

# Medidas de Similitud y Distancia



# Data

Medidas de similitud y Manejo de datos

# Data es esencial



# Data Categórica (Cualitativa)

## Nominal

- Género
- Color
- Ciudad

## Ordinal

- Rangos
- Likert
- Dureza de minerales

# Data Numérica (Cuántitativa)

## Intervalos

La diferencia entre un par de datos es significativa

- Temperaturas
- Fechas

## Ratios

El cero denota no existencia

- Peso
- Altura



# Adicionalmente



Continuos



Discretos



Binarios

# Similitud y Diferencia

- “Proximidad”
- Indica que tan parecido es un objeto o conjunto de objetos unos de otros.





## Similitud y Diferencia

Similitud mide que tan parecidos son dos objetos, mayormente son datos comprendidos entre 0 y 1.

Diferencia, mide que tan diferentes son dos objetos; su medida puede estar entre 0 y 1, o también entre 0 o  $\alpha$ .



# Ejemplo con valores nominal

- Ejemplo simple binario:

Diferencia

$$d = \begin{cases} 0 & \text{en caso de que } x = y \\ 1 & \text{en caso de que } x \neq y \end{cases}$$

Para similitud, se invierte

$$s = \begin{cases} 1 & \text{en caso de que } x = y \\ 0 & \text{en caso de que } x \neq y \end{cases}$$

$$d = 1 - s$$

# Ejemplo con valores ordinales

{mala calidad, baja calidad, normal, calidad buena, calidad superior}

Se transforma a:

**[0,1,2,3,4]**

Diferencia entre **calidad buena** y **calidad superior**:

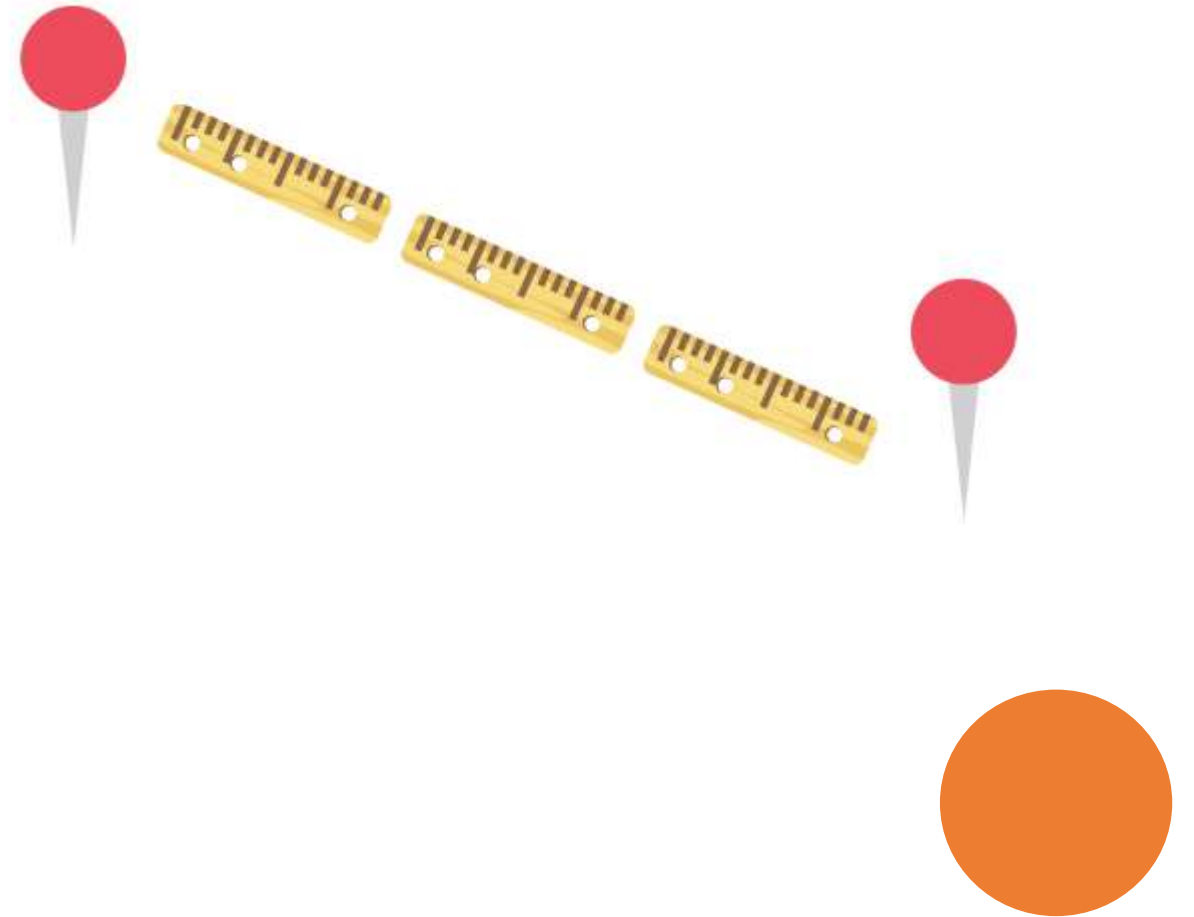
$$d = \frac{|x-y|}{n-1} \quad d = \frac{|3-4|}{5-1} \quad d = \frac{1}{4} \quad d = 0.25$$

Similitud

$$s = 1 - d \quad s = 1 - 0.25 \quad s = 0.75$$

# Distancias

Determina qué tan diferentes son unos objetos de otros





# Distancias

- Manhattan

Conocido como norma  $L_1$

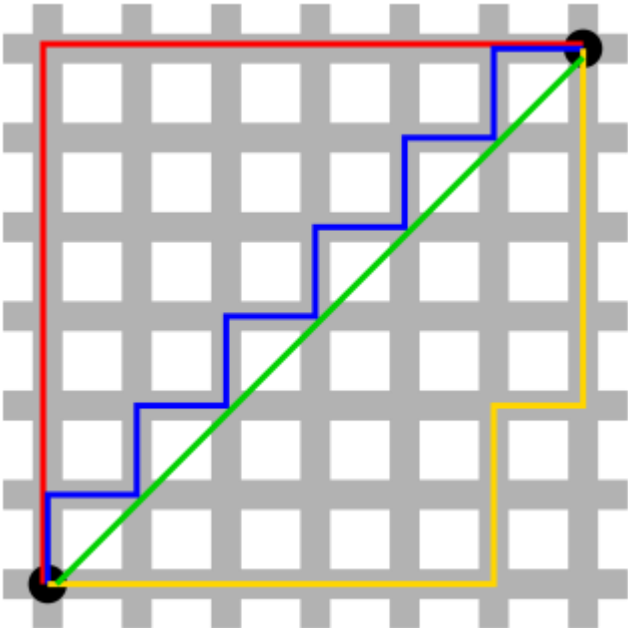
$$d(p, q) = |p_1 - q_1| + |p_2 - q_2| + \dots + |p_n - q_n|$$

$$d(p, q) = \sum_{i=1}^n |p_i - q_i|$$

Donde  $p$  y  $q$  son dos puntos y  $n$  es la dimensiones del vector

Ejemplo: Distancia de Hamming:

$$d(00111, 11001) = 4$$

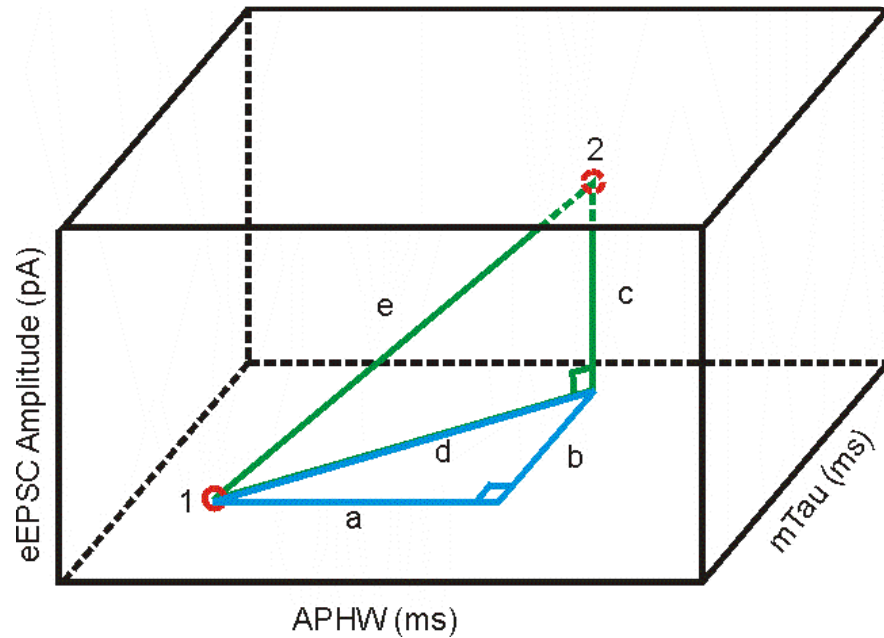



# Distancias

- Distancia Euclidiana  
Conocido como norma  $L_2$

$$d(p, q) = \sqrt{(p_1 - q_1)^2 + (p_2 - q_2)^2 + \dots + (p_n - q_n)^2}$$

$$\sqrt{\sum_{i=1}^n (p_i - q_i)^2}$$



	a	b	c	d	e	f	g	h	
8	5	4	3	2	2	2	2	2	8
7	5	4	3	2	1	1	1	2	7
6	5	4	3	2	1		1	2	6
5	5	4	3	2	1	1	1	2	5
4	5	4	3	2	2	2	2	2	4
3	5	4	3	3	3	3	3	3	3
2	5	4	4	4	4	4	4	4	2
1	5	5	5	5	5	5	5	5	1
	a	b	c	d	e	f	g	h	

# Distancias

- Chebyshev

Distancia del tablero de ajedrez

Norma  $L_\infty$

$$d(p, q) = \max_i |p_i - q_i|$$

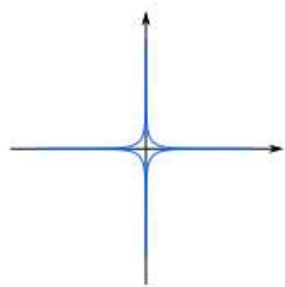
$$d(p, q) = \lim_{r \rightarrow \infty} \left( \sum_{i=1}^n |p_i - q_i|^r \right)^{\frac{1}{r}}$$

# Distancias

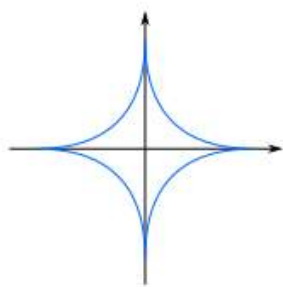
- Minkowski

Generalización de las distancias Manhattan, Euclidiana y Chebyshev

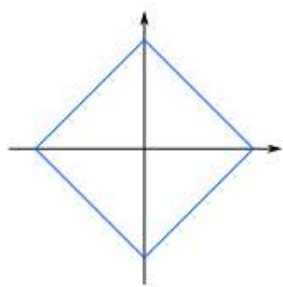
$$d(p, q) = \left( \sum_{i=1}^n |p_i - q_i|^r \right)^{\frac{1}{r}}$$



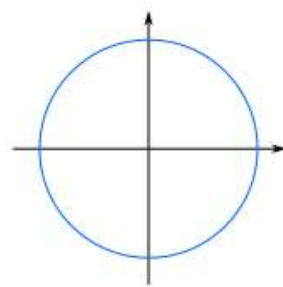
$$p = 2^{-2} \\ = 0.25$$



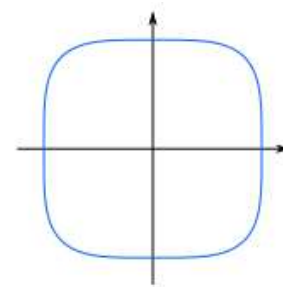
$$p = 2^{-1} \\ = 0.5$$



$$p = 2^0 \\ = 1$$

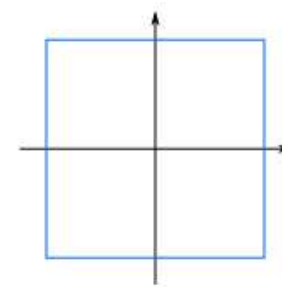


$$p = 2^1 \\ = 2$$



$$p = 2^2 \\ = 4$$

...



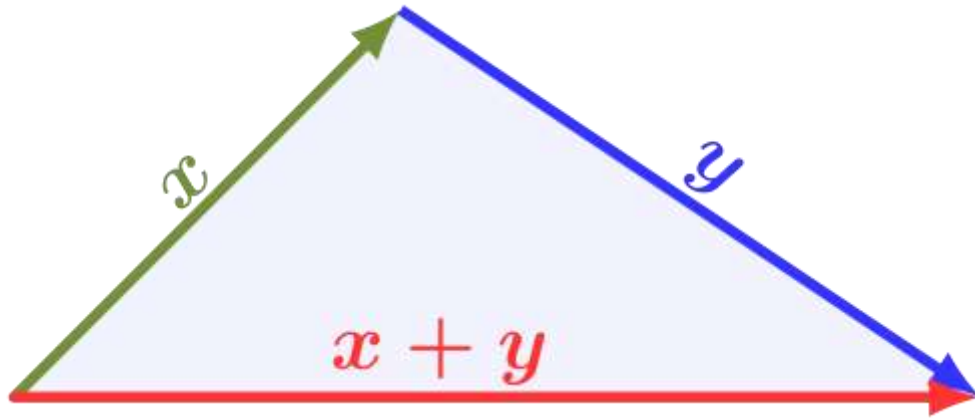
$$p = 2^{\infty} \\ = \infty$$



# Distancias

## Propiedades

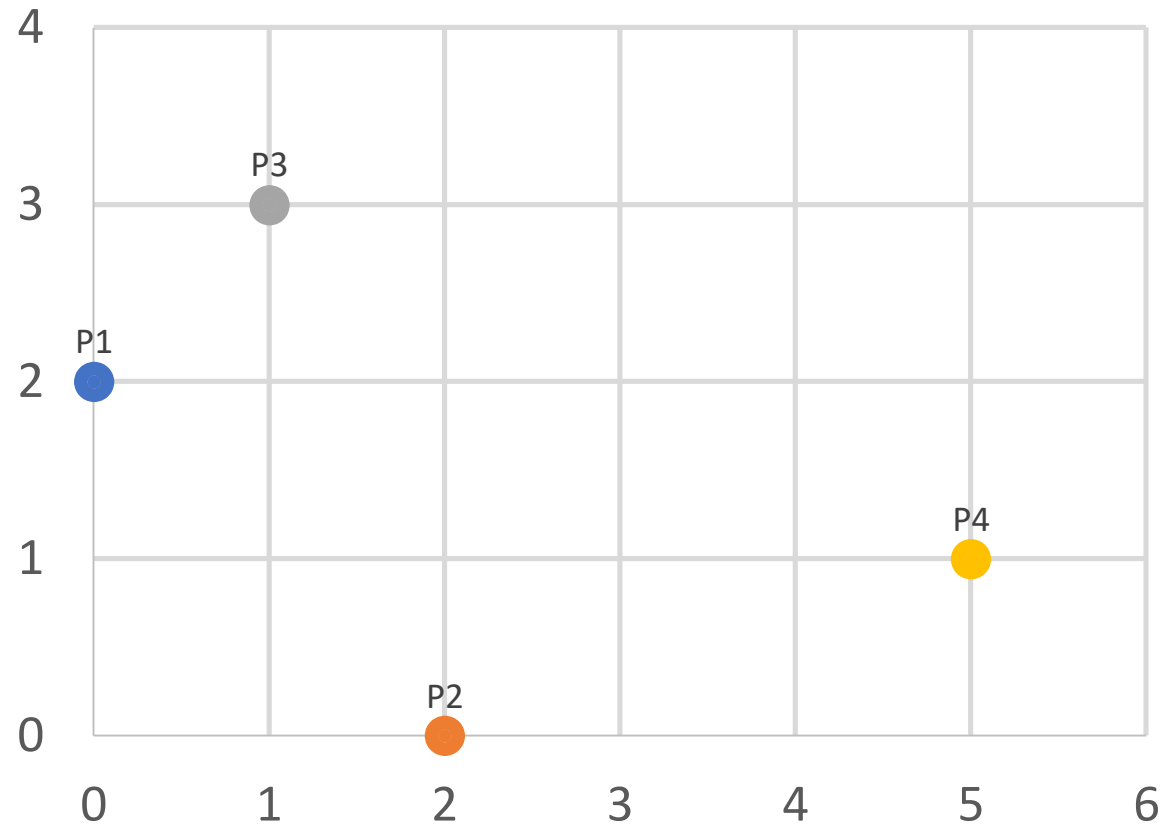
- **No-negatividad:**  $d(i,j) \geq 0$
- **Identidad:**  $d(i,i) = 0$
- **Simetría:**  $d(i,j) = d(j,i)$
- **Desigualdad triangular:**  $d(i,j) \leq d(i,k) + d(k,j)$



# Ejemplo (en grupos)

	x	y
P1	0	2
P2	2	0
P3	1	3
P4	5	1

Calcular la distancias L1, L2 y  $L_\infty$  para todos los puntos





Similitud

# Simple Matching Coefficient

Dado dos objetos A y B con n atributos binarios:

$$SMC = \frac{f_{00} + f_{11}}{f_{00} + f_{11} + f_{01} + f_{10}}$$

Donde:

$f_{00}$  es la frecuencia de atributos donde tanto A y B tienen el valor 0

$f_{11}$  es la frecuencia de atributos donde tanto A y B tienen el valor 1

$f_{01}$  es la frecuencia de atributos donde A tiene 0 y B tiene 1

$f_{10}$  es la frecuencia de atributos donde A tiene 1 y B tiene 0

# Simple Matching Coefficient

$$SMC = \frac{f_{00} + f_{11}}{f_{00} + f_{11} + f_{01} + f_{10}}$$



Fruta	Redonda?	Dulce?	Ácida?	Semilla?	Verde?
Manzana	1	1	1	1	0
Banana	0	1	0	0	0

Las coordenadas de la manzana es (1,1,1,1) mientras que de la banana es (0,1,0,0)

$$SMC = \frac{1 + 1}{1 + 1 + 0 + 3}$$

$$SMC = \frac{2}{5} = 0.4$$

# Coeficiente de Jaccard

- No considera transacciones negativas que no brindan información nueva

$$JC = \frac{f_{11}}{f_{11} + f_{01} + f_{10}} \quad JC = \frac{A \cup B}{A \cap B}$$

# Coeficiente de Jaccard

$$JC = \frac{f_{11}}{f_{11} + f_{01} + f_{10}}$$



Fruta	Redonda?	Dulce?	Ácida?	Semilla?	Verde?
Manzana	1	1	1	1	0
Banana	0	1	0	0	0

Las coordenadas de la manzana es (1,1,1,1) mientras que de la banana es (0,1,0,0)

$$JC = \frac{1}{1 + 0 + 3}$$

$$JC = \frac{1}{4} = 0.25$$



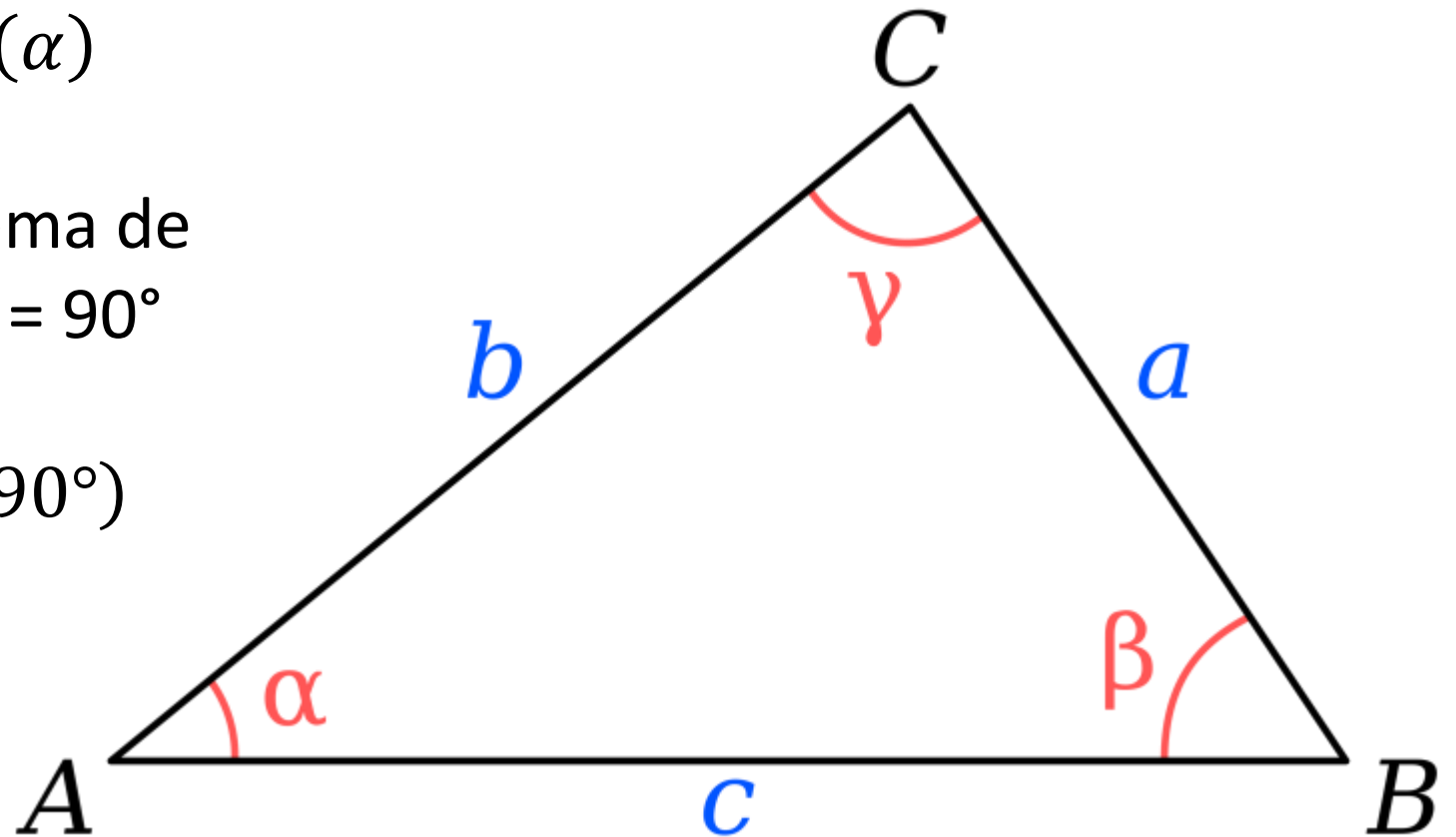
# Similitud coseno

- Recordando el teorema de Pitágoras y la ley del coseno

$$a^2 = b^2 + c^2 - 2bc \cos(\alpha)$$

Es la generalización del teorema de Pitágoras, si consideramos  $\alpha = 90^\circ$

$$a^2 = b^2 + c^2 - 2bc \cos(90^\circ)$$
$$a^2 = b^2 + c^2$$



# Similitud coseno

1)

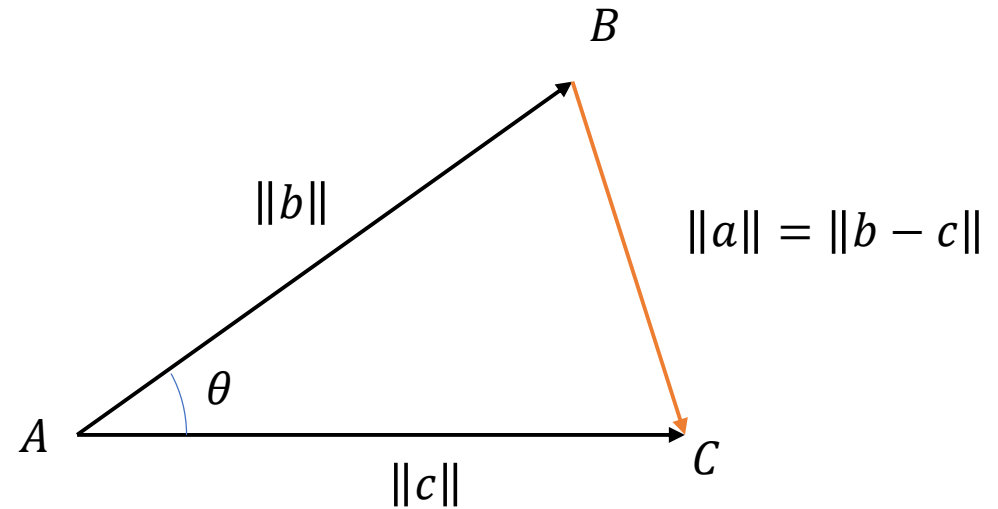
- $a^2 = b^2 + c^2 - 2bc \cos(\alpha)$
- $\|b - c\|^2 = \|b\|^2 + \|c\|^2 - 2\|b\|\|c\| \cos(\alpha) \quad (1)$

2)

- $\|b - c\|^2 = (b - c)^2$
- $= b^2 - 2bc + c^2$
- $= b \cdot b - 2b \cdot c + c \cdot c$
- $= \|b\|^2 - 2b \cdot c + \|c\|^2$

1 y 2

$$\|b\|^2 + \|c\|^2 - 2\|b\|\|c\| \cos(\alpha) = \|b\|^2 - 2b \cdot c + \|c\|^2$$



# Similitud coseno

$$\|b\|^2 + \|c\|^2 - 2\|b\|\|c\|\cos(\alpha) = \|b\|^2 - 2b \cdot c + \|c\|^2$$

$$-2\|b\|\|c\|\cos(\alpha) = -2b \cdot c$$

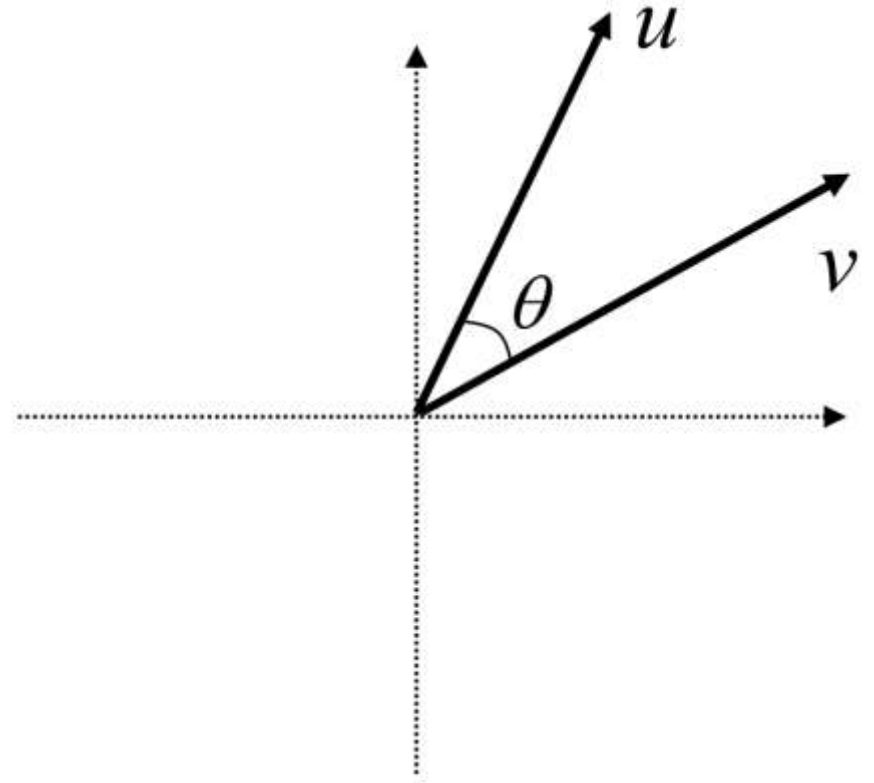
$$\|b\|\|c\|\cos(\alpha) = b \cdot c$$

$$\cos(\alpha) = \frac{b \cdot c}{\|b\|\|c\|}$$

# Similitud coseno

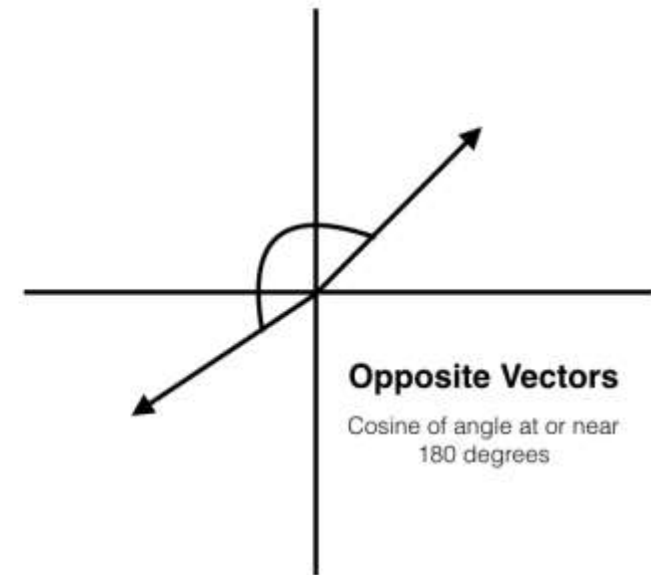
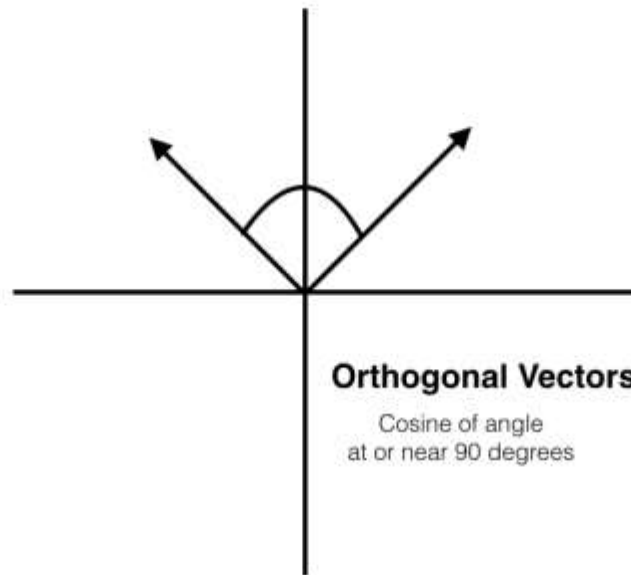
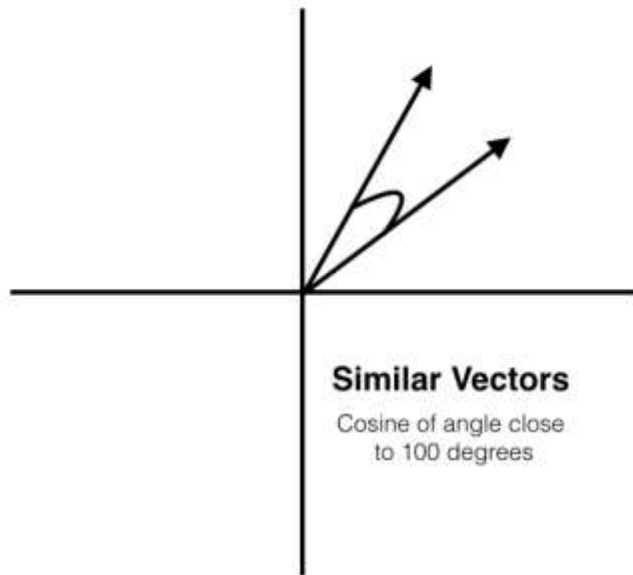
---

- Recordemos que si  $u = (x_1, y_1)$  y  $v = (x_2, y_2)$
- $$\cos(\theta) = \frac{x_1x_2 + y_1y_2}{\|u\|\|v\|} = \frac{u \cdot v}{\|u\|\|v\|}$$
- Este resultado es verdadero para N dimensiones
- El resultado es un número entre 0 y 1
- Mientras más cercano a 1, los vectores son mas similares



# Similitud Coseno

- <https://www.oreilly.com/library/view/mastering-machine-learning/9781785283451/ba8bef27-953e-42a4-8180-cea152af8118.xhtml>



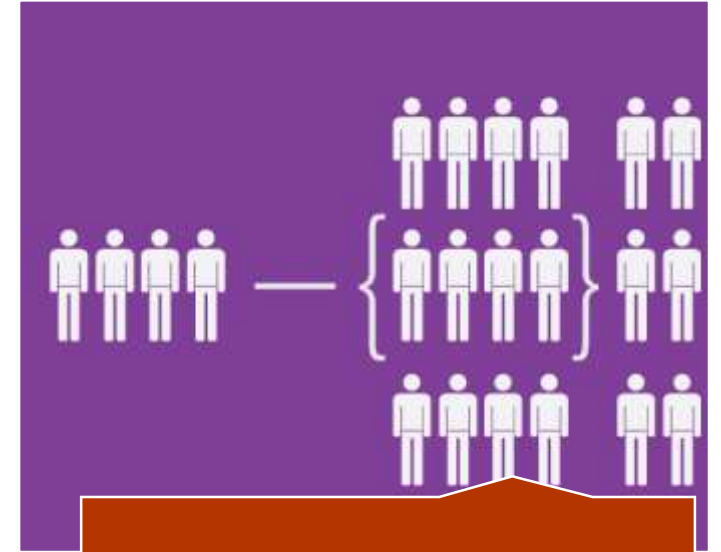
# Similitud Coseno (Ejemplos)



Análisis de Documentos



Poses



Usuarios similares



# Preparación de los Datos



# Tratamiento de Datos Categoricos

Genero

País


Color

Estado de Conservación

Preferencias

# Ordinal Encoding

- Usado cuando la data categórica tiene un orden natural

	edad	grado	Pais			edad	grado	Pais
0	30	bach	Francia		0	30	0	Francia
1	25	msc	Italia		1	25	1	Italia
2	22	nan	Peru		2	22	-1	Peru
3	40	doc	EEUU		3	40	2	EEUU
4	34	doc	nan		4	34	2	nan
5	50	doc	Francia		5	50	2	Francia

# One Hot Encoding

	edad	grado	Pais
0	30	bach	Francia
1	25	msc	Italia
2	22	nan	Peru
3	40	doc	EEUU
4	34	doc	nan
5	50	doc	Francia

	edad	grado	pais_EEUU	pais_Francia	pais_Italiapais_Peru	pais_nan
0	30	bach	0	1	0	0
1	25	msc	0	0	1	0
2	22	nan	0	0	0	1
3	40	doc	1	0	0	0
4	34	doc	0	0	0	0
5	50	doc	0	1	0	0



# Imputación de Datos

183.102

154.178

2455

# Datos faltantes



Eliminar



Promedio



Moda



Regresión




The image shows a close-up of a line graph on a grid. A pen is pointing to a sharp peak in the data series. The word "Outliers" is overlaid in white text. The background is a blue-tinted photograph of the graph.

Outliers



# Outlier o valores atípicos

- Dato que se encuentra por fuera del comportamiento general de una muestra de datos
  - Puede indicar variabilidad, errores de medición o novedades
- 



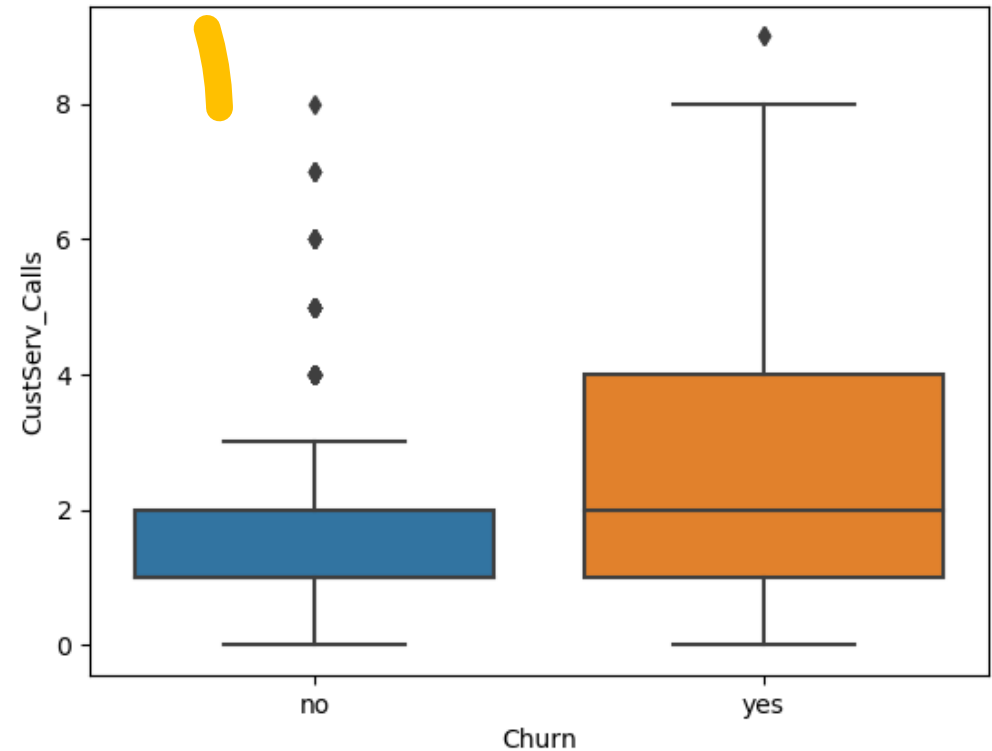
# Outliers

- Generan sesgos en los modelos
- Contienen información relevante
- Detección temprana de errores



# Formas de encontrarlos

- Z-Score
  - Más de  $3\sigma$  generalmente se lo considera outlier
- Boxplot
  - $Q1 - 1.5 \times IQR$
  - $Q3 + 1.5 \times IQR$
  - $IQR = Q3 - Q1$



# Normalización

```
for object to mirror_mod.mirror_object
operation == "MIRROR_X":
mirror_mod.use_x = True
mirror_mod.use_y = False
mirror_mod.use_z = False
operation == "MIRROR_Y":
mirror_mod.use_x = False
mirror_mod.use_y = True
mirror_mod.use_z = False
operation == "MIRROR_Z":
mirror_mod.use_x = False
mirror_mod.use_y = False
mirror_mod.use_z = True

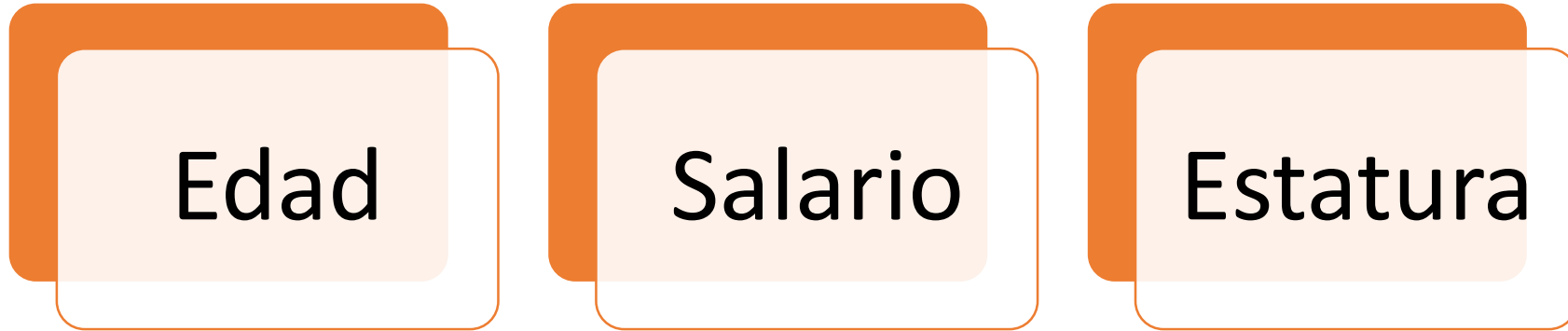
#selection at the end -add
mirror_ob.select= 1
mirror_ob.select=
context.scene.objects.active
("Selected" + str(modifier))
mirror_ob.select = 0
= bpy.context.selected_objects
data.objects[one.name].select

print("please select exactly

-- OPERATOR CLASSES -----

types.Operator):
X mirror to the selected
object.mirror_mirror_x"
mirror X"
```

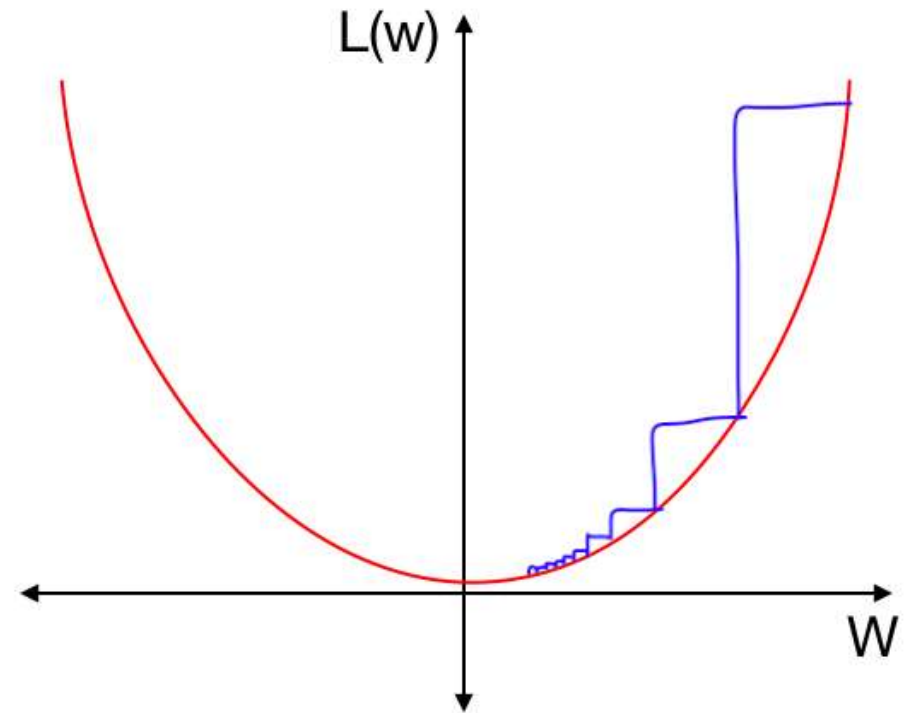
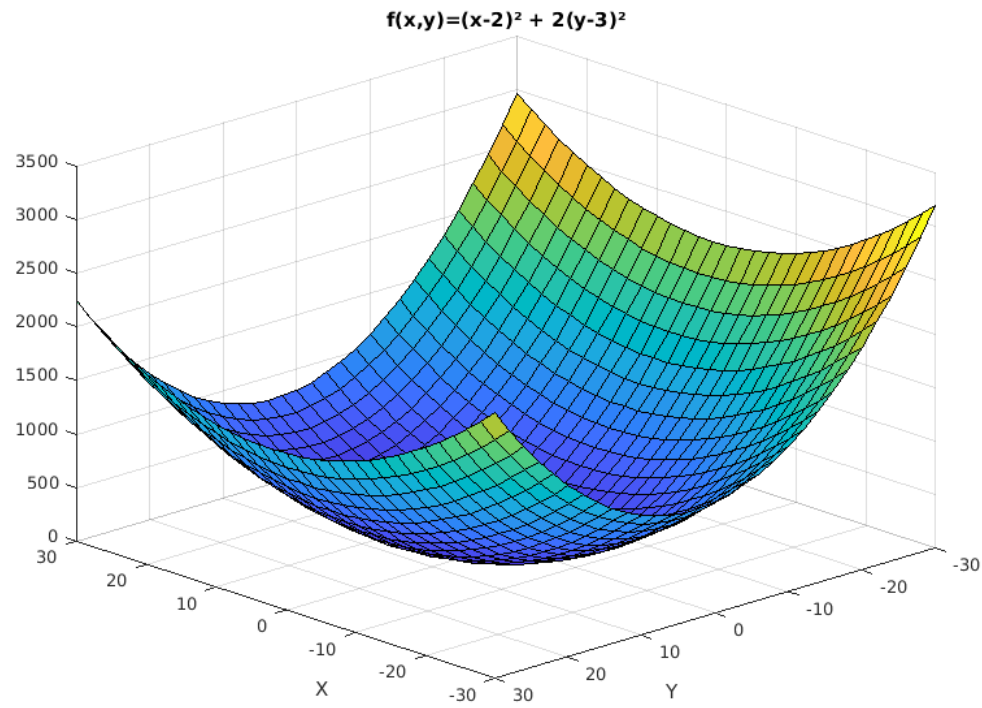
# Motivación



Diferentes órdenes de magnitud

## El problema

- Los modelos de ML no funcionan bien con datos de diferentes ordenes de magnitud
  - La gradiente demora en converger
- Se le da la misma importancia a todos



# Librería de Python para ML



<https://scikit-learn.org/>

## Limitaciones

---

No es para CV

---

No corre en GPU

---

No es muy flexible para  
redes neuronales

# Scikit-learn

## Clasificación

- SVM, K-NN, Random Forest

## Regresión

- SVR, K-NN, Random Forest, Lineal, Logaritmica

## Clustering

- K-Means, Gaussian Mixture

## Preprocesamiento

- Normalizar, Estandarizar, Imputación, Datos Categoricos





Train-Test

# ¿Por qué?

- Evaluar y validar modelos
- Ver que el modelo pueda generalizar
- Prevenir underfitting y overfitting

# Train-Validate-Test Split

- Cuando tengo mucha data



Entrenar:
<ul style="list-style-type: none"><li>• Usado para ajustar sus parámetros internos.</li><li>• (tiene variables de entrada y salida)</li></ul>

Validar
<ul style="list-style-type: none"><li>• Valida que está mejorando con cada iteración, pero no lo usa para “aprender”</li><li>• (tiene variables de entrada y salida)</li></ul>

Evaluar
<ul style="list-style-type: none"><li>• Esto es para nosotros. Con esto se valida que el modelo generaliza bien</li><li>• (la variable de salida es ocultada del modelo)</li></ul>

# Cross-Validation

- Cuando tengo poca data

