L6: Distances

Jeff M. Phillips

January 27, 2020

Distance : bivariate function

$$d : \mathcal{X} \times \mathcal{X} \to \mathbb{R}^+ \text{ or } \mathbb{R}_{>0}$$

$$d(a,b) =$$

metric

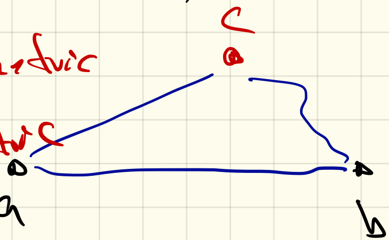
(M1) $d(a,b) \geq 0$ (non-negative)

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

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

(M4) $d(a,b) \leq d(a,c) + d(c,b)$ (triangle inequality)

- M1, M3, M4 pseudometric
- M1, M2, M4 quasimetric



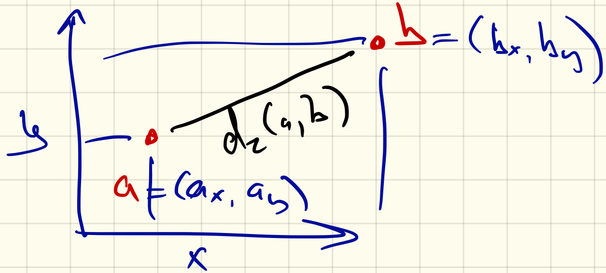
L_p Distances $\quad \mathcal{X} := \mathbb{R}^d$

$a, b \in \mathbb{R}^d \qquad a = (a_1, a_2, \dots, a_d)$

$$L_2(a,b) = d_2(a,b) = \|a-b\|_2 = \|a-b\|$$

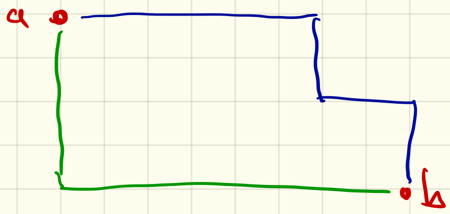
Euclidean

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



$$L_1(a,b) = d_1(a,b) = \|a-b\|_1$$

$$= \sum_{i=1}^{d} |a_i - b_i|$$



Manhattan dist.

SLC dist

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

Every L_p dist for $p \in [1, \infty)$

is a metric.

$$L_0 = \|a-b\|_0 = d - \sum_{i=1}^{d} \mathbb{1}(a_i = b_i)$$

if $a,b \in \{0,1\}^d$ bit strings

↳ Hamming dist

$$L_\infty = \|a-b\|_\infty = \lim_{p \to \infty} L_p(a,b)$$

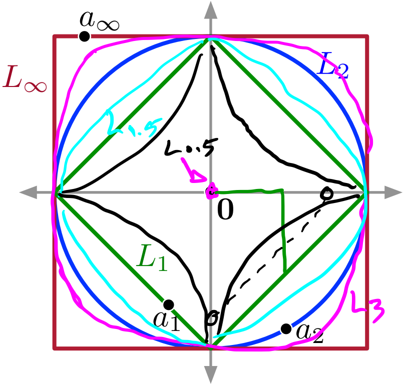
$$= \max_{i \in [1..d]} |a_i - b_i|$$

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



$L_{0.5}$

Lp Distances and Units

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

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

Rule
All coords must have same units.

non sense

Mahalanobis Dist $\quad a, b \in \mathbb{R}^d$

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

$M \in \mathbb{R}^{d \times d}$ $\qquad$ M p.d $\Rightarrow d_M$ metric

If $M = I$ (identity matrix)

$$d_I(a,b) = \|a-b\|_2$$

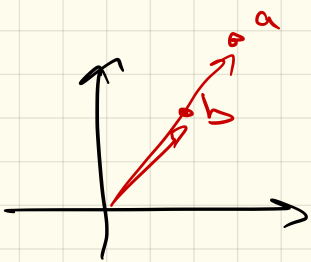
$$(a-b)^T(a-b) = \langle a-b, a-b \rangle = \|a-b\|^2$$

$$M = diag(m_1, m_2, \ldots m_d) = \begin{bmatrix} m_1 & & 0 \\ & m_2 & \\ 0 & & \ddots \\ & & & m_d \end{bmatrix}$$

Cosine dist

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

$$d_{cos}(a,b) = 1 - \underbrace{\frac{\langle a,b\rangle}{\|a\|\cdot\|b\|}}_{\le 1} = 1 - \frac{\sum_{i=1}^{d} a_i b_i}{\|a\|\cdot\|b\|}$$



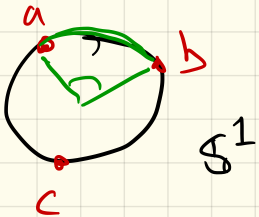
$$d_{cos}(a,b) = 0$$

$$a \neq b$$

only measure direction

$$a \to \bar{a} = \frac{a}{\|a\|} \qquad b \to \bar{b} = \frac{b}{\|b\|}$$

$$a, b \in S^{d-1} = \{x \in \mathbb{R}^d \mid \|x\| = 1\}$$

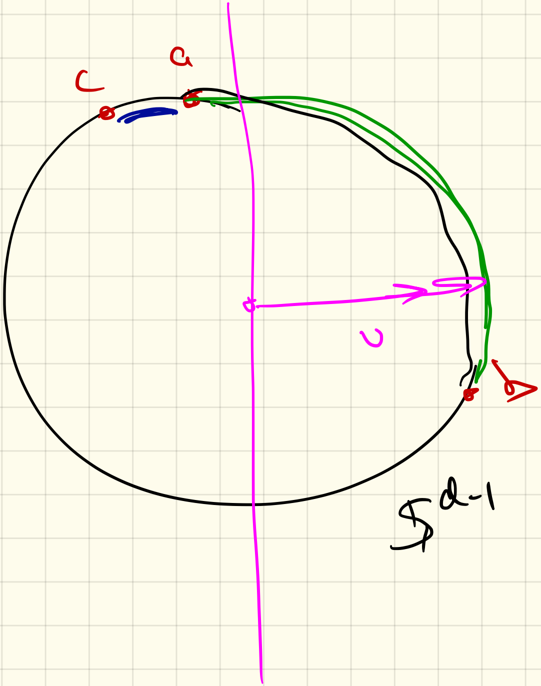


metric

↳ map to angular dist

$a, b \in S^{d-1}$ $\quad d_{ang}(a,b) = \cos^{-1}(\langle a,b\rangle) = \arccos(\langle a,b\rangle)$

Angular dist LSH



pick random unit vector $u \in S^{d-1}$

$$h_u(a) = \mathrm{sign}(\langle a, u \rangle)$$
$$= -1$$

$$h_u(b) = +1$$

LSH

$$\frac{d_{ang}(a,b)}{\pi} = \Pr\left[h_u(a) \neq h_u(b) \right]$$

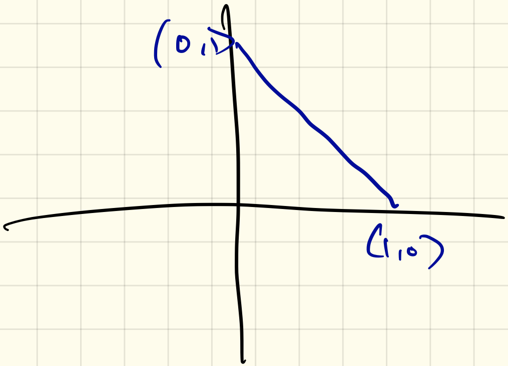
KL Divergence

$a, b \in \Delta_+^d$

$$d_{KL}(a \| b) = \sum_{i=1}^{d} a_i \ln(a_i / b_i)$$

$$\Delta^d = \{ x \in \mathbb{R}^d \mid \|x\|_1 = 1 \ \& \ \forall i \ x_i \geq 0 \}$$

$$\Delta_+^d = \{ x \in \mathbb{R}^d \mid \|x\|_1 = 1 \ \& \ \forall i \ x_i > 0 \}$$



Probability distribution