KNN

$\rho : X \times X \to [0, +\infty)$

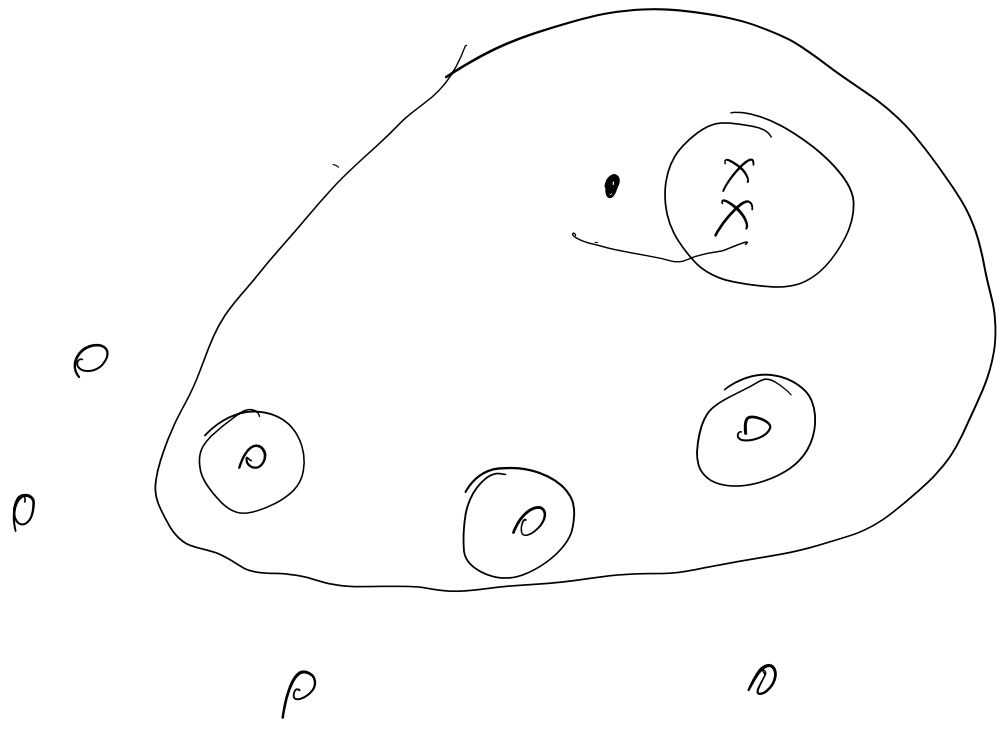
Обучение: Запоминаем X

Предсказание:

u - новый объект

$\rho(u, x_{(1)}) \le \rho(u, x_{(2)}) \dots \le \rho(u, x_\ell)$

$$a(u) = \underset{y \in Y}{\operatorname{argmax}} \sum_{k=1}^{K} w(k, u, x_{(k)}) [y_{(k)} = y]$$



$$w(k, u, x_{(k)}) = K\left(\frac{\rho(u, x_{(k)})}{h}\right)$$

$$\frac{1}{\rho(u, x_{(k)})}$$

регрессией:

$$a(u) = \frac{\sum_{k=1}^{k} K\left(\frac{\rho(u, x_{(k)})}{h}\right) y_{(k)}}{\sum_{k=1}^{k} K\left(\frac{\rho(u, x_k)}{h}\right)}$$

Зачем?

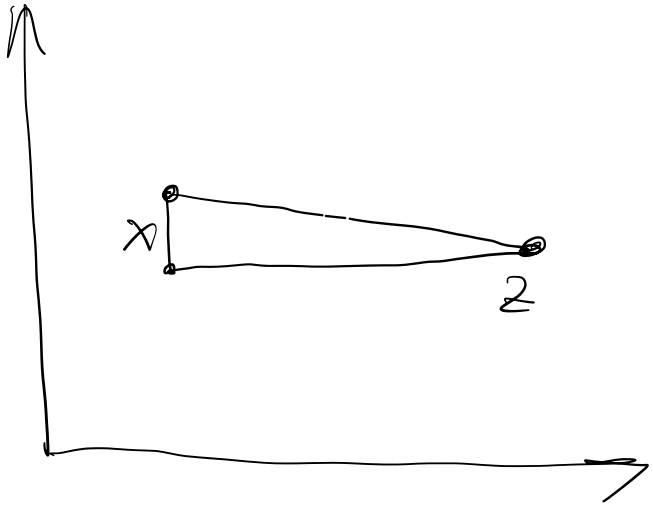
1) Легко задать расстояние, но сложно придумать признаки

2) если объектов одного класса мало

Метрика Минковского

$$\rho(x, z) = \left(\sum_{j=1}^{d} |x_j - z_j|^p \right)^{1/p}$$

1) $p = 2$ Евклидова метрика



2) $p = 1$
Манхэттенское расстояние

3) $p = \infty$ $\rho_\infty(x, z) = \max\limits_{j=1,2,\dots d} |x_j - z_j|$

метрика Чебышева

x x x x x x x

5) Косинусная метрика

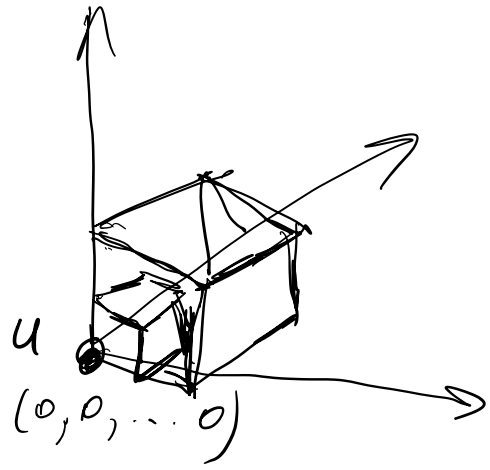
$$\langle x, z \rangle = \|x\| \cdot \|z\| \cdot \cos\theta$$

$$\rho_{\cos}(x, y) = \arccos\left(\frac{\langle x, y \rangle}{\|x\| \|y\|}\right)$$

6) Расстояние Джаккара

$$\rho(A, B) = 1 - \frac{|A \cap B|}{|A \cup B|}$$

Пространства размерности



$[0,1]^d$

5000 объектов

5 объектов

$[0,\varepsilon]^d \subset [0,1]^d \qquad \varepsilon \in (0,1)$

$\delta = \varepsilon^d$ найти δ

при котором в кубу $[0,\varepsilon]^d$ попадет хотя бы 5 объектов с вер-тью больше 0.95.

$$\min \delta : \sum_{k=5}^{5000} C_{5000}^{k} \delta^{k} (1-\delta)^{5000-k} \geq 0.95$$

$\delta = 0.0018$

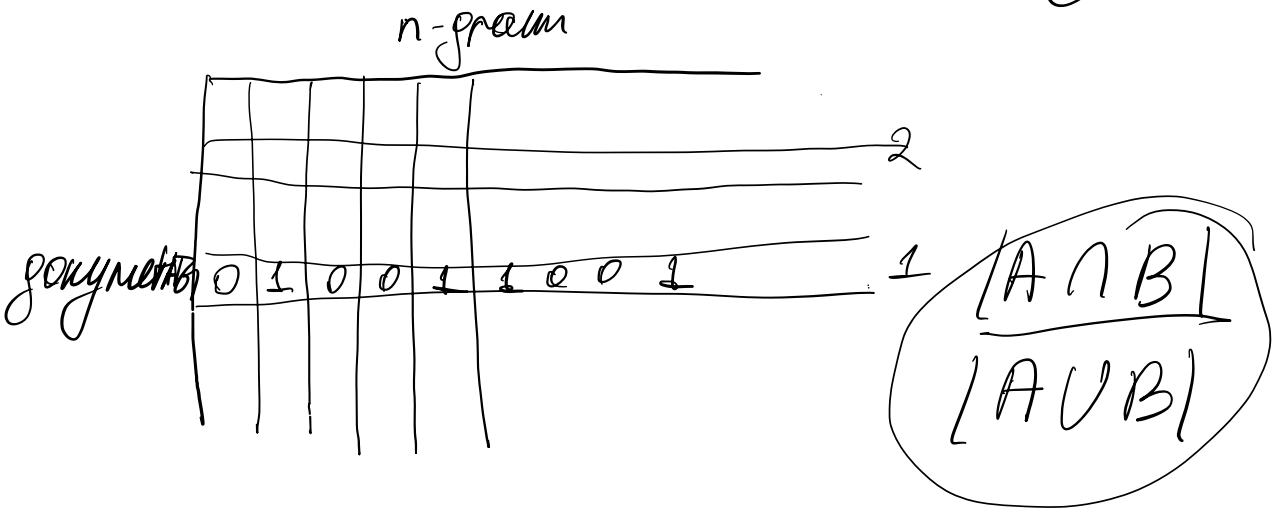
$d = 10, \quad \varepsilon = 0.53$

$d = 100 \quad \varepsilon = 0.94$

LSM

Min Hash LSH

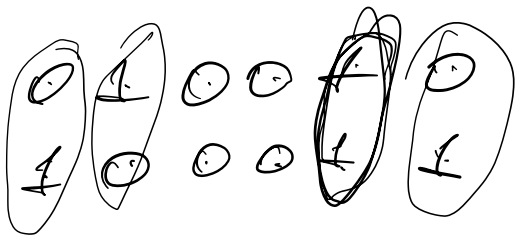
Задача: Поиск дублей документов

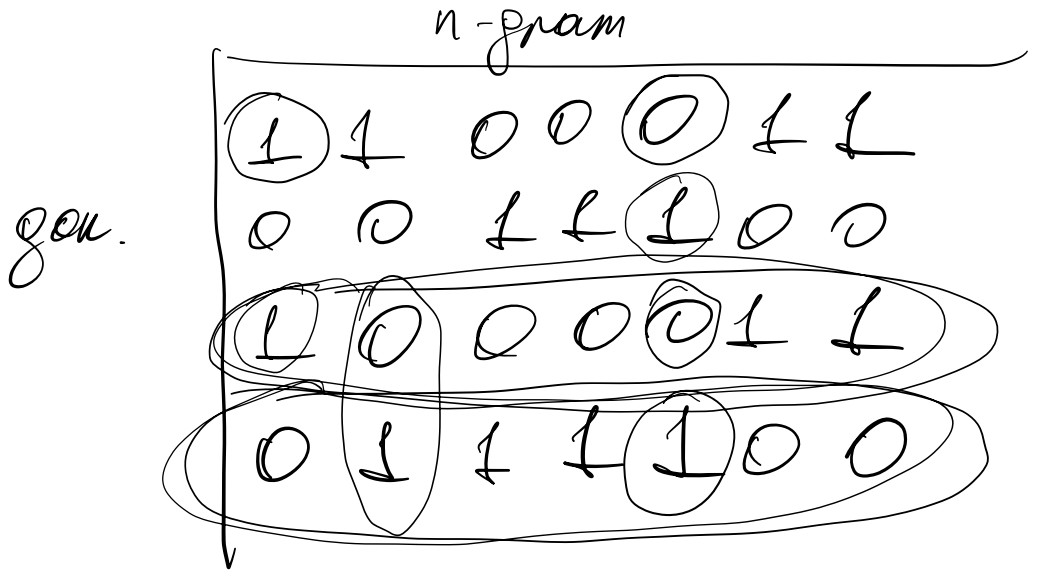


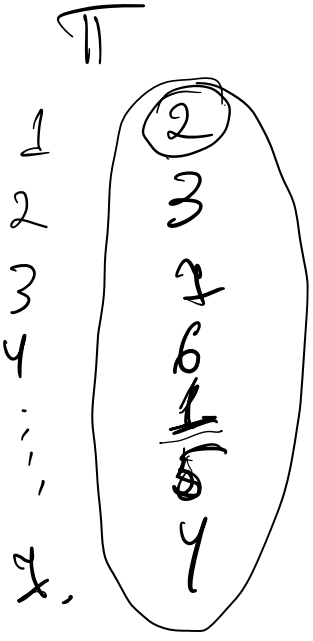
Bag of Words

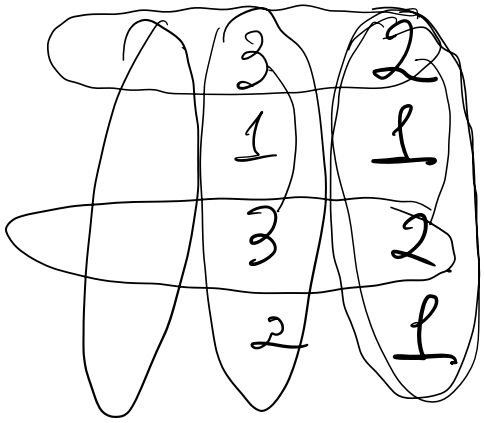
мама мыла раму

3-граммы мам, ама, ма␣, а␣м, ...









← Min hash док 1
для перестановки П.

Коэф Джак. $\frac{|A \cap B|}{|A \cup B|}$.

$\frac{A}{A+B+C.}$

A	1	1
B	1	0
C	0	1
D	0	0

---

1 - 1.

1 - 0.

0 - 1.

0 - 0

doc 1 0 0 1 1 0 ...

doc 2 0 1 0 1 0 ...

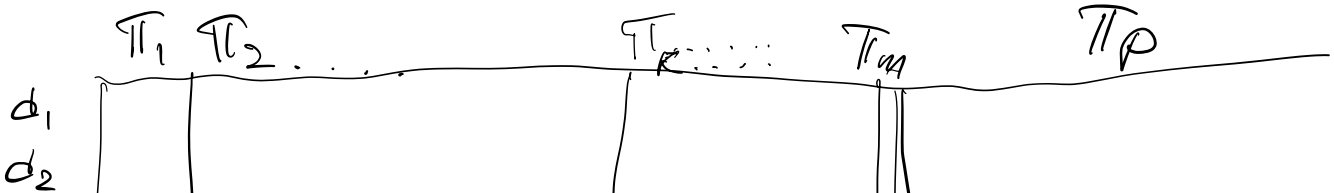
П - перестановка

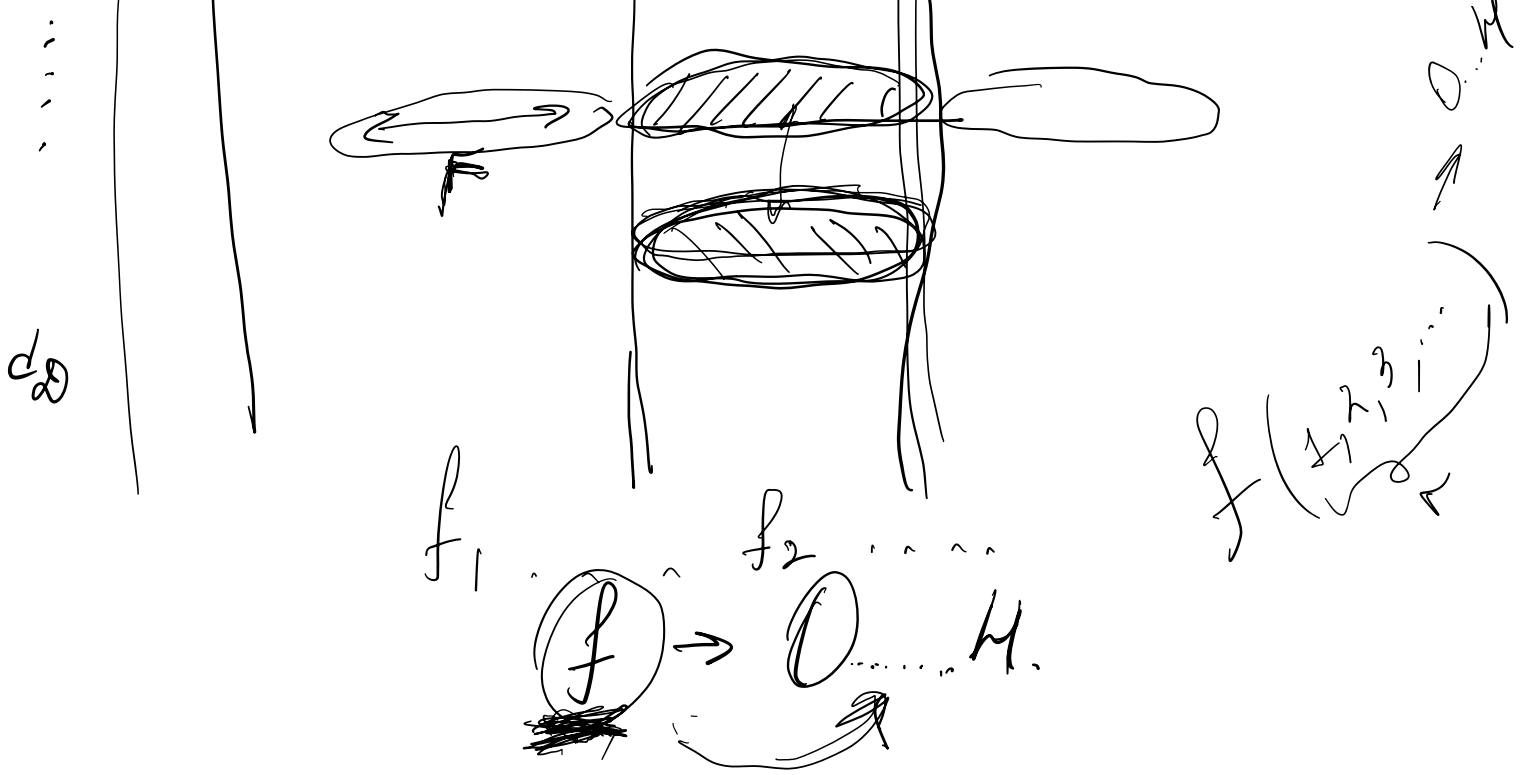
$\frac{A}{A+B+C.}$

0	1	0	1	←
0	1	1	0	←

1/3.

LSH




F
d_0
f_1
f_2
f
0
M