Distances

Note Title 2/1/2016

$D : X \times X \to \mathbb{R}^+$

$D(a,b) \to [0, \infty)$

- Nice Properties "metric"
- List of Distances
  - ↳ choice
- Unexpected Properties

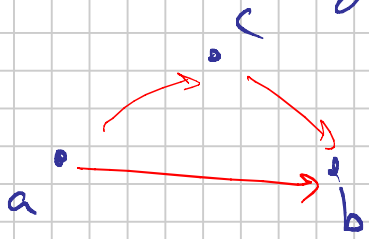
Metric

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

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

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

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



M1, M3, M4 → pseudo metric

M1, M2, M4 → quasi metric

L_p Distances

$p \in (0, \infty) + \{0\}, \{\infty\}$

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

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

$a = (a_1, a_2, \ldots, a_d)$

(M4) $\Leftarrow$ $p \geq 1$

- L_2 = Euclidean
  - ↳ metric

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

- L_1 = Manhattan
  = SLC Distance

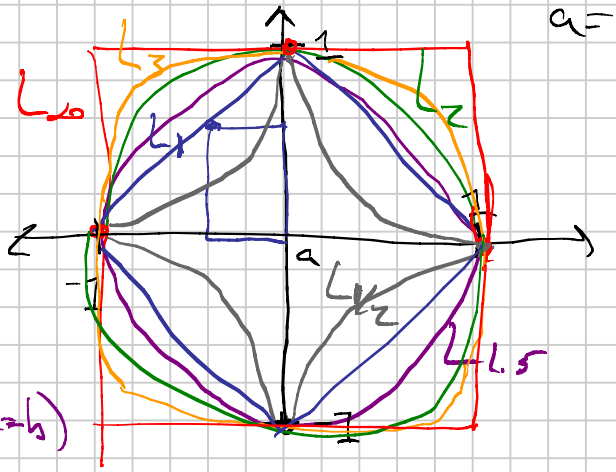
$D_1(a,b) = \|a-b\|_1 = \sum_{i=1}^{d} |a_i - b_i|$

$\left(2^{1000} + 1^{1000}\right)^{1/1000}$

$\|v\|_1$

- $L_\infty =$

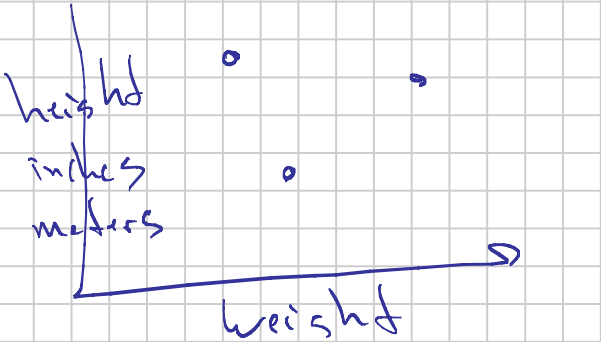
$$D_\infty(a,b) = \max_{i=1..d} |a_i - b_i|$$



- L_0 = Hamming

$$D_0(a,b) = \|a-b\|_0 = d - \sum_{i=1}^{d} \mathbb{1}(a=b)$$

WARNING ONLY Use for vectors in $\mathbb{R}^\phi$ where all coordinates have same unit.



→ "normalize" < all coords in $[0,1]$; mean = 0, var = 1

hacks!

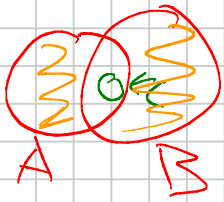
Distance Metric Learning (MDS)

Jaccard Distance D_J: $1 - JS(A,B) = 1 - \frac{|A \cap B|}{|A \cup B|} \in [0,1]$

metric

Triangle Inequality

Another set C



$$D_J(A,B) \leq D_J(A,C) + D_J(C,B)$$

$$= \frac{|A \setminus C|}{|A|} + \frac{|B \setminus C|}{|B|}$$

$$\geq \frac{|A \cup C| + |B \setminus C|}{|A \cup B|}$$

$$\geq \frac{|A \Delta B|}{|A \cup B|} = D_J(A,B)$$

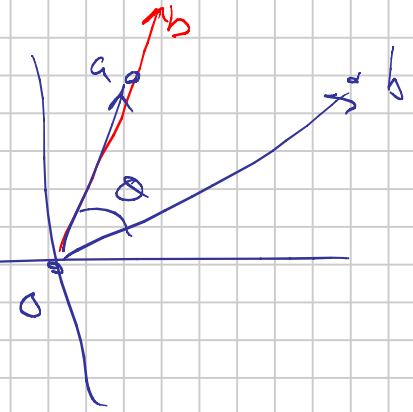
Assume $C \subset A$, $C \subset B$

$C \subset A \cap B$

Cosine Distance

$$D_{\cos}(a,b) = 1 - \frac{|\langle a,b\rangle|}{\|a\|_2 \|b\|_2} = 1 - \frac{\left|\sum_{i=1}^{d} a_i \cdot b_i\right|}{\|a\|_2 \|b\|_2}$$

$\cos(\theta)$



normalizing

"Bag-of-Words" data

(the and up it) (0, 1, 2, 0) ← "up and up"

$a = (1, 0, 0 \ldots) \Rightarrow a' = \frac{a}{\|a\|}$

$b = (2, 0, 0, 0 \ldots) \Rightarrow b' = \frac{b}{\|b\|}$

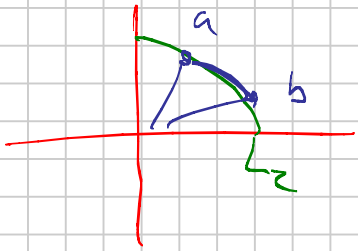
Metric

(M1) $\rightarrow \mathbb{R}^+$

NOT(M2) $D_{\cos}(a,b) = 0$ iff $a = b$

(M3) symmetric

(M4) Δ ineq.?

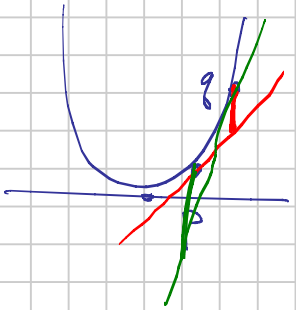


KL Divergence

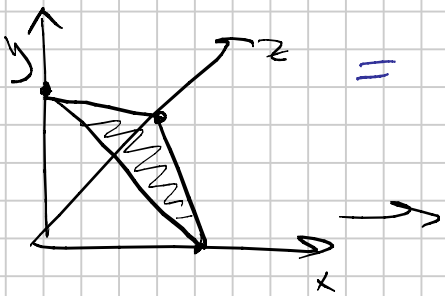
Kullback-Leibler

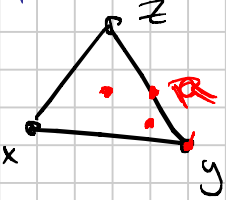
p, q discrete prob distributions

$p = (p_1, p_2, \ldots, p_d)$ s.t. $p_i \geq 0$, $\sum_{i=1}^{d} p_i = 1$



$$D_{KL}(p \| q) = \sum_{i=1}^{d} p_i \cdot \ln\left(\frac{p_i}{q_i}\right) = H(p,q) - H(p)$$





$$= \sum_{i=1}^{d} p_i e_i$$

$e_i = (0, 0, \ldots, \overset{i}{1}, 0, 0)$