L6: Distances

Jeff M. Phillips

January 29, 2018

Modeling

- mathematical properties
- math & data properties.
- efficiency

Distance

$$d : \mathcal{X} \times \mathcal{X} \to \mathbb{R}_{\geq 0}$$

space that data lies in

Range

ex: $\mathbb{R}^d$, $\{G=(V,E)\}$, text

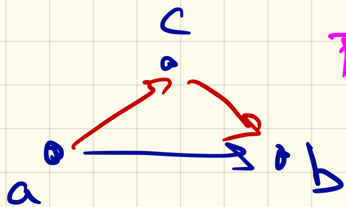
Metric

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



pseudometric M1, M3, M4

quasimetric M1, M2, M4

Data $\in \mathbb{R}^d$ $\quad a = (a_1, a_2, \dots a_d)$ $\quad b = (b_1, b_2, \dots b_d)$

L_p - distances

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

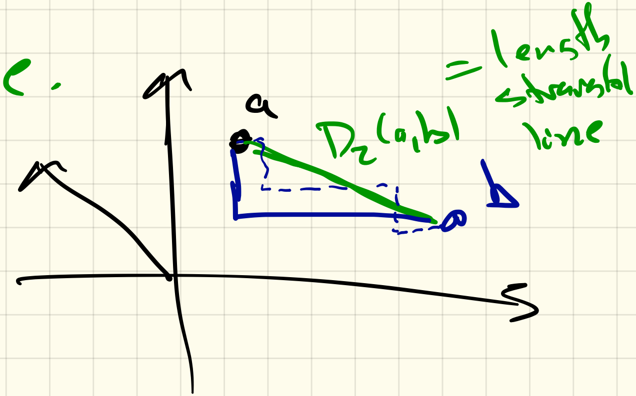
$p \in [1, \infty) \implies$ metric

consider $p = 2, 1, \infty, 0$

L_2 distance

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

Euclidean distance.



$(2, -2)$

$(-1, 1)$

$$\left(|3|^3 + |3|^3\right)^{1/3} = (27+27)^{1/3}$$

L_1-distance

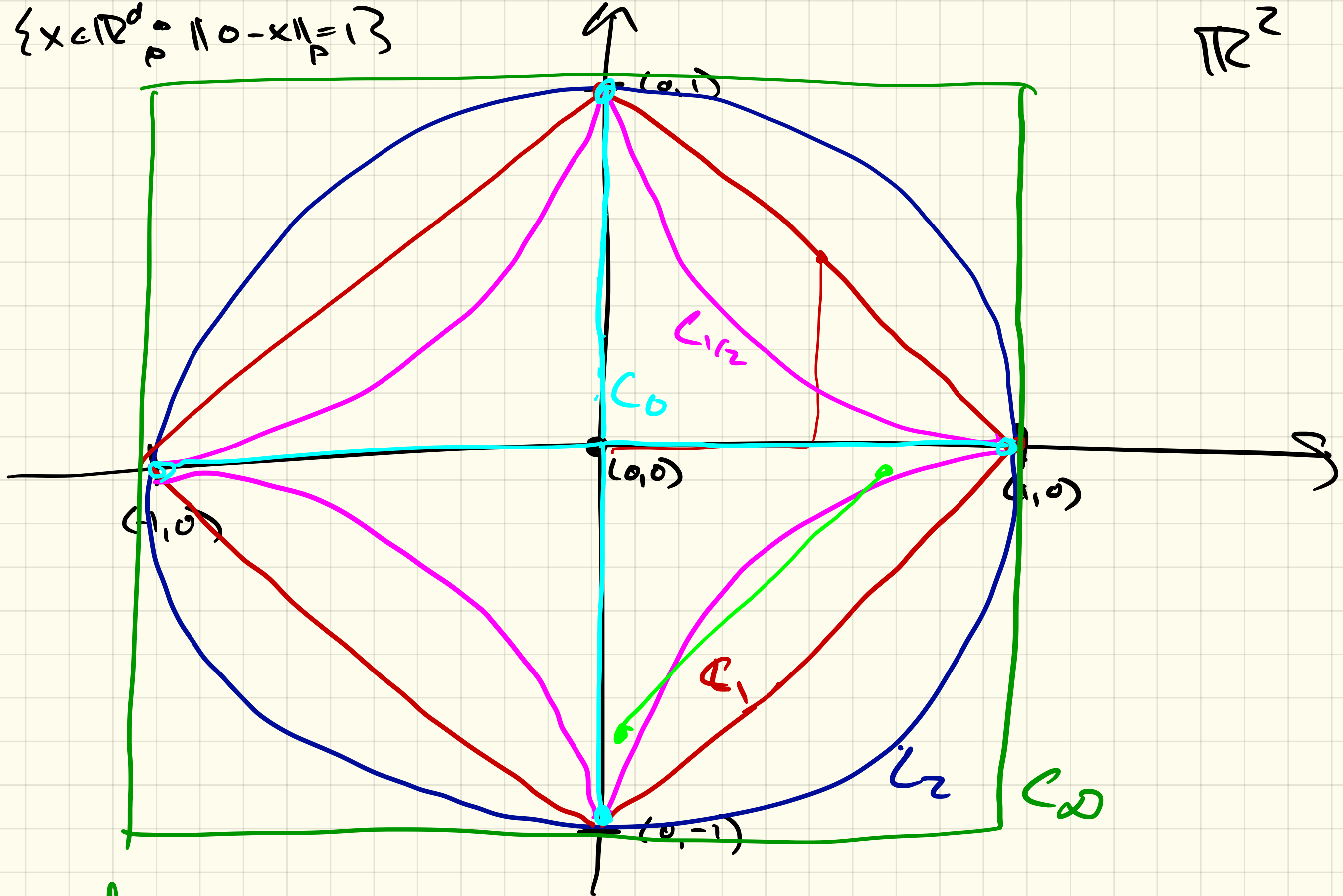
$$D_1 = \|a-b\|_1 = \sum_{i=1}^{d} |a_i - b_i|$$

Manhattan Distance

SLC - Distance

$C_p = \{ x \in \mathbb{R}^d : \|0 - x\|_p = 1 \}$

$\mathbb{R}^2$



$L_\infty = \max_{i=1}^{d} |a_i - b_i|$

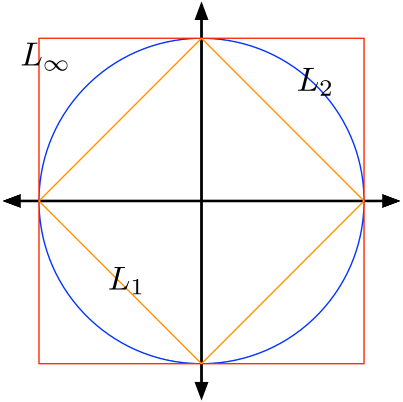
L_0 ≈ counting how many matches
= Hamming Distance

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 $\|a - b\|_p = 1$.



$L_0 = \frac{1}{d}(\#\text{ mismatches})$

Lp Distances and Units

Distance Metric Learning

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

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

Nonsense

w_1, w_2, w_3

$-250{,}000$
-4400
-1847

$\rightarrow$ 250,000
2250
104

Mahalanobis Distance

Requires a $d \times d$ matrix M

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

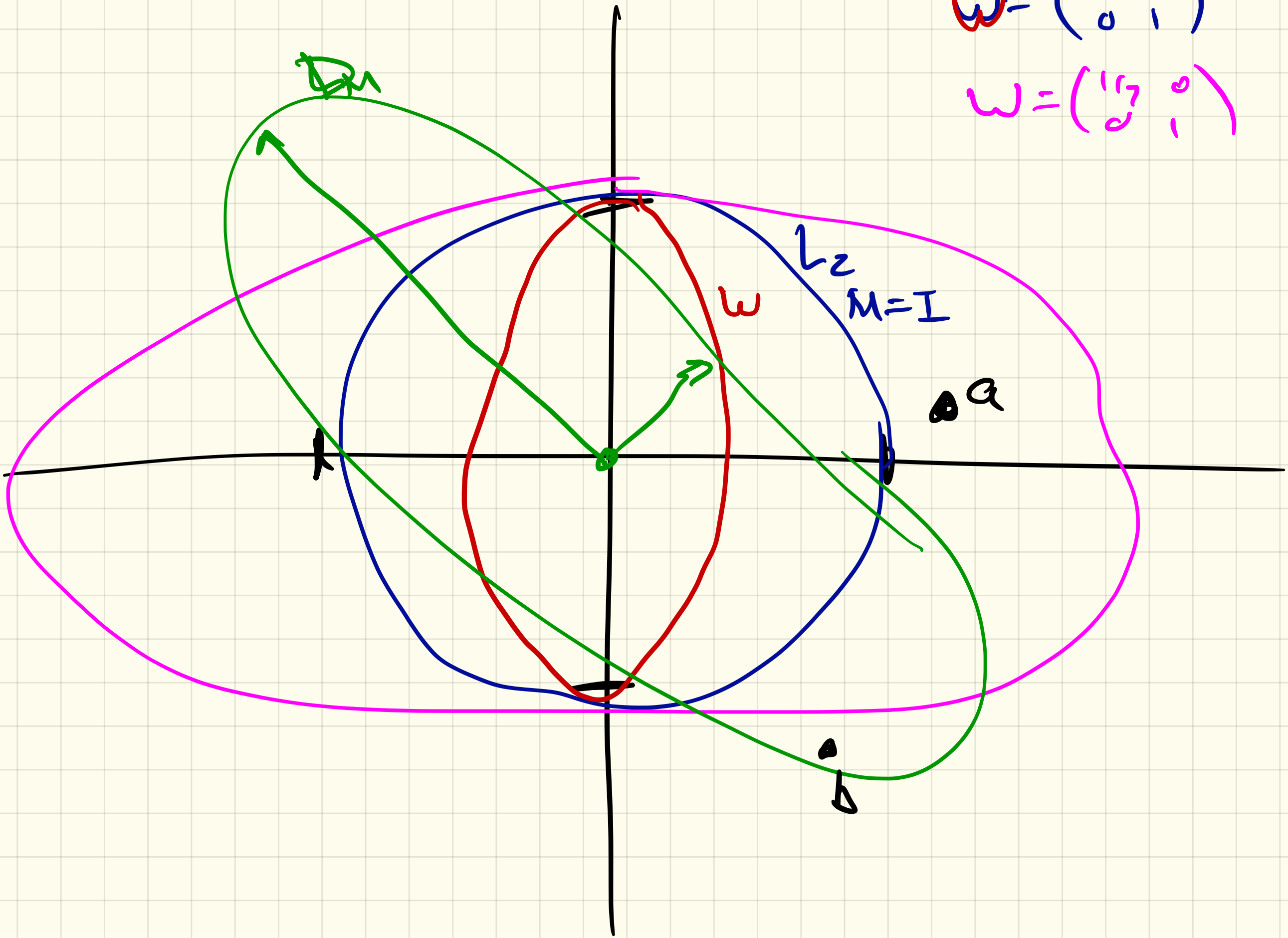
if $M = I = \begin{pmatrix} 1 & & 0 \\ & \ddots & \\ 0 & & 1 \end{pmatrix}$ $\quad D_M = D_2$

if $M = W = \begin{pmatrix} w_1 & & & 0 \\ & w_2 & & \\ & & \ddots & \\ 0 & & & w_d \end{pmatrix}$

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

$$W = \begin{pmatrix} 2 & 0 \\ 0 & 1 \end{pmatrix}$$

$$W = \begin{pmatrix} \frac{1}{2} & 0 \\ 0 & 1 \end{pmatrix}$$



Jaccard Distances

$$D_J(A,B) = 1 - JS(A,B)$$

metric

$$= 1 - \frac{|A \cap B|}{|A \cup B|} = \frac{|A \Delta B|}{|A \cup B|}$$

Cosine Distance

pseudometric

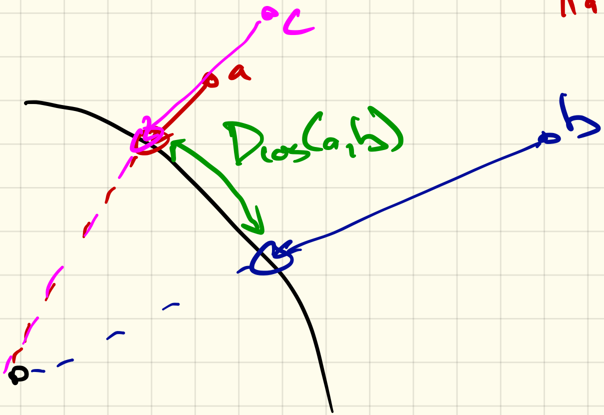
$\mathcal{X} \doteq \mathbb{R}^d$ $\quad a = (a_1, \ldots, a_d) \quad b = (b_1, \ldots, b_d)$

often each $a_i \geq 0$

common bag-of-words

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

$$= 1 - \left\langle \frac{a}{\|a\|}, \frac{b}{\|b\|} \right\rangle$$



LSHable.

KL - Divergence (Information Distance)

$a = (a_1 \dots a_d)$ comes from counts

$\downarrow$

$\hat{a} = \frac{a}{\|a\|_1}$ $\sum_{i=1}^{d} |\hat{a}_i| = 1$

probability distribution

$$d_{KL}(a,b) = \sum_{i=1}^{d} a_i \ln\left(\frac{a_i}{b_i}\right)$$

$$d_{H}(a,b) = \sqrt{\sum_{i=1}^{d} \left(\sqrt{a_i} - \sqrt{b_i}\right)^2}$$

Edit Distance

$D_{edit}(a, b)$ a = string b text

changes needed to get to a from b

a = cat → hats = b =D_{ed} = ?

c → h cat
+ s hat
hats

not 2 stable