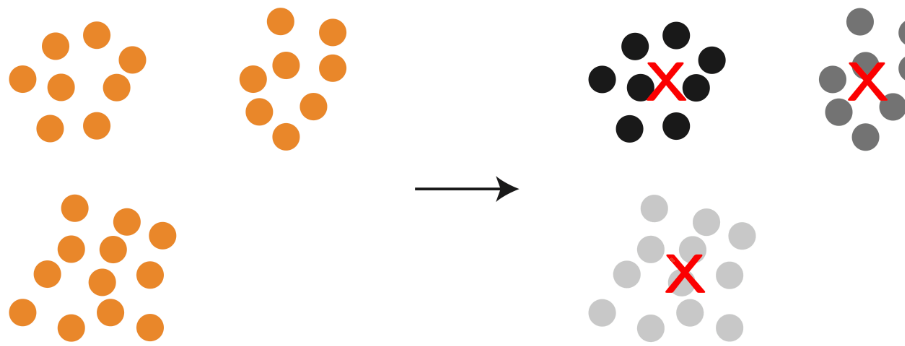
Задача кластеризации



Постановка:

Дано: $S \subset \mathbb{R}^d$ ← мн-во точек

размер n

$k \in \mathbb{N}$ ← число кластер

Выход: $T \subseteq \mathbb{R}^d$ $\quad |T| = k$ ← точки центров кластеров

$$\min_T \; cost(S,T) = \sum_{x \in S} \min_{z \in T} \|x - z\|_2^2$$

все точки, центры

евклидово расстояние

Обозначение:

$$cost(S,T) = \sum_{x \in S} \min_{z \in T} \|x-z\|_2^2 = \sum_{z \in T} \sum_{x \in C_z} \|x-z\|_2^2$$

кластер с объектами, ближ. к z

для множ. подмн. S

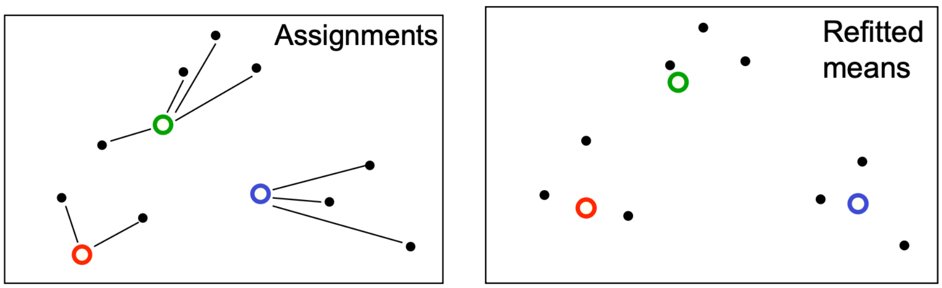
$$cost(C,T) = \sum_{x \in C} \min_{z \in T} \|z - x\|_2^2$$

$$\mathrm{cost}(C, z) = \sum_{x \in C} \|z - x\|_2^2$$

один центр

K-means

Идея:



Лемма $\forall C \subset \mathbb{R}^d \quad \forall z \in \mathbb{R}^d$

$$\mathrm{cost}(C, z) = \mathrm{cost}(C, \mathrm{mean}(C)) + |C| \cdot \|z - \mathrm{mean}(C)\|_2^2$$

$z = \mathrm{mean}(C)$ дает $\min \mathrm{cost}(C, z)$

Док-во: $$\mathrm{cost}(C, z) = \Bigg(\underbrace{\frac{1}{|C|} \sum_{x \in C} \|x - z\|_2^2}_{\mathbb{E}_{x \sim U(C)} \|x - z\|_2^2}\Bigg) \cdot |C|$$

$$\cancel{|C|} \cdot \mathbb{E}_x \|x - z\|_2^2 \overset{?}{=} \mathbb{E}\|X - \mathbb{E}X\|_2^2 \cdot \cancel{|C|} + \cancel{|C|} \cdot \|z - \mathbb{E}X\|_2^2$$

$$\mathbb{E}\|X\|_2^2 + \|\mathbb{E}X\|_2^2 - \underbrace{2\,\mathbb{E}(X \cdot \mathbb{E}X)}_{2\|\mathbb{E}X\|_2^2}$$

$$+ \|z\|_2^2 + \|\mathbb{E}X\|_2^2 - 2\, z \cdot \mathbb{E}X$$

$$= \mathbb{E}\|X\|_2^2 + \|z\|_2^2 - 2\, z \cdot \mathbb{E}X = \mathbb{E}\|X - z\|_2^2 \quad \blacksquare$$

Алгоритм K-mens:

Вход: $C_1^0 \dots C_k^0$ $\quad t=0$

while не выполнено условие останова

1) для $\forall i$: $z_i^{t+1} = \mathrm{mean}(C_i^t)$
2) для $\forall i$: $C_i^{t+1} = \{x \in S \mid i = \arg\min_j \|z_j^{t+1} - x\|_2^2\}$

<u>Теорема</u> K-means монотонно сходится

Док-во: $\sum_{z \in T} \sum_{x \in C_z} \|x - z\|_2^2 = \sum_{i=1}^{k} \sum_{x \in C_i} \|x - z_i\|_2^2$

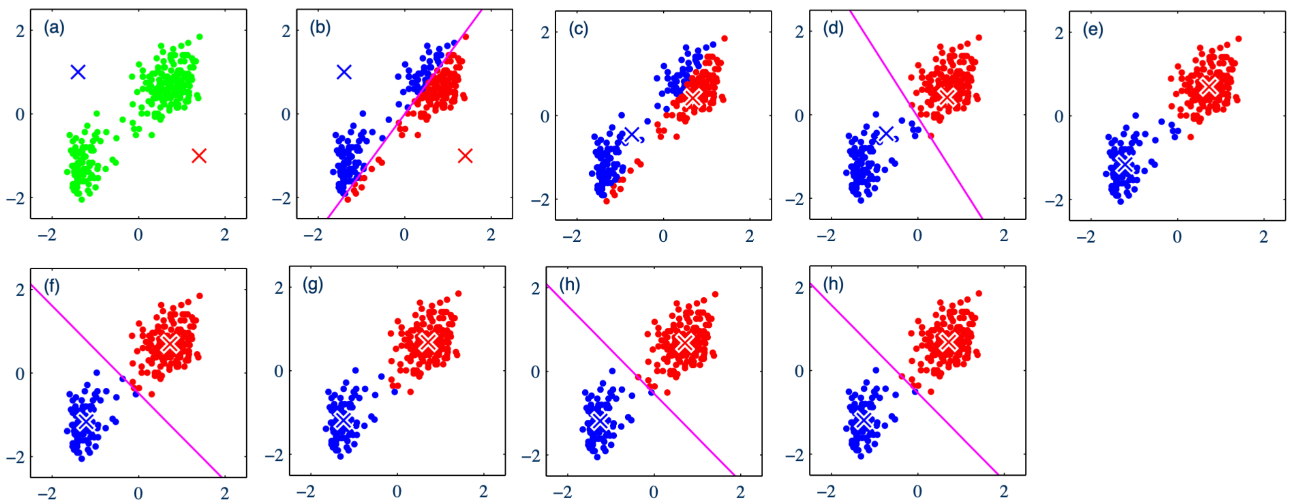
1) $\mathrm{cost}(C_i^t, z_i^{t+1}) \le \mathrm{cost}(C_i^t; z_i^t)$

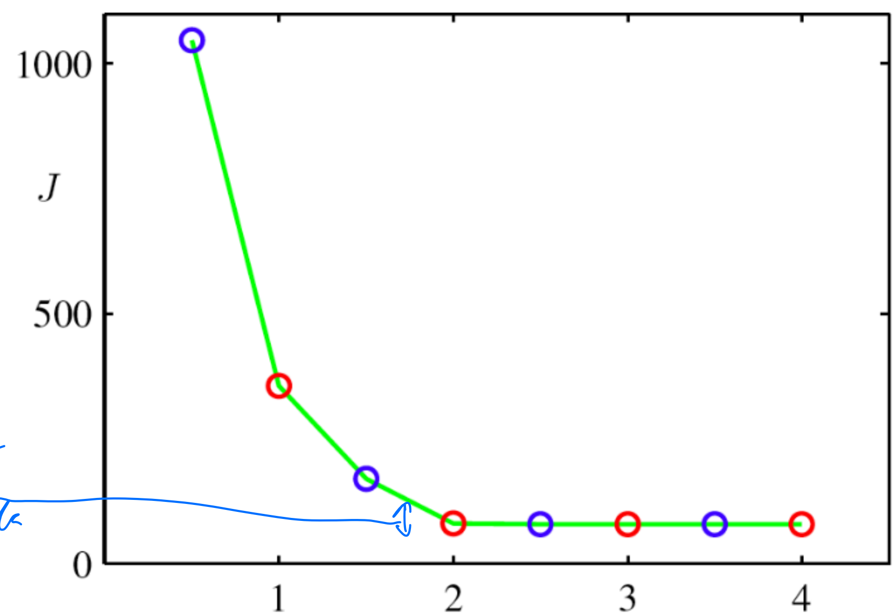
   $\sum_i \le$ по лемме 1

2) из сути задачи

   $\mathrm{cost}(C_i^{t+1}, z_i^{t+1}) \le \mathrm{cost}(C_i^t, z_i^{t+1})$

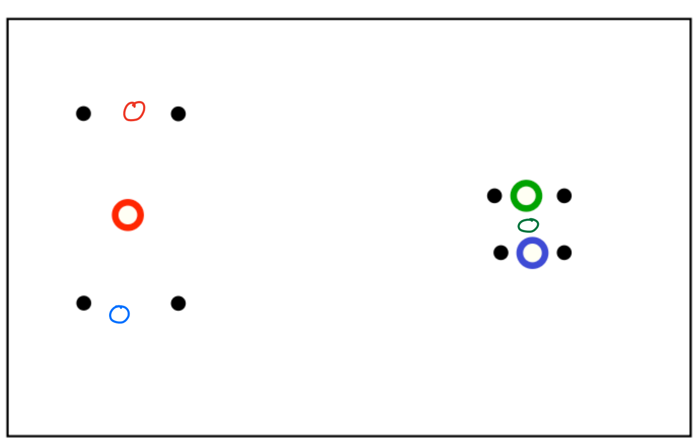
$$\mathrm{cost}(C_i^{t+1}; z_i^{t+1}) \le \mathrm{cost}(C_i^t; z_i^t) \quad \blacksquare$$





Критерий остановки

Примеры плохих результатов:



локальный минимум

А как инициализировать K-means?

K-means ++

Берем $x \in S$ — это первая точка z_1

for $i = 1 \ldots k-1$

берем $x \in S$ исходя

$$\mathbb{P}\{x\} \sim \mathrm{cost}(x, T_i) = \min_{z \in T_i} \|x - z\|_2^2$$

$T_i = \{z_1, \ldots z_i\}$

$$\mathbb{P}\{x\} = \frac{\min_{z \in T_i} \|x - z\|_2^2}{\sum_{x' \in S} \min_{z \in T_i} \|x' - z\|_2^2}$$

Берут 1 точку — в любом случае попал в кластер

Лемма 2

$\forall C \in \mathbb{R}^d$, $z \sim U(C)$, тогда

$$\mathbb{E}[\mathrm{cost}(C, z)] = 2\,\mathrm{cost}(C, \mathrm{mean}(C))$$

Док-во:

$$\mathbb{E}[\mathrm{cost}(C, z)] = \sum_{z' \in C} \frac{1}{|C|} \underbrace{\mathrm{cost}(C, z')}_{\text{лемма 1}}$$

$$= \frac{1}{|C|} \sum_{z' \in C} \Big(\mathrm{cost}(C, \mathrm{mean}(C)) + |C| \cdot \|z' - \mathrm{mean}(C)\|_2^2\Big)$$

$$= \operatorname{cost}(C, \operatorname{mean}(C)) + \sum_{z' \in C} \|z' - \operatorname{mean}(C)\|_2^2$$

$$= 2 \operatorname{cost}(C, \operatorname{mean}(C))$$

Лемма 3

T_i — текущие центры, и добавлено $z_{i+1} \in C^*_{i+1}$ — идеальный кластер

$$\mathbb{E}\left[\operatorname{cost}(C^*_{i+1}, T_i \cup \{z_{i+1}\})\right] \le 8 \operatorname{cost}(C^*_{i+1}, z^*_{i+1})$$

z^*_{i+1} — $\operatorname{mean}(C^*_{i+1})$

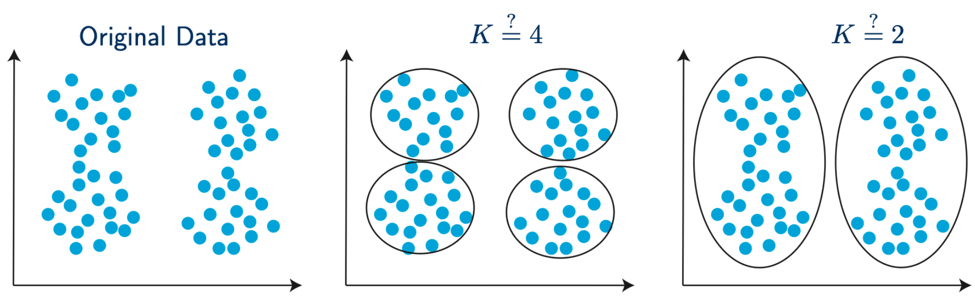
Теорема

T — центры для k-means++

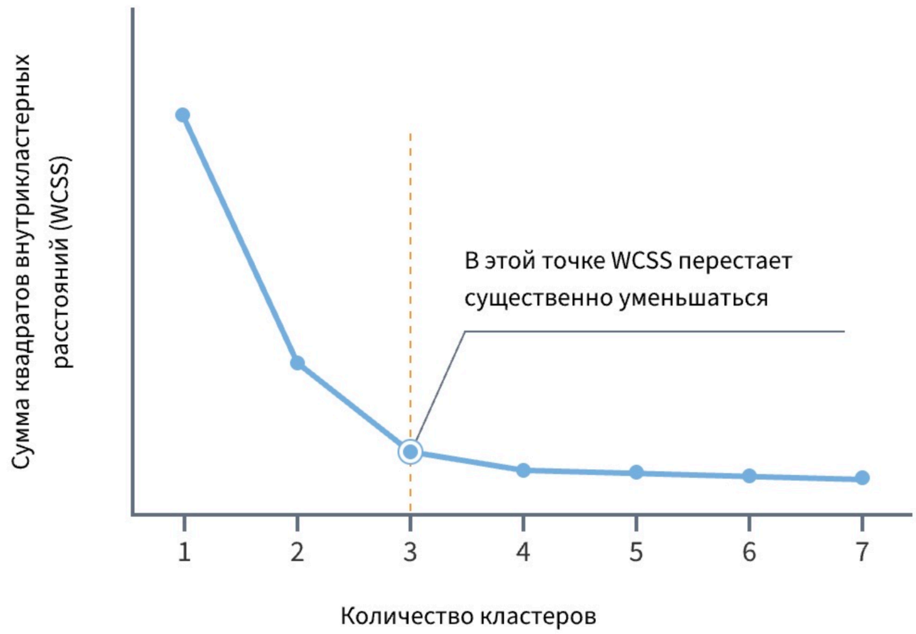
T^* — идеальные центры $(\operatorname{cost}(S, T^*))$ — цель

$$\mathbb{E}[\operatorname{cost}(S,T)] \le 8 \cdot (1 + \ln k) \operatorname{cost}(S, T^*)$$

Как выбрать k?



1) Метод локтя

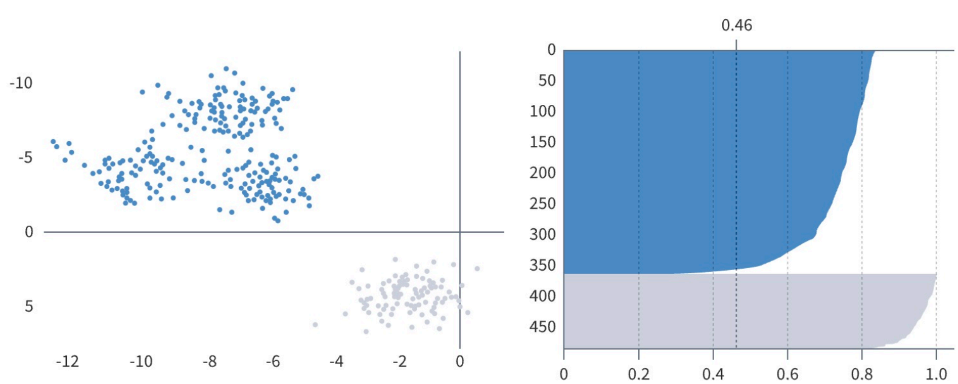


2) Метод силуэтов

для каждой $i \to a(i)$ – среднее расстояние от i точки до всех ост. в кластере

$\to b(i) =$ среднее расстояние от i точки до всех точек в „ближ." кластере

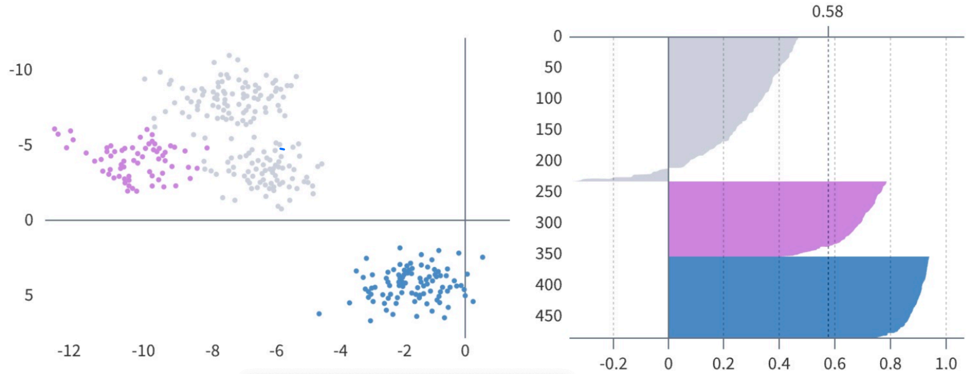
$$s(i) = \frac{b(i) - a(i)}{\max\{a(i); b(i)\}} \in [-1; 1]$$

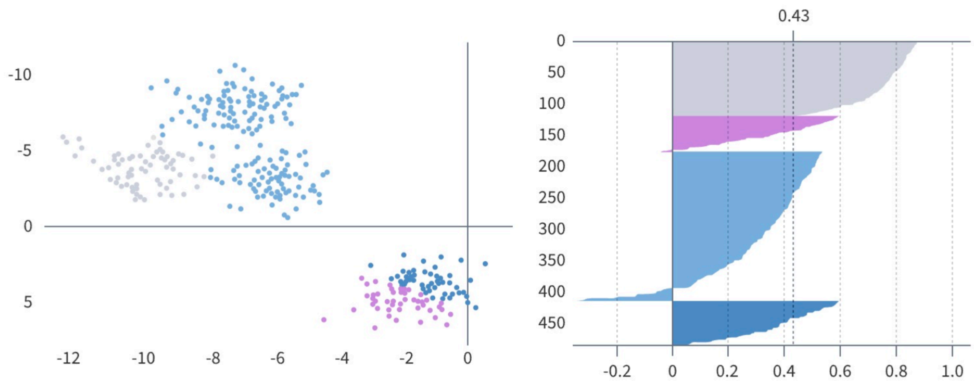


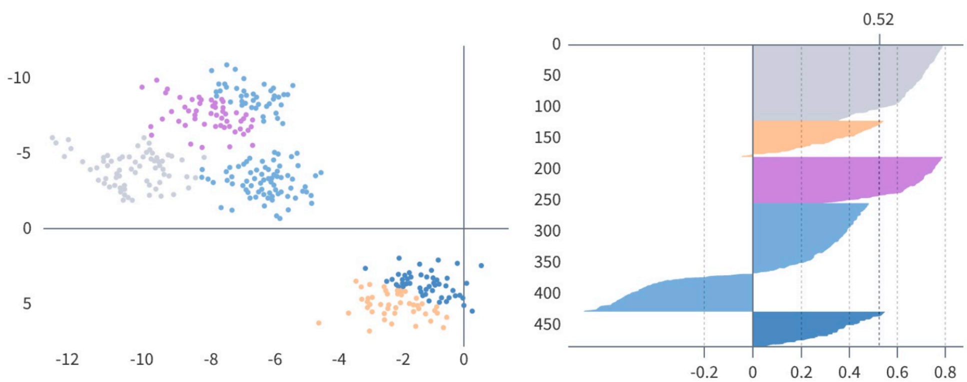
$s(i) \approx 1$ супер

$s(i) \approx 0$ неувер.

$s(i) \approx -1$ ошибка кластер.







3) Зазор статистика