Занятие 7
Кластеризация

Елена Кантонистова

ВШЭ, 2021

КЛАСТЕРИЗАЦИЯ

Даны объекты $x_1, \dots, x_l, x_i \in X$.

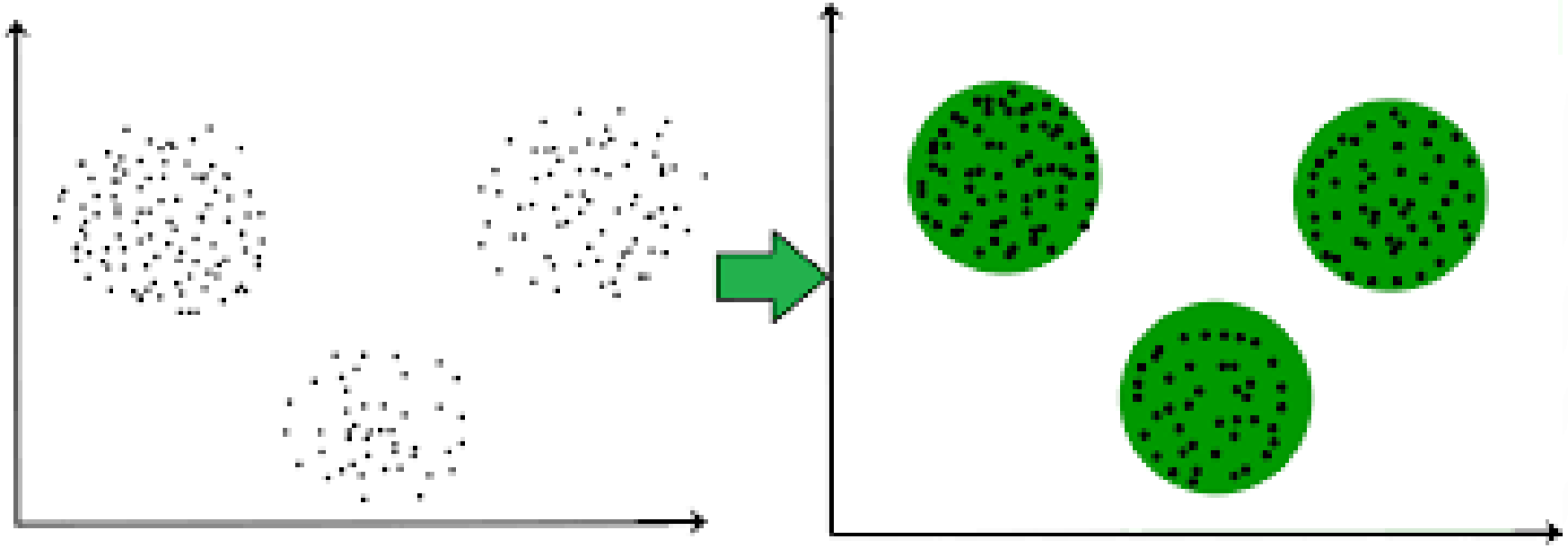
- Требуется выявить в данных K кластеров – таких областей, что объекты внутри одного кластера похожи друг на друга, а объекты из разных кластеров друг на друга не похожи.

КЛАСТЕРИЗАЦИЯ

Даны объекты $x_1, \dots, x_l, x_i \in X$.

- Требуется выявить в данных K кластеров – таких областей, что объекты внутри одного кластера похожи друг на друга, а объекты из разных кластеров друг на друга не похожи.
- Формализация задачи: необходимо построить алгоритм $a: X \to \{1, \dots, K\}$, сопоставляющий каждому объекту x номер кластера.

КЛАСТЕРИЗАЦИЯ



ОСНОВНЫЕ ПОНЯТИЯ В ЗАДАЧЕ КЛАСТЕРИЗАЦИИ

ЧТО ОПТИМИЗИРУЕМ?

- В любой задаче машинного обучения мы оптимизируем некоторую функцию, отвечающую за качество алгоритма.

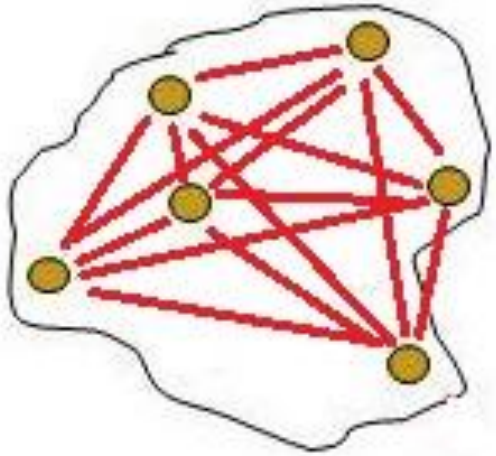
Что оптимизировать в задаче кластеризации?

ВНУТРИКЛАСТЕРНОЕ РАССТОЯНИЕ

Пусть c_k - центр k-го кластера

Внутри кластера все объекты максимально похожи, поэтому наша **цель – минимизировать внутрикластерное расстояние**:

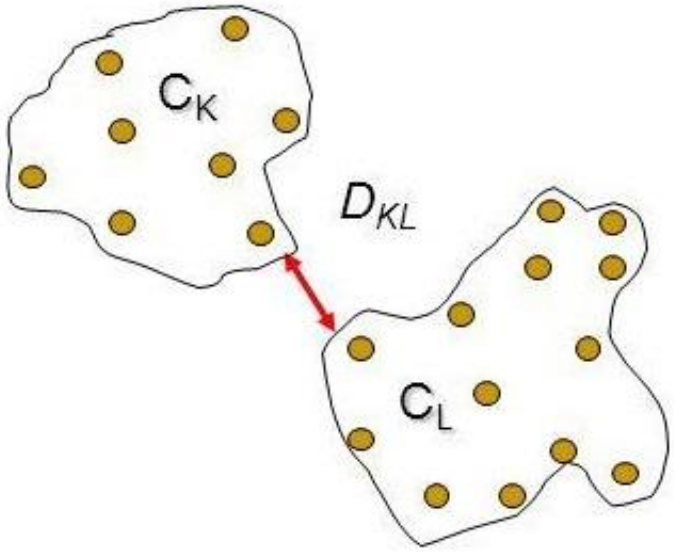
$$\sum_{k=1}^{K}\sum_{i=1}^{l}[a(x_i)=k]\rho(x_i,c_k)\to\min_a$$



МЕЖКЛАСТЕРНОЕ РАССТОЯНИЕ

Объекты из разных кластеров должны быть как можно менее похожи друг на друга, поэтому мы **максимизируем межкластерное расстояние**:

$$\sum_{i,j=1}^{l} [a(x_i) \neq a(x_j)]\rho(x_i, x_j) \to \max_a$$

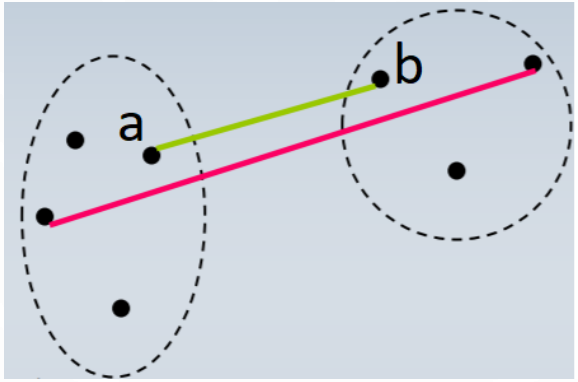


КАК СЧИТАТЬ РАССТОЯНИЯ МЕЖДУ ОБЪЕКТАМИ?

ВИДЫ РАССТОЯНИЙ МЕЖДУ ОБЪЕКТАМИ

- **Евклидово расстояние** – расстояние между точками в общепринятом понимании, то есть геометрическое расстояние между двумя точками.

$$\rho(a,b) = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2}$$



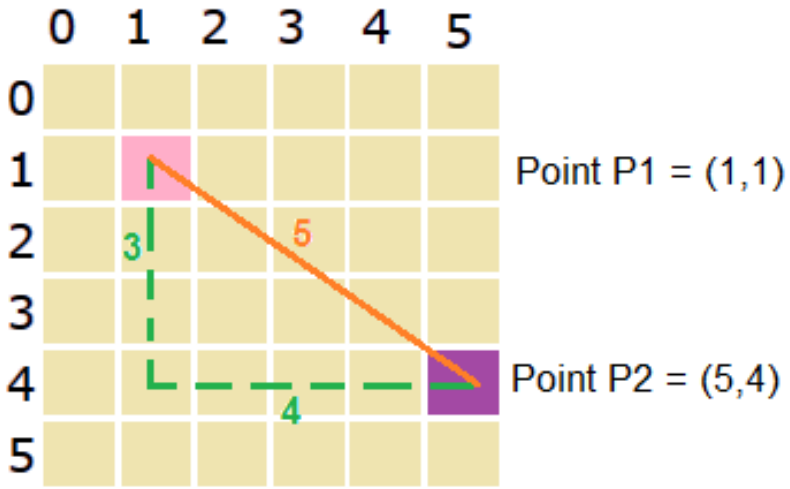
ВИДЫ РАССТОЯНИЙ МЕЖДУ ОБЪЕКТАМИ

- **Евклидово расстояние** – расстояние между точками в общепринятом понимании, то есть геометрическое расстояние между двумя точками.

$$\rho(a,b) = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2}$$

- **Манхеттенское расстояние** (расстояние городских кварталов):

$$\rho(a,b) = |x_1 - x_2| + |y_1 - y_2|$$



Euclidean distance = $\sqrt{(5-1)^2 + (4-1)^2} = 5$

Manhattan distance = $|5-1| + |4-1| = 7$

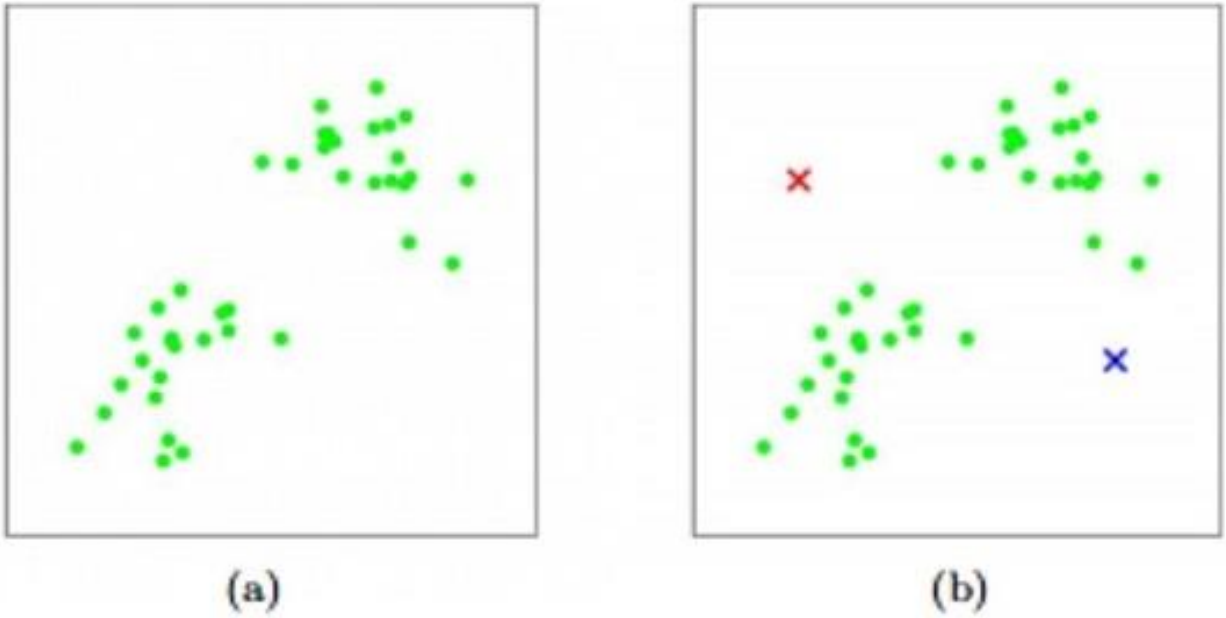
K-MEANS

K-MEANS

Дано: выборка $x_1, \dots, x_l$

Параметр: число кластеров K

Начало: **случайно выбрать центры кластеров $c_1, \dots, c_K$**



(a) (b)

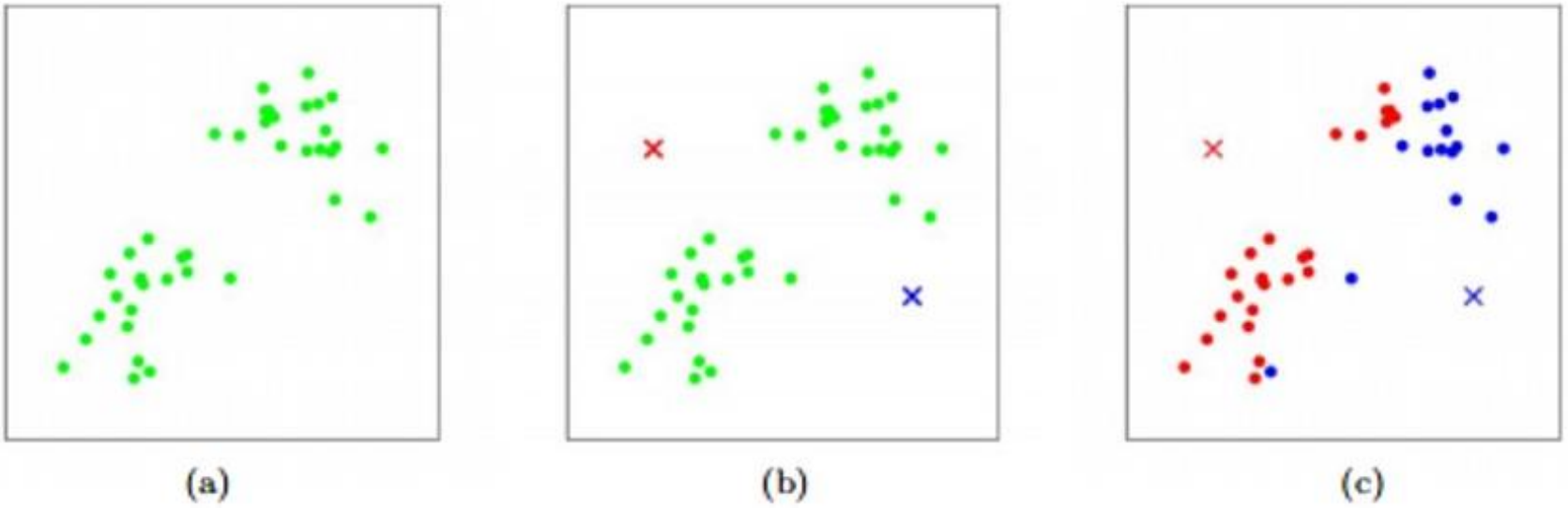
K-MEANS

Дано: выборка $x_1, \dots, x_l$

Параметр: число кластеров K

Начало: случайно выбрать центры кластеров $c_1, \dots, c_K$

1) **каждый объект отнести к ближайшему к нему центру кластера**



K-MEANS

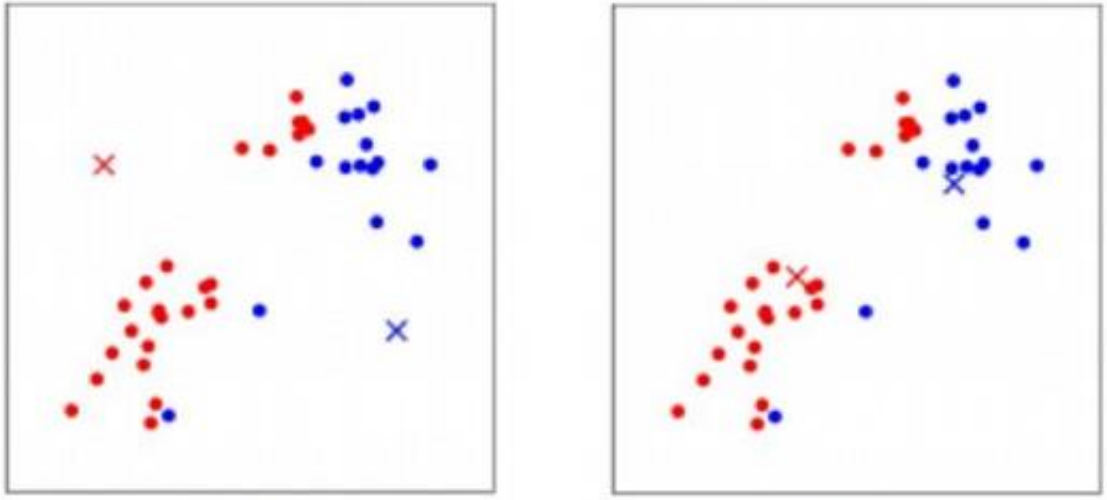
Дано: выборка $x_1, \dots, x_l$

Параметр: число кластеров K

Начало: случайно выбрать центры кластеров $c_1, \dots, c_K$

1) каждый объект отнести к ближайшему к нему центру кластера

2) пересчитать центры полученных кластеров



K-MEANS

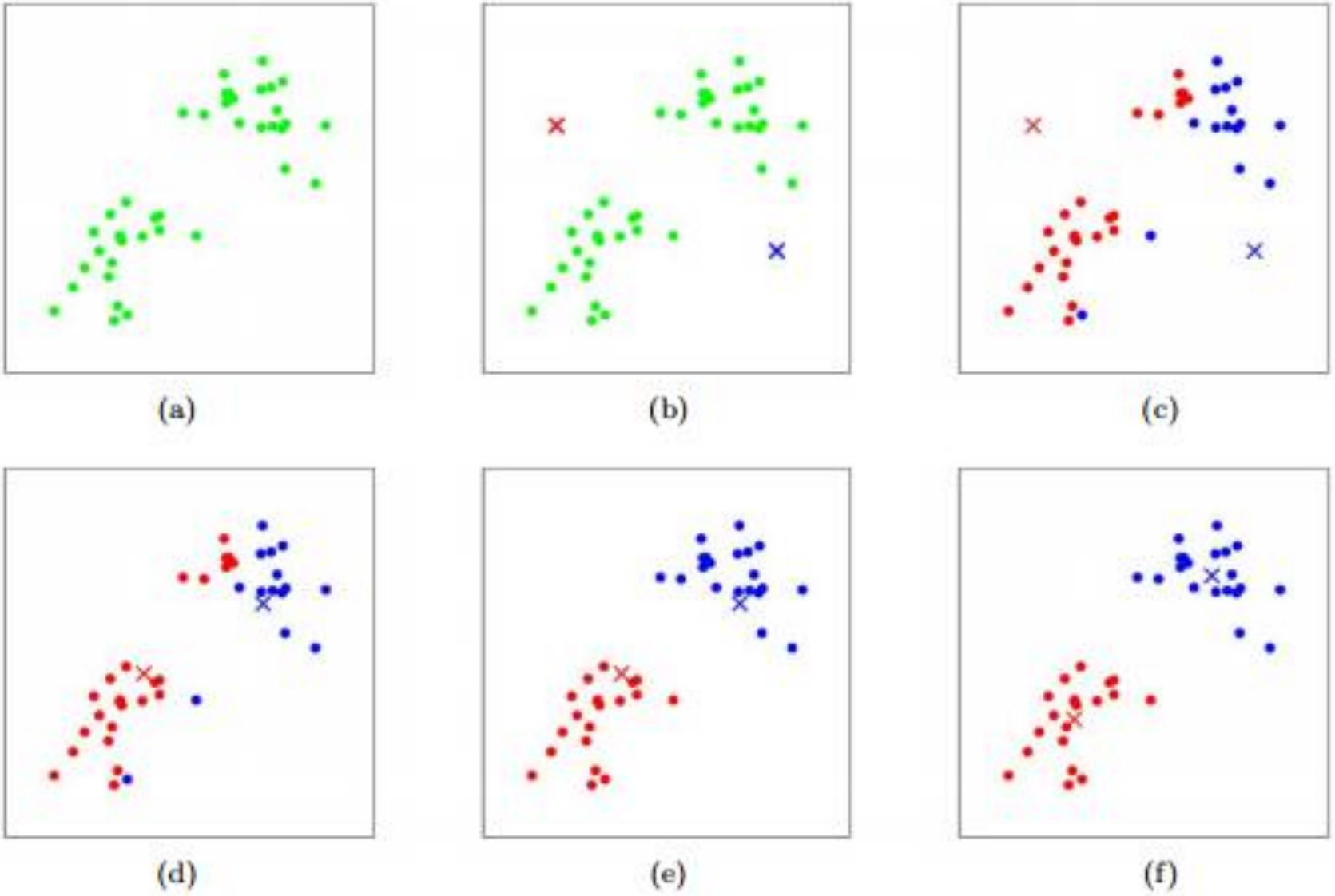
Дано: выборка $x_1, \ldots, x_l$

Параметр: число кластеров K

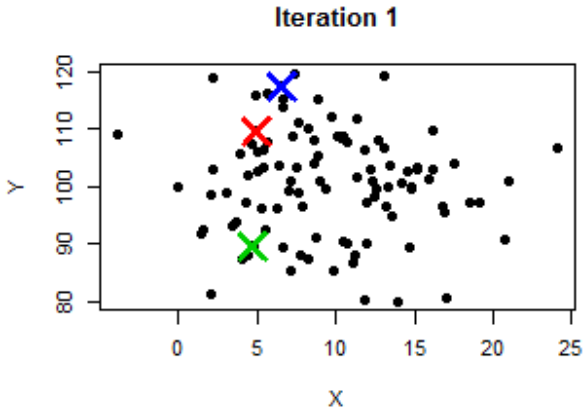
Начало: случайно выбрать центры кластеров $c_1, \ldots, c_K$

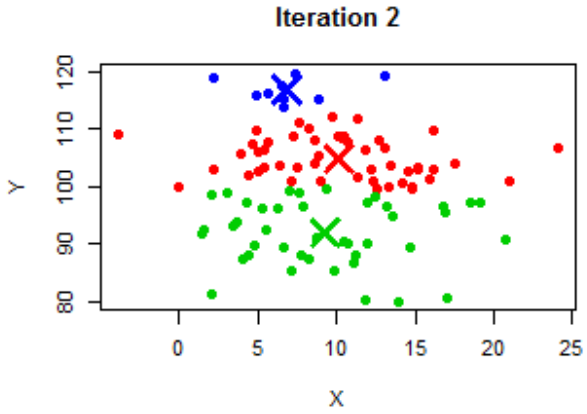
1) каждый объект отнести к ближайшему к нему центру кластера
2) пересчитать центры полученных кластеров
3) **повторить шаги 1 и 2 несколько раз до стабилизации кластеров**

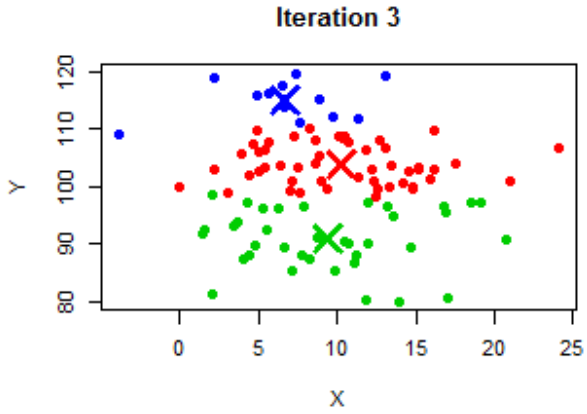
K-MEANS (ДВА КЛАСТЕРА)

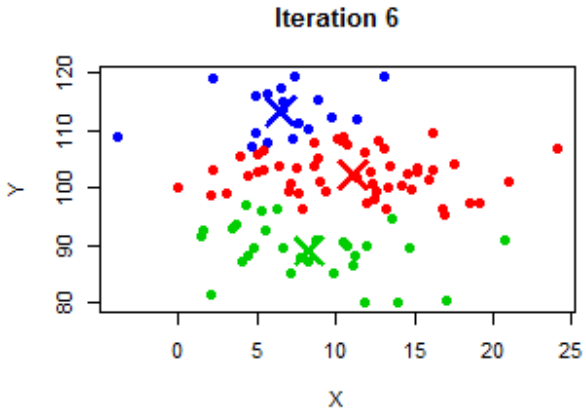


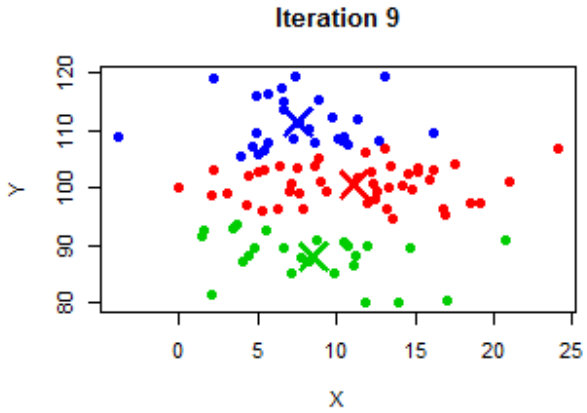
K-MEANS (ТРИ КЛАСТЕРА)

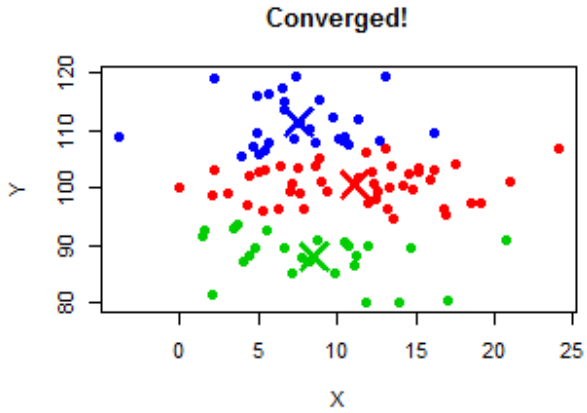












K-MEANS (МАТЕМАТИКА)

Дано: выборка $x_1, \dots, x_l$

Параметр: число кластеров K

Идея метода - минимизация внутрикластерного расстояния

$$\sum_{k=1}^{K} \sum_{i=1}^{l} [a(x_i) = k] \rho(x_i, c_k) \to \min_a$$

с $\rho(a, b) = (a - b)^2$, т.е.

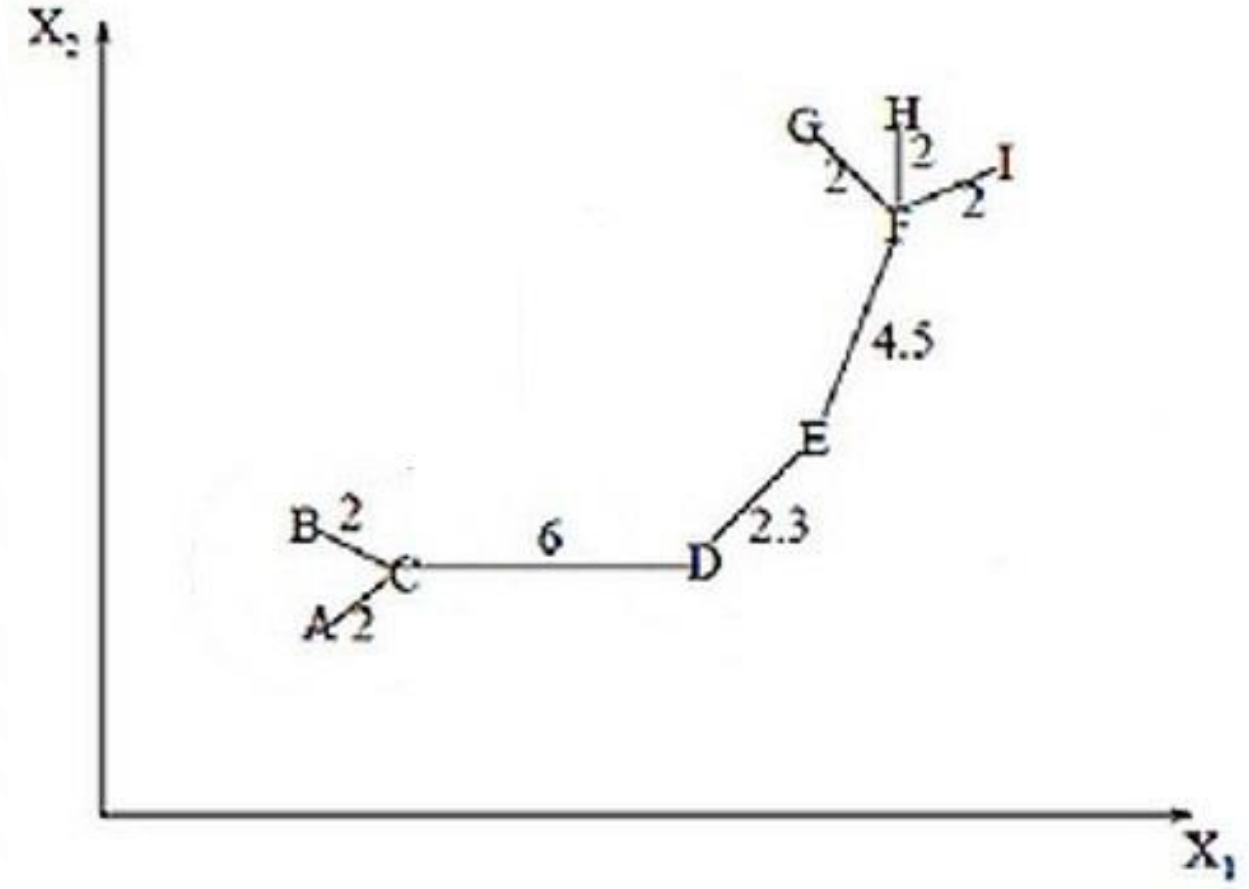
$$\sum_{k=1}^{K} \sum_{i=1}^{l} [a(x_i) = k] (x_i - c_k)^2 \to \min_a$$

K-MEANS ДЛЯ СЖАТИЯ ИЗОБРАЖЕНИЙ

ГРАФОВЫЕ МЕТОДЫ КЛАСТЕРИЗАЦИИ

ГРАФОВЫЕ МЕТОДЫ КЛАСТЕРИЗАЦИИ

- выборка представляется в виде графа, где в вершинах стоят объекты, а на рёбрах – расстояния между ними



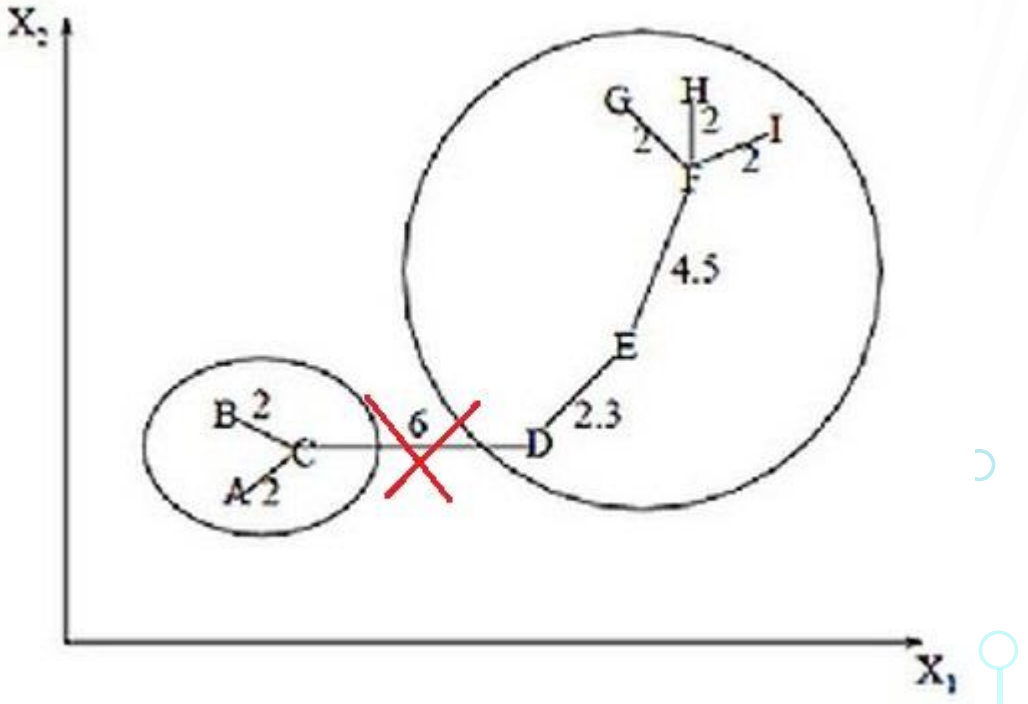
ГРАФОВЫЕ МЕТОДЫ КЛАСТЕРИЗАЦИИ

- выборка представляется в виде графа, где в вершинах стоят объекты, а на рёбрах – расстояния между ними

Алгоритм выделения связных компонент:

1) из графа удаляются все ребра, для которых расстояния больше некоторого значения R

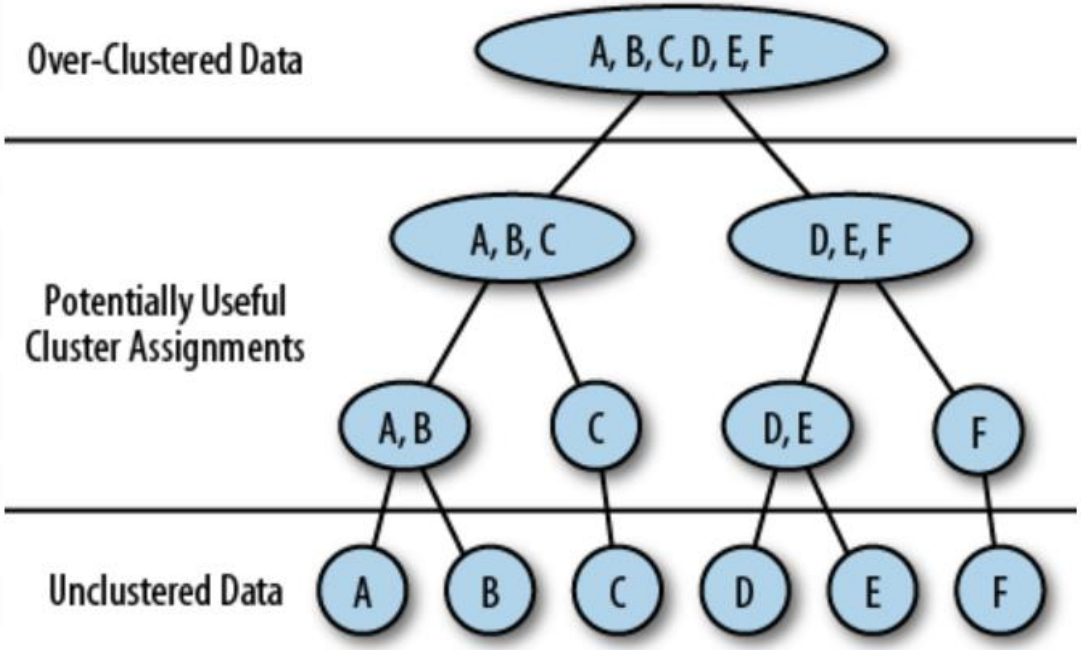
2) Кластеры – объекты, попадающие в одну компоненту связности



2) ИЕРАРХИЧЕСКАЯ КЛАСТЕРИЗАЦИЯ

Иерархия кластеров:

- на нижнем уровне - l кластеров, каждый из которых состоит из одного объекта
- на верхнем уровне – один большой кластер

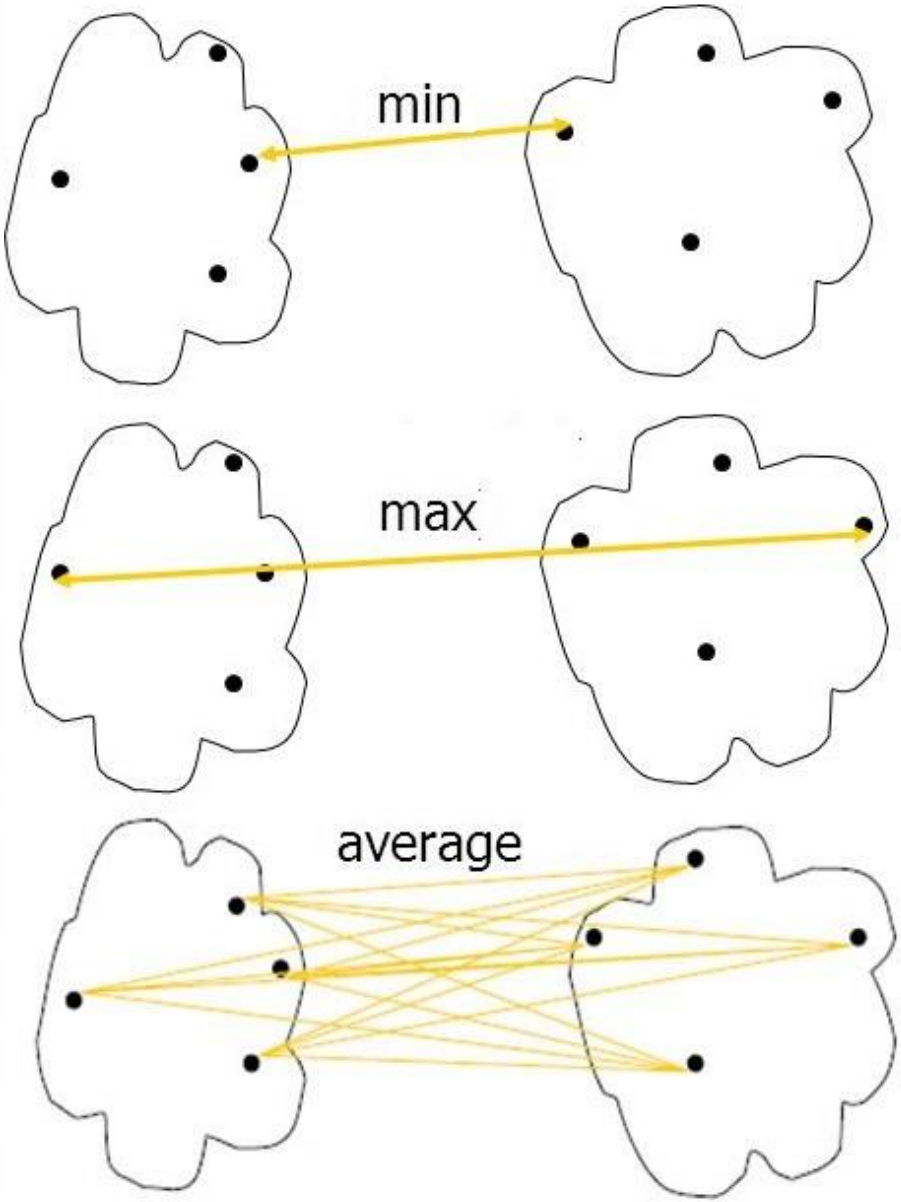


ИЕРАРХИЧЕСКАЯ КЛАСТЕРИЗАЦИЯ

Алгоритм Ланса-Уильямса:

- первый шаг: один кластер = один объект
- на каждом следующем шаге объединяем два наиболее близких друг к другу кластера с предыдущего шага

РАССТОЯНИЕ МЕЖДУ КЛАСТЕРАМИ

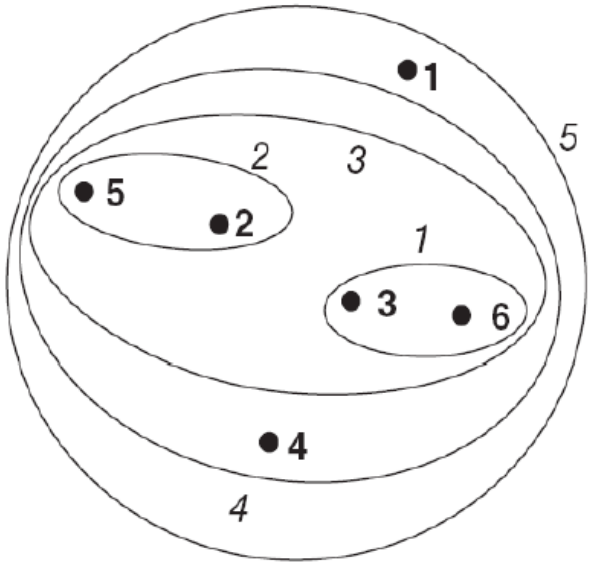


ИЕРАРХИЧЕСКАЯ КЛАСТЕРИЗАЦИЯ

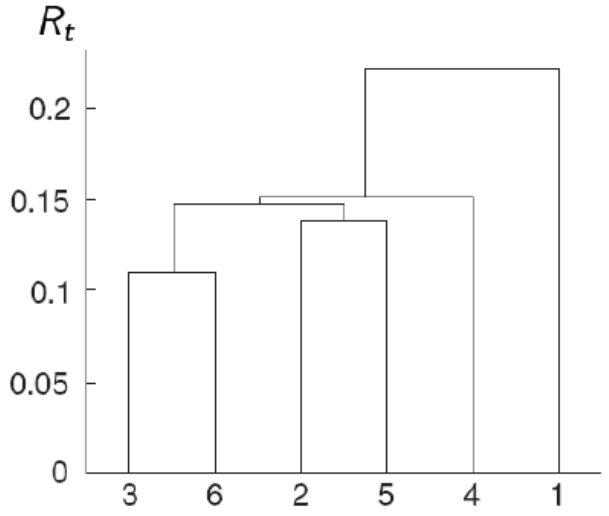
Алгоритм Ланса-Уильямса:

- первый шаг: один кластер = один объект
- на каждом следующем шаге объединяем два наиболее близких друг к другу кластера с предыдущего шага

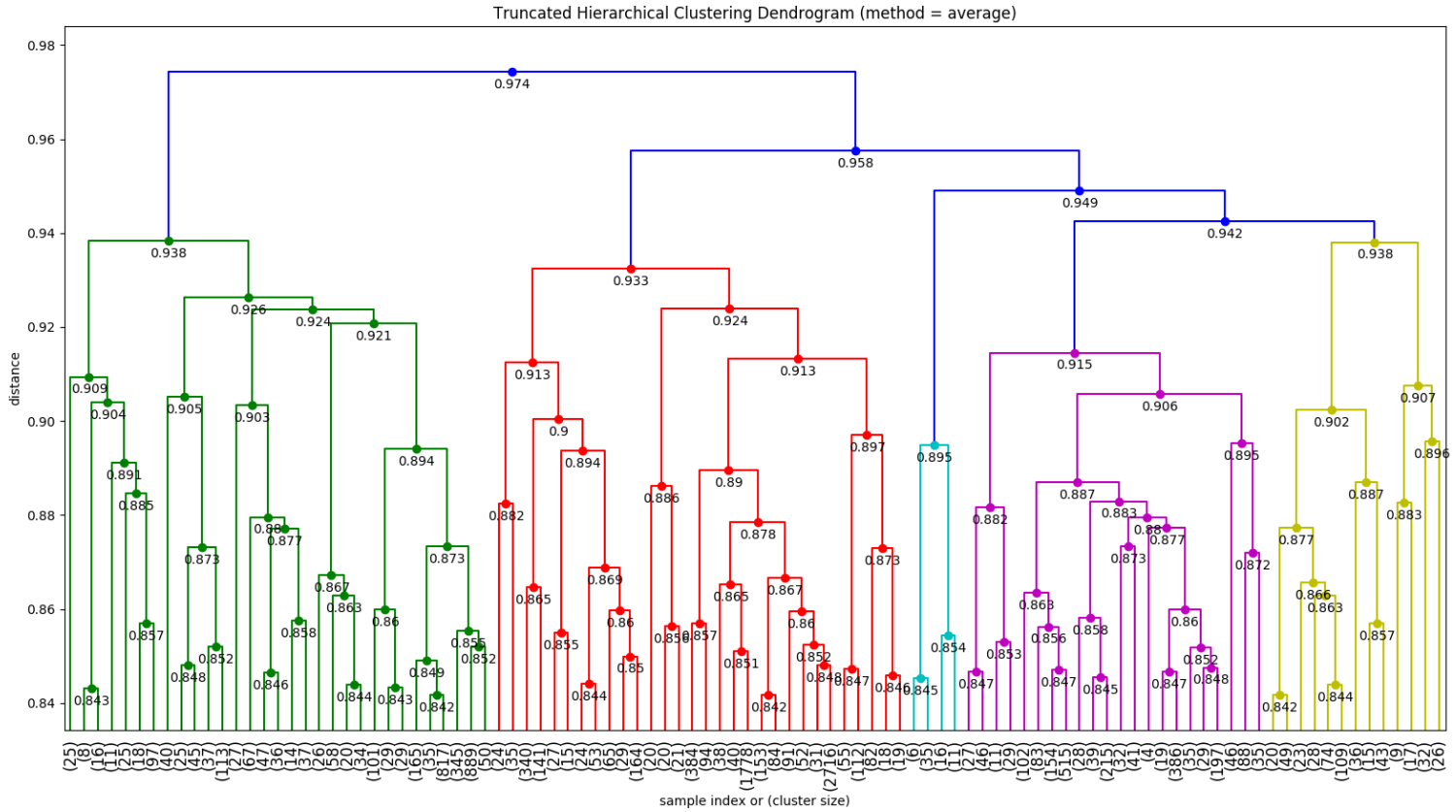
Диаграмма вложения



Дендрограмма



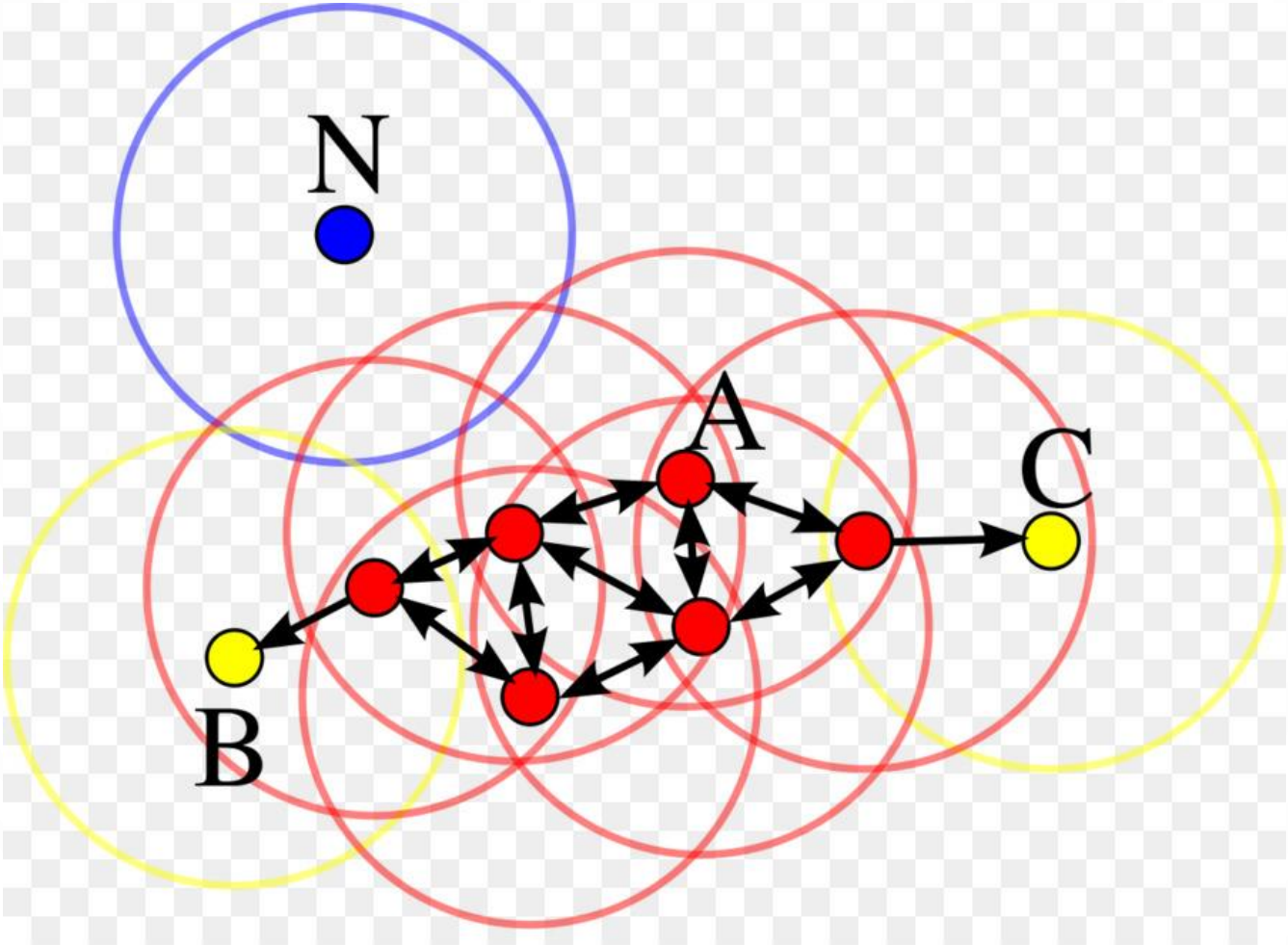
ВЫБОР ЧИСЛА КЛАСТЕРОВ



DENSITY-BASED CLUSTERING

ТИПЫ ОБЪЕКТОВ В DBSCAN

Объекты: основные, граничные, шумовые.



ПАРАМЕТРЫ МЕТОДА

- eps – размер окрестности
- min_samples – минимальное число объектов в окрестности (включая сам объект), для определения основных точек

АЛГОРИТМ DBSCAN

1. Выбрать точку без метки
2. Если в окрестности меньше, чем min_pts точек, то пометить её как шумовую
3. Создать кластер, поместить в него текущую точку (если это не шум, см. п.2)
4. Для всех точек из окрестности S:

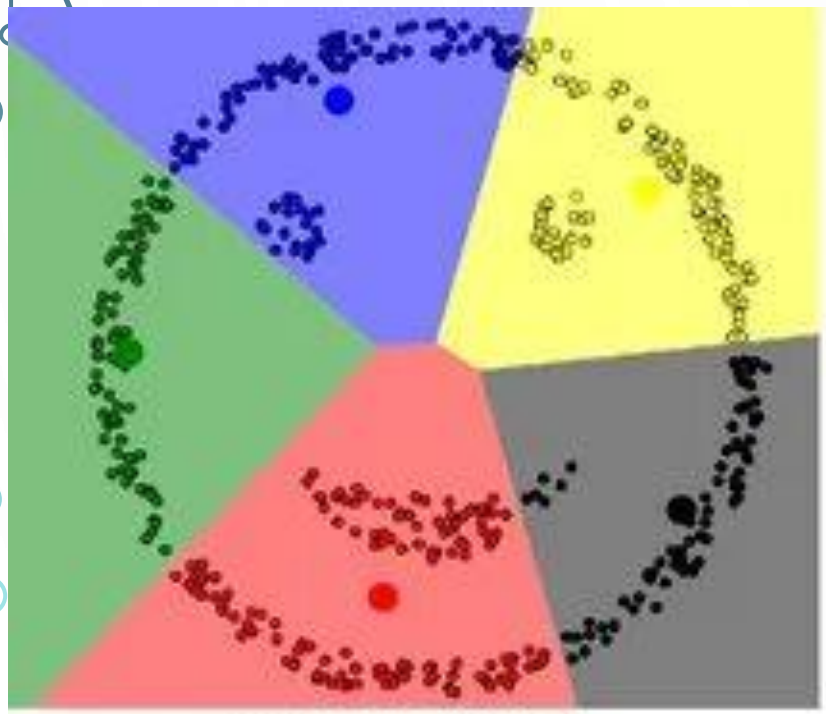
- если точка шумовая, то отнести к данному кластеру, но не использовать для расширения
- если точка основная, то отнести к данному кластеру, а её окрестность добавить к S

5. Перейти к шагу 1.

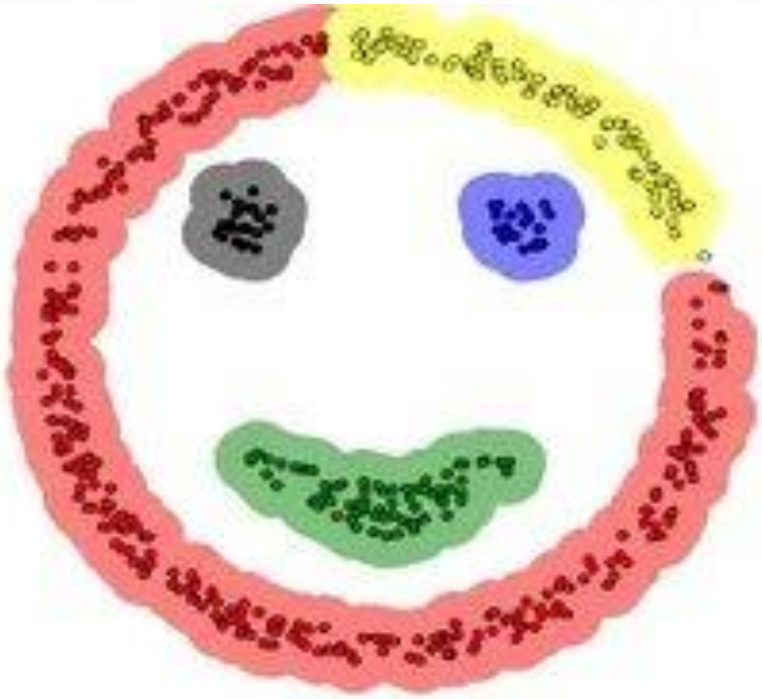
DBSCAN DEMO

https://www.naftaliharris.com/blog/visualizing-dbscan-clustering/

KMEANS AND DBSCAN



KMeans(K=5)



DBSCAN(MinPts=4, eps=1.0)