L4: Distances

Aug 27, 2025

Data Mining

Jeff M. Phillips

$\boxed{\text{Data } X} \subset \mathcal{X} = \mathbb{R}^d$ $\quad X = \{x_1, x_2, \ldots x_n\}$

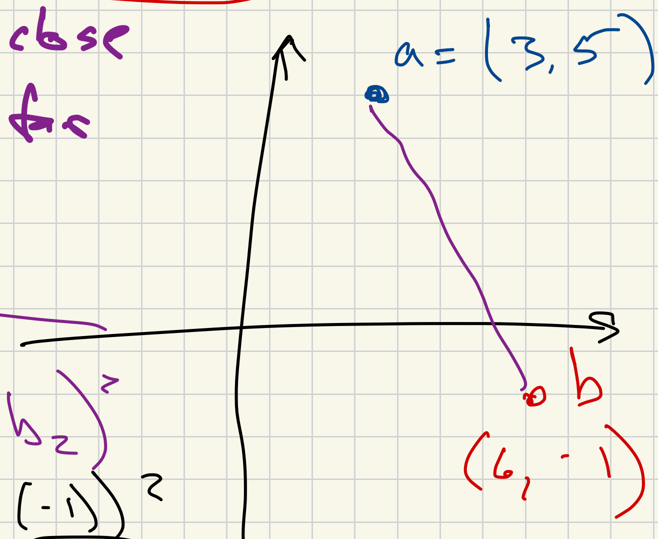
input

$x_i = (x_{i1}, x_{i2}, \ldots x_{id})$

$\boxed{\text{Distance } \; D : \mathcal{X} \times \mathcal{X} \to \mathbb{R}_{\geq 0}}$

D(a,b) small if a, b close

D(a,b) large if a, b far



$$D_{Euc}(a,b) = \sqrt{(a_1 - b_1)^2 + (a_2 - b_2)^2}$$

$$(3-6)^2 + (5-(-1))^2$$

$$3^2 + 6^2 = \sqrt{45}$$

Metric Distances

else $\Rightarrow$ quasimetric

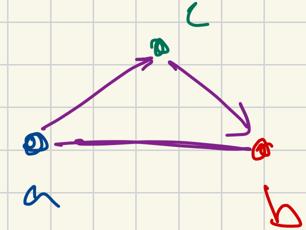
(M1) positivity $D(a,b) \ge 0$

(M2) identity $D(a,b) = 0$ if $\boxed{\text{and only if}}$ $a=b$

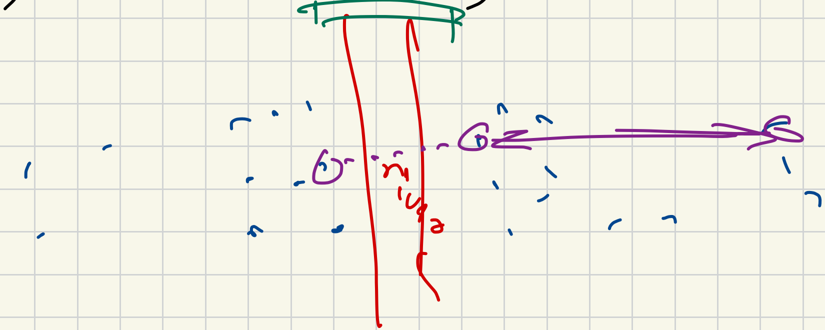
(M3) symmetry $D(a,b) = D(b,a)$

(M4) triangle inequality

$$D(a,b) \le D(a,c) + D(c,b)$$



else
$\Downarrow$
pseudometric



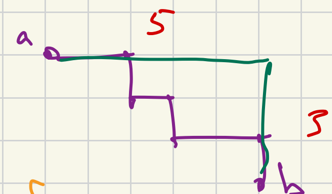
L_p - Distances

$D_p(a,b) \qquad a,b \in \mathbb{R}^d$

$$D_p(a,b) = \left(\sum_{i=1}^{d} |a_i - b_i|^p\right)^{1/p} = \|a-b\|_p$$

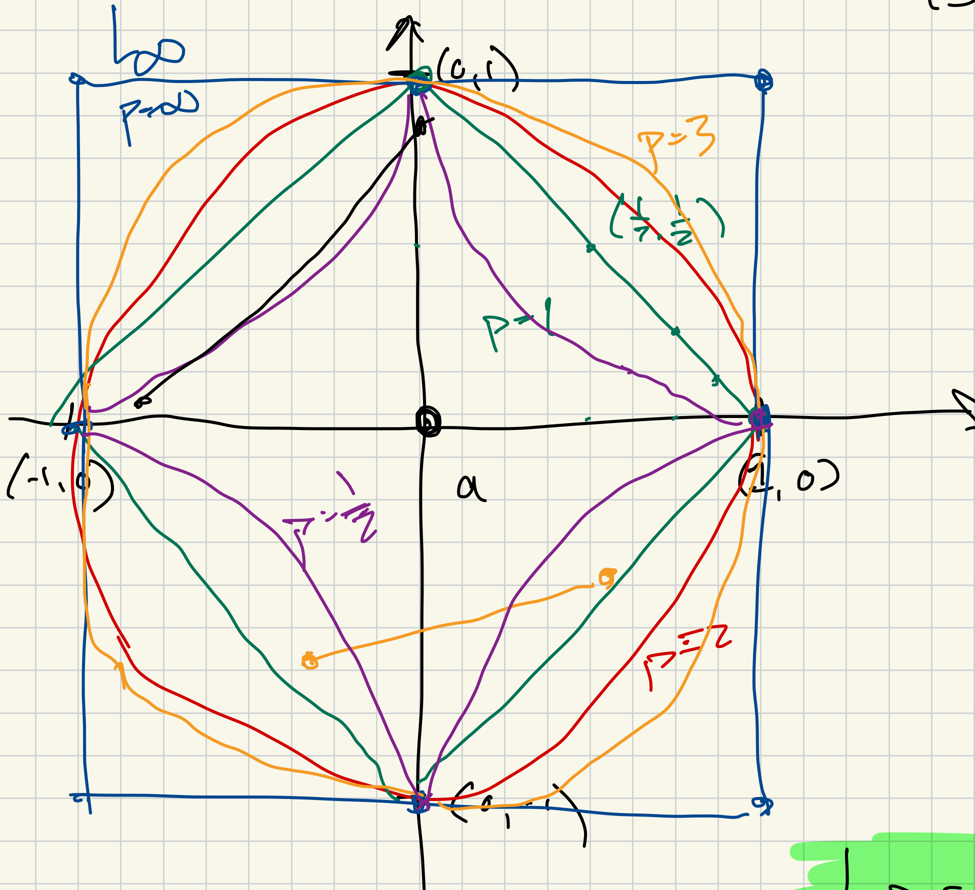
$$D_2(a,b) = \sqrt{\sum_{i=1}^{d} (a_i - b_i)^2} = \|a-b\|_2 = \|a-b\| = D_{Euc}(a,b)$$

$$D_1(a,b) = \sum_{i=1}^{d} |a_i - b_i|$$



$$D_\infty(a,b) = \left(\lim_{p\to\infty}\left(\sum_{i=1}^{d} |a_i - b_i|^p\right)^{1/p}\right) \Longrightarrow \max_{i\in[d]} |a_i - b_i|$$

$$= \|a-b\|_\infty$$



Draw all points $b \in \mathbb{R}^2$

$D_p(a,b) = 1$

L_p-ball $\subset$ $L_{p'}$-ball

subset

if $p \leq p'$

$p = 1/2$?

not convex

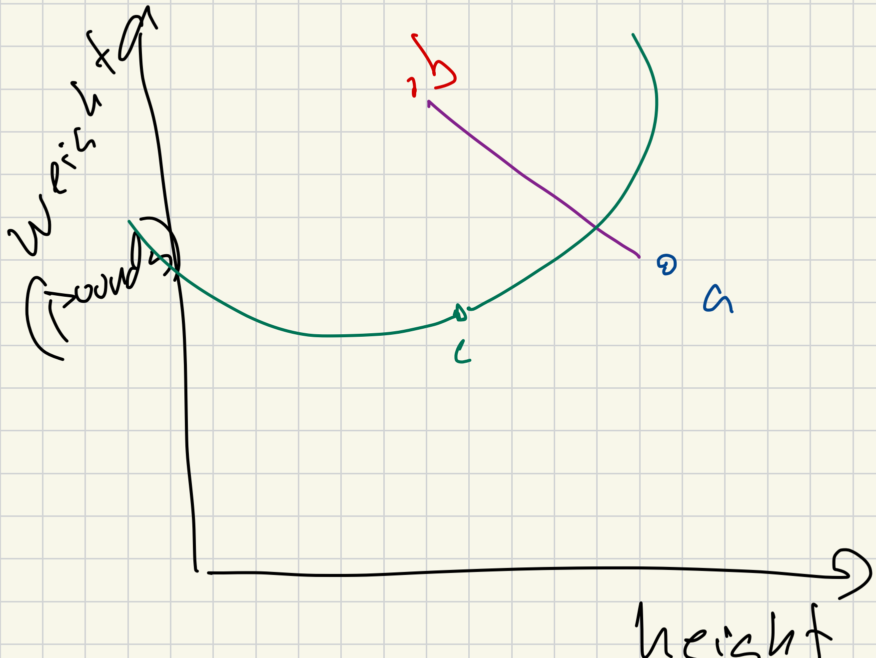
L_p-dist $\Rightarrow$ metric

$p \in [1, \infty) + \infty$


NEW CUYAMA
Population 562
Ft. above sea level 2150
Established 1951
TOTAL 4663
Don't do this!

$$D(a,b) = \sqrt{\sum_i^1 (a_i - b_i)^2 + (a_2 - b_2)^2}$$

Do not do this!



inches → meters

Mahalanobis Distance

$$D_M(a,b) = \sqrt{(a-b)^T M (a-b)}$$

$M \in \mathbb{R}^{d \times d}$

positive definite

$\Rightarrow$ metric

if $M = I = \begin{bmatrix} 1 & & 0 \\ & \ddots & \\ 0 & & 1 \end{bmatrix}$

$$= \sqrt{(a-b)^T (a-b)}$$

$$= \sqrt{\langle a-b, a-b \rangle}$$

$$= \sqrt{\sum_{i=1}^{d} (a_i - b_i)(a_i - b_i)}$$

$$= \sqrt{\sum_{i=1}^{d} (a_i - b_i)^2} = D_{Euc}$$

Cosine Distance

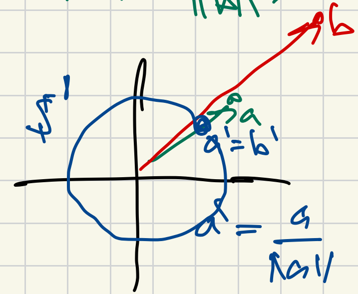
$a, b \in \mathbb{R}^d = \mathcal{X}$

$D_{cos} = 1 - \frac{\langle a,b\rangle}{\|a\|\cdot\|b\|} \in [0,2]$

cosine similarity

$S_{cos}(a,b) = \frac{\langle a,b\rangle}{\|a\|\cdot\|b\|}$

$= \left\langle \frac{a}{\|a\|}, \frac{b}{\|b\|} \right\rangle$



$a' = \frac{a}{\|a\|}$

$b' = \frac{b}{\|b\|}$

Metric

(M1) positivity

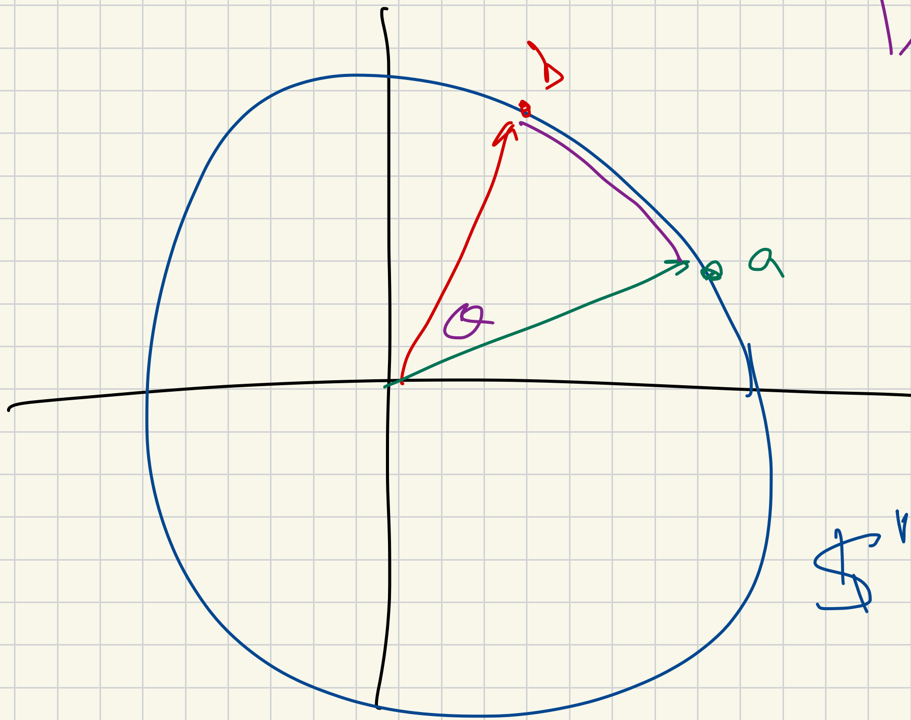
(M2) identity $D(a,b) = 0$ iff $a = b$

but it can if $\mathcal{X} = S^{d-1}$

(M3) symmetry

(M4) triangle. NO. $\Rightarrow D_{ang}(a,b) = \cos^{-1}(\langle a,b\rangle)$

$= \arccos(\langle a,b\rangle)$



$D_{ang}(a, b)$

$= \arccos(\langle a, b \rangle)$

$=$ angle between $\vec{a}, \vec{b}$ in radians

$\in [0, \pi]$