МО с учителем

$X = \{(x_i, y_i)\}_{i=1}^{\ell}$ - обучающая выборка

$a(x_i) \approx y_i$ $\qquad Q(a, X) \to \min\limits_{a \in A}$.

МО без учителя

Unsupervised learning.

$X = \{x_i\}_{i=1}^{\ell}$ нужно что-то сделать с X

Кластеризация

$X = \{x_i\}_{i=1}^{\ell}$ хотим $a : X \to \{1, \dots, K\}$,

модель кластеризации

x_i и x_j похожи $\Leftrightarrow$ $a(x_i) = a(x_j)$

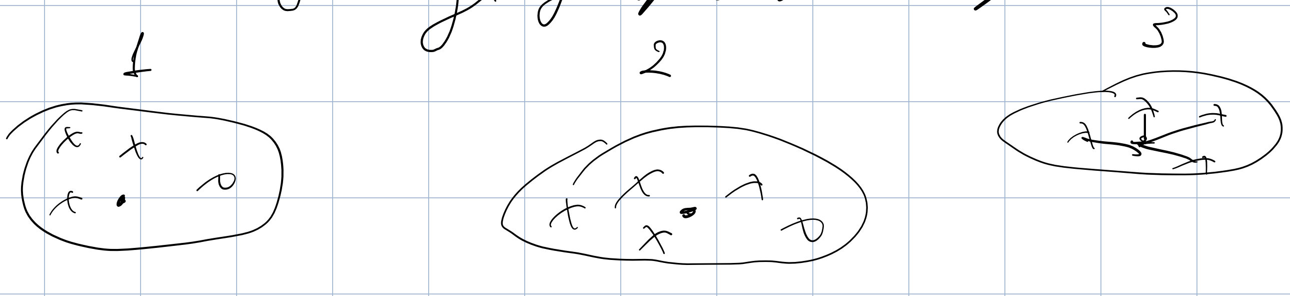
Зачем

1) Исследование данных

- социологический опрос.
- телефонный оператор
- кластеризация текстов

2) Генерация новых признаков

- добавить категориальный признак
- добавлять расстояния 1 .... к. до центров кластеров



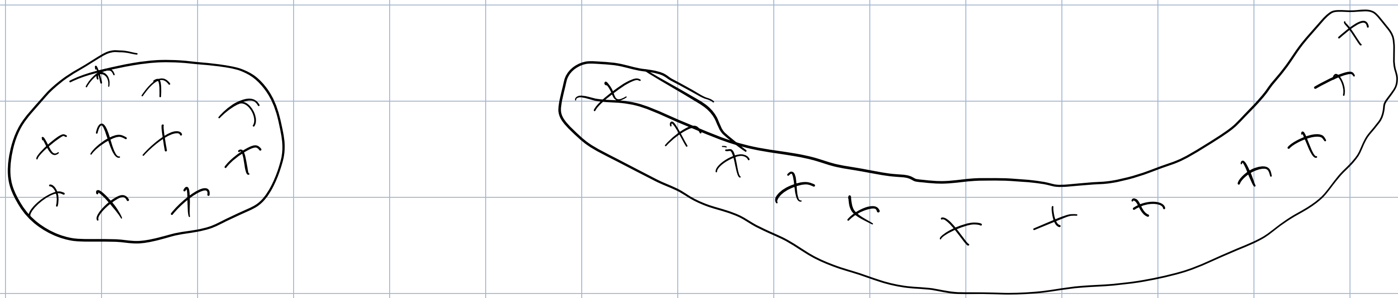
Метрики качества кластеризации

0) Посмотреть глазами

1) Внутрикластерное расстояние
требует "компактность кластеров"

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

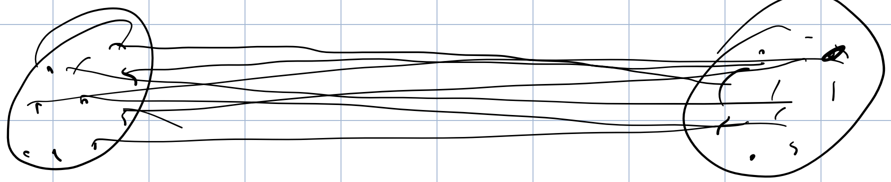
c_k - центр k-го кластера



2) Межкластерное расстояние

Требуем чтобы кластеры были "удалены друг от друга"

$$\sum_{i\neq j}[a(x_i)\neq a(x_j)]\,\rho(x_i, x_j)\to\max$$



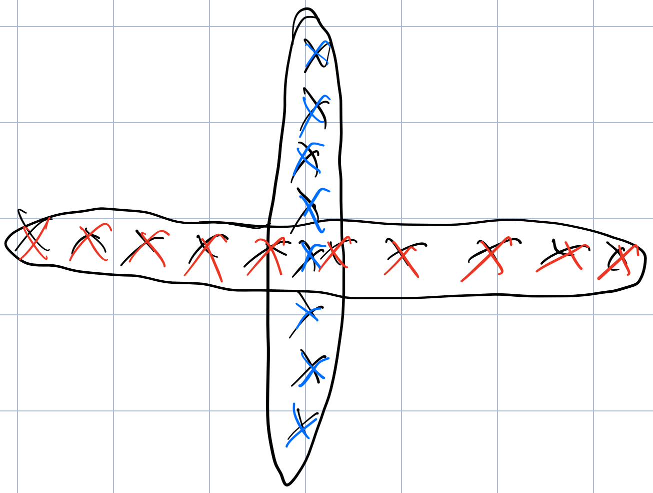
3) Комбинируем подходы

Индекс Данна

$d(k, k')$ — расстояние между кластерами k и k'

$d(k)$ - внутрикластерное расстояние

$$\frac{\min\limits_{1\le k\le k'\le K} d(k,k')}{\max\limits_{1\le k\le K} d(k)} \to \max$$



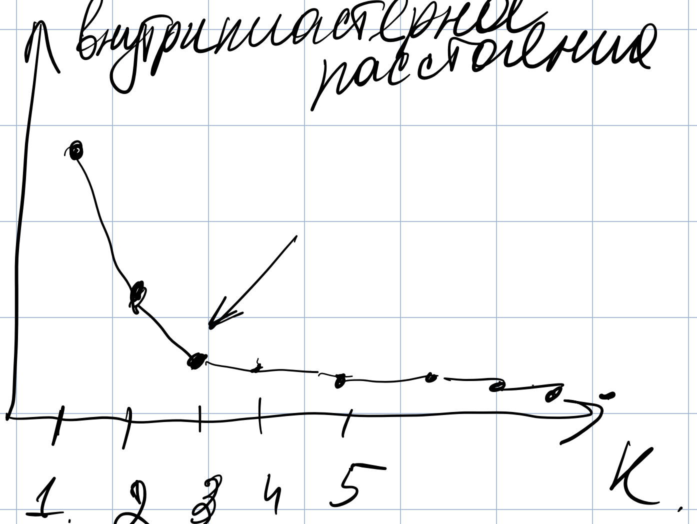
- метод выбора K - кол-во кластеров реализован внутри алгоритма кластеризации

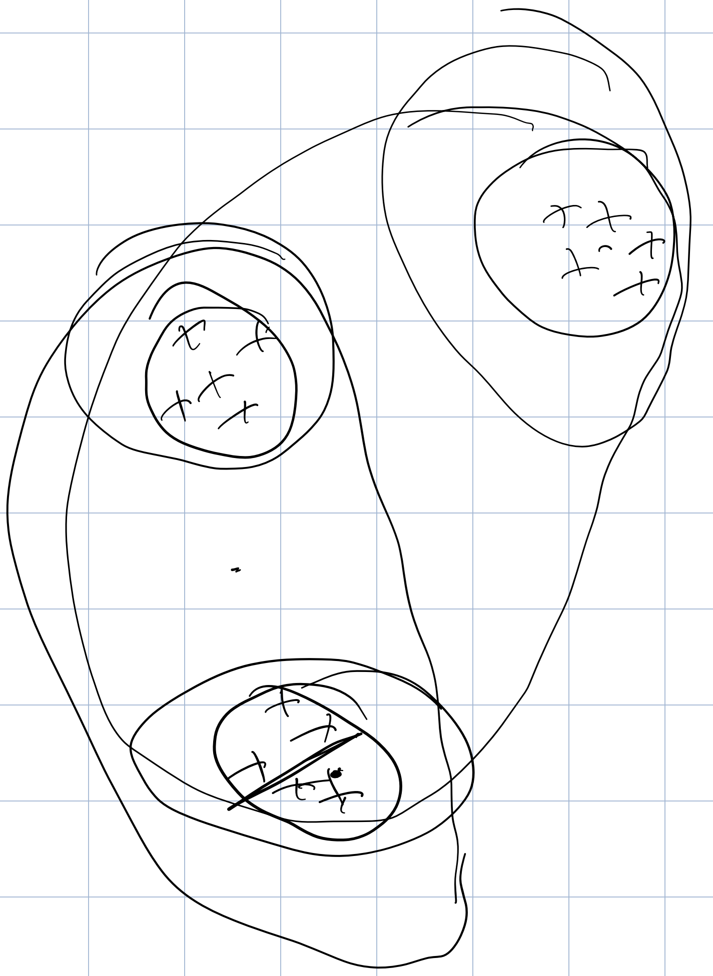
- к - гиперпараметр.

Как подобрать К?

1) Смотреть глазами

2) elbow method





K-means

ρ(x, z) - какое-то расстояние

K - гиперпараметр, число кластеров

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

c_k - центр кластера k.

Шаг 0 Инициализируем центры c_k

Шаг а Фиксируем c_k.

$$a(x_i)=\arg\min_{k=1,\dots K.}\rho(x_i,c_k)$$

Шаг б Фиксируем $a(x_i)$

Ищем центры масс

$$c_k=\arg\min_{c}\sum_{a(x_i)=k}\rho(x_i,c)$$

если $\rho_{(x,c_k)}=\|x-c_k\|^2$

$$c_k=\frac{1}{\sum_{i=1}^{\ell}[a(x_i)=k]}\sum_{i=1}^{\ell}[a(x_i)=k]\cdot x_i$$