L6: Distances

Jeff M. Phillips

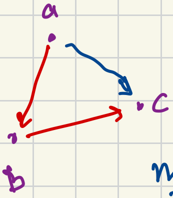
January 27, 2020

Distances

bivariate function

$$D : \underline{X} \times \underline{X} \to \mathbb{R}_{\geq} = [0, \infty)$$

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



$$\mathbb{R}^d = \{ x = (x_1, x_2, \dots, x_d) \}$$

metric $D(a,b)$ e.g. $a, b \in \mathbb{R}^d$

[M1]: $D(a,b) \geq 0$ non-negativity

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

[M3]: $D(a,b) = D(b,a)$ symmetry

[M4]: $D(a,b) + D(b,c) \geq D(a,c)$ triangle inequality

M_1, M_3, M_4 psuedo metric

$D(a,b) = 0$ and $a \neq b$.

Angular distance

M_1, M_2, M_4 quasimetric

$D(a,b) \neq D(b,a)$

Road Network.

drive time.

Sam I am $\rightarrow$ {[Sam], [I], [am]}

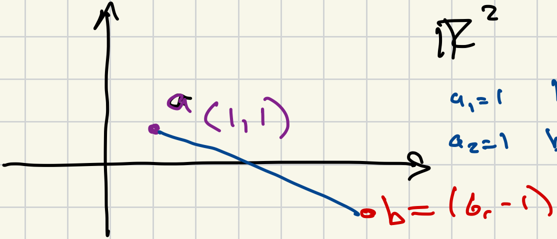
I am Sam $\rightarrow$

Distance between vectors in $\mathbb{R}^d$

L_p distances $\quad p \in [0, \infty], \infty$

$$L_p(a,b) = D_p(a,b) = \left(\sum_{j=1}^{d} (|a_j - b_j|)^p \right)^{1/p}$$

$$L_2(a,b) = \sqrt{\sum_{j=1}^{d} (a_j - b_j)^2} = \|a - b\|_2 = \|a-b\|$$



$\mathbb{R}^2$

$a_1 = 1 \quad b_1 = 6 \quad 1-6 = -5$

$a_2 = 1 \quad b_2 = -1 \quad 1--1 = 2$

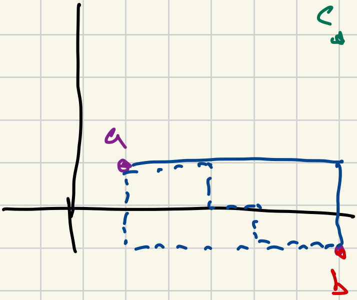
$\sqrt{(-5)^2 + (2)^2}$

$\sqrt{25+4} = \sqrt{29}$

$$L_1(a,b) = \sum_{j=1}^{d} |a_j - b_j|$$

Manhattan distance

SLC distance



$$L_\infty(a,b) = \max_{j \in [d]} |a_j - b_j|$$

$$L_0(a,b) = \sum_{j=1}^{d} \mathbb{1}(a_j = b_j)$$ ← counts number of same.

Hamming

$L_0(a,c) = 0$

$L_0(a,b) = 0$

$L_0(c,b) = 1$

All L_p distances are metrics

on $\mathbb{R}^d$ for $p \in [1, \infty), \infty$

Hamming

L_0 on $\{0,1\}^d = \mathcal{X}$

is a metric.

Lp Distances and Units

For $a = (a_1, a_2, \ldots, a_d)$ and $b = (b_1, b_2, \ldots, b_d) \in \mathbb{R}^d$,

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

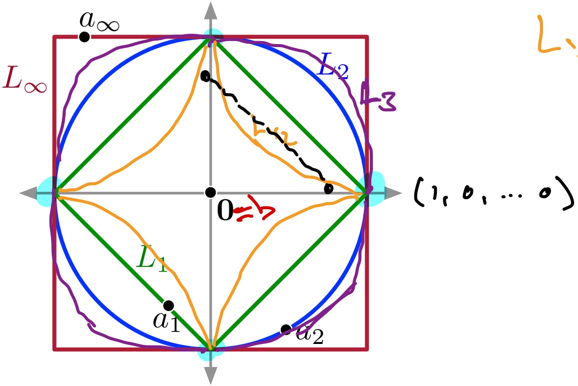
Do not do this in data science.

Lp Distances and Unit Balls

For $a = (a_1, a_2, \ldots, a_d)$ and $b = (b_1, b_2, \ldots, b_d) \in \mathbb{R}^d$,

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

Let $\cancel{a}\, b = (0, 0, \ldots, 0)$ and $\|a - b\|_p = 1$.



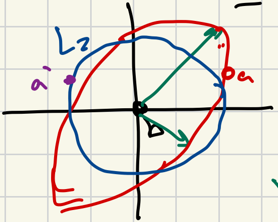
$L_{1/2}$

Mahalanobis Distance w.r.t. matrix $M \in \mathbb{R}^{d \times d}$

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

if $M = I$ identity matrix

$$D_I(a,b) = \sqrt{(a-b)^T I (a-b)} = \sqrt{\langle (a-b), (a-b) \rangle}$$
$$= \sqrt{\|(a-b)\|^2} = \|a-b\| = L_2(a,b)$$



$D_M(a,b) = 1$

metric if M p.d.

Jaccard Distance

Jaccard Similarity

$$D_J(S,T) = 1 - JS(S,T)$$

S, T are sets

do not need to define universe

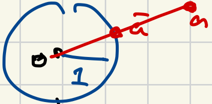
$S, T \in [U]$

$D_J(S,T) \in [0,1]$

D_J is a metric.

Most set distances $D = 1 - S$ are metrics, but not Sorensen-Dice

Cosine Distance



$\hookrightarrow$ Angular Distance

$$D_{cos}(a,b) = 1 - \frac{\langle a,b \rangle}{\|a\| \cdot \|b\|} = 1 - \frac{\sum_{j=1}^{d} a_j b_j}{\|a\| \cdot \|b\|}$$

$a \to \bar{a} = \frac{a}{\|a\|}$ ensures that $\bar{a} \in \mathbb{S}^{d-1}$

$\|\bar{a}\| = 1$

$$D_{cos}(a,b) = 1 - \langle \bar{a}, \bar{b} \rangle$$

$\mathbb{R}^d$

$\in [-1, 1]$

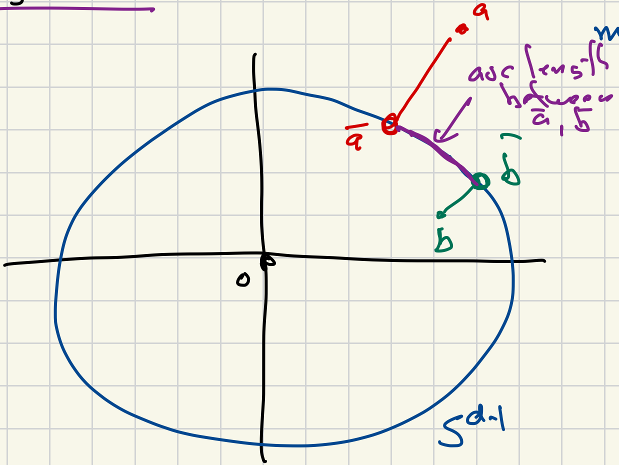
so not metric

Angular Distance

defined on $\mathbb{R}^d / \{0\}$

$$D_{ang}(a,b) = \arccos(\langle \bar{a}, \bar{b} \rangle) \in [0, \pi]$$

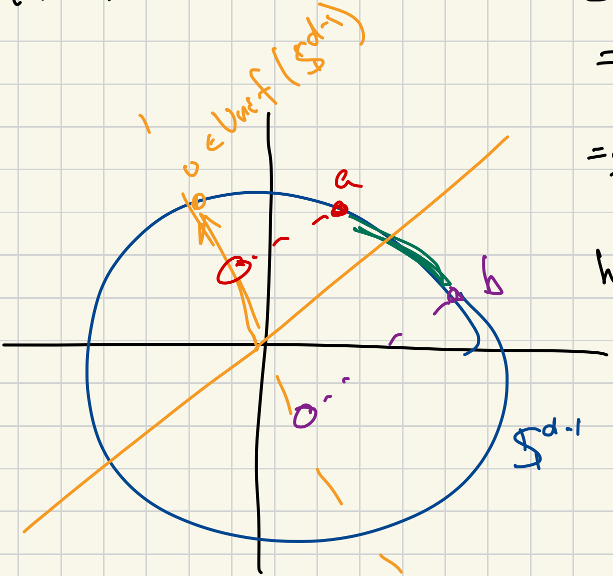
metric in S^{d-1}



LS Hash function for $S_{ang}(a,b)$

$$= \frac{1}{\pi}\left(1 - D_{ang}(a,b)\right)$$

$$= \frac{1}{\pi}\left(1 - \arccos(\bar{a},\bar{b})\right)$$



$h_u(a) = +1$ if $\langle a, u \rangle > 0$

-1 if $\langle a, u \rangle \leq 0$