

Кластеризация

Цели занятия

Разобрать методы машинного обучения в задачах кластеризации:

- Метод К-средних
- Иерархическая кластеризация
- Плотностные алгоритмы (DBSCAN)

Кластеризация –

- разделение исходного неразмеченного набора данных на несколько групп, состоящих из близких объектов (кластеров).

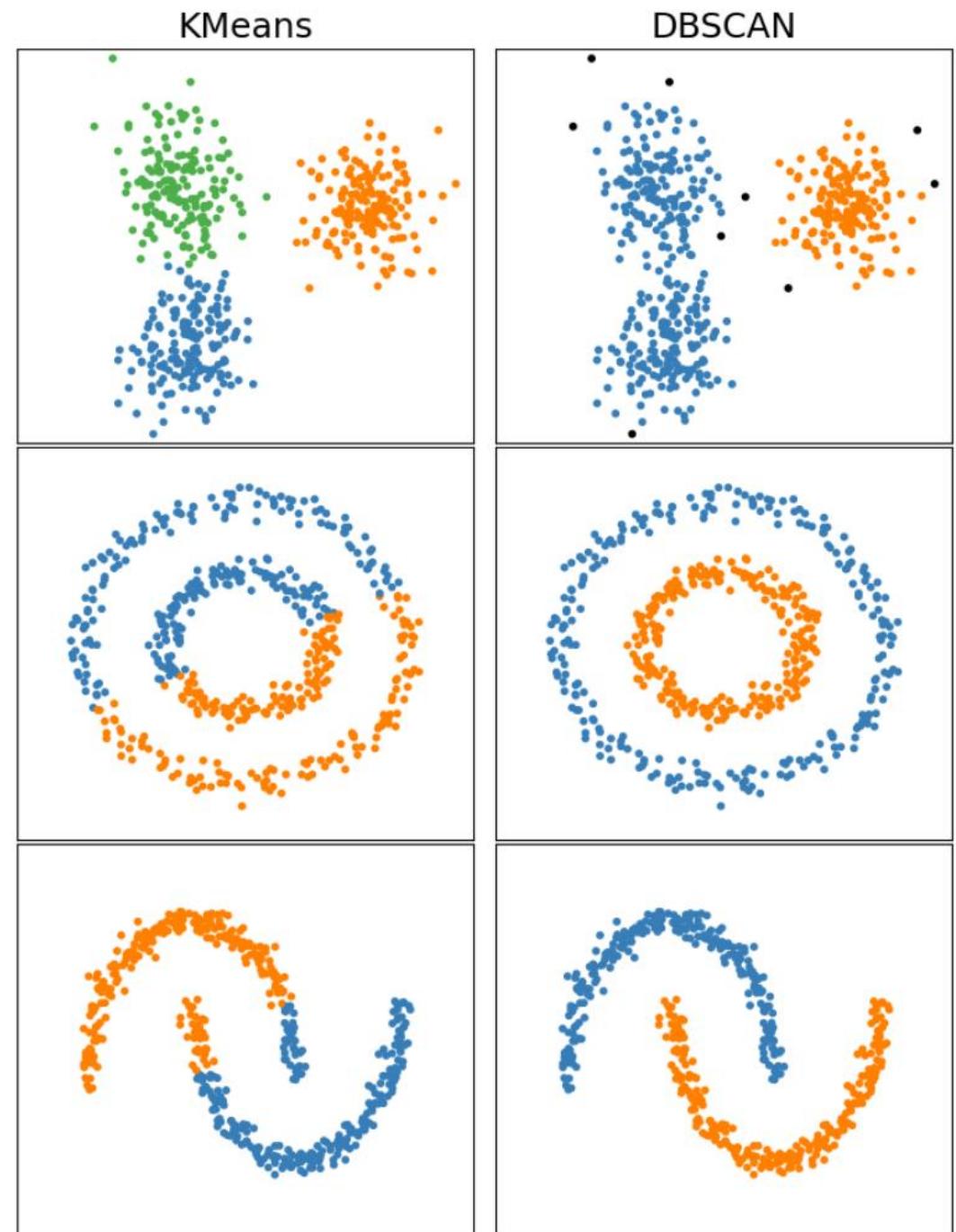
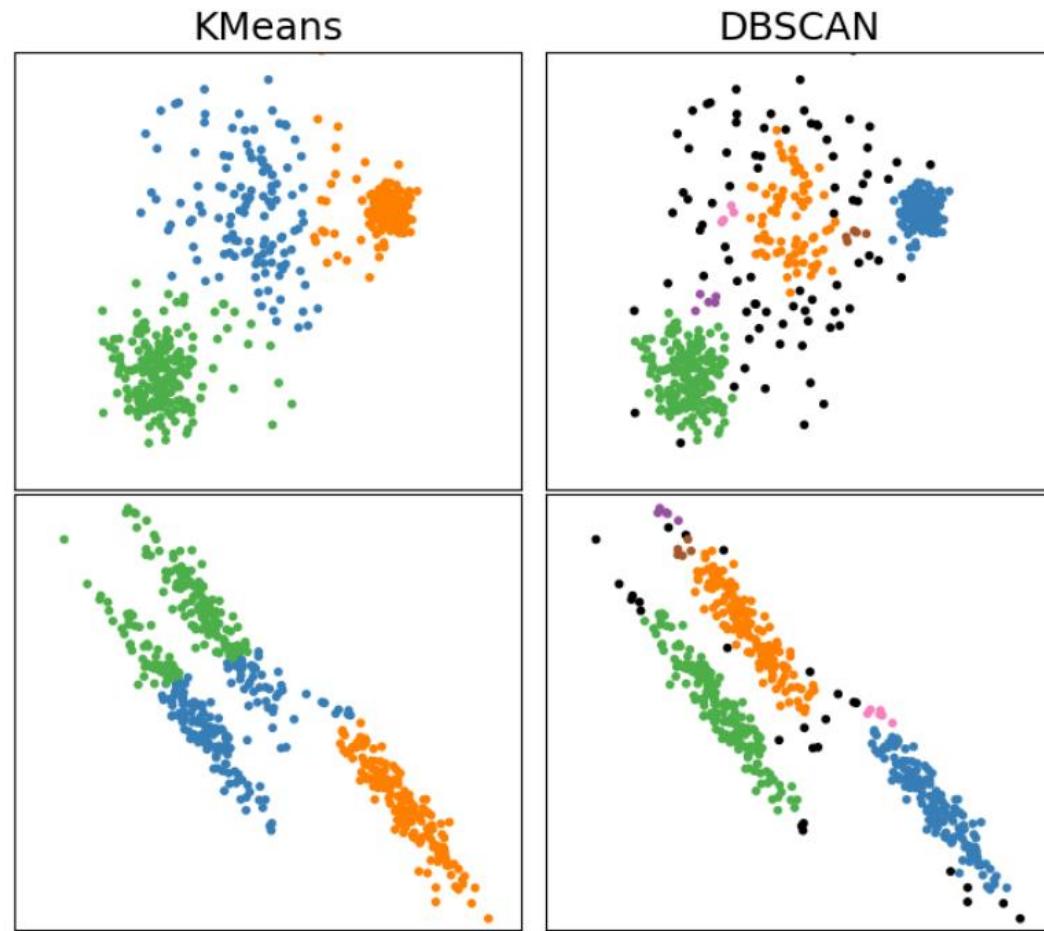
Проблемы постановки задачи:

- Не всегда понятно, сколько кластеров
- Сильная зависимость от метрики и нормализации
- Нет единого критерия качества кластеризации
- Даже точная постановка задачи не всегда есть

Цели

- Сжатие данных
 - Уменьшение кол-ва признаков
 - Уменьшение кол-ва принимаемых значений признака
 - Уменьшение кол-ва объектов
- Упростить обработку данных
 - Разбить данные на группы схожих объектов для дальнейшей работы с каждой в отдельности
- Поиск аномалий/выбросов

Примеры структур



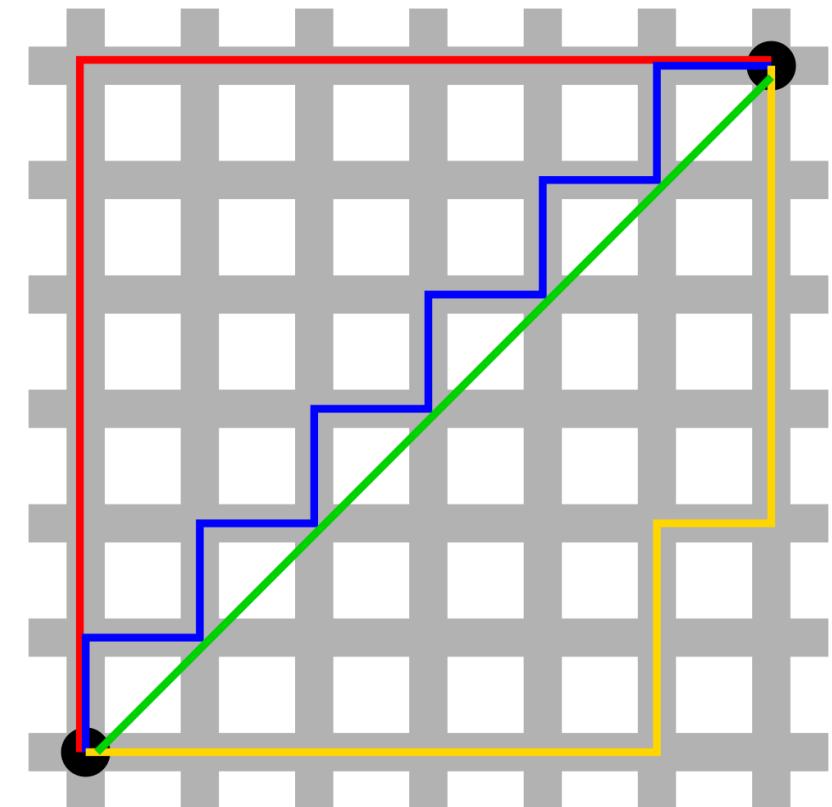
Метрика Минковского

- Евклидовая метрика (L2):

$$d = \sqrt{\sum_{k=0}^n (a_k - b_k)^2}$$

- Манхэттенское расстояние (L1):

$$d = \sum_{k=0}^n |a_k - b_k|$$



Проклятие размерности

- Чем больше размерность, тем больше необходимо объектов для покрытия пространства
- Рост экспоненциальный
- Частично лечится нормированием признаков

Качество кластеризации

- Минимизация среднего внутрикластерного расстояния S1
- Максимизация межкластерного расстояния S2
- Комбинация S1/S2

Другие опции:

- По ближайшему соседу
 - По дальнему соседу
-
- Долго -> можно использовать центры масс кластеров

Метод K-средних

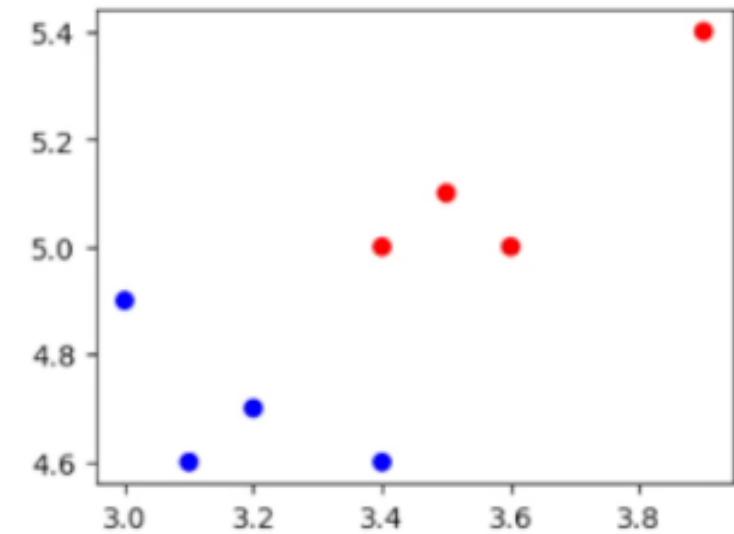
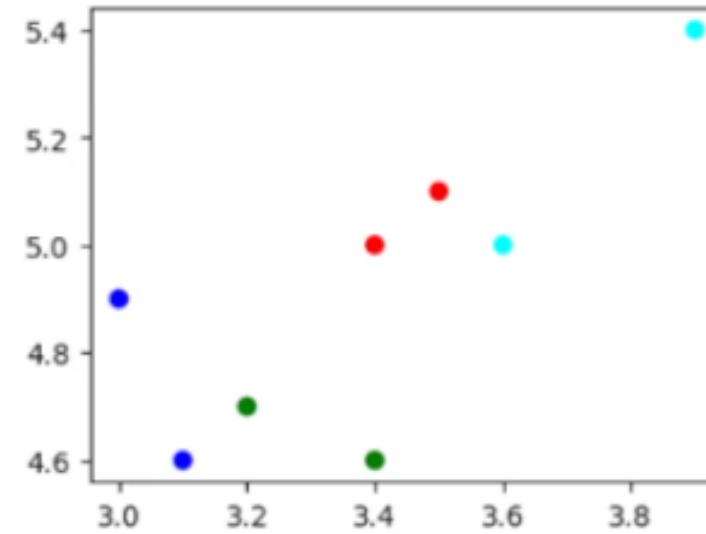
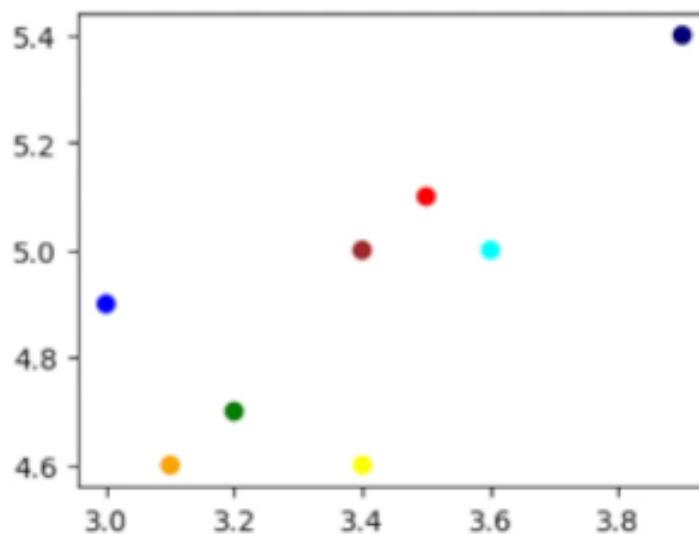
Число кластеров задается изначально.

Алгоритм:

- отнести каждый объект к ближайшему центру
- вычислить новые положения центров, пока они не перестанут меняться
- Работает со «сферическими» кластерами
- Сильно зависит от начального приближения

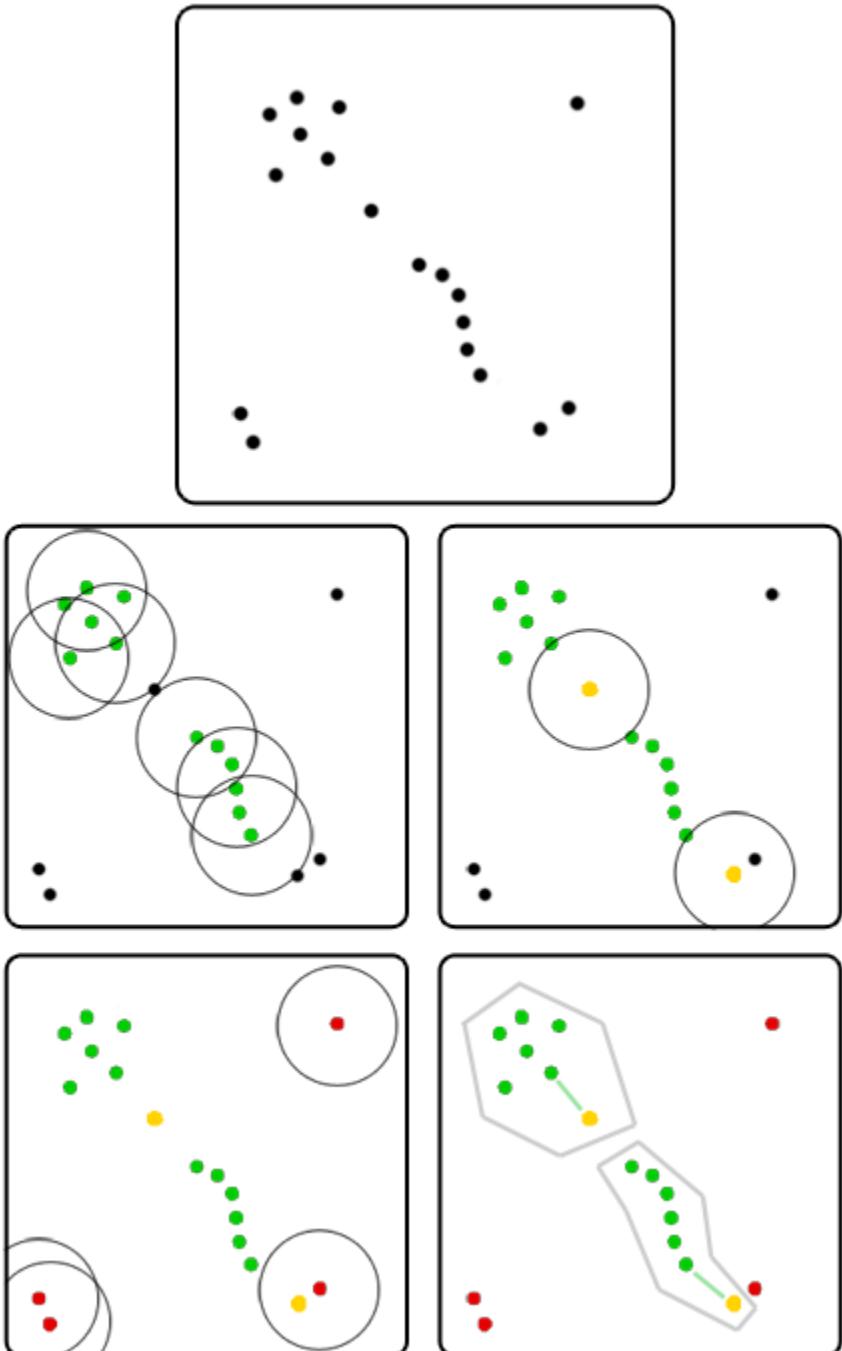
Иерархические методы

- Агломеративные
 - снизу-вверх, последовательное объединение
- Дивизимные
 - сверху-вниз, последовательное деление



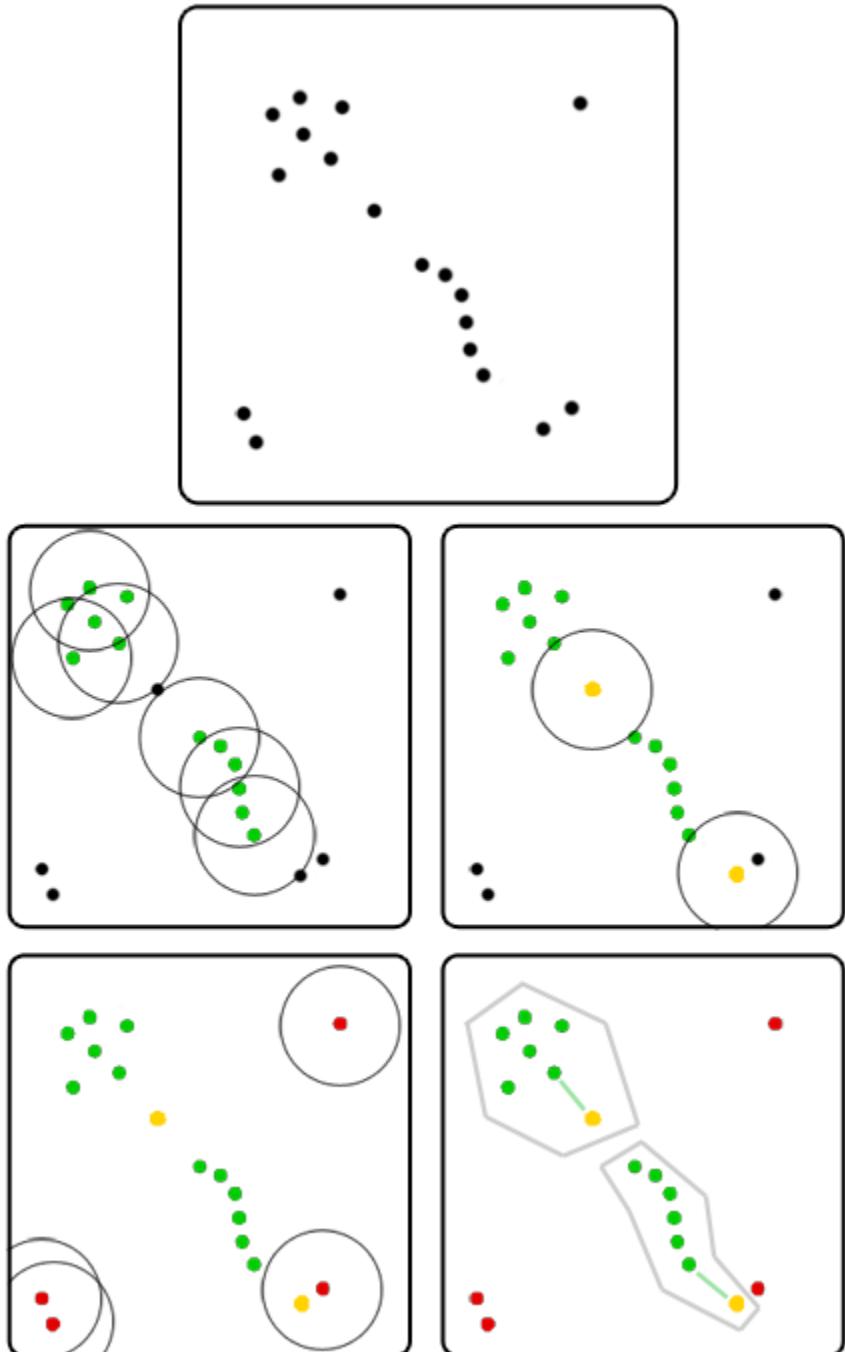
DBSCAN

- Density-Based Spatial Clustering of Applications with Noise
- Точки делятся на 3 вида: основная/корневая, граничная, шумовая
- Используется настраиваемый параметр R – радиус окрестности и p – плотность



DBSCAN

- Шумовые точки убираются из рассмотрения и не приписываются ни к какому кластеру
- Основные точки с общей окрестностью соединяются
- В полученном графе выделяются компоненты связности
- Каждая граничная точка относится к тому кластеру, в который попала ближайшая к ней основная точка



DBSCAN

- Быстрый алгоритм ($O(N \ln N)$ в среднем)
- Выделяет нетипичные объекты
- Не нужно подбирать кол-во кластеров
- Кластеры произвольной формы

Метрики качества

- Внешние — основаны на использовании известной информации (например, истинных меток), которая не задействовалась в процессе кластеризации.
- Внутренние — используют информацию только из структуры обучающего набора (нет информации об истинных метках).

Метрики на основе меток

- ARI (Adjusted Rand Index) – измеряет количество пар элементов, отнесённых к одинаковым и разным кластерам относительно общего количества возможных пар в данных. Симметрична, не зависит от перестановок меток и их значений.
- AMI (Adjusted Mutual Information) – измеряется мера статистической зависимости между двумя переменными, показывающая, сколько информации о значениях одной переменной можно получить, зная значения другой.

Силуэт

- не использует знаний о метках объектов
- показывает насколько в среднем объекты схожи внутри одного кластера и различны с объектами других кластеров.

$$s = \frac{b - a}{\max(a, b)}$$

, где а – среднее расстояние от объекта до объектов того же кластера, b – среднее расстояние от объекта до объектов другого ближайшего кластера

- диапазон [-1,1], где -1 – разрозненная кластеризация, 0 – кластеры пересекаются, 1 – хорошо выделенные кластеры