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

$$\rho : X \times X \to [0; \infty)$$

$u \in X$

$$\rho(u, x_u^{(1)}) \le \ldots \le \rho(u, x_u^{(l)})$$

$x_u^{(i)}$ – i-й ближайший сосед

$$a(u, X, k) = \underset{y \in Y}{\mathrm{argmax}} \sum_{i=1}^{k} w_i \, [y_u^{i} = y]$$

$k = 3$

$$\rho(u, x_u^{(1)}) = 1, \quad \rho(u, x_u^{(2)}) = 2 \quad \rho(u, x_u^{(3)}) = 100$$

$$w_i = K(\rho(u, x_u^{(i)}))$$

монотонно убывает

$$a(u, X, k) = \frac{\sum_{i=1}^{k} w_i y_u^{(i)}}{\sum_{i=1}^{k} w_i}$$

Проблемы

$$\rho(x, z) = \left(\sum_{j=1}^{d} (x_j - z_j)^2 \right)^{1/2}$$

Шумовые признаки

x_1	x_2	y
0,1	ξ_1	1
0,5	ξ_2	2

$\xi_i \sim U[0;1]$

$u = (0,0)$

$$P\left(\sqrt{(0{,}5-0)^2+(\xi_2-0)^2} \le \sqrt{(0{,}1-0)^2+(\xi_1-0)^2}\right) = ?$$

$$P\left(0{,}5^2+\xi_2^2 \le 0{,}1^2+\xi_1^2\right) = P\left(\underbrace{\xi_1^2}_{} \ge \underbrace{0{,}24+\xi_2^2}_{}\right) =$$

$$= \int_0^{\sqrt{0{,}76}} \int_{\sqrt{z_2^2+0{,}24}}^{1} dz_1\, dz_2 = \int_0^{\sqrt{0{,}76}} \left(1-\sqrt{z_2^2+0{,}24}\right) dz_2 =$$

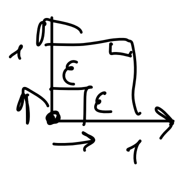
$$= 0{,}275$$

Проклятие размерности

d-мерный куб $[0,1]^d$

5000 объектов

$u = (0,0)$



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

$\delta = \varepsilon^d$, при котором в подкуб попадёт хотя бы 5 объектов, с вер-ю 0,95

$$\min\left\{\delta \,\middle|\, \sum_{k=5}^{5000} C_{5000}^{k}\, \delta^k (1-\delta)^{5000-k} \ge 0{,}95\right\} \approx 0{,}0018$$

Чтобы найти 5 соседей, нужно отступить

$0{,}0018 = \varepsilon^{d} \Rightarrow \varepsilon = 0{,}0018^{1/d}$

$d = 10 \Rightarrow \varepsilon = 0{,}53$, $d = 100 \Rightarrow \varepsilon = 0{,}94$

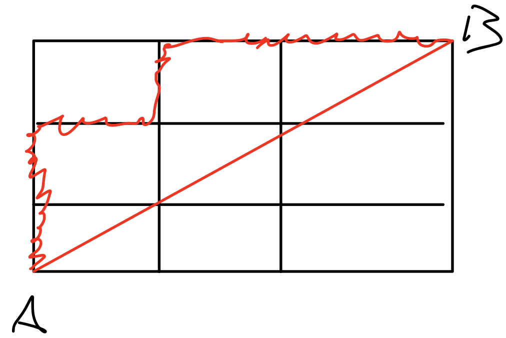
Ф-ции расстояния

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

$p(0;1)$ – не метрика,

Евклидова: $p = 2$

Манхэттенское: $p = 1$



Чебышёв $p = \infty$

$$\rho_\infty(x, z) = \max_{j=1,d} |x_j - z_j|$$

„Считающее" расстояние $p = 0$

$$\rho_0(x, z) = \sum_{j=1}^{d} [x_j \neq z_j]$$

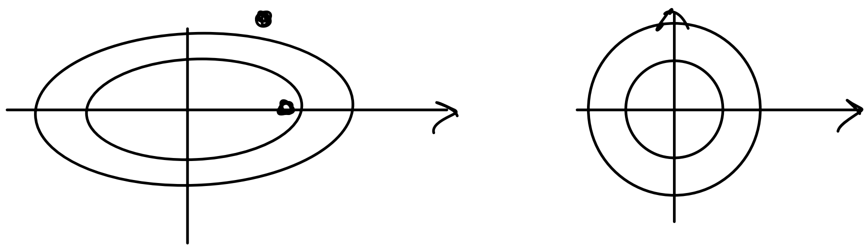
$$\rho_p(x, z, w) = \left(\sum_{j=1}^{d} w_j |x_j - z_j|^p \right)^{1/p}$$

$w_j \geq 0$

$$f(x) = \rho_2(x, 0, w)$$

$$f^2(x) = \sum_{j=1}^{d} w_j x_j^2 , \quad x_j = \frac{x'_j}{\sqrt{w_j}}$$

$$f^2(x') = \sum_{j=1}^{d} x'^2_j = c$$



$$w_j = \left| \frac{\sum_i x_{ij} y_i}{\left(\sum_i x_{ij}^2\right)^{1/2} \left(\sum_i y_i\right)^{1/2}} \right|$$

Расст Махаланобиса

$$\rho(x, z) = \sqrt{(x-z)^T S^{-1} (x-z)} \qquad f = \rho(x, 0)$$

$S = S^T$ S - положительно определена

$$-1 = T$$

$$S = Q^{-1} \Lambda Q = Q^T \Lambda Q \Rightarrow \Lambda = Q^T S Q$$

ортонормированная матрица

$$f^2(x) = \rho^2(x, 0) = x^T S^{-1} x \; (=)$$

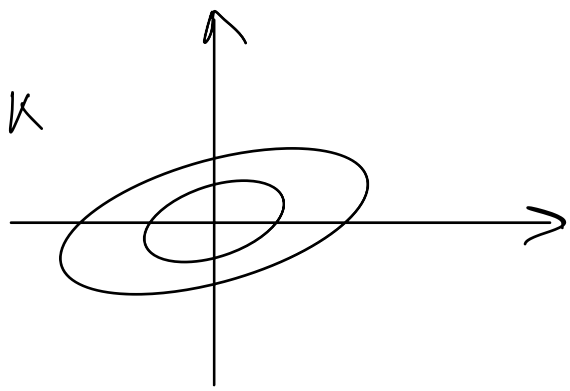
$$x' = Q^T x$$

$$\Rightarrow \; x'^T Q^T S^{-1} Q x' = x'^T (Q^{-1} S Q)^{-1} x' = \underset{1\times d}{x'^T} \, \underset{d\times d}{\Lambda^{-1}} \, \underset{d\times 1}{x'}$$

$$= \sum_{j=1}^{d} \frac{x'^2}{\lambda_j}$$

$$\hat{S} = \frac{1}{n-1} X^T X \; - \text{МНК}$$

$$\frac{1}{n} X^T X \rightarrow ML$$



Косинусная мера.

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

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

d - размер словаря

$x_i \in R^d$ $\quad x_{ij} = 1$, если j-е слово встречается в x_i

Рассм. Жаккара

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

$$\mathcal{U} = \{u_1, \dots, u_N\}$$

$d \in R^N$, $d_i = 1$, если $u_i \in$

$$|X \wedge Y| = \sum_{j=1}^{N} x_j y_j = \langle X, Y \rangle$$

$$|X \cup Y| = \sum_{j=1}^{N} x_j^2 + \sum_{j=1}^{N} y_j^2 - \sum_{j=1}^{N} x_j y_j =$$

$$= \|X\|^2 + \|Y\|^2 - \langle X, Y \rangle$$

$$\rho_j(X, Y) = 1 - \frac{\langle X, Y \rangle}{\|X\|^2 + \|Y\|^2 - \langle X, Y \rangle}$$

1) Считать не на всех признаках
2) LSH, хэш-функции