



Конспект > 21 урок > Кластеризация

>Введение

>K-means

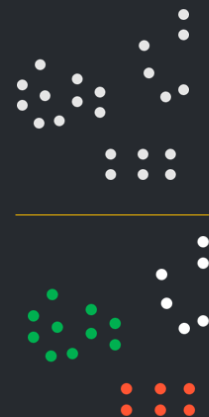
>DBSCAN

>Введение

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

ОБЩАЯ ИДЕЯ

- Пусть дан набор объектов $X = \{x_i\}_i^n$
- Решить задачу кластеризации – значит выделить в данных K кластеров – по факту, областей в пространстве признаков
- $a(x): X \rightarrow \{1, \dots, K\}$
- Главное требование к кластерам – хотим получить похожие друг на друга объекты внутри!
- И при этом непохожие на объекты другого кластера
- Почти классификация, просто без явного таргета!



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

МОТИВАЦИЯ

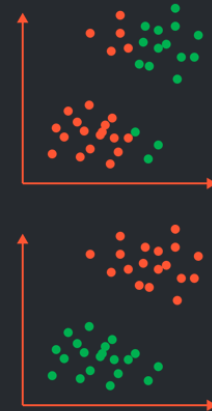
- Выявление структуры данных и, например, использование кластеров как категорий для будущего анализа
- Сжатие данных: выделяем часть представителей каждого полученного кластера
- Обнаружение аномалий, шума



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.

МЕТРИКИ

- Как оценить качество кластеризации?
- Пусть у каждого кластера k есть некоторый центр c_k !
- Внутрикластерное расстояние:
- $\sum_{k=1}^K \sum_{i=1}^n [a(x_i) = k] \cdot \rho(x_i, c_k) \rightarrow \min$
- Межкластерное расстояние:
- $\sum_{i,j=1}^n [a(x_i) \neq a(x_j)] \cdot \rho(x_i, x_j) \rightarrow \max$



>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.

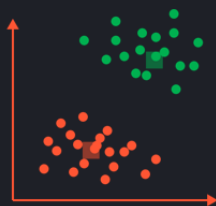
K-MEANS

- 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/параллелепипеды
- 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 подход к кластеризации.

>DBSCAN

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

— Density-based spatial clustering of applications with noise

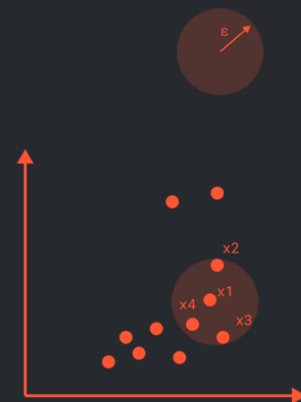
— У метода два гиперпараметра: ε, n

— Введем некоторые определения объектов для DBSCAN:

— Окрестность точки: $N_\varepsilon(x) = \{x_i \in X: |x - x_i|\}$

— $N_\varepsilon(x_1) = \{x_2, x_3, x_4\}$

— $|N_\varepsilon(x_1)| = 3$



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

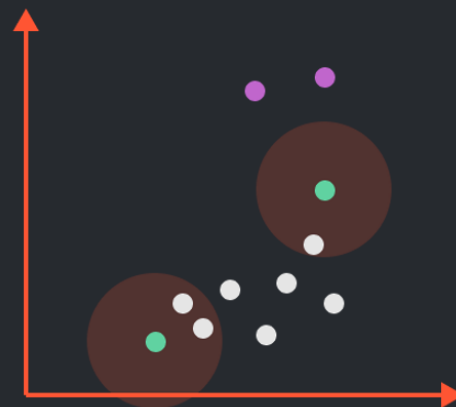
— Точка ядровая: $|N_\varepsilon(x)| \geq n$

— Точка **пограничная**: не ядровая, при этом

$\exists x_i: \rho(x, x_i) \leq \varepsilon$ и x_i — ядровая

— Точка **шумовая**: не ядровая и не пограничная

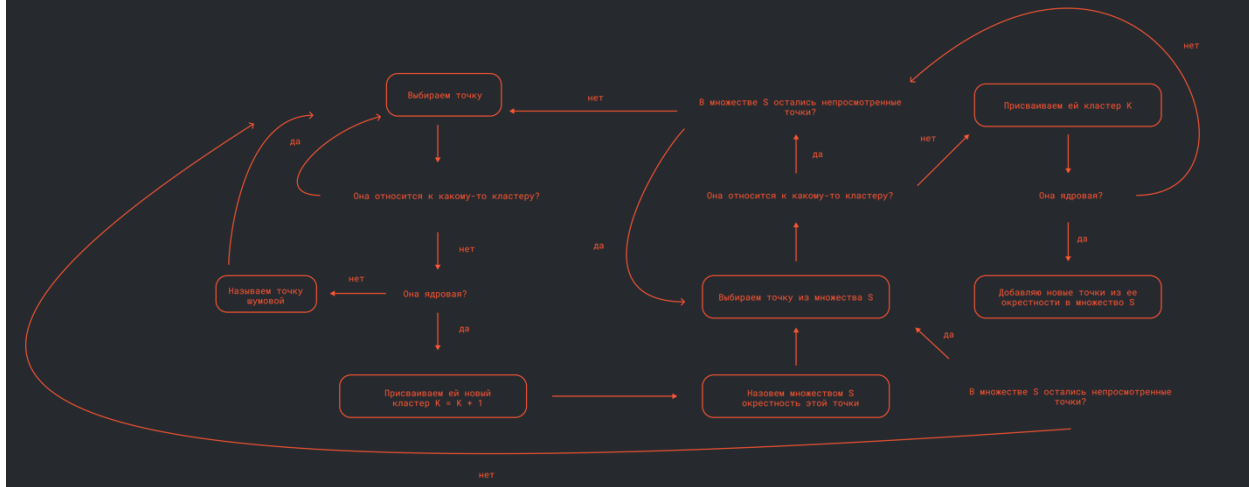
— Пусть $n = 3$



Теперь когда определили 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.

DBSCAN

Скажем, в начале все точки не принадлежат никакому классу, и $K = 0$



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