Clase 1 (10/11/2025) TDA - EAFIT

Análisis Topológico de Datos

* https://github.com/joperea/TDA-Fall2025

Conjuntos de Datos (X, d_X) :

a) Un conjunto X (finito o no)

ejemplos: Vectores, imágenes, sonidos, videos, documentos (texto), moléculas, etc.

b) Una función (distancia)

$$d_X : X \times X \to [0, \infty] \quad \text{tal que}$$

i) $d_X(x,x) = 0$ para todo x en X ($\forall x \in X$)

ii) $d_X(x,x') = d_X(x',x)$ $\forall x, x' \in X$ (simétrica)

Si además

iii) $d_X(x,x'') \leq d_X(x,x') + d_X(x',x'')$ $\forall x, x', x'' \in X$

(

desigualdad triangular)

diremos que d_X es una pseudo métrica extendida

($d_X(x,x')=0$, $x \neq x'$) ($d_X(x,x') = \infty$)

iv) Si i) — iii) y $d_X(x,x') < \infty \ \forall x,x' \in X$,

d_X es una <u>pseudo métrica</u>

v) Si i) — iv) y $d_X(x,x') = 0$ solo si $x = x'$,

entonces d_X es una métrica y (X, d_X)

un <u>espacio</u> <u>métrico</u>.

<u>Ejemplos</u>

Distancias entre vectores:

distancia Euclidea

norma Euclidea

1) $X = \mathbb{R}^n$ y $d_2(x,y) = \|x - y\|_2$

donde $\underset{\underset{x}{\parallel}}{\|(x_1,\dots,x_n)\|_2} = \left(x_1^2 + \dots + x_n^2 \right)^{1/2}$.

Por tanto $d_2(x,y) = \left((x_1 - y_1)^2 + \dots + (x_n - y_n)^2 \right)^{1/2}$.

<u>Ejercicio</u>:

a) Muestre que $(\mathbb{R}^n, d_2(\cdot,\cdot))$ es un espacio métrico.

b) Sea $p \in \mathbb{N} = \{1, 2, 3, \ldots\}$ y para $x, y \in \mathbb{R}^n$ defina

$$d_p(x,y) = \| x - y \|_p \leftarrow \text{p-norma}$$

$$= \left(|x_1 - y_1|^p + \cdots + |x_n - y_n|^p \right)^{1/p}$$

Muestre que $(\mathbb{R}^n, d_p)$ es un espacio métrico.

c) Para $(x_1, \ldots, x_n) \in \mathbb{R}^n$ sea

$$\| (x_1, \ldots, x_n) \|_\infty = \max \{ |x_1|, |x_2|, \ldots, |x_n| \}$$

norma del supremo

y defina $d_\infty(x, y) = \| x - y \|_\infty$

$$= \max \{ |x_1 - y_1|, \ldots, |x_n - y_n| \}$$

Muestre que $(\mathbb{R}^n, d_\infty)$ es un espacio métrico,

y que $\lim_{p \to \infty} d_p(x,y) = d_\infty(x,y) \quad \forall x, y \in \mathbb{R}^n$

$d_p(\cdot, \cdot)$:

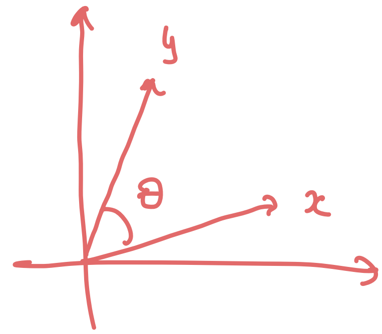
- $p = 1$: Distancia Manhattan (del taxista)
- $p = 2$: Distancia Euclidea
- $p = \infty$: Distancia del supremo.

2) Para $x, y \in \mathbb{R}^n$ sea $x \cdot y = x_1 y_1 + x_2 y_2 + \dots + x_n y_n$

$\uparrow$ producto punto (interno)

Si $x, y \neq \vec{0}$, defina:

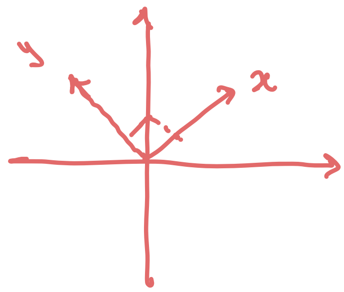
$$d_{\cos}(x,y) = \frac{x \cdot y}{\|x\|_2 \cdot \|y\|_2} = \cos\theta$$



$\uparrow$ distancia del coseno.

<u>Note</u>: $d_{\cos}(\cdot,\cdot)$ <u>no</u> es una métrica

i) $x = (1,1)$, $y = (-1, 1)$



$\|x\|_2 = \|y\|_2 = \sqrt{2}$

$x \cdot y = 1 \cdot (-1) + 1 \cdot 1 = 0$

$$d_{\cos}\big((1,1), (-1,1)\big) = \frac{0}{\sqrt{2} \cdot \sqrt{2}} = 0.$$

ii) <u>Ejercicio</u>: Muestre que $d_{\cos}(\cdot,\cdot)$ <u>no</u> satisface la desigualdad triangular.

<u>Note</u>: $d_{\cos}(\cdot,\cdot)$ es útil para comparar textos.

Diccionario:	palabra 1,	palabra 2,	...,	palabra n
Texto 1 $\longrightarrow$	$(t_{11}$,	t_{12},	...,	$t_{1n}) = t_1$
$\vdots$		$\vdots$		
Texto i $\longrightarrow$	$(t_{i1}$,	t_{i2},	...,	$t_{in}) = t_i$
$\vdots$		$\vdots$		
Texto m $\longrightarrow$	$(t_{m1}$,	t_{m2},	...,	$t_{mn}) = t_m$

t_{ik} := # de veces que la palabra k aparece en el Texto i

similitud(Texto i, Texto j) $= d_{cos}(t_i, t_j)$

Visualización y Reducción de Dimensión

Problema: Dado (X, d_X), $X = \{x_1, \dots, x_L\}$,

encontrar $n \in \mathbb{N}$ (usualmente $1 \leq n \leq 5$)

y $Y = \{y_1, \dots, y_L\} \subseteq \mathbb{R}^n$ tal que

$$d_X(x_i, x_j) \approx \|y_i - y_j\|_2, \quad 1 \leq i, j \leq L$$

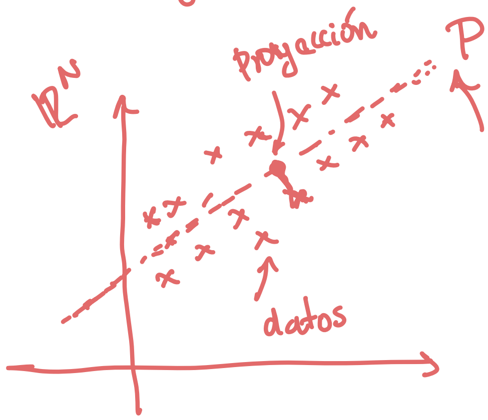
Análisis de Componentes Principales (PCA)

En el caso $X = \{x_1, \dots, x_L\} \subseteq \mathbb{R}^N$ $(N >> 1)$

mucho mas grande que

$$d_X(x_i, x_j) = \| x_i - x_j \|_2$$

idea: "regresion lineal" (mínimos cuadrados)



mayor "plano" n-dimensional:

$\underbrace{p_1, \dots, p_n}$ $\in \mathbb{R}^N$, $b \in \mathbb{R}^N$

Base ortonormal (componentes principales)

traslación

$$P = \{ b + \lambda_1 p_1 + \dots + \lambda_n p_n \mid \lambda_i \in \mathbb{R}, \ 1 \le i \le n \}$$

$proy_P : \mathbb{R}^N \to P$ proyección orthogonal

$$y_\ell = proy_P(x_\ell), \quad 1 \le \ell \le L$$

proyección

Datos

$(\lambda_{\ell 1}, \lambda_{\ell 2}, \dots, \lambda_{\ell n}) \in \mathbb{R}^n$

$(x_{\ell 1}, x_{\ell 2}, \dots, x_{\ell N}) \in \mathbb{R}^N$.