LOCALITY SENSITIVE HASHING FUNCTION FAMILY

Klasyczne haszowanie: $h \leftarrow MD5$

$\boxed{D}$ (dokument) $\xrightarrow{h} h(D) \rightarrow \boxed{\begin{array}{c} D \\ \hline h(D) \end{array}}$ podpis cyfrowy

$D' \leftarrow D +$ drobna zmiana $\Rightarrow h(D')$ bardzo mocno się różni od $h(D)$

$$d_H(h(D), h(D')) \approx \frac{1}{2} L$$

$$h : \Sigma^* \rightarrow \{0,1\}^L$$

CEL:

$D \approx D' \rightarrow h(D) \approx h(D')$

podobne

Przestrzenie metryczne: (X, d)

1. $d : X^2 \to [0, \infty)$
2. $d(x,y) = d(y,x)$
3. $d(x,y) = 0 \iff x = y$
4. $d(x,z) \leq d(x,y) + d(y,z)$

Uwagi:

- zał. że $\neg(3)$: są x, y, $x \neq y$ i $d(x,y) = 0$

  na X def. rel. równow.:

  $$x \approx y \equiv d(x,y) = 0$$

  na $X/\approx$ def:

  $$\tilde{d}([x],[y]) = d(x,y) \qquad (X/\approx, \tilde{d}) \leftarrow \text{prz. metr}$$

• zał. że ¬(2) ¬$(d(x,y) = d(y,x))$ dla pewnych x, y

$$\tilde{d}(x,y) = d(x,y) + d(y,x)$$

P1 $X = \mathbb{R}^n$; $p \geq 1$

$$\|\bar{x}\|_p = \sqrt[p]{\sum_{i=1}^{n} |x_i|^p}$$ ← norma L_p

$$d_p(\bar{x}, \bar{y}) = \|x - y\|_p$$

$$\bar{B}(x, r) = \{y : d(x,y) \leq r\}$$

$$\bar{B}_p(\bar{0}, 1) = \{y : \|y\|_p \leq 1\}$$

B_1  taksówkowa

B_2 

B_5 

B_∞ 

$$\|\bar{x}\|_\infty = \max\{|x_i| : i = 1..n\}$$

(P2) $X = \{0,1\}^n \rightsquigarrow X = P(\{1, \dots n\})$

$\bar{x} \in \{0,1\}^n \longrightarrow A_x = \{i : \bar{x}(i) = 1\}$

odl. Hamminga: $h(x,y) = \sum_{i=1}^{n} |x_i - y_i| \quad (= d_1(\bar{x}, \bar{y}))$

$= \sum_{i=1}^{n} [\![x_i \neq y_i]\!] = |\{i : x_i \neq y_i\}| = |A_x \triangle A_y|$

$(P(X), d_H) : \quad d_H(A, B) = |A \triangle B|$

$\begin{cases} P_{fin}(X) = \{D \subseteq X : |D| < \infty\} \\ d(A, B) = |A \triangle B| . \end{cases}$

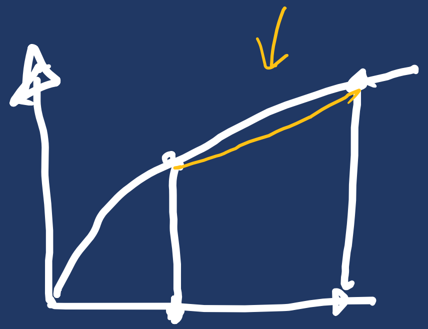
(PP) $X = \Sigma^*$

$D \subseteq X \leftarrow$ dokumenty

CEL: przekszt. metrykę d w taką metrykę d' aby
$$0 \leq d'(x,y) \leq 1$$

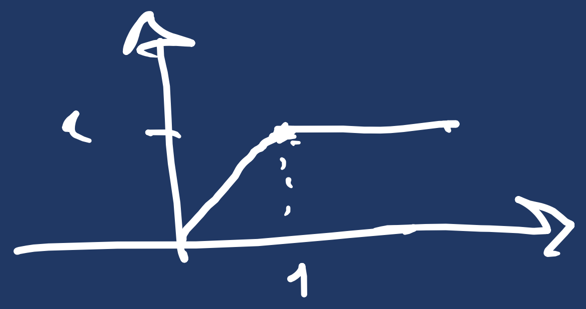
FAKT. Zał., że (X, d) jest p. metr. Niech $f: [0,\infty) \to [0,\infty)$ będzie taka, że

- $f(0) = 0$
- f jest rosnąca
- f jest wklęsła

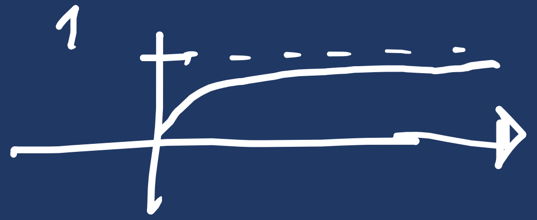


Niech $\tilde{d} = f \circ d$ (czyli $\tilde{d}(x,y) = f(d(x,y))$)

Wtedy $(X, \tilde{d})$ jest pr. metryczną.

P $\quad f(x) = \min\{x, 1\}$ 

$\tilde{d}(x,y) = \min\{\|x-y\|_2, 1\}$

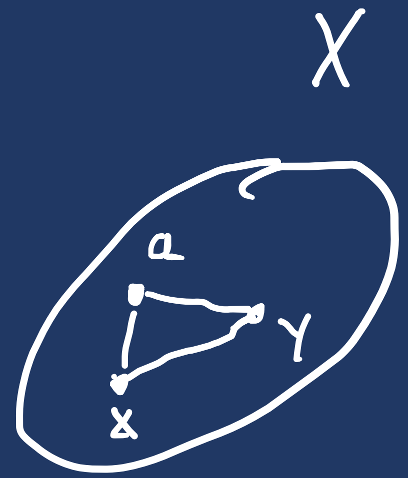
P $\quad f(x) = \dfrac{x}{1+x}$ 

$$f'(x) = \frac{(1+x)-x}{(1+x)^2} = \frac{1}{(1+x)^2} \geq 0$$

Z $\quad f''(x) < 0 \leftarrow$ wklęsłość

Tw (Steinhaus) Niech (X,d) będzie p. metr. Niech $a \in X$. Niech

$$\rho(x,y) = \frac{2\cdot d(x,y)}{d(x,a)+d(y,a)+d(x,y)}$$

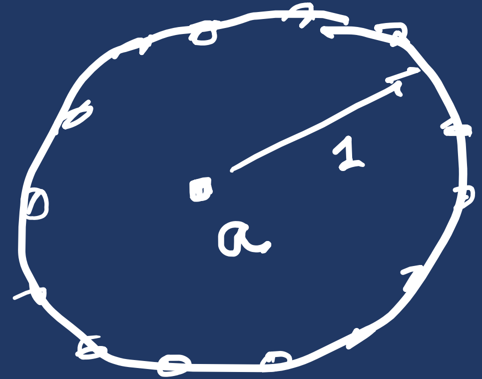


Wtedy ρ jest metryką na X.

U1. $d(x,a)+d(a,y)+d(x,y) \geq d(x,y)+d(x,y)$

$$\rho(x,y) \leq \frac{2d(x,y)}{2d(x,y)} = 1.$$

U2. $$d(x,a) = \frac{2d(x,a)}{d(x,a)+\underbrace{d(a,a)}_{0}+d(x,a)} = 1$$



ZADANIE

(P) $A, B \subseteq X$, X-skończ.

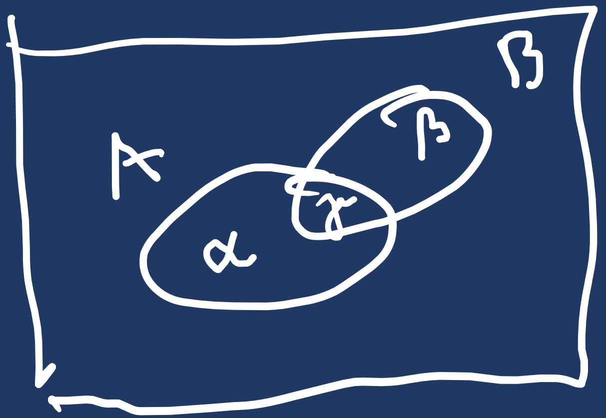
Tw. Steinh

$a = \phi$

$$d(A,B) = |A \triangle B|$$

$$\rho(A,B) = \frac{2|A \triangle B|}{|A \triangle \phi| + |B \triangle \phi| + |A \triangle B|}$$

$$= \frac{2|A \triangle B|}{|A| + |B| + |A \triangle B|} = \frac{2(\alpha+\beta)}{(\alpha+\gamma) + (\beta+\gamma) + (\alpha+\beta)}$$



$$= \frac{2(\alpha+\beta)}{2(\alpha+\beta+\gamma)} = \frac{\alpha+\beta}{\alpha+\beta+\gamma}$$

$$= \frac{|A \triangle B|}{|A \cup B|}$$

$$d_J(A,B) = \frac{|A \div B|}{|A \cup B|}$$

odległość Jaccarda

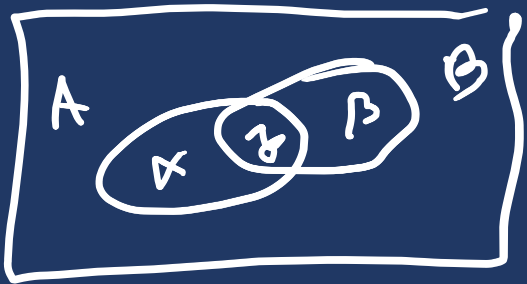
- $0 \leq d_J \leq 1$
- $d_J(A,B) = 1 \ : \ A \cap B = \emptyset$
-

Podobnost:

$$s : X \times X \longrightarrow [0,1]$$

- $s(x,x) = 1$
- $s(x,y) = s(y,x)$

$$s(x,y) = 1 \equiv (x = y)$$



Dosk. podobnost:

$$s(x,y) = 1 - d(x,y)$$

d : metr. t.ž. $0 \leq d \leq 1$

$$J(A,B) = 1 - d_J(A,B) = 1 - \frac{|A \triangle B|}{|A \cup B|} = \frac{|A \cup B| - |A \triangle B|}{|A \cup B|}$$

$$= \frac{\alpha + \beta + \gamma - (\alpha + \beta)}{\alpha + \beta + \gamma} = \frac{\gamma}{\alpha + \beta + \gamma} = \frac{|A \cap B|}{|A \cup B|}$$

Pod. Jaccarda:

$$\boxed{J(A,B) = \frac{|A\cap B|}{|A\cup B|}}$$

$\Omega = \{1, 2, \ldots, n\}$ $\quad A \subseteq \Omega, \; A \neq \emptyset$

$S_n = Sym_n =$ zbiór wszyst. perm. Ω

S_n - pr. probab. z jedn. prawd.

$$P(\{\sigma\}) = \frac{1}{n!}$$

$$h_A(\sigma) = \min\{k : \sigma(k) \in A\}, \quad \sigma \in S_n$$

$$\sigma : \{1..n\} \xrightarrow[\text{na}]{1-1} \{1..n\}$$

Q: $\Pr_\sigma[h_A(\sigma) = h_B(\sigma)] = ?$

$$\Pr[h_A(\sigma) = h_B(\sigma)] = \sum_{k=1}^{n} P[h_A(\sigma) = h_B(\sigma) \mid h_A(\sigma) = k]\, P[h_A(\sigma) = k]$$

$$= \sum_{k=1}^{n} P[h_A(\sigma) = h_B(\sigma) = k \mid h_A(\sigma) = k] \cdot P[h_A(\sigma) = k]$$

spoza $A \cup B$

$$= \sum_k \frac{|A \cap B|}{|A \cup B|} P[h_A(\sigma) = k] =$$

$$= J(A,B) \sum_k P[h_A(\sigma) = k]$$

$$= J(A,B)$$

$$\boxed{\Pr_\sigma\left(h_A(\sigma)=h_B(\sigma)\right) = J(A,B)}$$

$G_{A,B} = \{\sigma \in S_n : h_A(\sigma) = h_B(\sigma)\}$

$P(G_{A,B}) = J(A,B)$

- $\sigma_1, \sigma_2, \dots, \sigma_K$ ← niezal. wybr. permutacje z S_n

$\mathbb{N}^K \ni s(A) = [h_A(\sigma_1), \dots, h_A(\sigma_K)]$ ← szkic A

- A, B : fix

$$S(A,B) = \sum_{i=1}^{K} [\![h_A(\sigma_i) = h_B(\sigma_i)]\!]$$

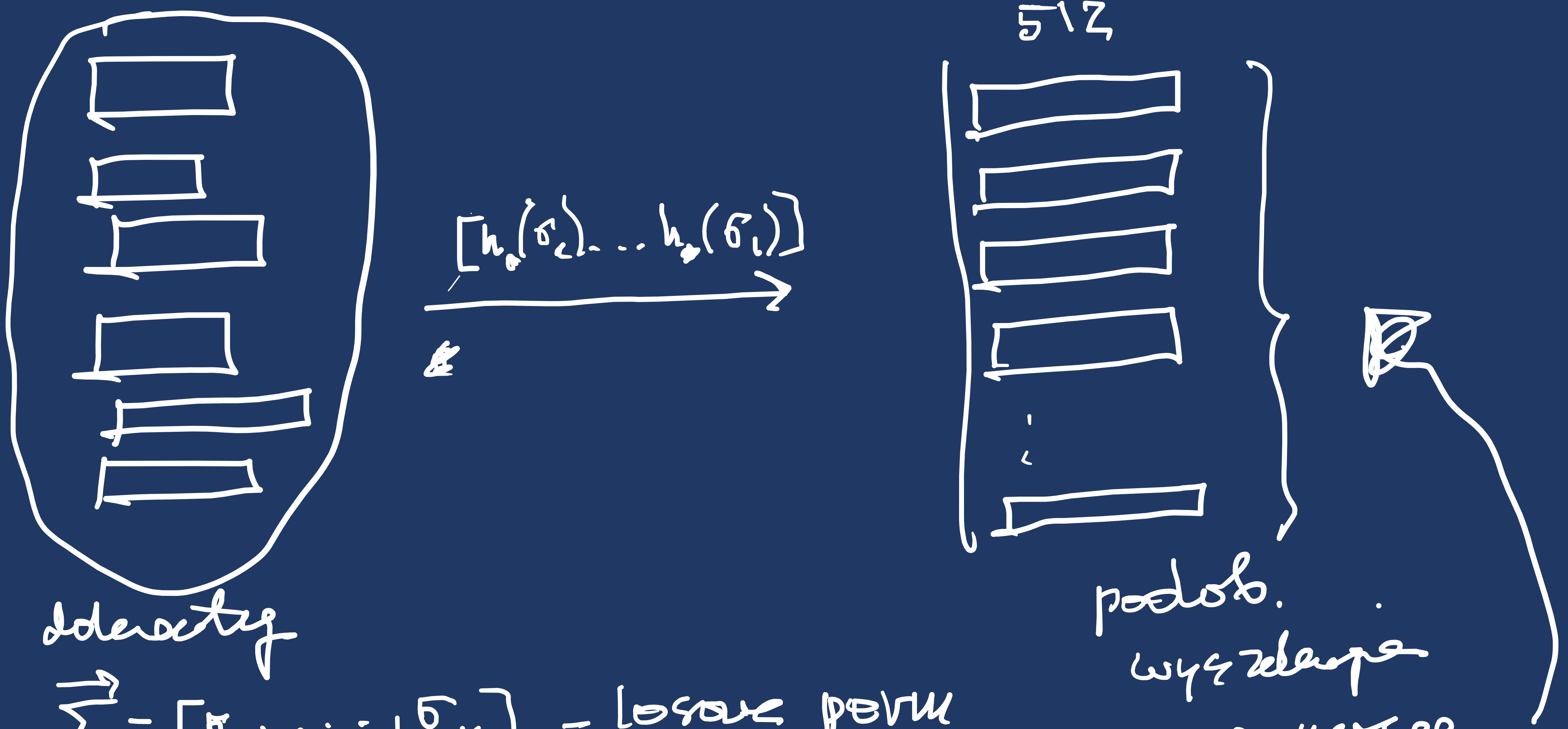
$$E[s(A,B)] = \sum_{l=1}^{k} E\left(\left[\!\left[h_A(\sigma_l) = h_B(\sigma_l) \right]\!\right]\right)$$

$$= \sum_{l=1}^{K} \Pr\left[h_A(\sigma_l) = h_B(\sigma_l)\right] = \sum_{l=1}^{k} s(A,B) =$$

$$= k \cdot s(A,B),$$

$$s(A,B) = \frac{1}{K} \sum_{l=1}^{K} \left[\!\left[h_A(\sigma_l) = h_B(\sigma_l) \right]\!\right]$$

$$= \frac{1}{K} d_H\left(s(A), s(B)\right)$$



dokumenty

$\vec{\Sigma} = [\sigma_1, \ldots, \sigma_K]$ – losowe perm

podob. wyszukiwanie za pomocą szkiców