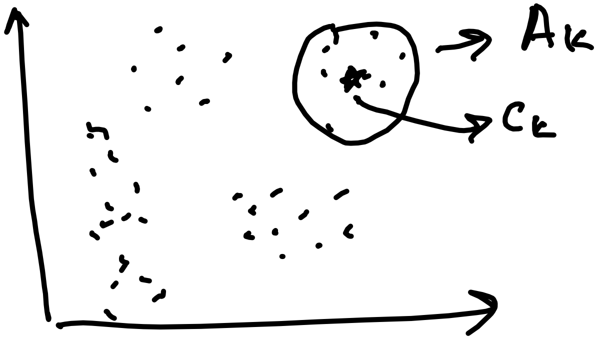
Кластеризация



k-means

$$Q = \sum_{k=1}^{K} \sum_{i: x_i \in A_k} \|x_i - c_k\|^2 \to \min_{c_1,\dots,c_K}$$

$$A_k = \{x_i \mid \rho(x_i, c_k) \le \rho(x_i, c_j) \cancel{\forall k \neq j}\} \quad \forall j \neq k$$

$i = 1, \dots, n$

$j = 1, \dots, K$

$$Q \to \min_{c_k}$$

$$\sum_{i: x_i \in A_k} \|x_i - c_k\|^2 \to \min_{c_k}$$

$$\|x_i - c_k\|^2 = (x_i^T - c_k^T)(x_i - c_k) =$$

$$= (x_i^T x_i - 2x_i^T c_k + c_k^T c_k)$$

$\swarrow$

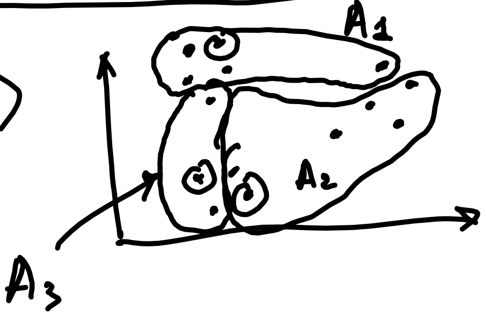
$$\nabla_{c_k} = -2x_i + 2c_k$$

$$\nabla_{c_k} \sum_{i:x_i\in A_k} \|x_i - c_k\|^2 = \sum_{i:x_i\in A_k} (-2x_i + 2c_k) = 0$$

$$-\sum_{i:x_i\in A_k} x_i + |A_k|\cdot c_k = 0$$

$$\boxed{c_k = \frac{1}{|A_k|}\cdot\sum_{i:x_i\in A_k} x_i}$$

1)



2) Пересчитываем центры.

Критерии останова:

1) Число набл. (доля), сменивших кластер.

2) $\| c_k^{old} - c_k^{new} \| < \varepsilon$

3) Изменение Q

$$\left| \frac{Q_{new} - Q_{old}}{Q_{old}} \right|$$

---

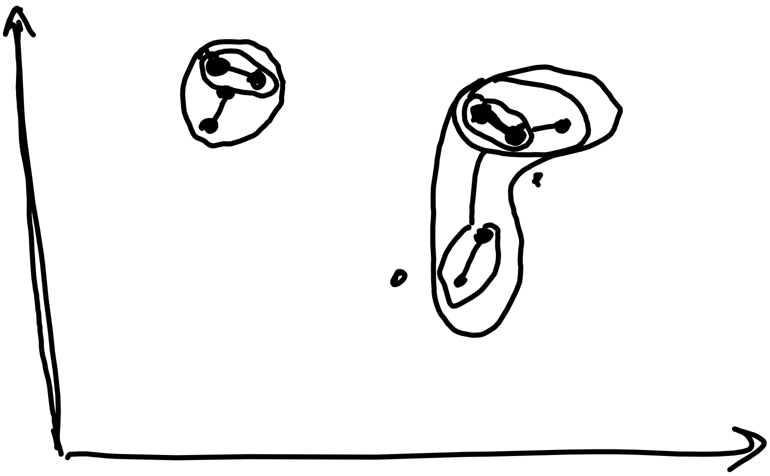
0) $\bar{\rho}_{внутр.}$, $\bar{\rho}_{межкл.}$

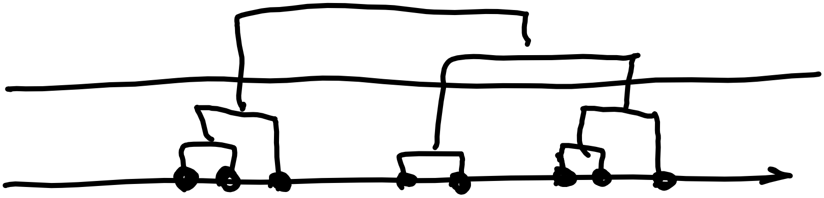
1) Индекс. Дана $= \dfrac{\bar{\rho}_{внутр.}}{\bar{\rho}_{межкл.}} \to \min$

Duna Indx

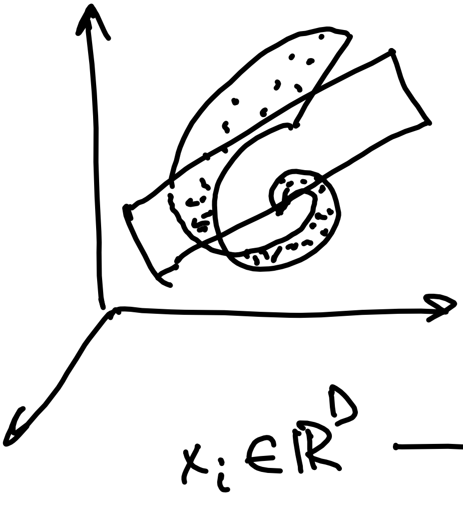
2) $$\text{Silouette}_i = \frac{-\big(d(x_i, c_k) - d(x_i, c_j)\big)_{j\,k+1}}{d(x_i, c_{k+1})}$$

$x_i : d(x_i, c_k) < d(x_i, c_{k+1})$





TSNE



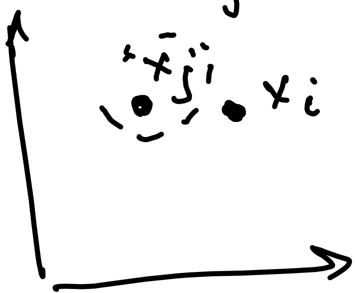
$x_i \in \mathbb{R}^D \longrightarrow y_i \in \mathbb{R}^d$

$$\sum_{i=1}^{n}\sum_{j=1}^{n}\left(\rho(x_i,x_j)-\rho(y_i,y_j)\right)^2 \to \min_{y_1,\dots,y_n}$$

$\boxed{\text{M}\cancel{\text{DS}}}$

$$\frac{\rho(x_i,x_j)}{\rho(x_i,x_k)} = \alpha \approx \frac{\rho(y_i,y_j)}{\rho(y_i,y_k)}$$

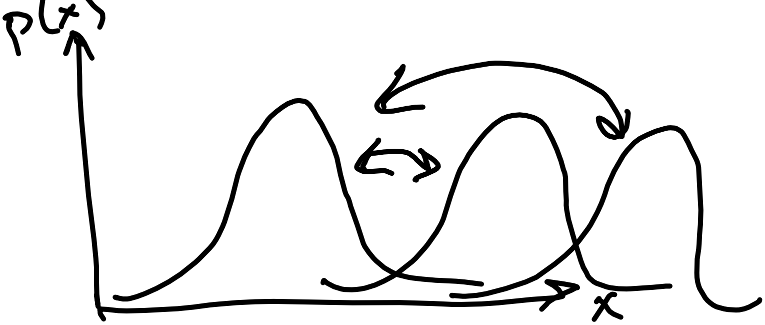
$$p_{j|i} = \frac{\exp\left(-\frac{\|x_i-x_j\|^2}{2\delta^2}\right)}{\sum\limits_{k\neq j}\exp\left(-\frac{\|x_i-x_k\|^2}{2\delta^2}\right)}$$



$$\boxed{p_{ij} = \frac{p_{i|j}+p_{j|i}}{2n}} \quad , \quad \boxed{q_{ij} = \frac{q_{i|j}+q_{j|i}}{2n}}$$

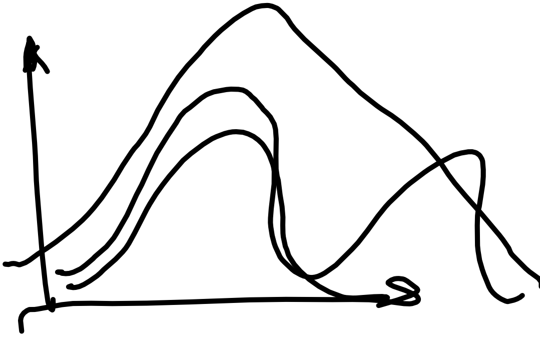
$$q_{j|i} = \frac{\left(1+\|y_i-y_j\|^2\right)^{-1}}{\sum\limits_{k\neq j}\left(1+\|y_i-y_k\|^2\right)^{-1}}$$

$q_{ij} \approx p_{ij}$



$$KL(p \| q) = \int_{\mathbb{R}} p(x) \log \frac{p(x)}{q(x)} dx =$$

$$= E_p\left(\log \frac{p}{q}\right)$$



$$KL(N_1 \| N_2)$$

$$N_1 = N(a_1, \sigma^2)$$

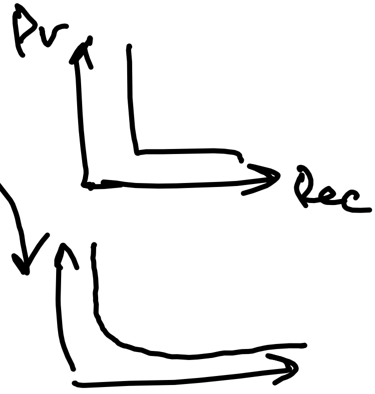
$$N_2 = N(a_2, \sigma^2)$$

$$Q = \sum_i \sum_j p_{ij} \log \frac{p_{ij}}{q_{ij}} \to \min_{y_1,\dots,y_n}.$$

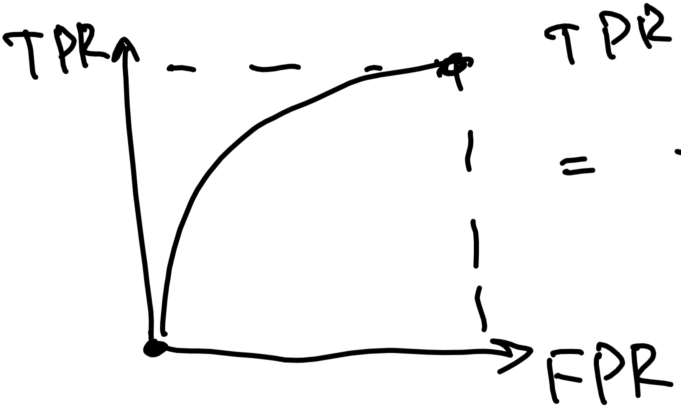
$\boxed{\text{Семинар}}$:

1) $f_1 = \dfrac{2Pr \cdot Rec}{Pr + Rec}$

$\max(\min(Pr, Rec))$



2) AUC-ROC $= TPR(FPR)$



$TPR = Rec =$

$$= \frac{\sum_i [y_i = 1] \cdot [a_i = 1]}{\sum_i [y_i = 1]}$$

$$FPR = \frac{\sum_i [y_i = -1] \cdot [a_i = 1]}{\sum_i [y_i = -1]}$$

$$\text{AUC-ROC} = \frac{\sum_i \sum_j [y_i < y_j] \cdot [a_i < a_j]}{\sum_i \sum_j [y_i < y_j] = n_+ \cdot n_-}$$

3) $\mathbb{E}(\text{AUC-ROC}) = 0.5$

4)

$$\begin{cases} a_k^T \underset{D\times D}{S} a_k \to \max\limits_{a_k} \\ \|a_k\|^2 = 1 \end{cases}$$

$x_i \in \mathbb{R}^D$

$$S = \frac{1}{n} \sum_{i=1}^{n} x_i \cdot x_i^T$$

$a_k \in \mathbb{R}^D$

5) $x_i \in \mathbb{R}^D$, $A = \underset{D\times d}{\begin{bmatrix} | & & | \\ a_1 & \dots & a_d \\ | & & | \end{bmatrix}}$

$$\underset{\mathbb{R}^d}{\underset{\uparrow}{\hat{x}_i}} = (\underbrace{A^T A}_{I_d})^{-1} \underset{d\times D}{A^T} \underset{D\times 1}{x_i} = A^T x_i$$