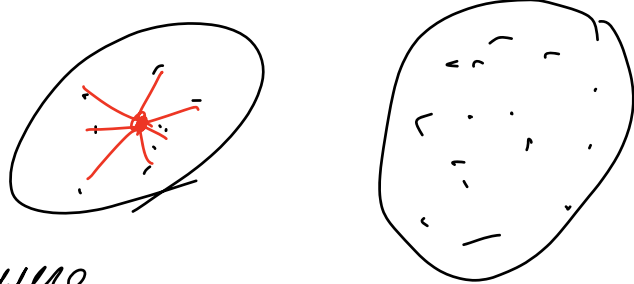# Обучение без учителя

$$X = (x_i)_{i=1}^l$$

Что можно кластеризовать:
- табличные
- тексты — тематика
  - окрас
  - Стиль
- Картинки

Метрики качества:

Внутренние

1) Внутрикластерное расстояние

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

c_k — центр кластера (Не обязательно геом. центр.)

2) Межкластерное расстояние

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

3) Индекс Данна (Дини)

$$\frac{\min_{1 \le K \le K' \le K} d(K, K')}{\max_{1 \le K \le K} d(K)} \rightarrow max$$

$d(K, K')$ — межкластерное
$d(K)$ — внутрикластерное

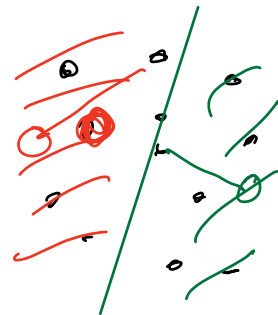Если знаем реальные метки класса,
то используем любую метрику
классификации

$$K - means.$$

Оптимизируем внутрикластерное
расстояние

1) Выбираем (случайно) позиции k
центроидов C_k, $k \in \overline{1, K}$

2) Фиксируем центроиды
$a(x_i) = argmin_{1 \le K \le K} \rho(x_i, C_k)$

3) Фиксируем распределение по кластерам.

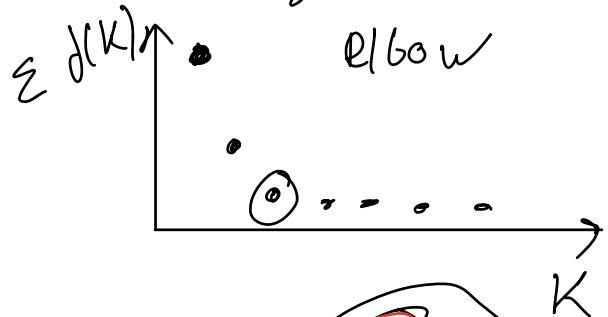$$c_k = \frac{\sum_{i=1}^{\ell} [a(x_i) = K] x_i}{\sum_{i=1}^{\ell} [a(x_i) = K]}$$

1) Стартовое кол-во и положение кластеров

K means ++

2) Способ пересчёта центроидов
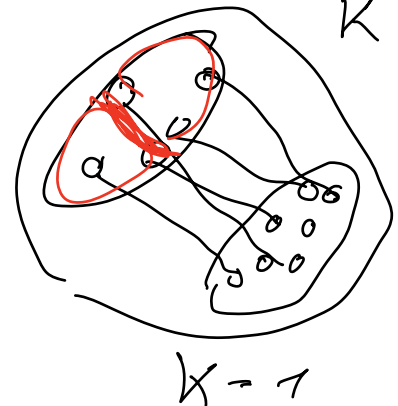
mean

median

$\xi \, J(K)$ elbow

K

DBSCAN

Графовые методы } след,

Спектральная.   пара

K = 1

Иерархическая

Восходящая   Aglomerative

Нисходящая   Descentive

1) Каждое наблюдение это отдельный кластер

2) На каждом шаге объединяем два
наиболее близких кластера (Расст.
Уорда)

3) Завершаем когда остался один кластер.