Методы машинного обучения

План лекции

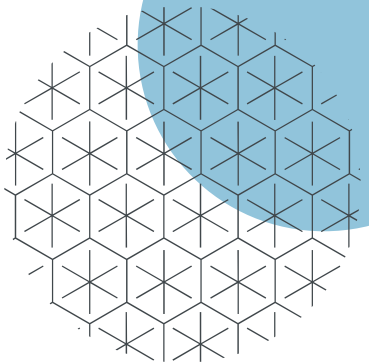
01 Что и зачем кластеризовать?

02 Как учить без учителя?

03 Что по метрикам?

01

Что и зачем кластеризовать?

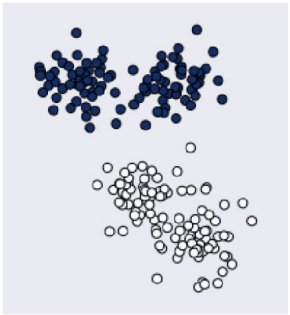


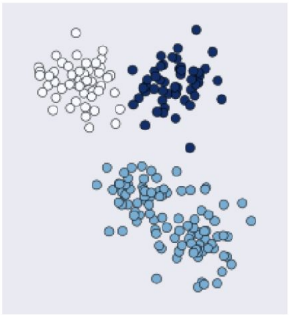
Кластеризация

Кластеризация - разбиение множества объектов на группы похожих.

Что можно подать на вход?

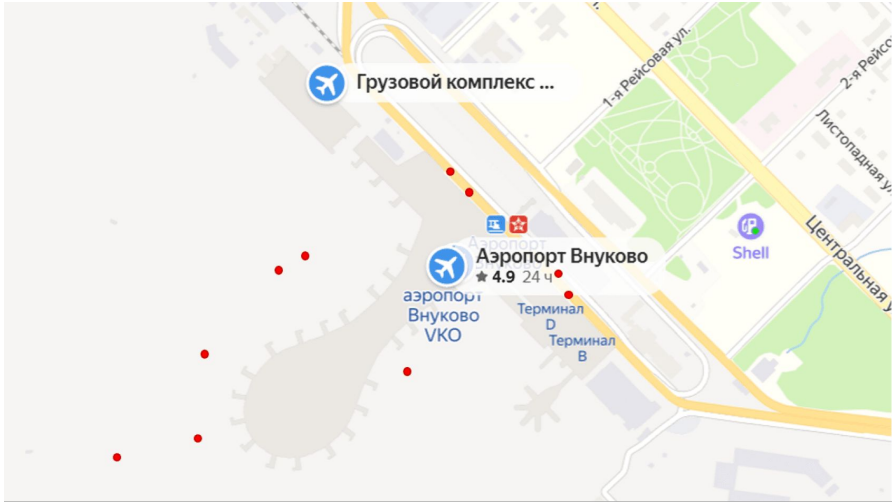
- признаковые описания
- попарные расстояния

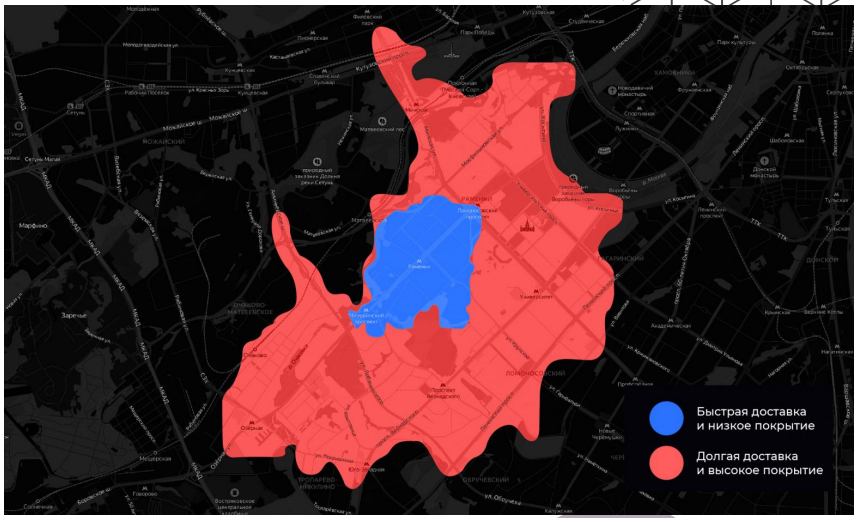

n_clusters=2


n_clusters=3


n_clusters=4

Где можно встретить?





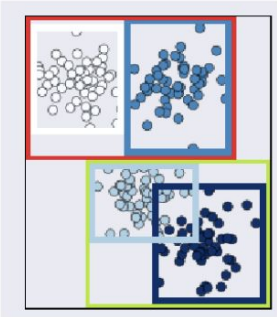
02

Как учить без учителя?

Методы кластеризации

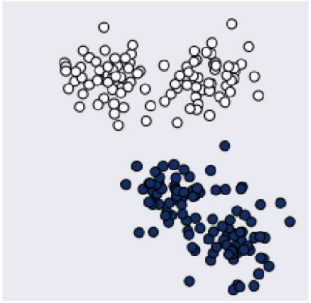
Плоские/разделяющие - кластеризация на n непересекающихся кластеров.

Иерархические - делим данные сначала на один большой кластер, далее рекурсивно делим на подкластеры

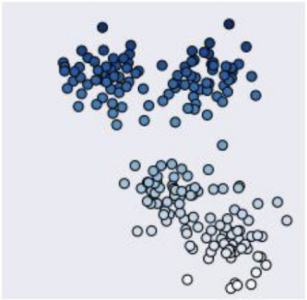


Виды кластеризации

Четкая (hard) - разбиение на непересекающиеся кластеры

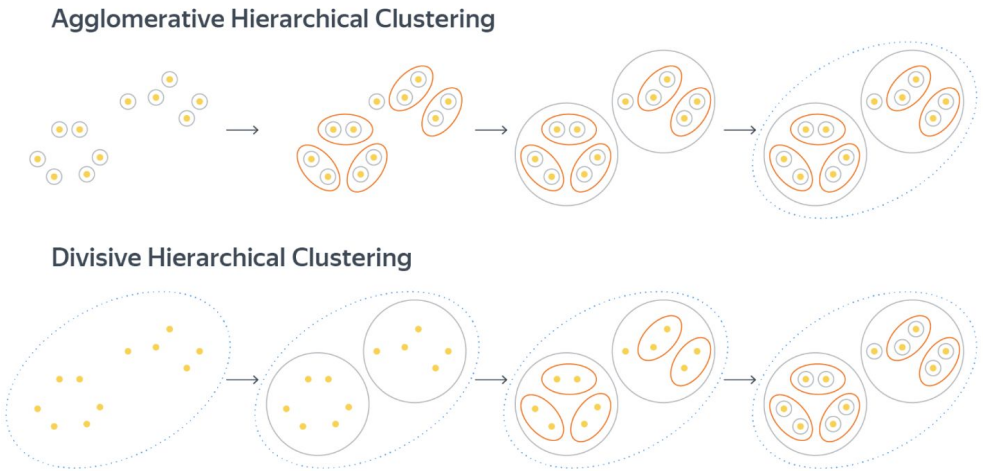


Нечеткая - определение степени принадлежности к кластерам



Иерархическая кластеризация

Во-первых, есть деление алгоритмов кластеризации на **агломеративные (agglomerative)** и **дивизивные (divisive)**.
Агломеративные алгоритмы начинают с **небольших** кластеров (обычно с кластеров, состоящих из одного объекта) и постепенно объединяют их в кластеры побольше.
Дивизивные начинают с **больших** кластеров (обычно – с одного единственного кластера) и постепенно делят на кластеры поменьше.
Не используют на больших данных!!!



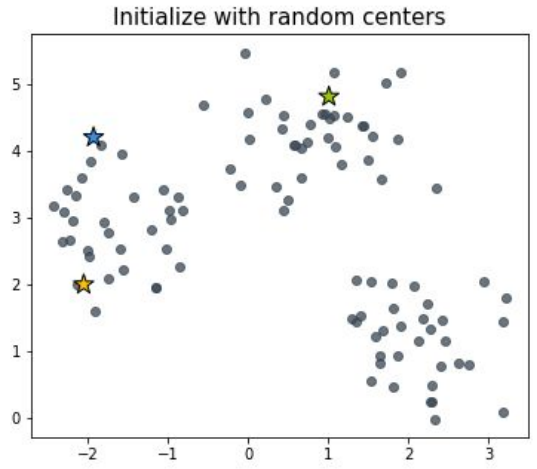
K-means

Концептуально k-means состоит из двух шагов:

- распределение объектов выборки по кластерам;
- пересчёт центров кластеров.

На самом деле:

- выбираем k случайных центров в пространстве признаков.
- объекты относим к ближайшему.
- находим центры масс кластеров и заново перераспределяем объекты по ним для уточнения центров.
- процесс продолжаем до тех пор, пока центры кластеров не перестанут меняться.



K-means - как выбрать центр?

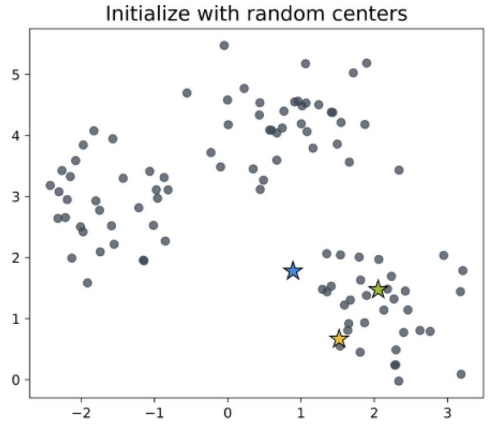
Потенциальная проблема:

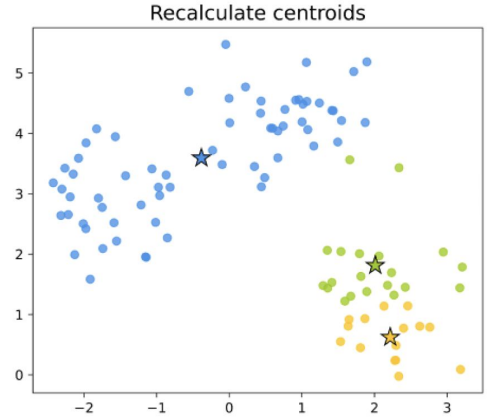
- кучное размещение центров.

Чтобы избежать - выгодно брать максимально удаленные друг от друга центры.

На практике (K-means++):

- первый центр выбираем случайно из равномерного распределения на точках выборки;
- каждый следующий центр выбираем из случайного распределения на объектах выборки, в котором вероятность выбрать объект пропорциональна квадрату расстояния от него до ближайшего к нему центра кластера.





K-means - что оптимизируем?

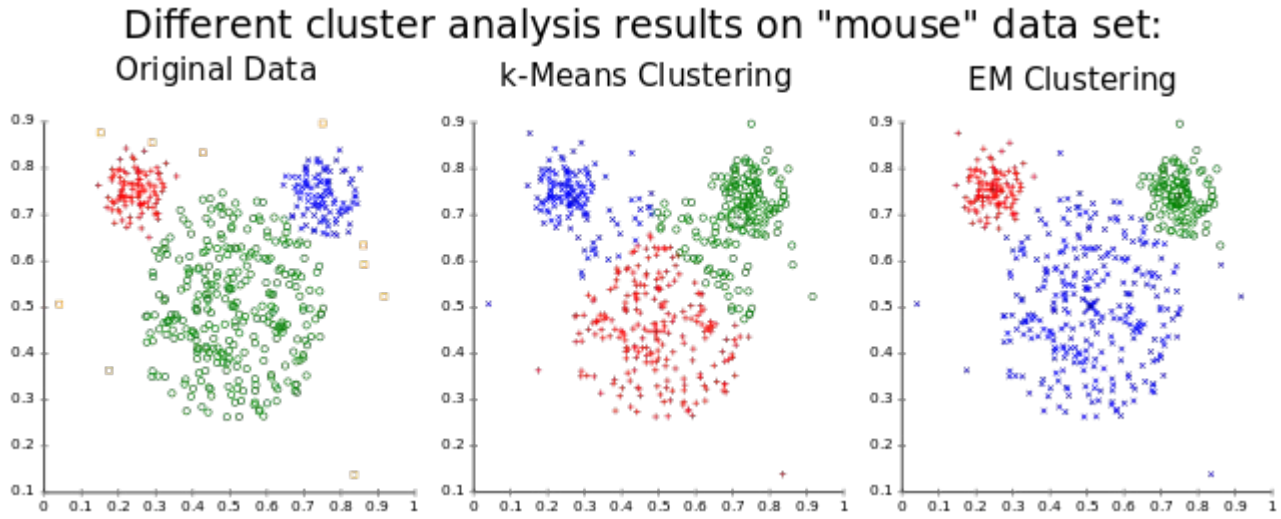
Потенциальная проблема:

Оба шага алгоритма работают на уменьшение среднего квадрата евклидова расстояния от объектов до центров их кластеров

$$\Phi_0 = \frac{1}{nK} \sum_{k=1}^{K} \sum_{i=1}^{n} (\mu_k - x_i)^2 \mathbb{I}[a(x_i) = k]$$

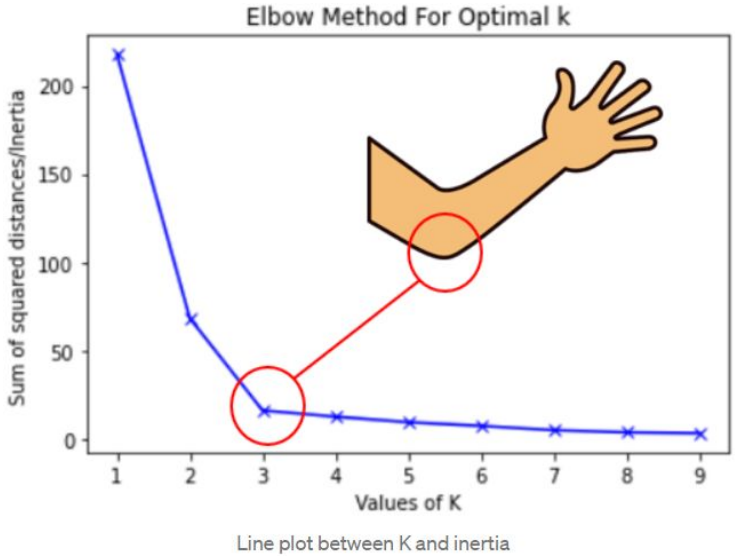
В чем проблема с k-means?

- ожидает, что размер кластеров примерно одинаков
- выстраивает сферические формы (оптимизируем квадрат расстояния)



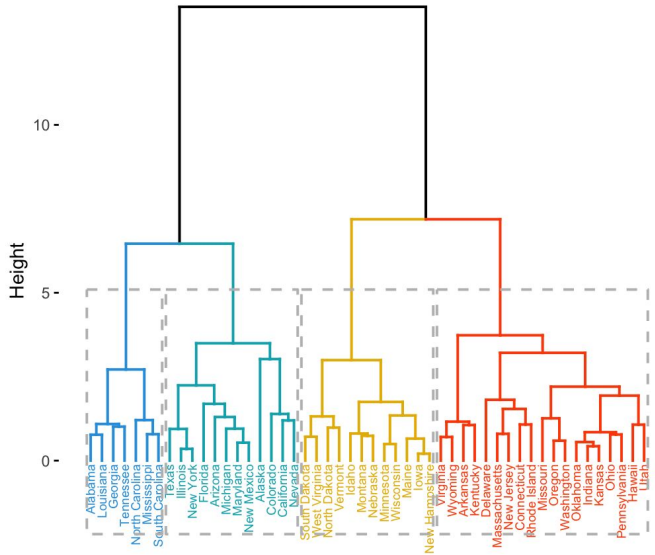
Как выбрать количество кластеров?

Метод локтя



Line plot between K and inertia

Дендрограммы



DBSCAN

Алгоритм **DBSCAN (Density-based spatial clustering of applications with noise)** развивает идею кластеризации с помощью выделения связных компонент.

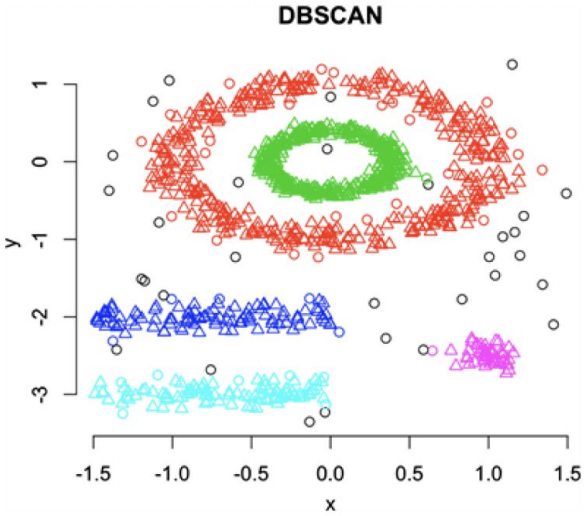
Виды точек:

- **внутренние / основные точки (core points)** - вокруг больше N объектов выборки
- **граничные (border points)** - рядом есть основные, но в окрестности объектов меньше
- **шумовые точки (noise points)** - рядом нет основных и мало объектов выборки.

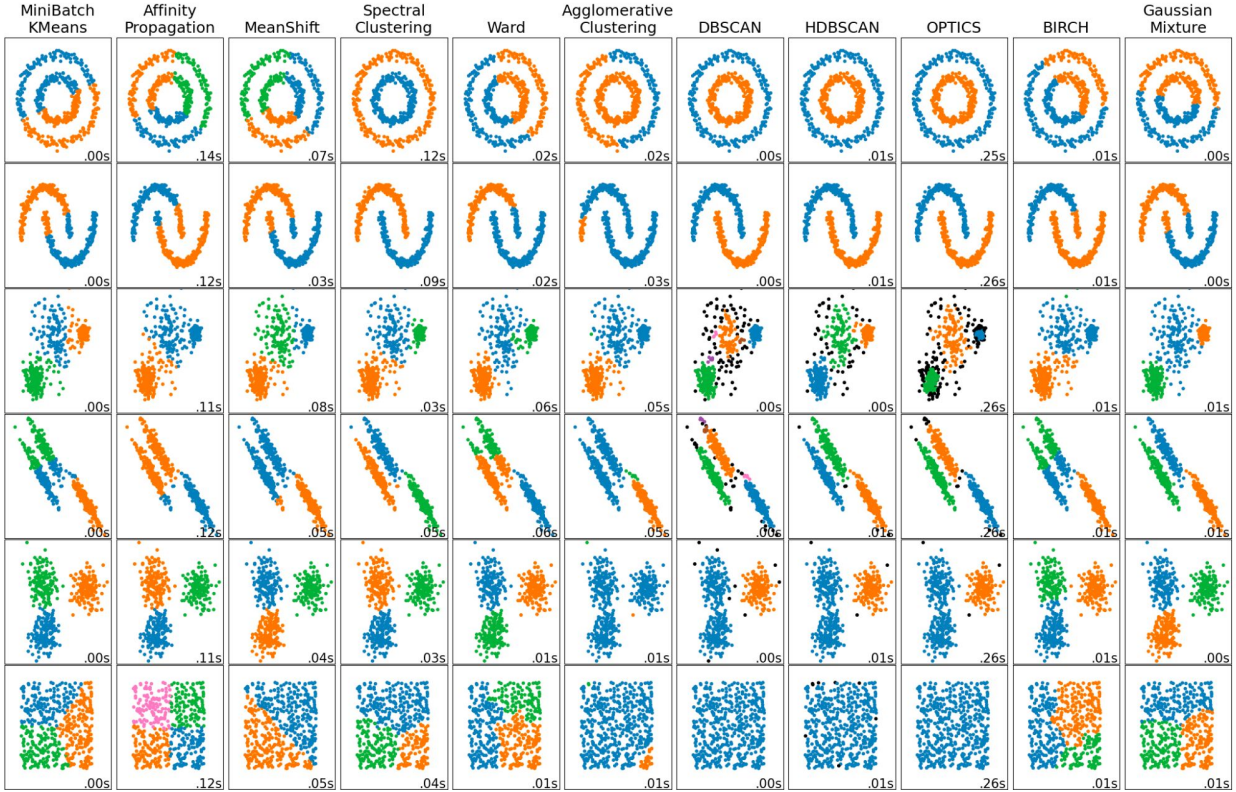
DBSCAN

Алгоритм кластеризации:

- Шумовые точки убираются из рассмотрения и не приписываются ни к какому кластеру.
- Основные точки, у которых есть общая окрестность, соединяются ребром.
- В полученном графе выделяются компоненты связности.
- Каждая граничная точка относится к тому кластеру, в который попала ближайшая к ней основная точка.



And so on


MiniBatch KMeans
Affinity Propagation
MeanShift
Spectral Clustering
Ward
Agglomerative Clustering
DBSCAN
HDBSCAN
OPTICS
BIRCH
Gaussian Mixture
.00s .14s .07s .12s .02s .02s .00s .01s .25s .01s .00s
.00s .12s .03s .09s .02s .03s .00s .01s .26s .01s .00s
.00s .11s .08s .03s .06s .05s .00s .00s .26s .01s .01s
.00s .12s .05s .05s .06s .05s .00s .01s .26s .01s .01s
.00s .11s .04s .03s .01s .01s .00s .01s .26s .01s .00s
.00s .12s .05s .04s .01s .01s .00s .01s .26s .01s .01s

03

Что по метрикам?

Метрики, кроме тех, что про посмотреть

Среднее внутрикластерное расстояние

$$F_0 = \frac{\sum_{i=1}^{n}\sum_{j=i}^{n}\rho(x_i,x_j)\mathbb{I}[a(x_i)=a(x_j)]}{\sum_{i=1}^{n}\sum_{j=i}^{n}\mathbb{I}[a(x_i)=a(x_j)]}$$

Среднее межкластерное расстояние

$$F_1 = \frac{\sum_{i=1}^{n}\sum_{j=i}^{n}\rho(x_i,x_j)\mathbb{I}[a(x_i)\neq a(x_j)]}{\sum_{i=1}^{n}\sum_{j=i}^{n}\mathbb{I}[a(x_i)\neq a(x_j)]}$$

Метрики, кроме тех, что про посмотреть

Гомогенность

$$Homogeneity = 1 - \frac{H_{class|clust}}{H_{class}}$$

Полнота

$$Completeness = 1 - \frac{H_{clust|class}}{H_{clust}}$$

Коэффициент силуэта

$$S(x_i) = \frac{B(x_i) - A(x_i)}{\max(B(x_i), A(x_i))}$$

Кластеризация по мнению Kandinsky