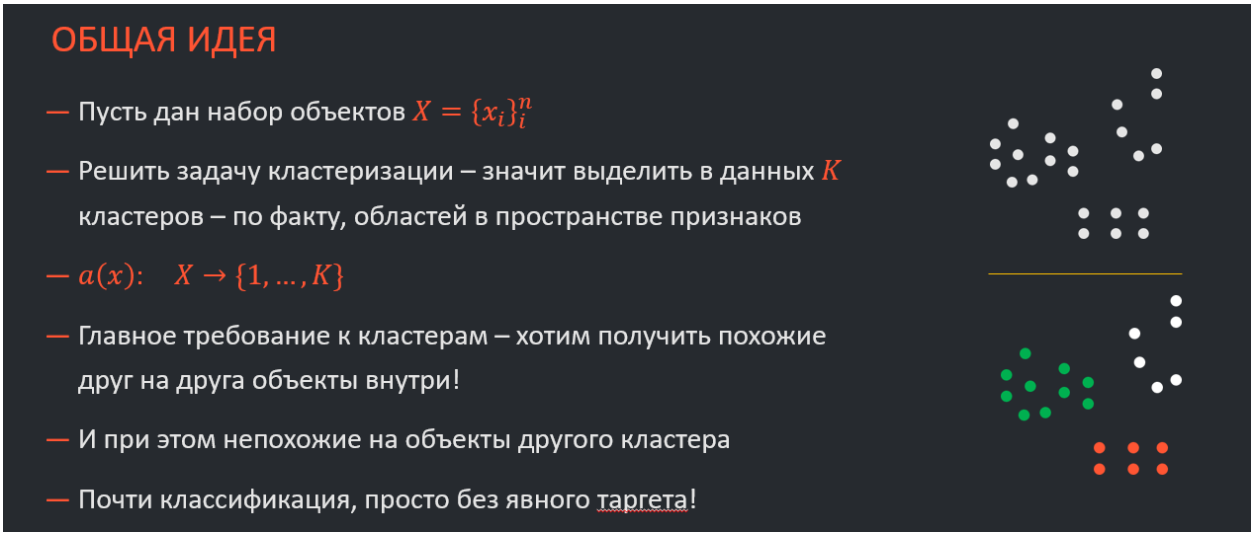
KARPOV.COURSES >>>
КОНСПЕКТ

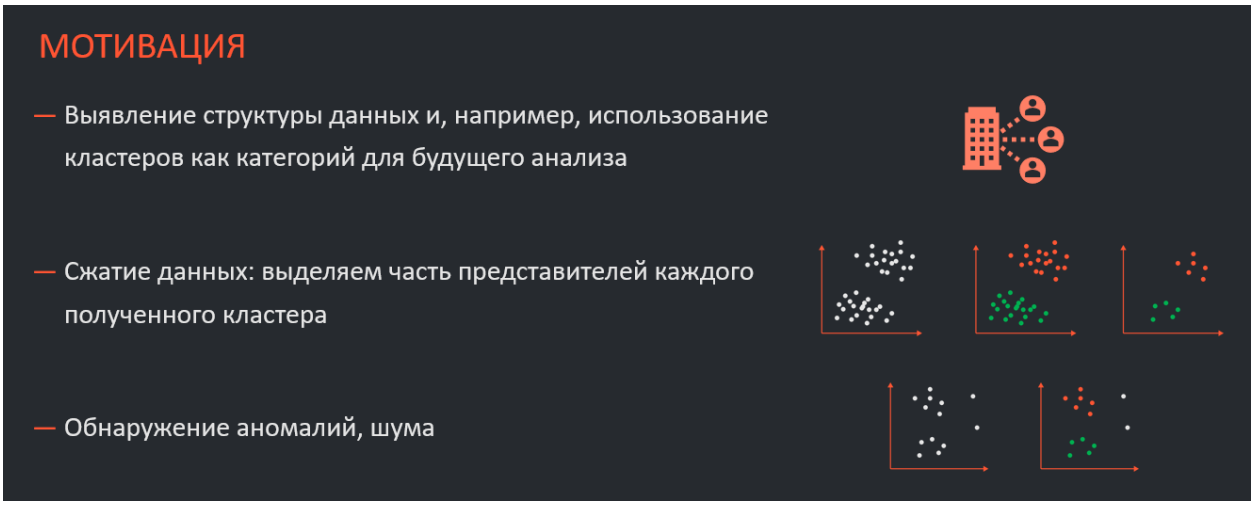
Конспект > 21 урок > Кластеризация

>Введение

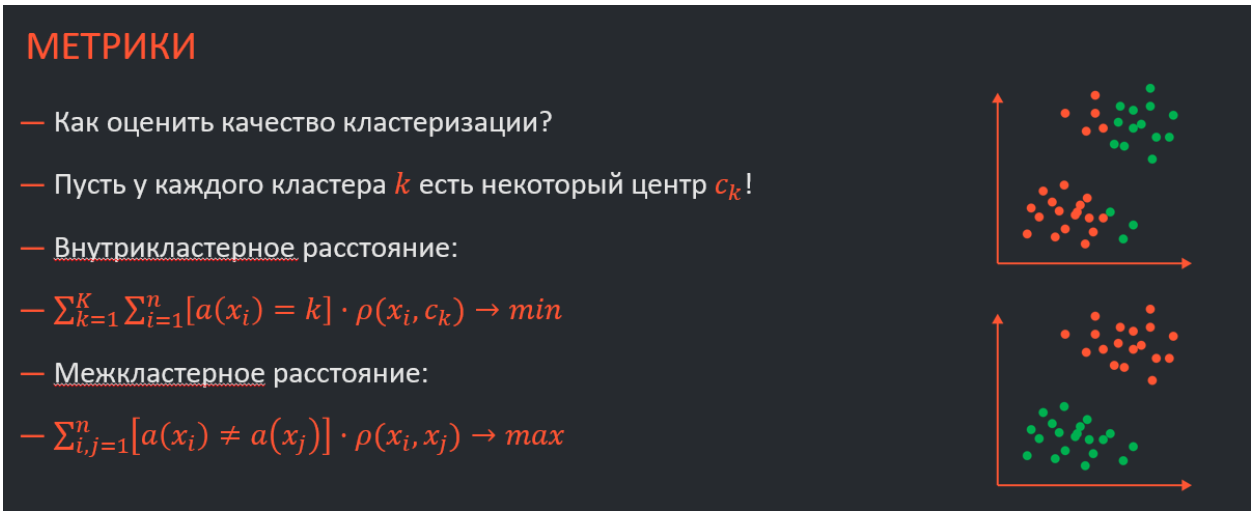
Задача кластеризации одна из самых популярных задач машинного обучения и относится к обучению без учителя (Unsupervised learning).



Давайте рассмотрим мотивацию, т.е. зачем нам это нужно.



Возникает важный вопрос, а как нам вообще понять, что кластеризация, которую мы получили является хорошей. Например на картинке ниже мы можем увидеть, что сверху и снизу мы получили какие-то кластеры. Нам нужно понять какая модель сделала разбиение лучше, для этого нужно ввести метрики, которые позволят нам оценить качество кластеризации. Мы можем использовать внутрикластерное расстояние и межкластерное расстояние. Важно сделать пояснение, что есть и другие метрики качества, но на данном этапе нам достаточно этих двух.



>K-means

K-means является одним из базовых и очень простых алгоритмов кластеризации. Давайте попробуем понять как он устроен и сразу рассмотрим сам алгоритм.

K-MEANS

— Заранее решим, какое количество кластеров K хотим выделить

— Случайным образом инициализируем центры $c_1, \dots, c_K$ кластеров

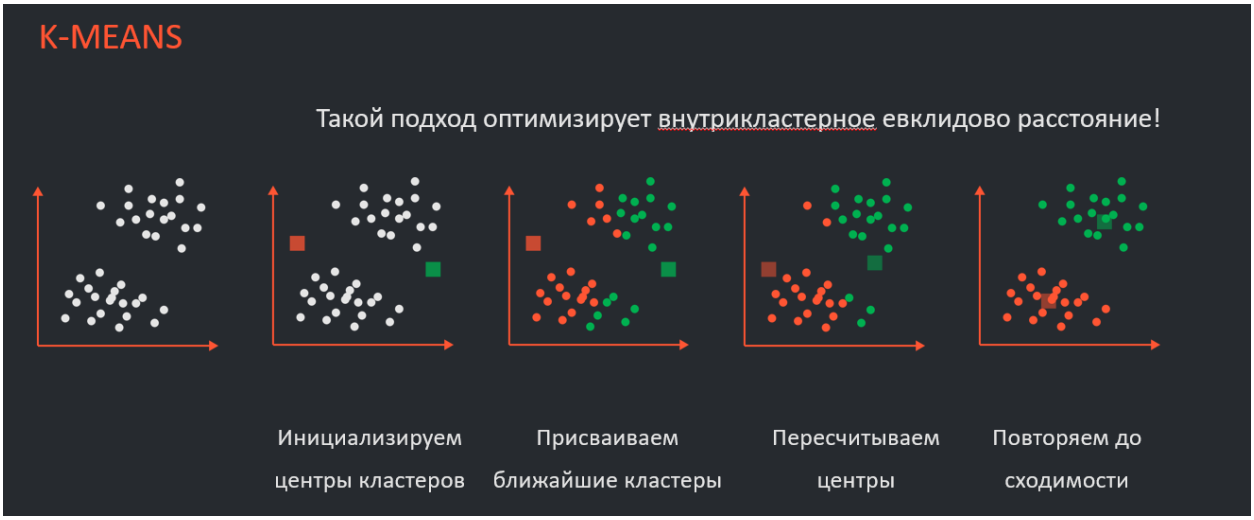
— Повторяем до сходимости:

— *Каждой точке присваиваем тот кластер, чей центр ближе*

— *Пересчитываем центры кластеров:* $c_k = \frac{1}{\sum_i^n [a(x_i)=k]} \sum_i^n [a(x_i) = k] \cdot x_i$

— В качестве критерия останова можно выбрать фиксацию центров

Давайте визуализируем шаги алгоритма, которые были указаны на картинке ниже.



Также мы можем оптимизировать используя не только евклидово расстояние.

K-MEANS

— Можно, вообще говоря, оптимизировать и не евклидовое расстояние между кластерами

— Например, Манхэттана!

— В таком случае, центры кластеров на каждой итерации будут пересчитываться по иной формуле, и метод будет носить уже другое название

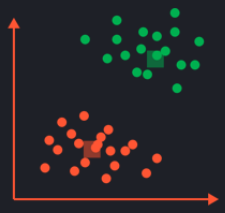
— Например, K-medians

После рассмотрения k-means является важным сделать небольшое резюме, рассмотреть преимущества и недостатки метода.

ПРЕИМУЩЕСТВА

— Прост в реализации

— Можно параллелить



НЕДОСТАТКИ

— Получаем кластеры крайне простой формы: сферы/параллелипипиды

— Нужно задавать число кластеров

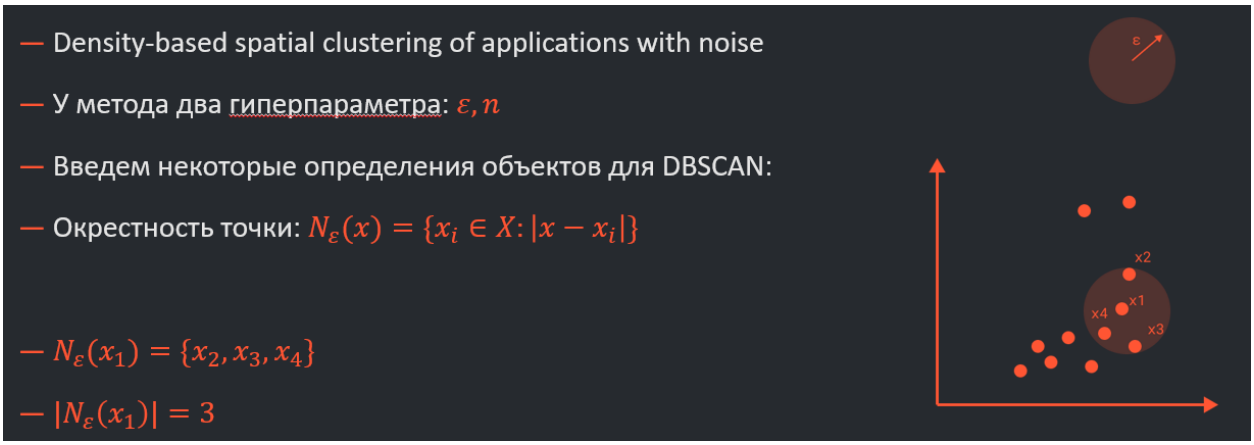
— Никак не учитываем шумовые объекты

— Неустойчив к инициализации

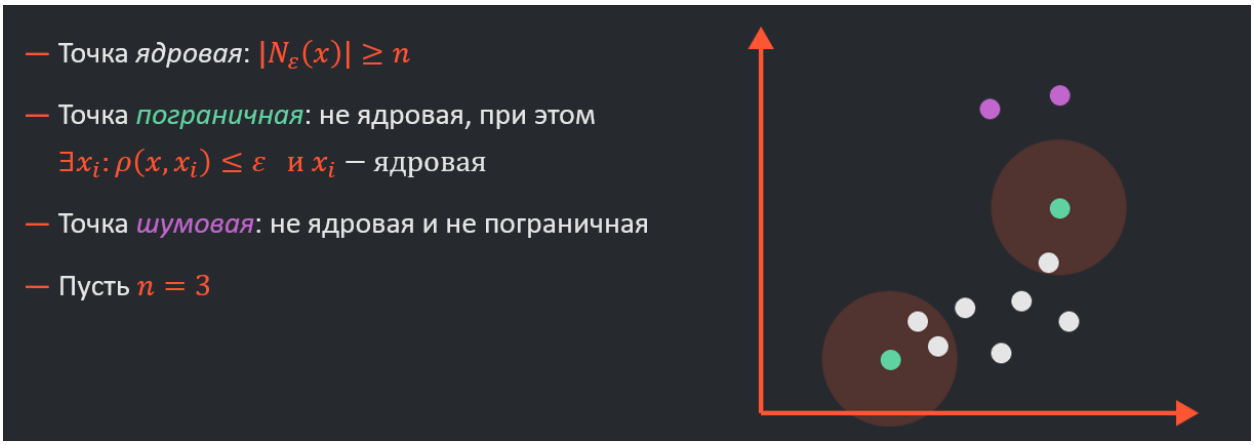
Как можно увидеть k-means имеет мало преимуществ и много недостатков и поэтому далее будет рассмотрен другой поход к кластеризации.

>DBSCAN

Перед тем как непосредственно перейти к механике и интуиции работы DBSCAN, давайте рассмотрим некоторые определения и установим гиперпараметры.



Рассмотрим еще несколько определений, которые помогут нам в понимании метода DBSCAN



Теперь кода определи необходимые понятия и вводные, можно непосредственно перейти к рассмотрению алгоритма. В большем размере рассмотреть эту картинку можно в презентации к лекции. Она очень детально показывает работу алгоритма.

После того как рассмотрели DBSCAN, является важным сделать некоторое резюме, рассмотреть преимущества и недостатки метода.

ПРЕИМУЩЕСТВА

- Нет необходимости выбирать количество кластеров
- Может выделять сколь угодно сложные формы кластеров
- Указывает на шум в данных

НЕДОСТАТКИ

- Необходимо делать предположения о ε, n
- Сложно параллелить
- Без дополнительных модификаций не получится разделить кластеры разной плотности