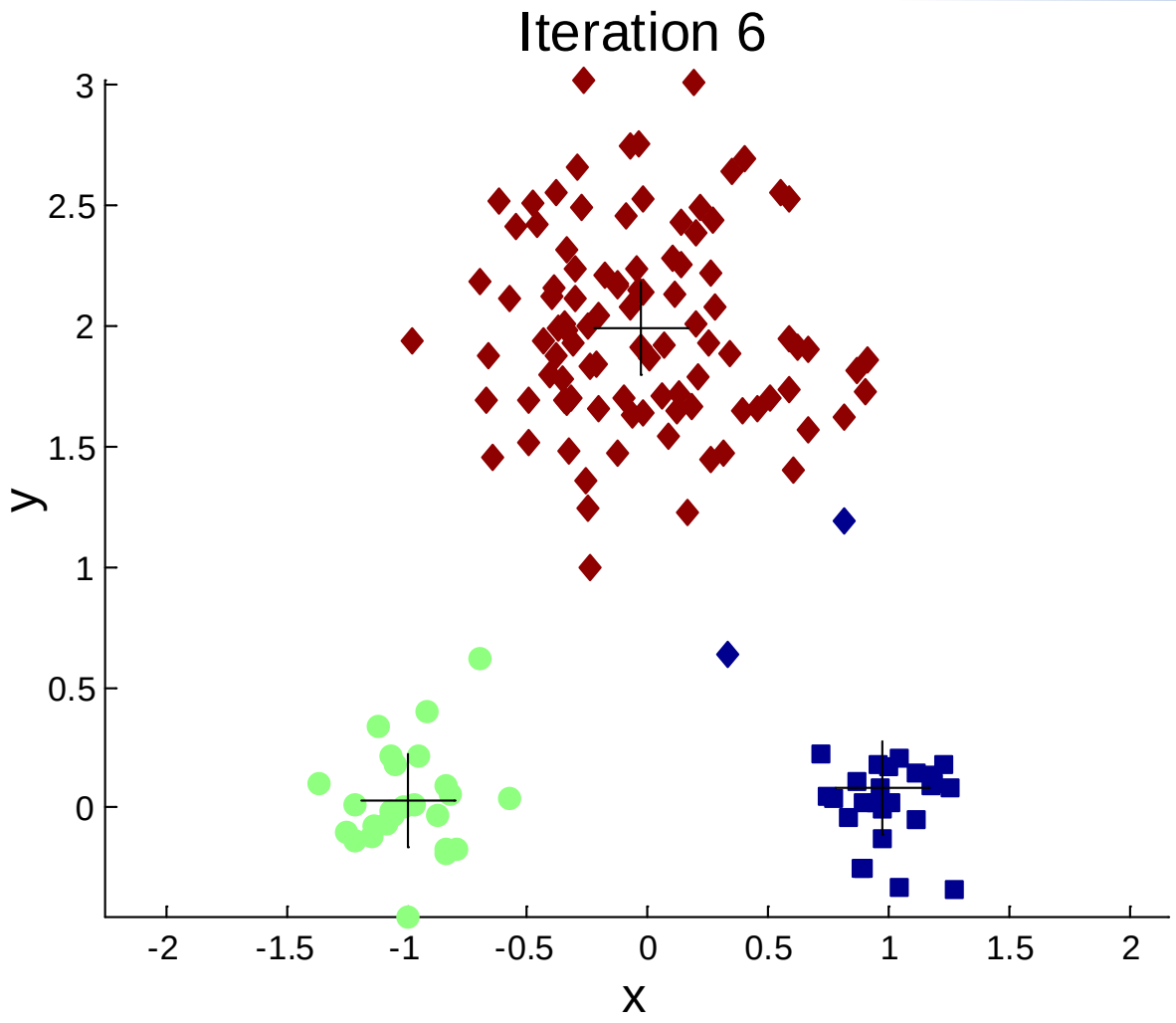
Разделительный алгоритм *k*-means (*k*-средних)

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

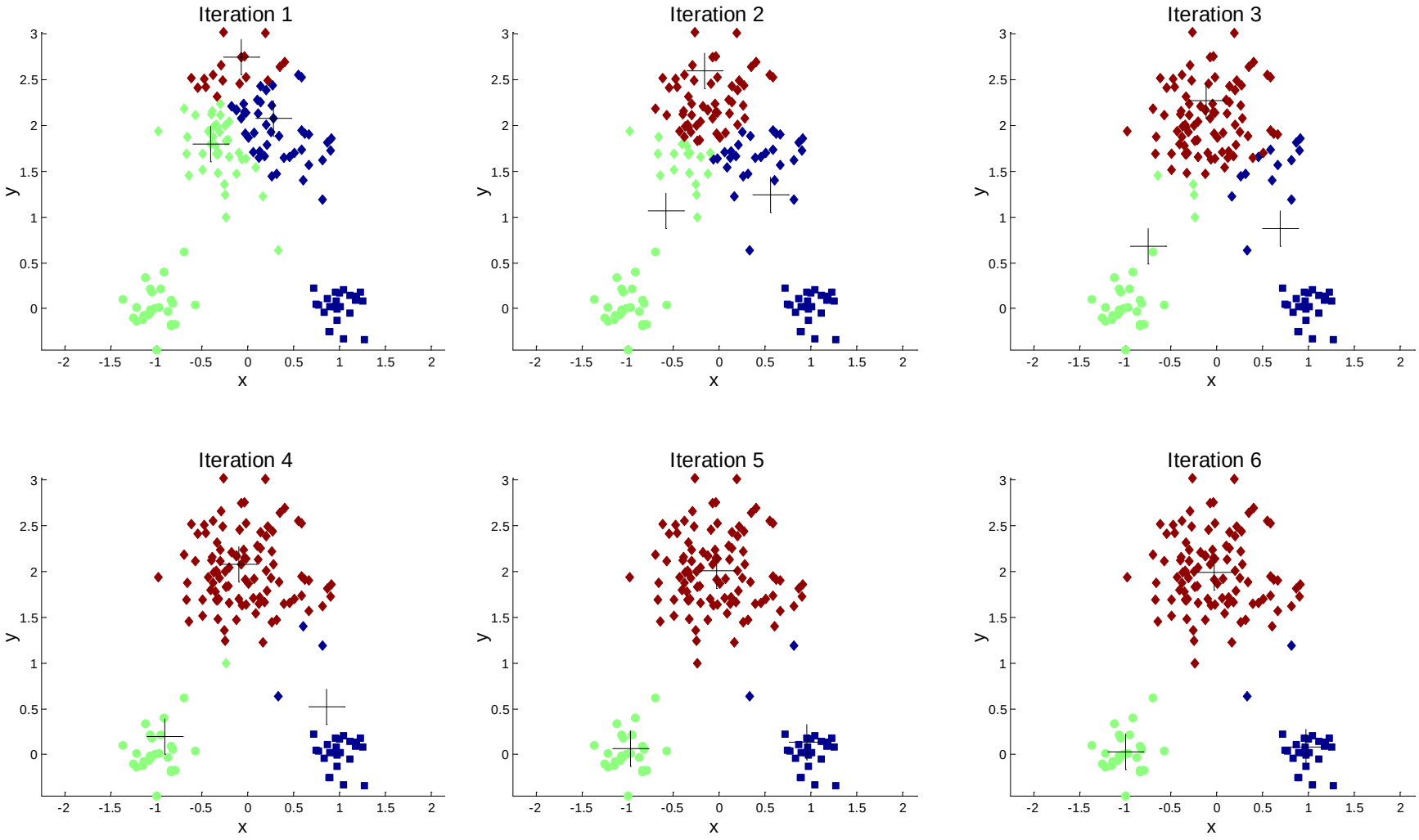
Гуго Штейнгауз (Hugo Steinhaus) (1887-1972)

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

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
  - Может приводить к пустым кластерам в процессе работы
  - Неприменимость к категориальным данным (для них нужно использовать k-modes)
  - Неприменимость для кластеров невыпуклой формы
  - Чувствительность к размеру, плотности, шумам и выбросам в данных

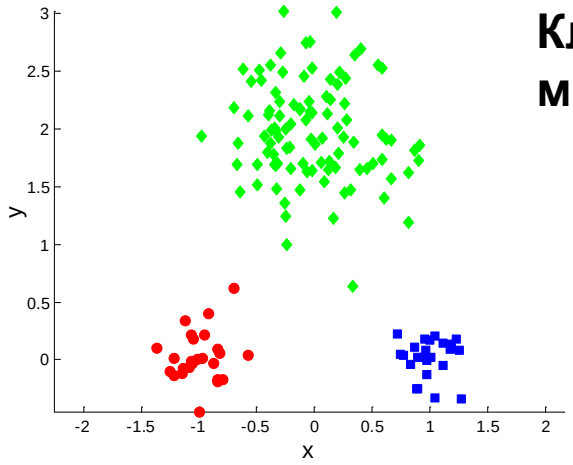
Мера для выявления кластеров в k-means

- Сумма квадратов ошибок, Sum of Squared Errors

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

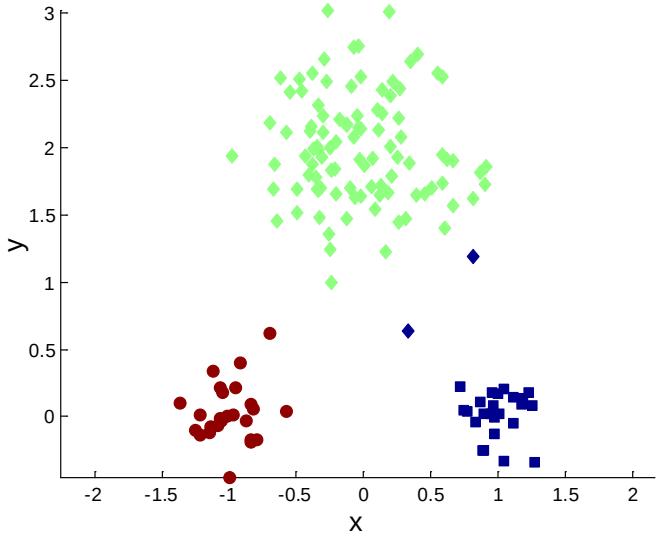
  - C_i $(1 < i \leq k)$ – кластер
  - x – точка кластера
  - m_i – центроид (mean) кластера
- Из двух вариантов кластеризации выбирается имеющий меньшую суммарную ошибку
  - Увеличение k уменьшает ошибку. Хороший вариант кластеризации с меньшим k может иметь меньшее SSE, чем плохой вариант с б*о*льшим k

Различные варианты кластеризации k-means

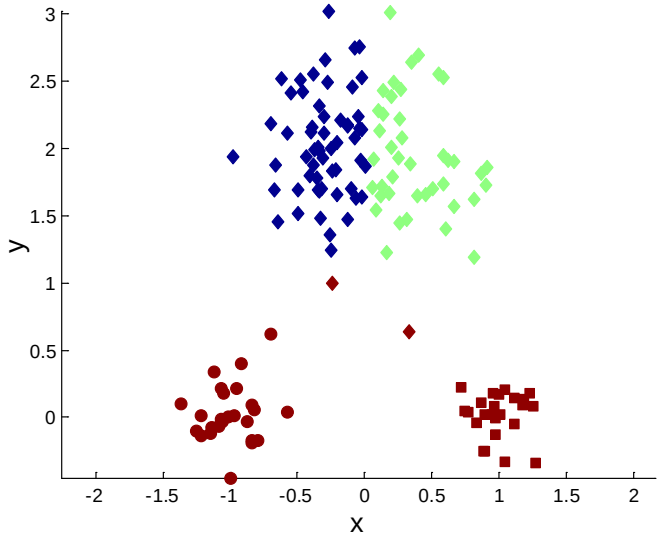


Кластеризуемое множество

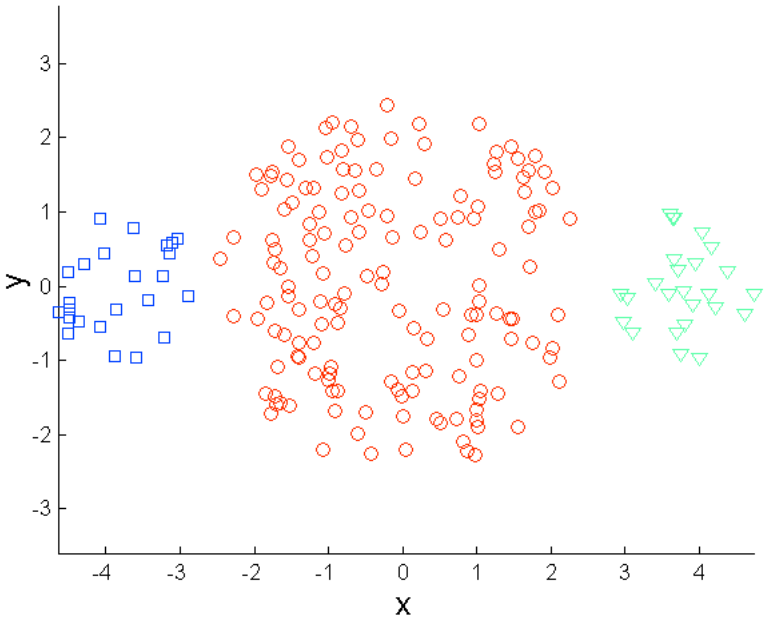
Оптимальная кластеризация



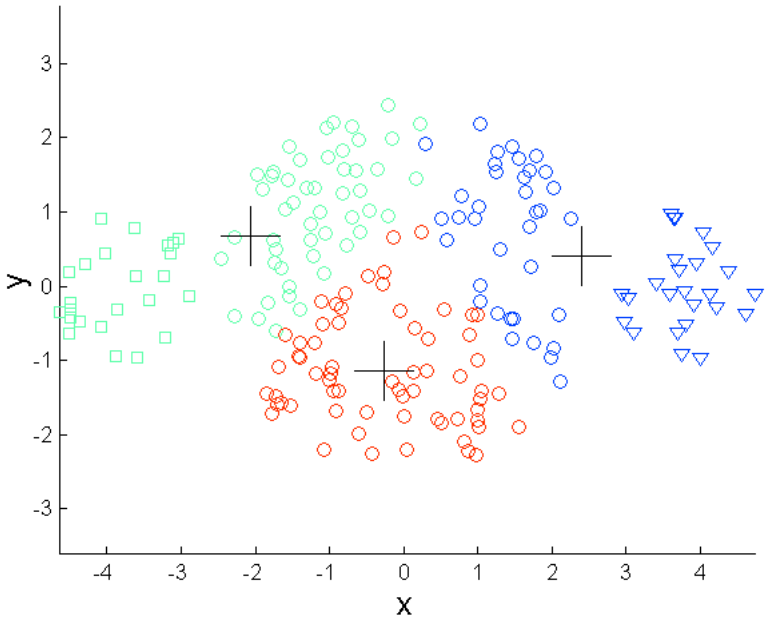
Суб-оптимальная кластеризация



Влияние размеров кластеров

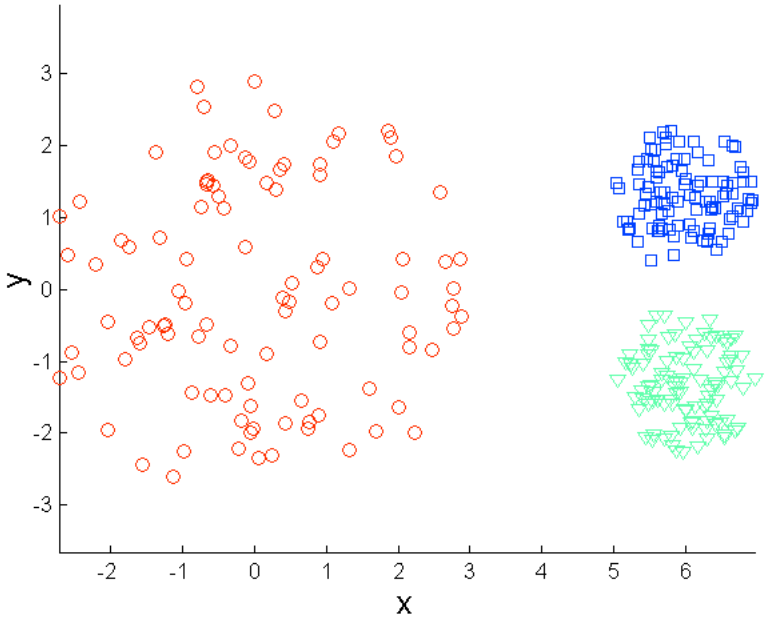


Кластеризуемое множество

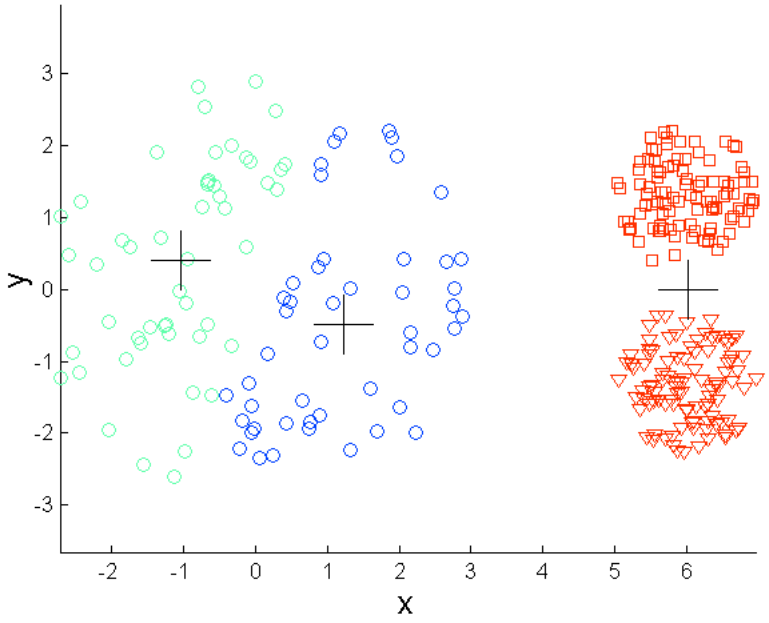


3-means

Влияние плотности кластеров

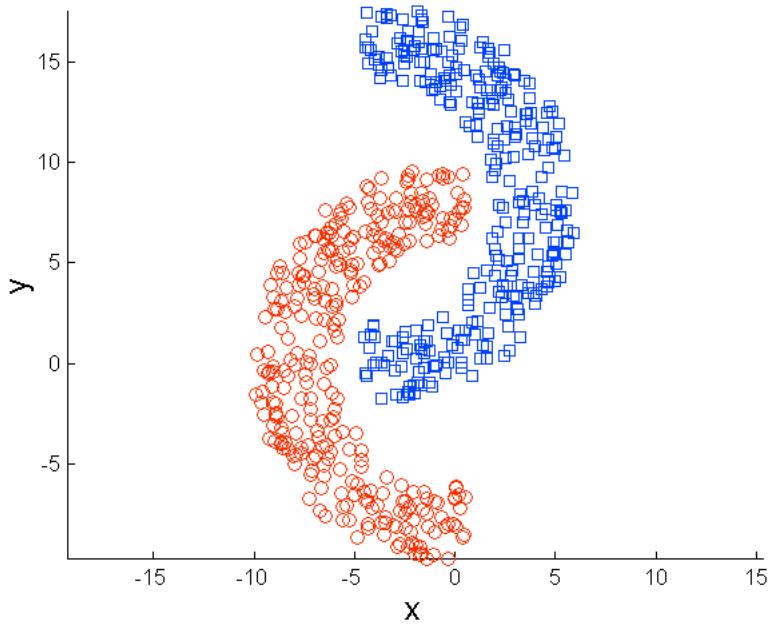


Кластеризуемое множество

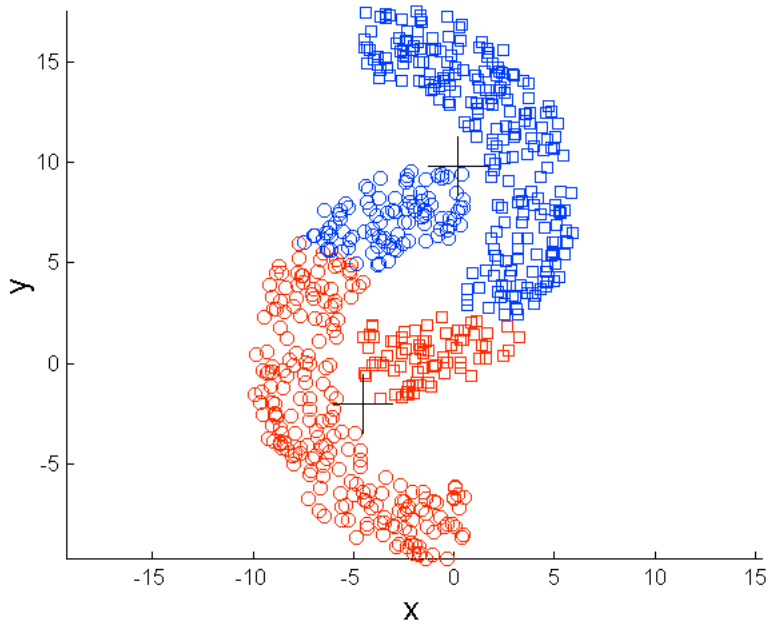


3-means

Влияние формы кластеров

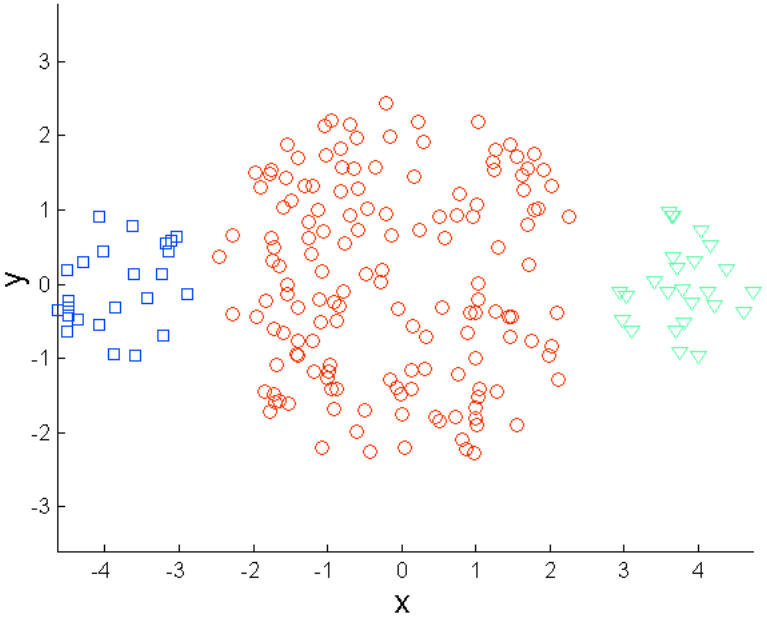


Кластеризуемое множество

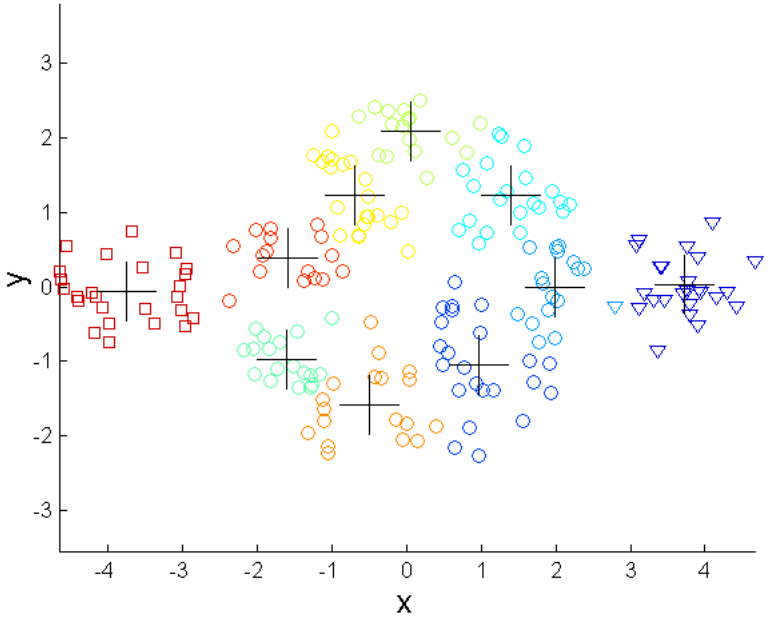


2-means

Увеличение количества кластеров



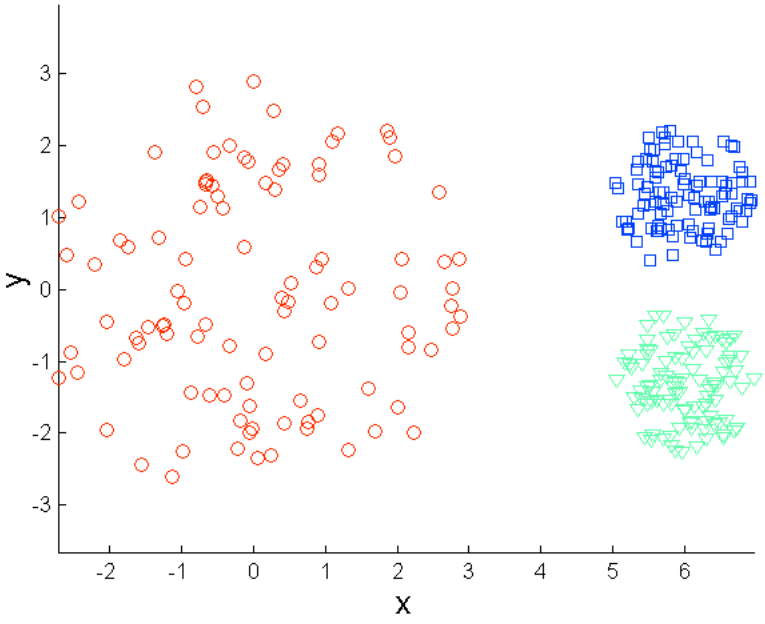
Кластеризуемое множество



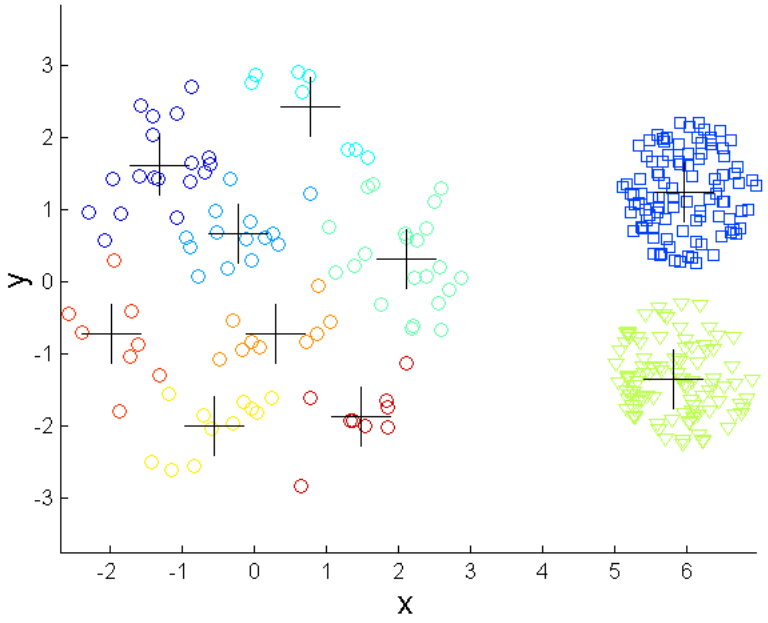
10-means

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

Увеличение количества кластеров



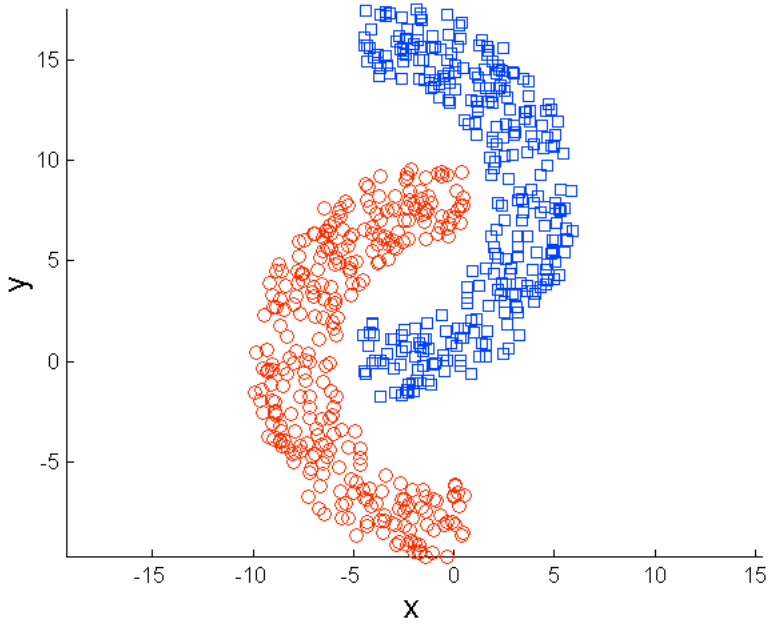
Кластеризуемое множество



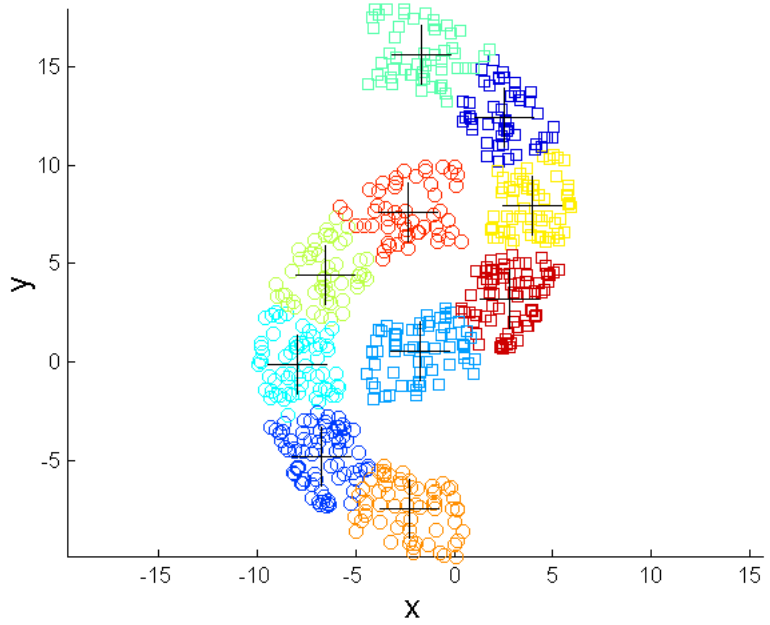
10-means

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

Увеличение количества кластеров



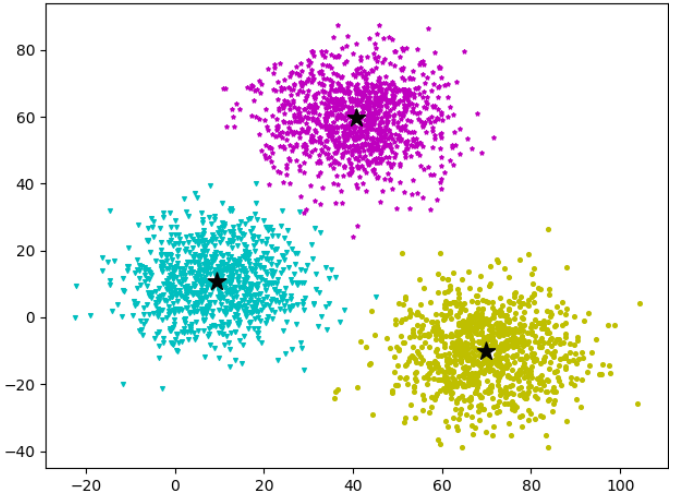
Кластеризуемое множество



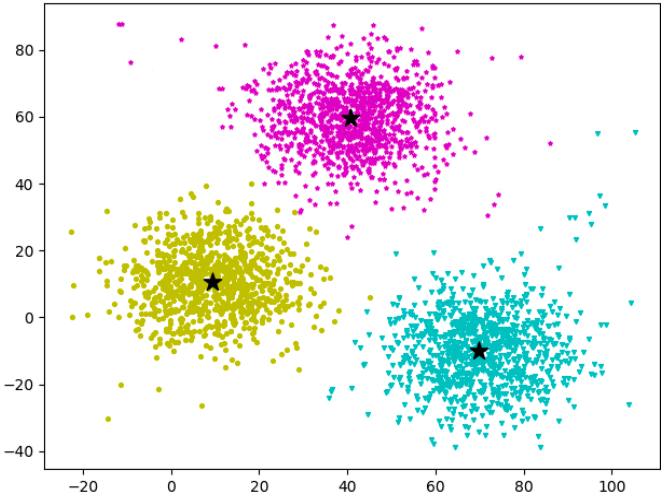
10-means

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

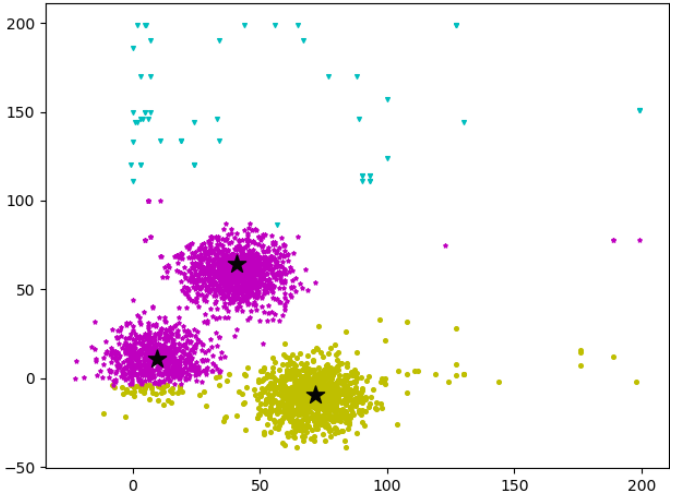
Влияние шумов и выбросов



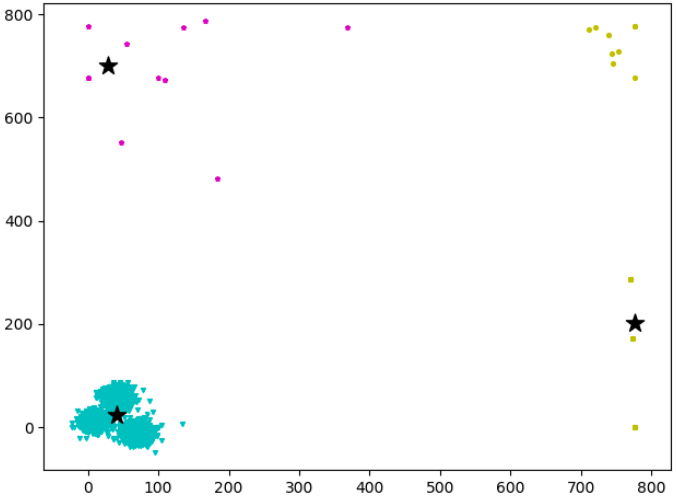
В данных
нет шума
S=0.88



В 3% данных
есть шум
S=0.62

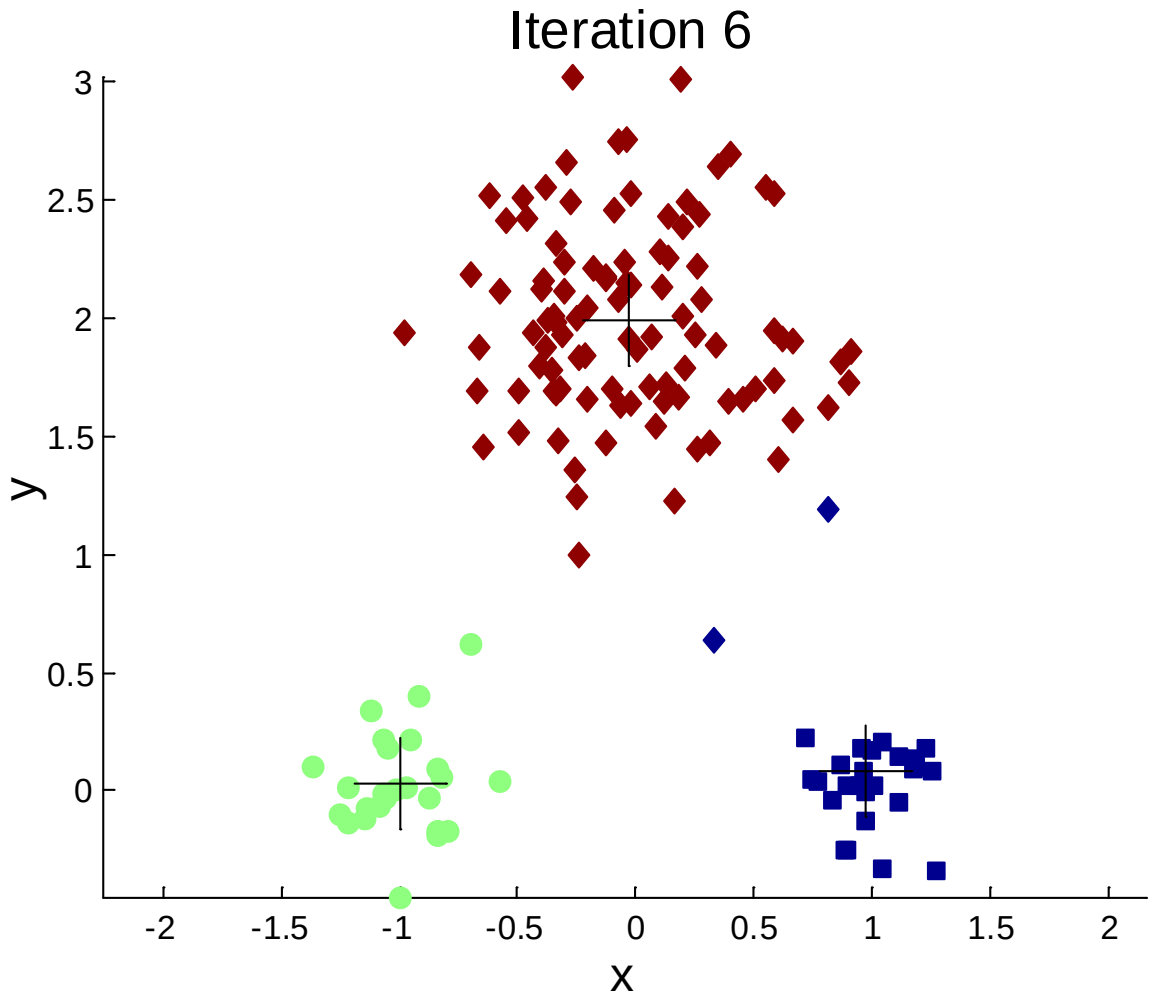


В 5% данных
есть шум
S=0.39

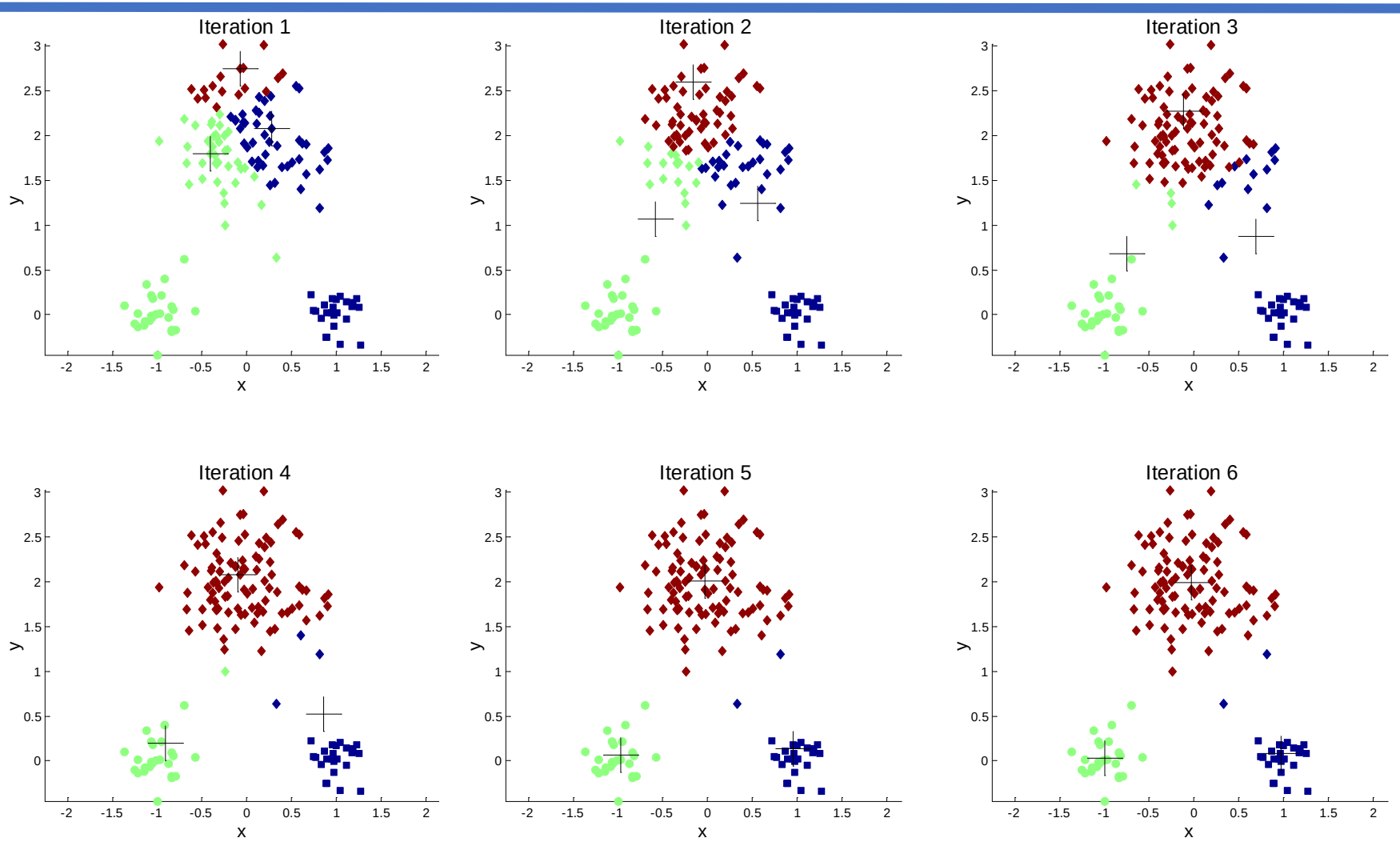


В 10% данных
есть шум
S=-0.05

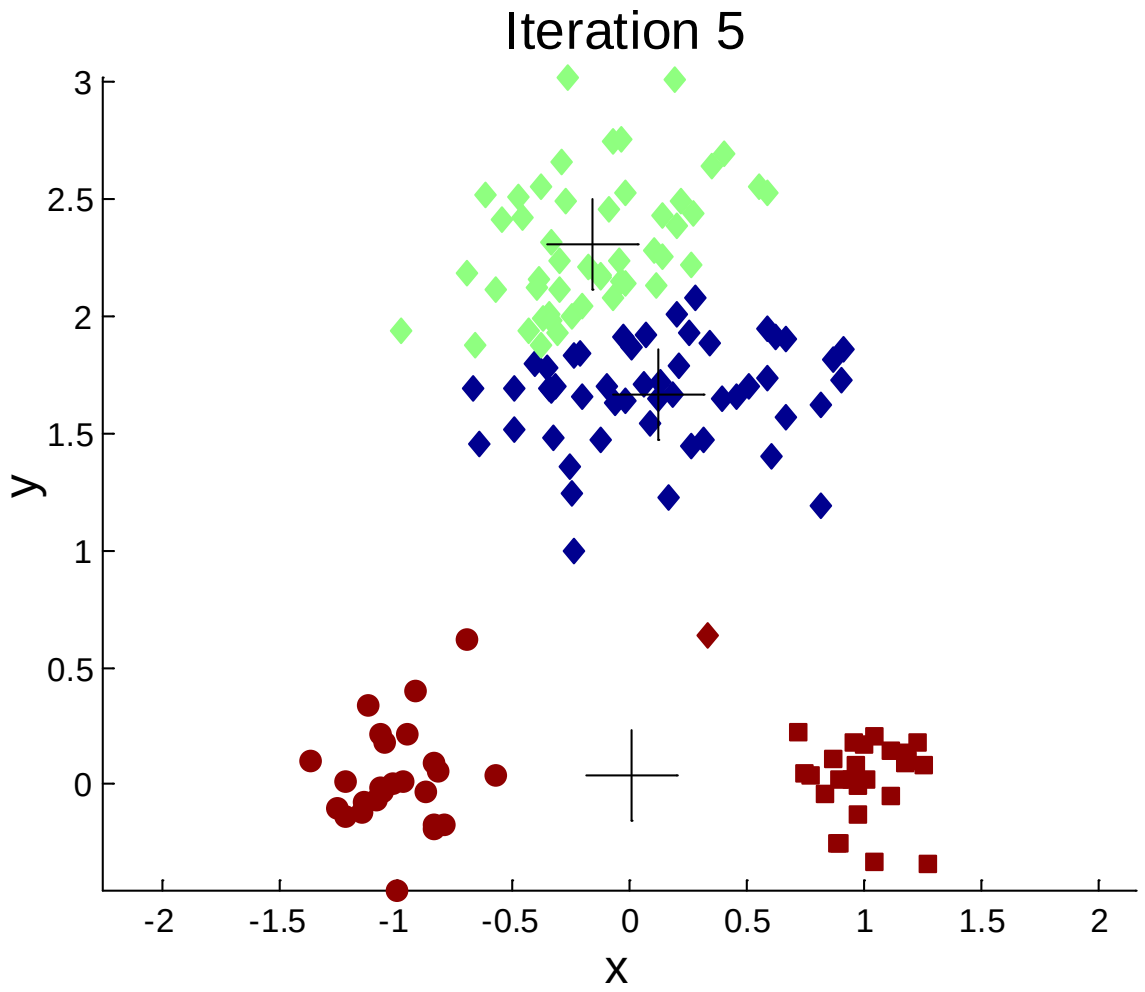
Влияние начального выбора центроидов (1)



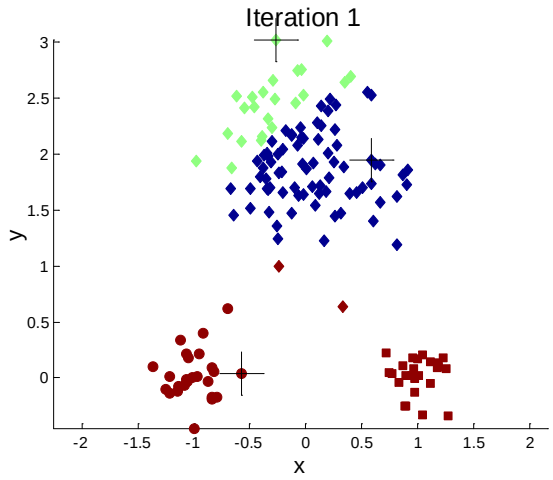
Влияние начального выбора центроидов (1)

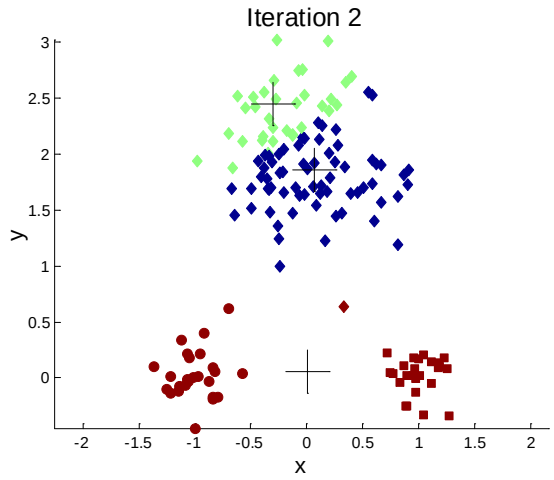


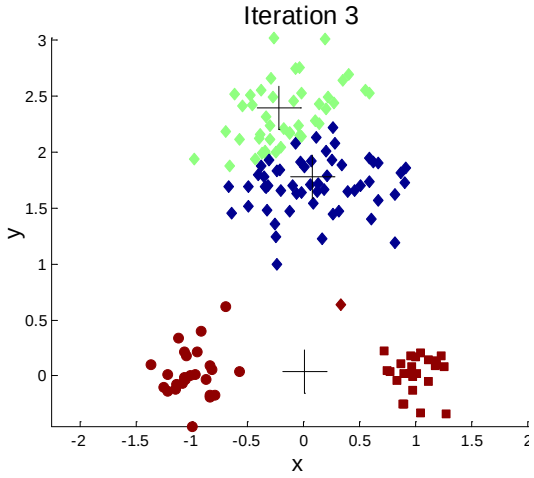
Влияние начального выбора центроидов (2)

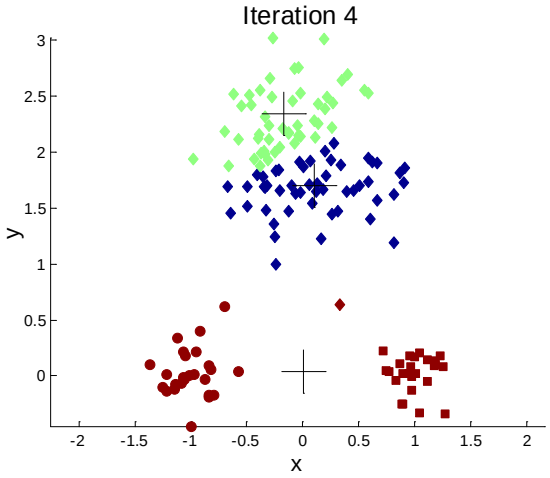


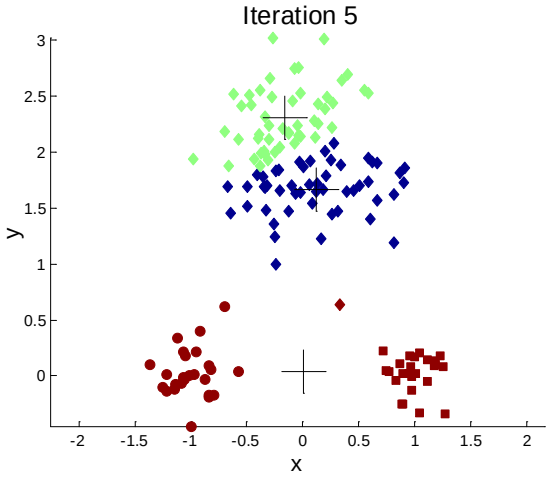
Влияние начального выбора центроидов (2)



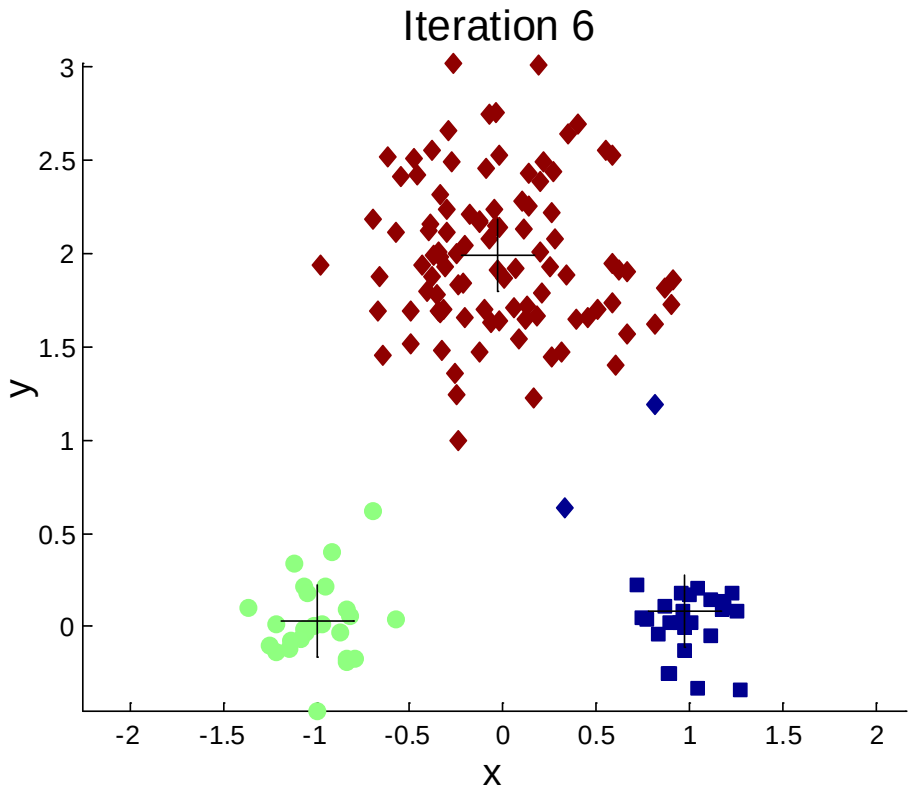


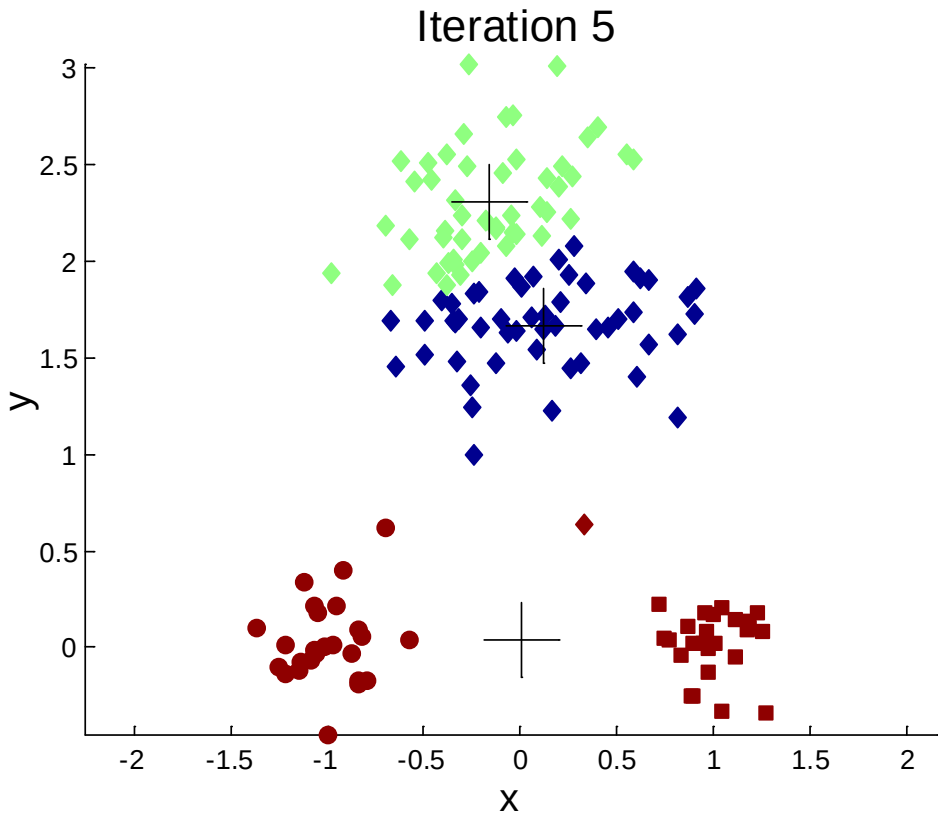






Влияние начального выбора центроидов (1 vs. 2)





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 точек
  - Наиболее удаленная точка может оказаться выбросом
  - Высокая трудоемкость
- Применить предыдущий подход для случайного подмножества точек
- Взять более k точек, выполнить k-means, выполнить иерархическую кластеризацию результата

Алгоритм k-means++

- Медленнее, чем случайный выбор начальных центроидов (сложность $O(\log k)$), но лучше по SSE
- Выбор центроидов

Выбрать случайным образом первый центроид C_1

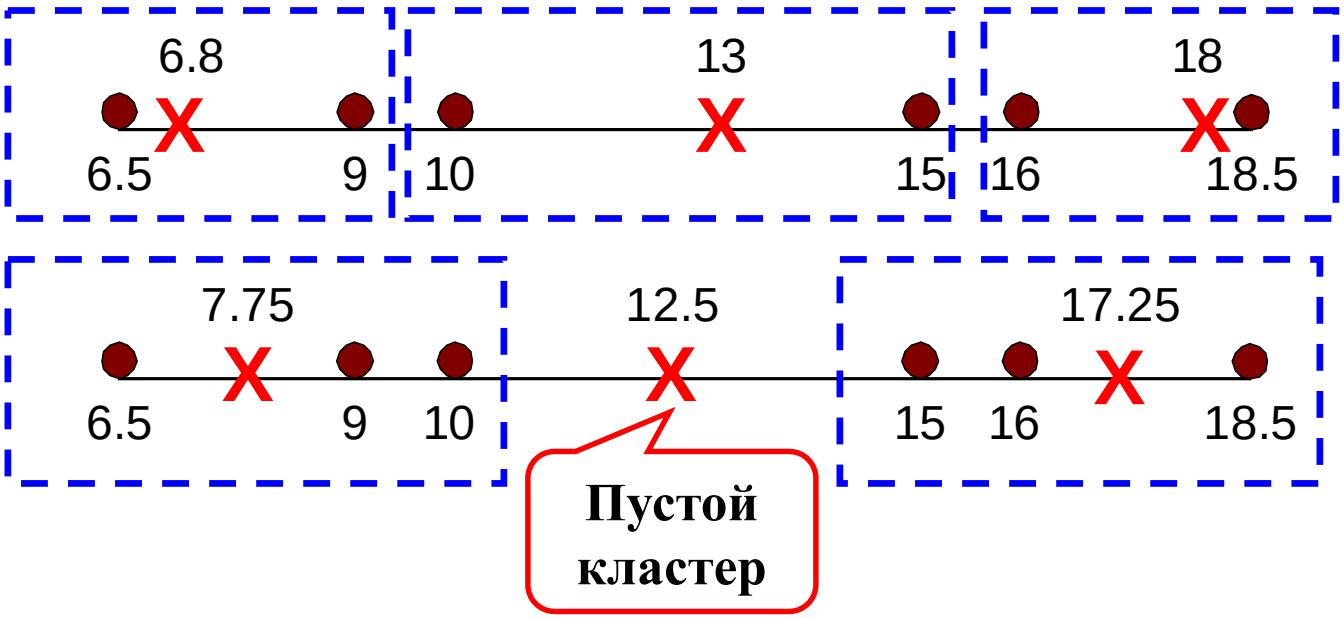
for $i \coloneqq 2$ to k **do**

for each x_i

найти вес x_i – квадрат мин расстояния от точки до уже выбранных центроидов $C_1, \dots, C_j$ $(1 \le j < k)$, $w_i = \min\limits_{1 \le j < k} dist^2(C_j, x_i)$

Выбрать случайно новый центроид C_i с вероятностью $p_i = \frac{w_i}{\sum_i w_i}$

Пустые кластеры и стратегии их обработки



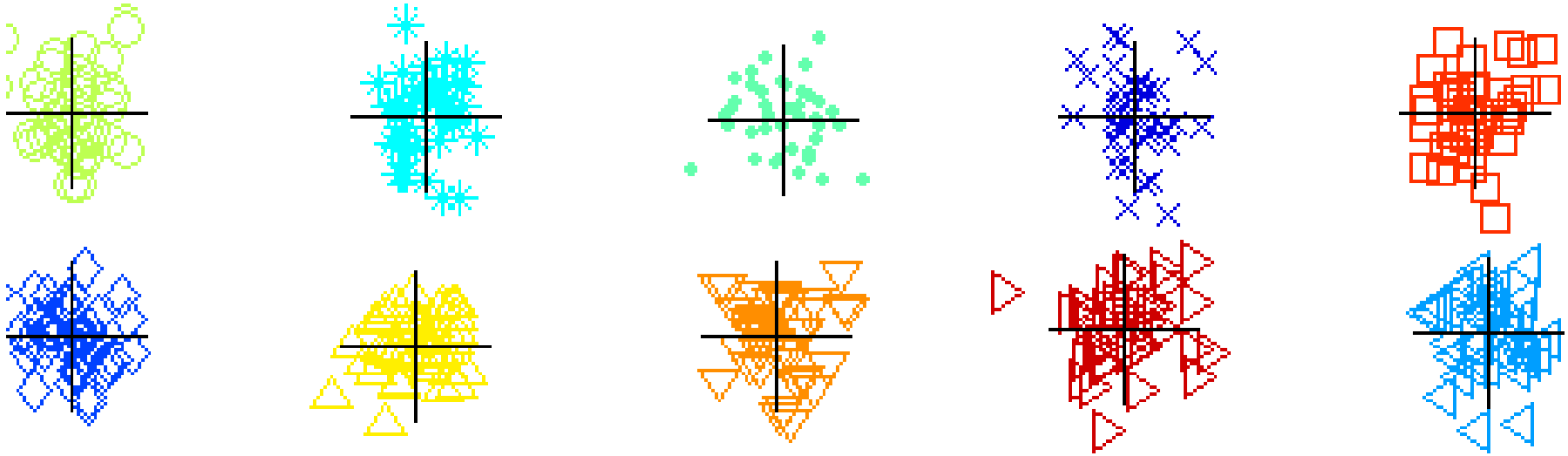
- Выбирать точку с наибольшим вкладом в SSE
- Выбирать точку с из кластера с наибольшим SSE
- Если несколько пустых кластеров, повторять

Техники улучшения k-means

- Предобработка
  - Нормализация данных
  - Поиск и удаление выбросов
- Постобработка
  - Удаление маломощных кластеров (выбросы/аномалии)
  - Разбиение «слабых» кластеров с высоким SSE
  - Слияние «близких» кластеров с низким SSE
- Обработка
  - Применение разбиения и слияния в процессе кластеризации
  - Инкрементное обновление центроидов: не после назначения всех объектов центроиду, а после каждого назначения
    - Каждое назначение обновляет 0-2 центроида
    - Исключаются пустые кластеры

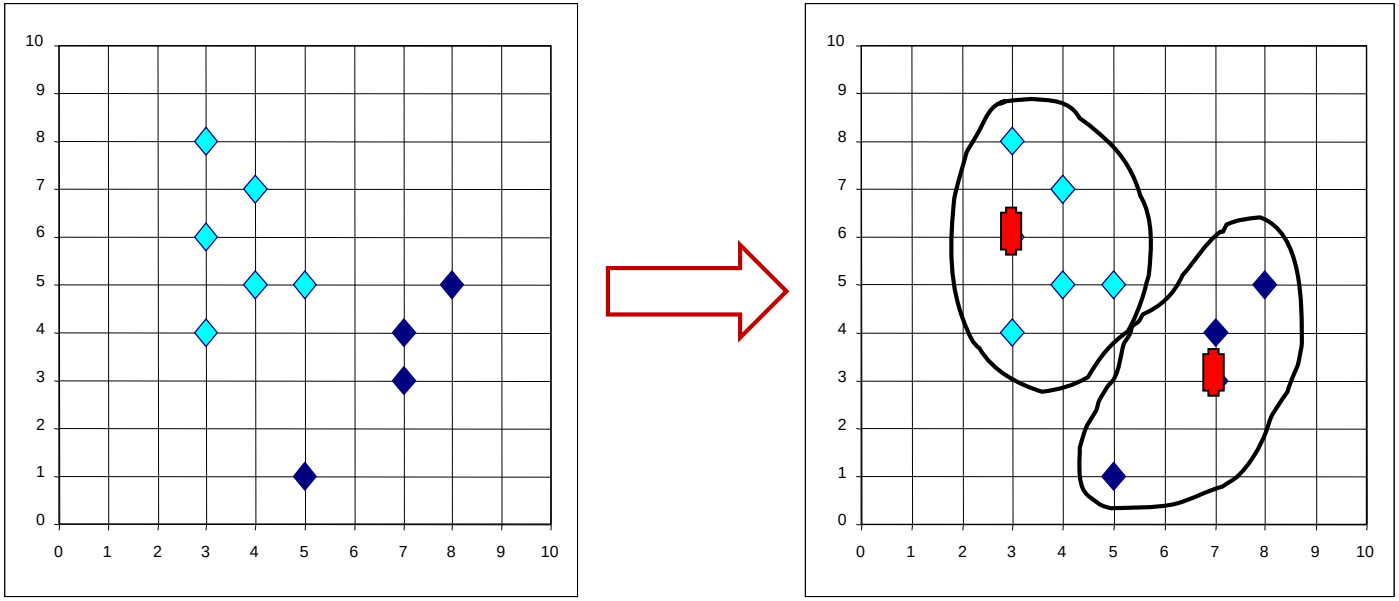
Бисекционный *k*-means (иерархическое разбиение)

1: Initialize the list of clusters to contain the cluster containing all points.
2: **repeat**
3: Select a cluster from the list of clusters
4: **for** $i = 1$ to *number_of_iterations* **do**
5: Bisect the selected cluster using basic K-means
6: **end for**
7: Add the two clusters from the bisection with the lowest SSE to the list of clusters.
8: **until** Until the list of clusters contains K clusters



k-medoids (PAM, Partitioning Around Medoids)

- Медоид – репрезентативный объект кластера

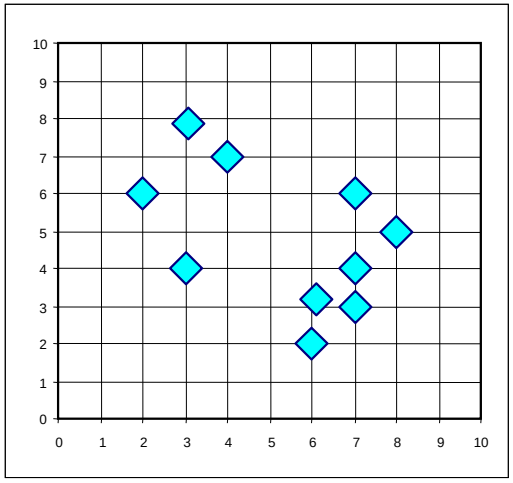


- Менее чувствителен к шумам и выбросам, чем *k*-means (медоид – объект кластера, центроид – искусственный объект)

k-medoids (PAM, Partitioning Around Medoids)

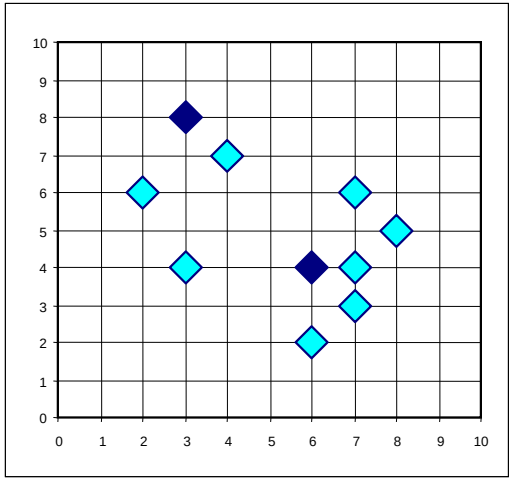
взять k случайных объектов в качестве медоидов $\breve{o}_1, \ldots, \breve{o}_k$
repeat
  назначить объектам кластер с ближайшим медоидом;
  выбрать случайный объект не-медоид o;
  вычислить цену обмена o и $\breve{o}$ как разность их SSE
  $E = \sum_{i=1}^{k} \sum_{p \in C_i} dist(p, o)^2 - \sum_{i=1}^{k} \sum_{p \in C_i} dist(p, \breve{o})^2$;
  if $E < 0$ **then**
    обменять местами объект o и медоид $\breve{o}$
until нет изменений

Пример работы *k*-medoids

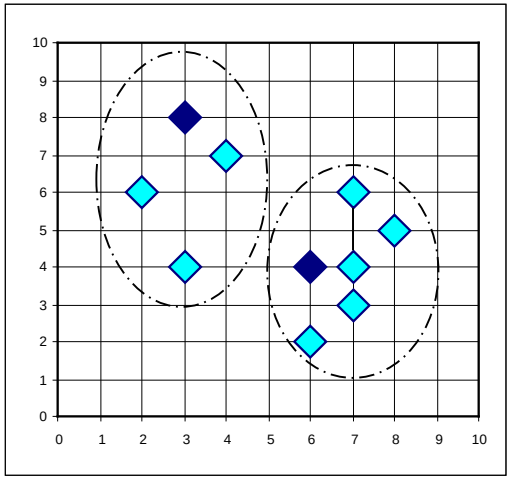


k=2

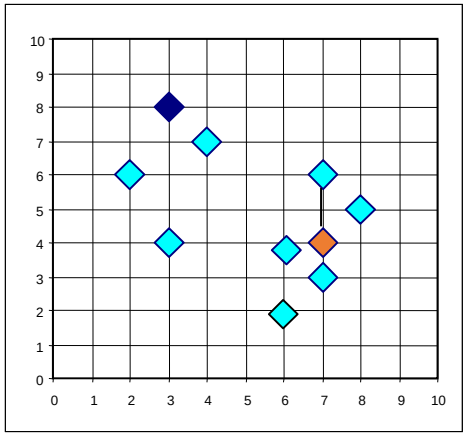
Выбрать k случайных объектов в качестве медоидов



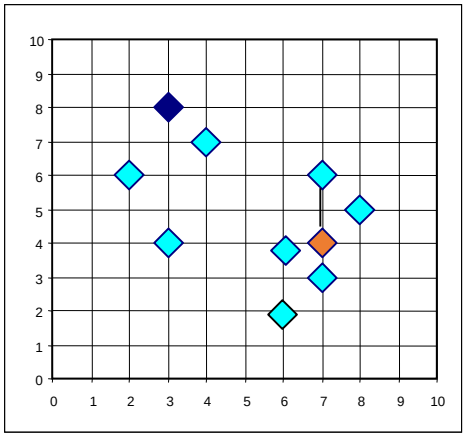
Присвоить оставшимся объектам кластеры по ближайшему медоиду



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

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