Universidad Nacional Autónoma de México
Facultad de Ingeniería

Estimación de similitudes

Distancias y métricas de aprendizaje automático

Guillermo Molero-Castillo
guillermo.molero@ingenieria.unam.edu

Noviembre, 2020

Métricas en aprendizaje automático

- En varios algoritmos de aprendizaje automático es necesario medir la separación o similitud entre diferentes registros (elementos).

ID	Salario	Casado	Coche	Hijos	Vivienda	Sindicato	Faltas/Año	Antigüedad	Sexo
E1	10000	Sí	No	0	Alquiler	No	7	15	M
E2	20000	No	Sí	1	Alquiler	Sí	3	3	F
E3	15000	Sí	Sí	2	Prop	Sí	5	10	M
E4	30000	Sí	Sí	1	Alquiler	No	15	7	F
E5	10000	Sí	Sí	0	Prop	Sí	1	6	M
E6	40000	No	Sí	0	Alquiler	Sí	3	16	F
E7	25000	No	No	0	Alquiler	Sí	0	8	M
E8	20000	No	Sí	0	Prop	Sí	2	6	F
E9	20000	Sí	Sí	3	Prop	No	7	5	M
E10	30000	Sí	Sí	2	Prop	No	1	20	M
E11	45000	No	No	0	Alquiler	No	2	12	F
E12	8000	Sí	Sí	2	Prop	No	3	1	M
E13	20000	No	No	0	Alquiler	No	27	5	F
E14	10000	No	Sí	0	Alquiler	Sí	0	7	M
E15	8000	No	Sí	0	Alquiler	No	3	2	M

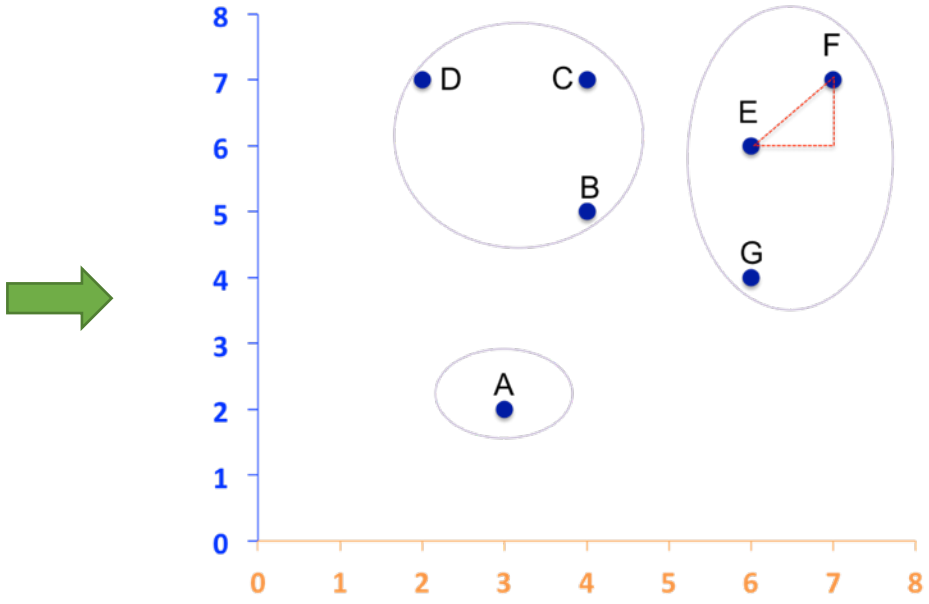
La similitud no se puede medir a ojo de buen cubero.

¿Algunas ideas para medir similitudes entre elementos?

Métricas en aprendizaje automático

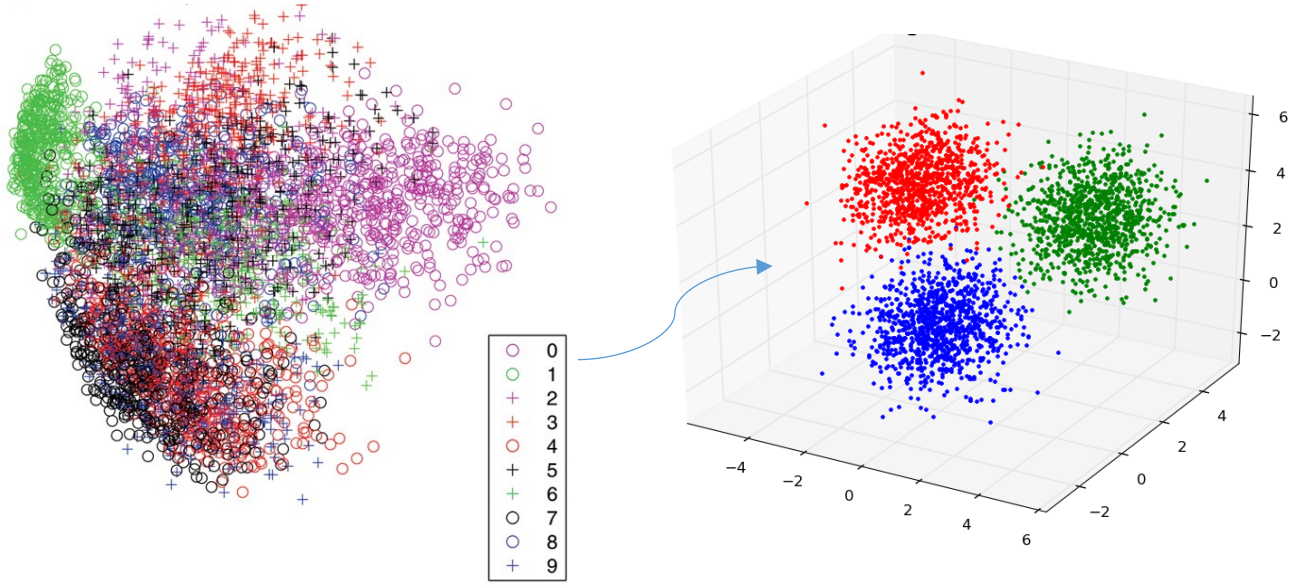
- Por ejemplo, en el **análisis de cluster** es necesario saber el **grado de similitud** entre los registros. La forma de hacer esto es utilizando las distancias. Asumiendo que los datos son puntos en un espacio de n dimensiones.

Sujeto	Lealtad a la tienda (x)	Lealtad a la marca (y)
A	3	2
B	4	5
C	4	7
D	2	7
E	6	6
F	7	7
G	6	4



Métricas en aprendizaje automático

- El objetivo del **análisis de cluster (agrupamiento)** es dividir un conjunto de datos (población de datos heterogénea) en un **número de grupos con elementos similares**, de acuerdo a la semejanza de sus elementos.



Los grupos nacen a partir de los datos y se descubren patrones ocultos en éstos.

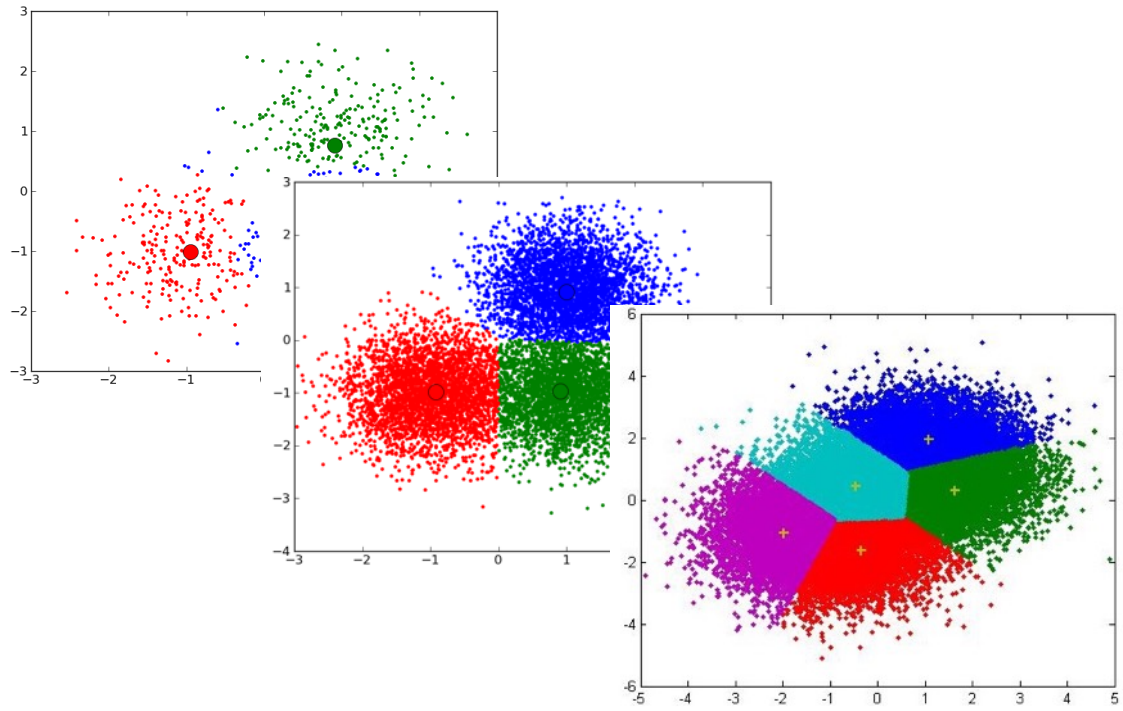
Métricas en aprendizaje automático

La complejidad en el **análisis** se da cuando se tiene un amplio número de variables.

Estacion	Altitud	EneP	EneTO	EneTM	EneTMin	FebP	FebTO	FebTM	FebTMin	MarP	MarTO	MarTM	MarTMin
16006	360	28.24	19.64	33.37	14.03	1.26	20.85	34.66	15.27	1.3	23.22	36.31	18.22
16007	682	21.48	17.37	33.84	15.49	2.6	18.59	35.44	16.7	2.08	20.44	37.17	18
16014	1708	18.94	7.08	25.02	4.26	5.88	8.24	27.01	5.27	4.34	10.04	29.43	6.95
16016	1840	19.47	6.32	21.31	3.76	6.36	7.92	23.29	5.16	7.07	10.02	25.27	6.93
16017	1694	18.04	7.04	24.59	4.78	6.55	8.39	26.61	5.99	6.82	10.79	29.29	7.84
16020	2020	23.9	5.24	23.45	3.07	9.13	5.97	25.07	3.91	7.29	7.18	26.98	5.01
16023	1500	13.76	5.24	22.09	1.44	5.18	6.41	23.46	2.58	5.43	8.58	25.52	4.36
16024	1693	14.54	7.17	23.39	5.62	2.86	8.72	25.47	7.22	3.19	11.22	27.93	9.64
16027	1831	22.26	9.81	23.3	5.86	4.73	10.9	24.92	6.87	5.05	13.22	27.81	9.26
16031	1632	22.97	8.42	25.06	7.08	3.92	9.55	27.14	8.21	3.44	11.69	29.42	9.84
16033	2415	24.03	5.97	19.32	4.34	6.22	6.6	20.6	5.02	7.04	7.7	22.11	6.1
16043	1581	35.22	6.88	23.91	5.23	3.3	7.89	25.59	6.1	3.22	9.72	28.03	7.55
16045	2240	30.16	11.5	20.54	8.9	7.48	12.17	22.19	9.27	8.74	13.64	24.3	10.54
16048	1567	20.52	9.43	25.88	6.99	6.22	10.55	27.69	8.3	3.75	12.26	29.89	9.93
16050	1950	20.37	8.06	22.49	6.72	6.5	9.27	24.28	7.8	5.98	11.66	27.1	9.57

...

...



Métricas en aprendizaje automático

Ejemplo. Análisis de grupos de empleados

ID	Salario	Casado	Coche	Hijos	Vivienda	Sindicato	Faltas/Año	Antigüedad	Sexo
E1	10000	Sí	No	0	Alquiler	No	7	15	M
E2	20000	No	Sí	1	Alquiler	Sí	3	3	F
E3	15000	Sí	Sí	2	Prop	Sí	5	10	M
E4	30000	Sí	Sí	1	Alquiler	No	15	7	F
E5	10000	Sí	Sí	0	Prop	Sí	1	6	M
E6	40000	No	Sí	0	Alquiler	Sí	3	16	F
E7	25000	No	No	0	Alquiler	Sí	0	8	M
E8	20000	No	Sí	0	Prop	Sí	2	6	F
E9	20000	Sí	Sí	3	Prop	No	7	5	M
E10	30000	Sí	Sí	2	Prop	No	1	20	M
E11	45000	No	No	0	Alquiler	No	2	12	F
E12	8000	Sí	Sí	2	Prop	No	3	1	M
E13	20000	No	No	0	Alquiler	No	27	5	F
E14	10000	No	Sí	0	Alquiler	Sí	0	7	M
E15	8000	No	Sí	0	Alquiler	No	3	2	M

Métricas en aprendizaje automático

Ejemplo. Análisis de grupos de empleados

ID	Salario	Casado	Coche	Hijos	Vivienda	Sindicato	Faltas/Año	Antigüedad	Sexo
E1	10000	1	0	0	0	0	7	15	1
E2	20000	0	1	1	0	1	3	3	0
E3	15000	1	1	2	1	1	5	10	1
E4	30000	1	1	1	0	0	15	7	0
E5	10000	1	1	0	1	1	1	6	1
E6	40000	0	1	0	0	1	3	16	0
E7	25000	0	0	0	0	1	0	8	1
E8	20000	0	1	0	1	1	2	6	0
E9	20000	1	1	3	1	0	7	5	1
E10	30000	1	1	2	1	0	1	20	1
E11	45000	0	0	0	0	0	2	12	0
E12	8000	1	1	2	1	0	3	1	1
E13	20000	0	0	0	0	0	27	5	0
E14	10000	0	1	0	0	1	0	7	1
E15	8000	0	1	0	0	0	3	2	1

Métricas en aprendizaje automático

Cluster 1: 5 casos	
Salario :	21000
Casado :	No = 0.8
	Sí = 0.2
Coche :	No = 0.8
	Sí = 0.2
Hijos :	1
Vivienda :	Alquiler = 0.6
	Propiet = 0.4
Sindicato :	No = 0.4
	Sí = 0.6
Faltas/Año :	8
Antigüedad :	5
Sexo :	M = 0.6
	F = 0.4

Cluster 2: 6 casos	
Salario :	10167
Casado :	No = 0.33
	Si = 0.67
Coche :	No = 0.17
	Sí = 0.83
Hijos :	2 (algunos)
Vivienda :	Alquiler = 0.5
	Propiet = 0.5
Sindicato :	Sí = 0.5
	No = 0.5
Faltas/Año :	3
Antigüedad :	7
Sexo :	M = 1

Cluster 3: 4 casos	
Salario :	36250
Casado :	No = 0.5
	Sí = 0.5
Coche :	Sí = 1.0
Hijos :	2
Vivienda :	Alquiler = 0.75
	Propiet = 0.25
Sindicato :	No = 0.75
	Sí = 0.25
Faltas/Año :	5
Antigüedad :	14
Sexo :	M = 0.25
	F = 0.75

- **Grupo 1**. Empleados con salario promedio de $21000, con un hijo en promedio, solteros en su mayoría (80%) y sin coche (80%). No tienen vivienda propia en su mayoría (60%), son sindicalizados (60%), con muchas faltas (8 al año), una antigüedad promedio de 5 años y en su mayoría varones (60%).
- **Grupo 2**. Empleados con salario promedio de $10167, algunos con hijos, casados en su mayoría (67%) y con coche (83%). La mitad no tiene vivienda propia (50%), la mitad no son sindicalizados (50%), pocas faltas (3 al año), una antigüedad de 7 años y todos de sexo masculino (100%).
- **Grupo 3**. Empleados con salario promedio de $36250, solo uno con 2 hijos, la mitad casados (50%) y en su mayoría con coche (75%). La mayoría sin vivienda propia (75%), no sindicalizadas en su mayoría (75%), con 5 faltas en promedio y en su mayoría de sexo femenino (75%).

Métricas en aprendizaje automático

- Matemáticamente, una distancia es una función, $\boldsymbol{d(a,b)}$, que asigna un valor positivo a cada par de puntos de un espacio n-dimensional. Tiene las siguientes propiedades:

  - **No negativa**, el valor puede ser mayor o igual a cero: $\boldsymbol{d(a,b) \geq 0}$
  - **Simétrica**, la distancia entre *a* y *b* es la misma que entre *b* y *a*: $\boldsymbol{d(a,b) = d(b,a)}$
  - La distancia con el mismo punto es cero: $\boldsymbol{d(a,a) = 0}$

Métricas en aprendizaje automático

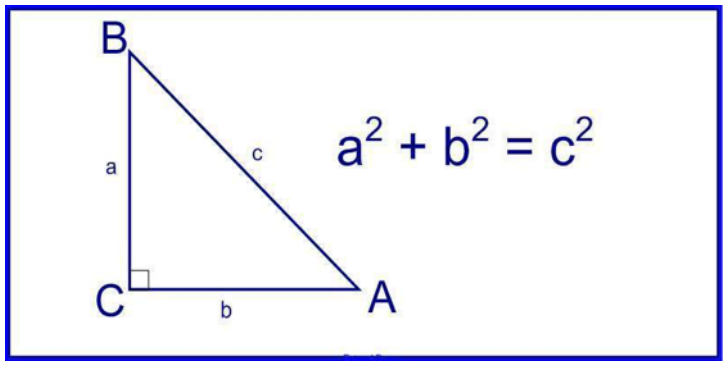
- Algunas métricas conocidas:

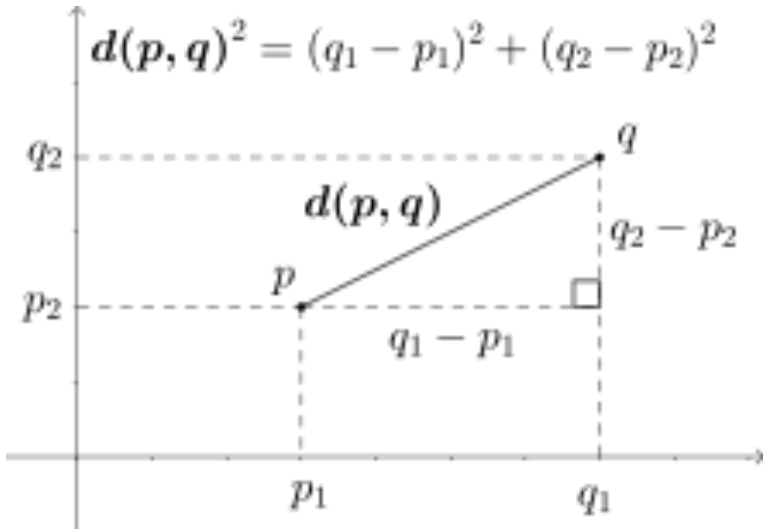
  - Distancia Euclidiana o Euclídea
  - Distancia de Chebyshev
  - Distancia de Manhattan o Geometría del taxista
  - Distancia de Minkowsky

Métodos para medir la similitud de elementos

1. Distancia Euclidiana

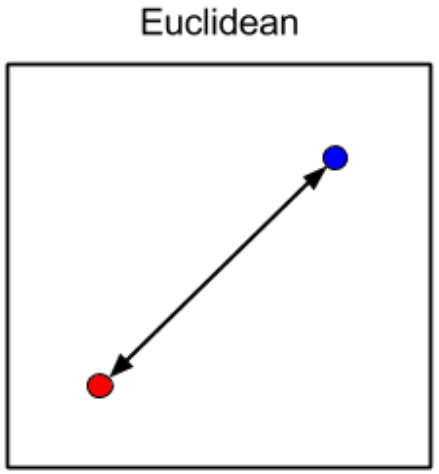
- **Distancia euclidiana** (euclídea, por Euclides) es una función usada para calcular la distancia entre dos puntos, conocida también como **espacio euclidiano**.
- Sus bases se encuentran en la aplicación del **Teorema de Pitágoras (métrica Pitagórica).**
- Donde la distancia euclidiana viene a ser la longitud de la **hipotenusa**.





1. Distancia Euclidiana

Dimensiones:



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

$$dist(p,q) = \sqrt{(p_1 - q_1)^2}$$ **1 dimensión**

$$dist(p,q) = \sqrt{(p_1 - q_1)^2 + (p_2 - q_2)^2}$$ **2 dimensiones**

$$dist(p,q) = \sqrt{(p_1 - q_1)^2 + (p_2 - q_2)^2 + (p_3 - q_3)^2}$$ **3 dimensiones**

$$dist(p,q) = \sqrt{(p_1 - q_1)^2 + (p_2 - q_2)^2 + (p_3 - q_3)^2 + \ldots + (p_n - q_n)^2}$$ ***n* dimensiones**

1. Distancia Euclidiana

Para el cálculo de las distancia euclidiana se utiliza: $dist(p,q) = \sqrt{\sum_{i=1}^{n}(p_i - q_i)^2}$

Ejemplo:

Sujeto	Lealtad a la tienda (x)	Lealtad a la marca (y)
A	3	2
B	4	5
C	4	7
D	2	7
E	6	6
F	7	7
G	6	4

$$dist(p,q) = d_{ij} = D_{(A,B)} = \sqrt{(3-4)^2 + (2-5)^2} = \sqrt{(-1)^2 + (-3)^2} = \sqrt{10} = \mathbf{3.16}$$

$$D_{(E,F)} = \sqrt{(6-7)^2 + (6-7)^2} = \sqrt{2} = 1.41$$

Matriz de similitudes
(distancias euclidianas)

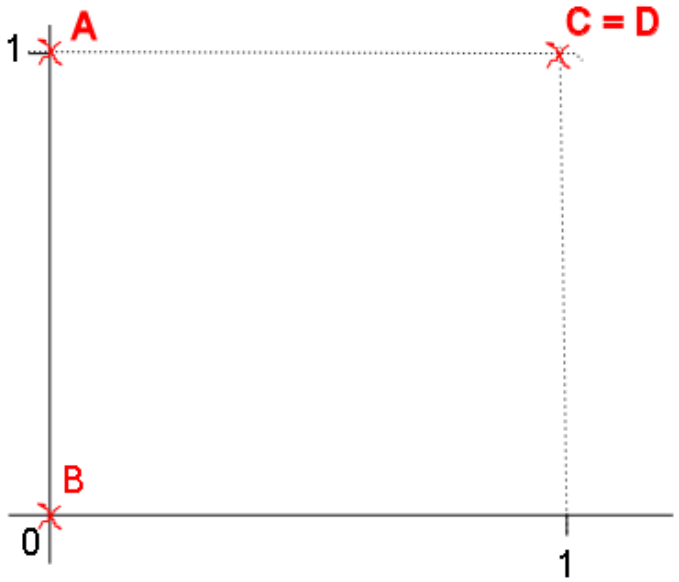
Sujetos	A	B	C	D	E	F	G
A	---						
B	3.16	---					
C	5.10	2.00	---				
D	5.10	2.83	2.00	---			
E	5.00	2.24	2.24	4.12	---		
F	6.40	3.61	3.00	5.00	1.41	---	
G	3.61	2.24	3.61	5.00	2.00	3.16	---

1. Distancia Euclidiana

¿Qué pasa cuando dos elementos iguales?

$$dist(p,q) = dij = \sqrt{\sum_{i=1}^{n}(p_i - q_i)^2}$$

ID	X	Y
A	0	1
B	0	0
C	1	1
D	1	1

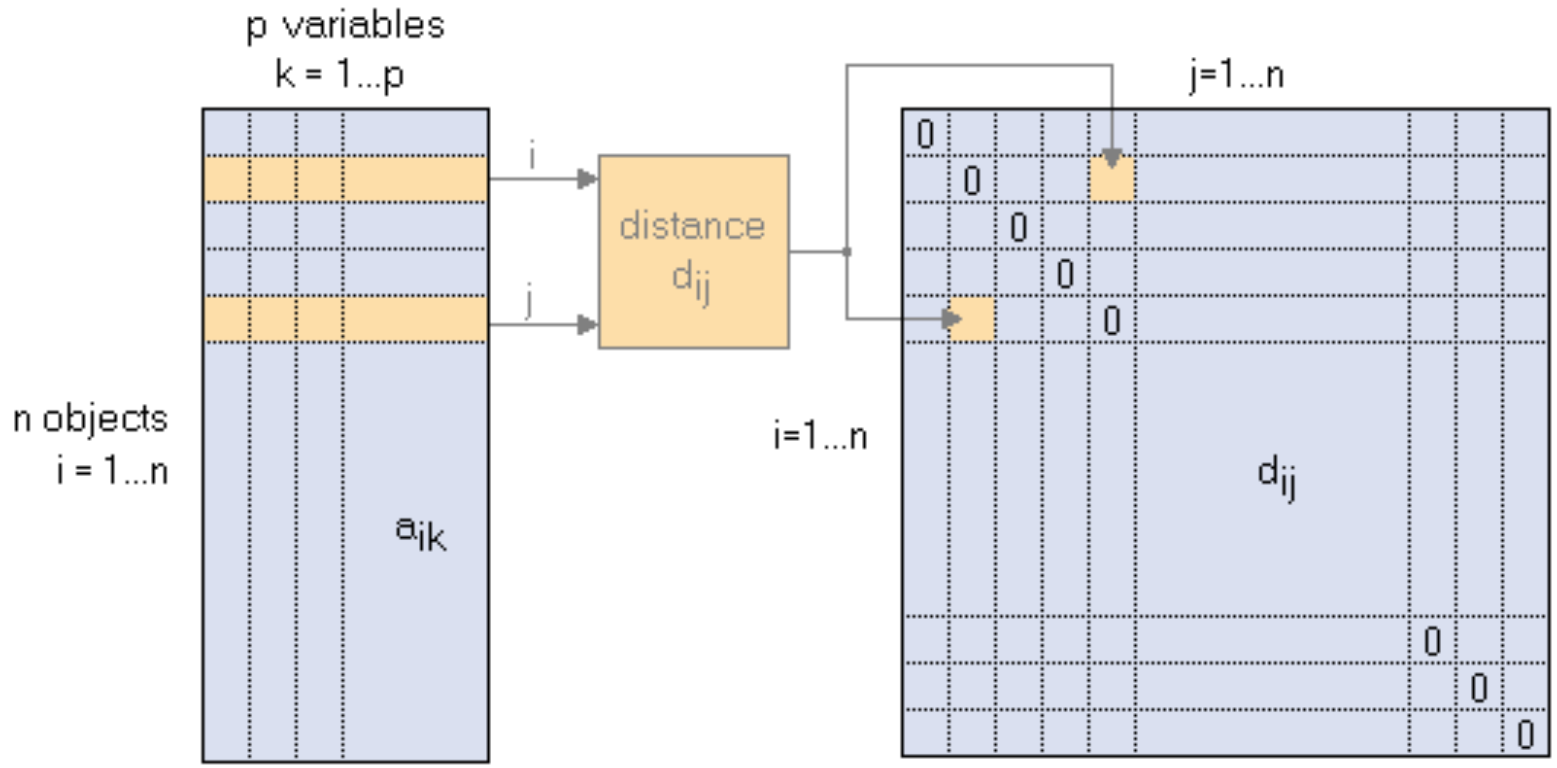


$$D_{(C,D)} = \sqrt{(1-1)^2 + (1-1)^2} = \sqrt{0} = \mathbf{0}$$

1. Distancia Euclidiana

Matriz de similitudes

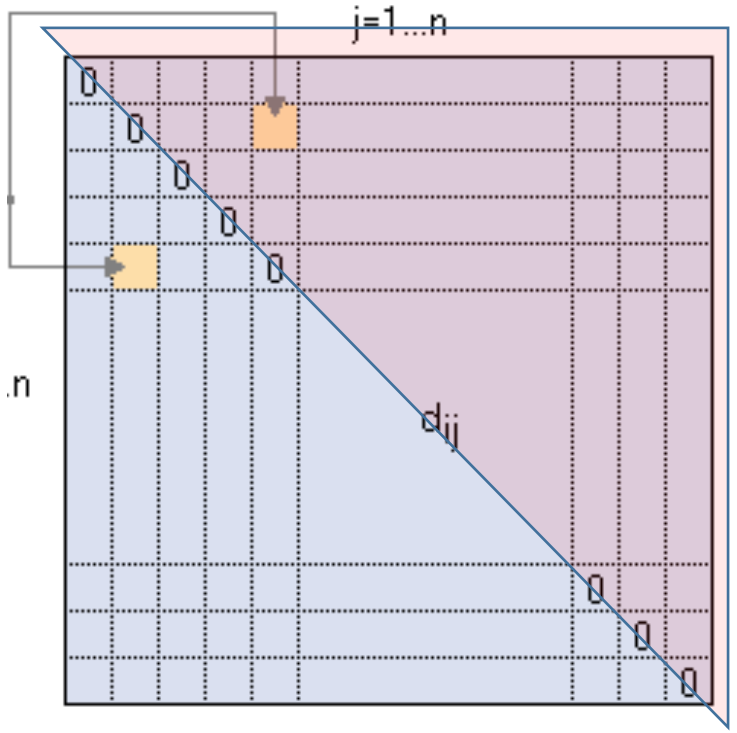
$$dist(p,q) = dij = \sqrt{\sum_{i=1}^{n}(p_i - q_i)^2}$$



1. Distancia Euclidiana

Matriz de similitudes

$$dist(p,q) = dij = \sqrt{\sum_{i=1}^{n}(p_i - q_i)^2}$$



1. Distancia Euclidiana

ID	Salario	Casado	Coche	Hijos	Vivienda	Sindicato	Faltas/Año	Antigüedad	Sexo
E1	10000	1	0	0	0	0	7	15	1
E2	20000	0	1	1	0	1	3	3	0
E3	15000	1	1	2	1	1	5	10	1
E4	30000	1	1	1	0	0	15	7	0
E5	10000	1	1	0	1	1	1	6	1
E6	40000	0	1	0	0	1	3	16	0
E7	25000	0	0	0	0	1	0	8	1
E8	20000	0	1	0	1	1	2	6	0
E9	20000	1	1	3	1	0	7	5	1
E10	30000	1	1	2	1	0	1	20	1
E11	45000	0	0	0	0	0	2	12	0
E12	8000	1	1	2	1	0	3	1	1
E13	20000	0	0	0	0	0	27	5	0
E14	10000	0	1	0	0	1	0	7	1
E15	8000	0	1	0	0	0	3	2	1

Obtener la matriz de distancias

$$dist(p,q) = dij = \sqrt{\sum_{i=1}^{n}(p_i - q_i)^2}$$

1. Distancia Euclidiana

ID	Salario	Casado	Coche	Hijos	Vivienda	Sindicato	Faltas/Año	Antigüedad	Sexo
E1	10000	1	0	0	0	0	7	15	1
E2	20000	0	1	1	0	1	3	3	0
E3	15000	1	1	2	1	1	5	10	1
E4	30000	1	1	1	0	0	15	7	0
E5	10000	1	1	0	1	1	1	6	1
E6	40000	0	1	0	0	1	3	16	0
E7	25000	0	0	0	0	1	0	8	1
E8	20000	0	1	0	1	1	2	6	0
E9	20000	1	1	3	1	0	7	5	1
E10	30000	1	1	2	1	0	1	20	1
E11	45000	0	0	0	0	0	2	12	0
E12	8000	1	1	2	1	0	3	1	1
E13	20000	0	0	0	0	0	27	5	0
E14	10000	0	1	0	0	1	0	7	1
E15	8000	0	1	0	0	0	3	2	1

Obtener la matriz de distancias

$$dist(p,q) = dij = \sqrt{\sum_{i=1}^{n}(p_i - q_i)^2}$$

$$dist_{(E1,\,E2)} = \sqrt{(10000-20000)^2 + (1-0)^2 + (0-1)^2 + (0-1)^2 + (0-0)^2 + (0-1)^2 + (7-3)^2 + (15-3)^2 + (1-0)^2}$$

$$dist_{(E1,\,E2)} = \sqrt{(-10000)^2 + (1)^2 + (-1)^2 + (-1)^2 + (0)^2 + (-1)^2 + (4)^2 + (12)^2 + (1)^2}$$

$$dist_{(E1,\,E2)} = \sqrt{100000000 + 1 + 1 + 1 + 0 + 1 + 16 + 144 + 1} = \sqrt{100000165} = 10000.008$$

1. Distancia Euclidiana

Obtener la matriz de distancias

ID	Salario	Casado	Coche	Hijos	Vivienda	Sindicato	Faltas/Año	Antigüedad	Sexo
E1	10000	1	0	0	0	0	7	15	1
E2	20000	0	1	1	0	1	3	3	0

$$dist(p,q) = dij = \sqrt{\sum_{i=1}^{n}(p_i - q_i)^2}$$

```
from math import sqrt
E1 = (10000,1,0,0,0,0,7,15,1) #datos del punto 1
E2 = (20000,0,1,1,0,1,3,3,0)  #datos del punto 2
#La función zip() es un iterador de tuplas
dst1 = sqrt(sum((E1-E2)**2 for E1, E2 in zip(E1, E2)))
dst1
```

```
10000.008249996597
```

1. Distancia Euclidiana

Obtener la matriz de distancias

ID	Salario	Casado	Coche	Hijos	Vivienda	Sindicato	Faltas/Año	Antigüedad	Sexo
E1	10000	1	0	0	0	0	7	15	1
E2	20000	0	1	1	0	1	3	3	0

$$dist(p,q) = dij = \sqrt{\sum_{i=1}^{n}(p_i - q_i)^2}$$

```
import numpy as np
import matplotlib as plt
E1 = np.array([10000,1,0,0,0,0,7,15,1])
E2 = np.array([20000,0,1,1,0,1,3,3,0])
dst2 = np.sqrt(np.sum((E1-E2)**2))
dst2
```

```
10000.008249996597
```

1. Distancia Euclidiana

Obtener la matriz de distancias

ID	Salario	Casado	Coche	Hijos	Vivienda	Sindicato	Faltas/Año	Antigüedad	Sexo
E1	10000	1	0	0	0	0	7	15	1
E2	20000	0	1	1	0	1	3	3	0

$$dist(p,q) = dij = \sqrt{\sum_{i=1}^{n}(p_i - q_i)^2}$$

```
from scipy.spatial import distance
E1 = (10000,1,0,0,0,0,7,15,1)
E2 = (20000,0,1,1,0,1,3,3,0)
dst3 = distance.euclidean(E1,E2)
dst3
```

```
10000.008249996597
```

1. Distancia Euclidiana

Obtener la matriz de distancias

	0	1	...	13	14
0	0.000000	10000.008250	...	10.770330	2000.046749
1	10000.008250	0.000000	...	10000.001350	12000.000167
2	5000.003600	5000.005700	...	5000.004000	7000.005357
3	20000.003275	10000.008100	...	20000.005725	22000.003909
4	10.954451	10000.000850	...	2.000000	2000.005750
5	30000.000350	20000.004250	...	30000.001517	32000.003094
6	15000.003333	5000.003700	...	15000.000067	17000.001382
7	10000.005550	3.464102	...	10000.000350	12000.000833
8	10000.005550	5.291503	...	10000.003250	12000.001500
9	20000.001675	10000.014900	...	20000.004425	22000.007591
10	35000.000514	25000.001700	...	35000.000457	37000.001392
11	2000.054499	12000.000375	...	2000.013000	2.645751
12	10000.025100	24.145393	...	10000.036800	12000.024458
13	10.770330	10000.001350	...	0.000000	2000.008750
14	2000.046749	12000.000167	...	2000.008750	0.000000

1. Distancia Euclidiana

Obtener la matriz de distancias

	1	2	3	4	5	6	7	8	9	10	11	12	13	14
2	10000.008250													
3	5000.003600	5000.005700												
4	20000.003275	10000.008100	15000.003767											
5	10.954451	10000.000850	5000.003600	20000.005025										
6	30000.000350	20000.004250	25000.000940	10000.011400	30000.001783									
7	15000.003333	5000.003700	10000.001800	5000.023100	15000.000267	15000.002500								
8	10000.005550	3.464102	5000.003100	10000.008700	10000.000150	20000.002550	5000.001100							
9	10000.005550	5.291503	5000.003100	10000.003700	10000.002350	20000.003750	5000.007100	6.164414						
10	20000.001675	10000.014900	15000.003900	19.183326	20000.005025	10000.001400	5000.015300	10000.010200	10000.013100					
11	35000.000514	25000.001700	30000.000367	15000.006567	35000.000600	5000.001900	20000.000550	25000.000780	25000.001740	15000.002433				
12	2000.054499	12000.000375	7000.006143	22000.004159	2000.008500	32000.003641	17000.001941	12000.001375	12000.001375	22000.008295	37000.001757			
13	10000.025100	24.145393	5000.051800	10000.007550	10000.034100	20000.017475	5000.073999	25.079872	20.322401	10000.045450	25000.013480	12000.025000		
14	10.770330	10000.001350	5000.004000	20000.005725	2.000000	30000.001517	15000.000067	10000.000350	10000.003250	20000.004425	35000.000457	2000.013000	10000.036800	
15	2000.046749	12000.000167	7000.005357	22000.003909	2000.005750	32000.003094	17000.001382	12000.000833	12000.001500	22000.007591	37000.001392	2.645751	12000.024458	2000.008750

1. Distancia Euclidiana

La distancia euclidiana, a pesar de su sencillez de cálculo, tiene un inconveniente:

- Si las variables utilizadas están correlacionadas, entonces estas variables darán información, en gran medida, redundante.
- Como consecuencia, la distancia euclidiana aumentará la diferencia entre los elementos.

2. Distancia de Chebyshev

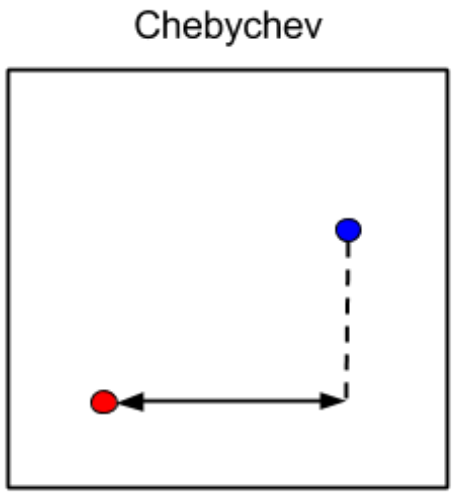
2. Distancia de Chebyshev

- La **distancia de Chebyshev** entre puntos es el valor máximo absoluto de las diferencias entre las coordenadas de un par de elementos.
- Lleva el nombre del matemático ruso Pafnuty Chebyshev, conocido por su trabajo la geometría analítica y teoría de números.
- Otros nombres para la distancia de Chebyshev son métrica máxima.

ID	X	Y	Z
A	2	3	4
B	5	9	11

$$d_{Cheb}(p,q) = \max|p_i - q_i|$$

$$d_{Cheb}(A,B) = \max\{|2-5|, |3-9|, |4-11|\} = \max\{3,6,7\} = \mathbf{7}$$



Se utiliza en la programación de movimientos de robots industriales.

2. Distancia de Chebyshev

Datos de empleados

ID	Salario	Casado	Coche	Hijos	Vivienda	Sindicato	Faltas/Año	Antigüedad	Sexo
E1	10000	1	0	0	0	0	7	15	1
E2	20000	0	1	1	0	1	3	3	0

$$d_{Cheb}(p,q) = \max |p_i - q_i|$$

```
from scipy.spatial import distance
E1 = (10000,1,0,0,0,0,7,15,1)
E2 = (20000,0,1,1,0,1,3,3,0)
dst = distance.chebyshev(E1,E2)
dst
```

```
10000
```

3. Distancia de Manhattan

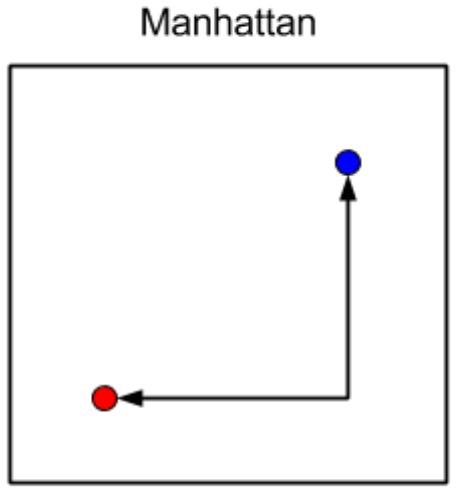
3. Distancia de Manhattan

- La **distancia euclidiana** es una buena métrica. Sin embargo, en la vida real, por ejemplo en una ciudad, es imposible moverse de un punto a otro de manera directa.
- La **distancia de Manhattan** es útil para calcular la distancia entre dos puntos en una enorme cuadrícula.
- Se llama Manhattan debido al diseño de cuadrícula de la mayoría de las calles de la isla de Manhattan.

3. Distancia de Manhattan

- La **distancia de Manhattan** también se conoce como geometría del taxi, distancia de la manzana de la ciudad, y distancia rectilínea.

ID	X	Y	Z
A	2	3	4
B	5	9	11



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

$$d_{Manh}(A,B) = |x_2 - x_1| + |y_2 - y_1| + |z_2 - z_1|$$

$$d_{Manh}(A,B) = |5-2| + |9-3| + |11-4| = 3 + 6 + 7 = \mathbf{16}$$

3. Distancia de Manhattan

Datos de empleados

ID	Salario	Casado	Coche	Hijos	Vivienda	Sindicato	Faltas/Año	Antigüedad	Sexo
E1	10000	1	0	0	0	0	7	15	1
E2	20000	0	1	1	0	1	3	3	0

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

```
from scipy.spatial import distance
E1 = (10000,1,0,0,0,0,7,15,1)
E2 = (20000,0,1,1,0,1,3,3,0)
dst = distance.cityblock(E1,E2)
dst
```

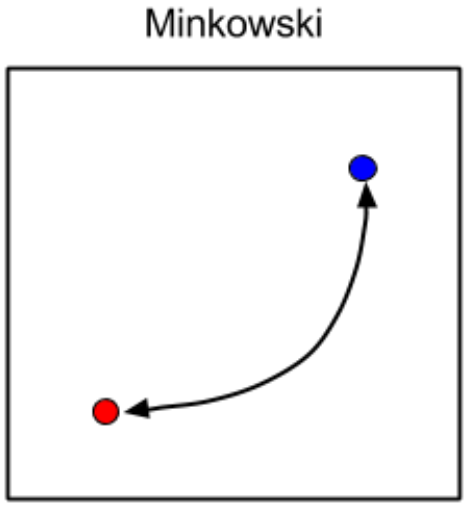
```
10021
```

4. Distancia de Minkowski

4. Distancia de Minkowski

- La **distancia de Minkowski** es una distancia entre dos puntos en el espacio n-dimensional. Es una generalización de las distancias Euclidiana, Manhattan y Chebyshev.
- Se llama Manhattan debido al diseño de cuadrícula de la mayoría de las calles de la isla de Manhattan.

$$d_{Mink}(q,p) = \lambda\sqrt{\sum_{i=1}^{n}(q_i - p_i)^{\lambda}} = \left(\sum_{i=1}^{n}(q_i - p_i)^{\lambda}\right)^{1/\lambda}$$



donde λ es el orden para calcular la distancia de tres formas diferentes:

- $\lambda = 1$, distancia de Manhattan (métrica L^1)
- $\lambda = 2$, distancia Euclidiana (métrica L^2)
- $\lambda = \infty$, distancia de Chebyshev (métrica L)
- Los valores intermedios de λ, por ejemplo, $\lambda = 1.5$, proporcionan un equilibrio entre las medidas.

4. Distancia de Minkowski

Datos de empleados

ID	Salario	Casado	Coche	Hijos	Vivienda	Sindicato	Faltas/Año	Antigüedad	Sexo
E1	10000	1	0	0	0	0	7	15	1
E2	20000	0	1	1	0	1	3	3	0

$$d_{Mink}(q,p) = \lambda \sqrt{\sum_{i=1}^{n} (q_i - p_i)^{\lambda}}$$

```
from scipy.spatial import distance
E1 = (10000,1,0,0,0,0,7,15,1)
E2 = (20000,0,1,1,0,1,3,3,0)
dst = distance.minkowski(E1,E2)
dst
```

```
10000.008249996597
```

Distancias

```
from scipy.spatial import distance
E1 = (10000,1,0,0,0,0,7,15,1)
E2 = (20000,0,1,1,0,1,3,3,0)
dst3 = distance.euclidean(E1,E2)
dst3
```

```
10000.008249996597
```

```
from scipy.spatial import distance
E1 = (10000,1,0,0,0,0,7,15,1)
E2 = (20000,0,1,1,0,1,3,3,0)
dst = distance.chebyshev(E1,E2)
dst
```

```
10000
```

```
from scipy.spatial import distance
E1 = (10000,1,0,0,0,0,7,15,1)
E2 = (20000,0,1,1,0,1,3,3,0)
dst = distance.cityblock(E1,E2)
dst
```

```
10021
```

```
from scipy.spatial import distance
E1 = (10000,1,0,0,0,0,7,15,1)
E2 = (20000,0,1,1,0,1,3,3,0)
dst = distance.minkowski(E1,E2)
dst
```

```
10000.008249996597
```

Ejemplo

Obtención de la matriz de distancia

Ejemplo

```
DatosEmp <- read.table("/Users/guille/Documents/1 FI-UNAM/1 Cursos/2021-1/1 IA2021-1/2 CasosPracticos/3 Similitudes/Empleados.txt", header=T, sep="\t")

DatosEmp
```

	ID	Salario	Casado	Coche	Hijos	Vivienda	Sindicato	FaltasAno	Antiguedad	Sexo
1	E1	10000	1	0	0	0	0	7	15	1
2	E2	20000	0	1	1	0	1	3	3	0
3	E3	15000	1	1	2	1	1	5	10	1
4	E4	30000	1	1	1	0	0	15	7	0
5	E5	10000	1	1	0	1	1	1	6	1
6	E6	40000	0	1	0	0	1	3	16	0
7	E7	25000	0	0	0	0	1	0	8	1
8	E8	20000	0	1	0	1	1	2	6	0
9	E9	20000	1	1	3	1	0	7	5	1
10	E10	30000	1	1	2	1	0	1	20	1
11	E11	45000	0	0	0	0	0	2	12	0
12	E12	8000	1	1	2	1	0	3	1	1
13	E13	20000	0	0	0	0	0	27	5	0
14	E14	10000	0	1	0	0	1	0	7	1
15	E15	8000	0	1	0	0	0	3	2	1

Ejemplo

Distancias <- dist(DatosEmp[2:10], method = "euclidean”)

Distancias

```
             1           2           3           4           5           6           7           8           9          10          11          12          13          14
2  10000.008250
3   5000.003600  5000.005700
4  20000.003275 10000.008100 15000.003767
5     10.954451 10000.000850  5000.003600 20000.005025
6  30000.000350 20000.004250 25000.000940 10000.011400 30000.001783
7  15000.003333  5000.003700 10000.001800  5000.023100 15000.000267 15000.002500
8  10000.005550     3.464102  5000.003100 10000.008700 10000.000150 20000.002550  5000.001100
9  10000.005550     5.291503  5000.003100 10000.003700 10000.002350 20000.003750  5000.007100     6.164414
10 20000.001675 10000.014900 15000.003900    19.183326 20000.005025 10000.001400  5000.015300 10000.010200 10000.013100
11 35000.000514 25000.001700 30000.000367 15000.006567 35000.000600  5000.001900 20000.000550 25000.000780 25000.001740 15000.002433
12  2000.054499 12000.000375  7000.006143 22000.004159  2000.008500 32000.003641 17000.001941 12000.001375 12000.001375 22000.008295 37000.001757
13 10000.025100    24.145393  5000.051800 10000.007550 10000.034100 20000.017475  5000.073999    25.079872    20.322401 10000.045450 25000.013480 12000.025000
14    10.770330 10000.001350  5000.004000 20000.005725     2.000000 30000.001517 15000.000067 10000.000350 10000.003250 20000.004425 35000.000457  2000.013000 10000.036800
15  2000.046749 12000.000167  7000.005357 22000.003909  2000.005750 32000.003094 17000.001382 12000.000833 12000.001500 22000.007591 37000.001392     2.645751 12000.024458  2000.008750
```

Ejemplo

```
Distancias <- round(dist(DatosEmp[2:10], method = "euclidean"), 2)

Distancias
```

```
          1        2        3        4        5        6        7        8        9       10       11       12       13       14
2  10000.01
3   5000.00  5000.01
4  20000.00 10000.01 15000.00
5     10.95 10000.00  5000.00 20000.01
6  30000.00 20000.00 25000.00 10000.01 30000.00
7  15000.00  5000.00 10000.00  5000.02 15000.00 15000.00
8  10000.01     3.46  5000.00 10000.01 10000.00 20000.00  5000.00
9  10000.01     5.29  5000.00 10000.00 10000.00 20000.00  5000.01     6.16
10 20000.00 10000.01 15000.00    19.18 20000.01 10000.00  5000.02 10000.01 10000.01
11 35000.00 25000.00 30000.00 15000.01 35000.00  5000.00 20000.00 25000.00 25000.00 15000.00
12  2000.05 12000.00  7000.01 22000.00  2000.01 32000.00 17000.00 12000.00 12000.00 22000.01 37000.00
13 10000.03    24.15  5000.05 10000.01 10000.03 20000.02  5000.07    25.08    20.32 10000.05 25000.01 12000.02
14    10.77 10000.00  5000.00 20000.01     2.00 30000.00 15000.00 10000.00 10000.00 20000.00 35000.00  2000.01 10000.04
15  2000.05 12000.00  7000.01 22000.00  2000.01 32000.00 17000.00 12000.00 12000.00 22000.01 37000.00     2.65 12000.02  2000.01
```

Ejemplo

```
Distancias <- round(dist(DatosEmp[2:10], method = "maximum"), 2)

Distancias
```

```
      1     2     3     4     5     6     7     8     9    10    11    12    13    14
2  10000
3   5000  5000
4  20000 10000 15000
5      9 10000  5000 20000
6  30000 20000 25000 10000 30000
7  15000  5000 10000  5000 15000 15000
8  10000     3  5000 10000 10000 20000  5000
9  10000     4  5000 10000 10000 20000  5000     5
10 20000 10000 15000    14 20000 10000  5000 10000 10000
11 35000 25000 30000 15000 35000  5000 20000 25000 25000 15000
12  2000 12000  7000 22000  2000 32000 17000 12000 12000 22000 37000
13 10000    24  5000 10000 10000 20000  5000    25    20 10000 25000 12000
14     8 10000  5000 20000     1 30000 15000 10000 10000 20000 35000  2000 10000
15  2000 12000  7000 22000  2000 32000 17000 12000 12000 22000 37000     2 12000  2000
```

Ejemplo

```
Distancias <- round(dist(DatosEmp[2:10], method = "manhattan"), 2)

Distancias
```

```
       1     2     3     4     5     6     7     8     9    10    11    12    13    14
2  10021
3   5012  5013
4  20019 10018 15017
5     18 10009  5010 20019
6  30009 20014 25013 10024 30015
7  15016  5011 10012  5021 15006 15013
8  10019     6  5011 10018 10003 20012  5007
9  10015    12  5009 10014 10011 20022  5017    12
10 20015 10024 15015    30 20017 10012  5019 10020 10022
11 35010 25013 30012 15021 35012  5007 20008 25009 25019 15015
12  2022 12007  7012 22021  2010 32021 17016 12011 12009 22021 37018
13 10032    29  5034 10017 10032 20037  5032    29    27 10047 25032 12034
14    18 10009  5012 20019     4 30013 15002 10005 10015 20019 35010  2014 10032
15  2019 12004  7015 22020  2009 32016 17011 12008 12012 22024 37013     5 12029  2009
```

Ejemplo

```
Distancias <- round(dist(DatosEmp[2:10], method = "minkowski"), 2)

Distancias
```

```
          1        2        3        4        5        6        7        8        9       10       11       12       13       14
2  10000.01
3   5000.00  5000.01
4  20000.00 10000.01 15000.00
5     10.95 10000.00  5000.00 20000.01
6  30000.00 20000.00 25000.00 10000.01 30000.00
7  15000.00  5000.00 10000.00  5000.02 15000.00 15000.00
8  10000.01     3.46  5000.00 10000.01 10000.00 20000.00  5000.00
9  10000.01     5.29  5000.00 10000.00 10000.00 20000.00  5000.01     6.16
10 20000.00 10000.01 15000.00    19.18 20000.01 10000.00  5000.02 10000.01 10000.01
11 35000.00 25000.00 30000.00 15000.01 35000.00  5000.00 20000.00 25000.00 25000.00 15000.00
12  2000.05 12000.00  7000.01 22000.00  2000.01 32000.00 17000.00 12000.00 12000.00 22000.01 37000.00
13 10000.03    24.15  5000.05 10000.01 10000.03 20000.02  5000.07    25.08    20.32 10000.05 25000.01 12000.02
14    10.77 10000.00  5000.00 20000.01     2.00 30000.00 15000.00 10000.00 10000.00 20000.00 35000.00  2000.01 10000.04
15  2000.05 12000.00  7000.01 22000.00  2000.01 32000.00 17000.00 12000.00 12000.00 22000.01 37000.00     2.65 12000.02  2000.01
```

Lecturas complementarias de posible interés

1) Mathematics and Artificial Intelligence, two branches of the same tree
URL: https://www.sciencedirect.com/science/article/pii/S1877042810002004

2) Why is Mathematics Vital to Thrive In Your AI Career
URL: https://towardsdatascience.com/why-is-mathematics-vital-to-thrive-in-your-ai-career-c11bd8446ddc