5 Tipos de Datos y Medidas de disimilaridad

5.1 Distancia o Métrica

Una función de distancia debe cumplir los siguientess requisitos matemáticos, para todo i, j, h

(D1) $d(i, j) \geq 0$

(D2) $d(i, i) = 0$

(D3) $d(i, j) = d(j, i)$

(D4) $d(i, j) \leq d(i, h) + d(h, j)$

5.2 Disimilaridades

Un índice de disimilaridad debe por lo menos cumplir **(D1)**, **(D2)**, **(D3)**, si además cumple la **desigualdad triangular (D4)** se dice que es una métrica

5.2.1 Similaridad

$$d(i,j) = 1 - s(i,j)$$

Un coeficiente de similaridad vale 0 si i y j NO están próximos o se parecen.

5.2.2 Cálculo de Disimilaridades

Pueden ser obtenidas de varias maneras. Desde las variables: cuantitativas, binarias, nominales u ordinales.

También pueden provenir de *rankings* subjetivos acerca de cuanto los objetos difieren unos de otros.

5.2.3 Ejemplo

14 estudiantes de posgrado en economía(de diferentes partes del mundo) llenan una matriz con coeficientes entre 0(identicas) y 10(muy diferentes) que expresan diferencias subjetivas entre 11 disciplinas científicas.

Astronomía	0.00										
Biología	7.86	0.00									
Química	6.50	2.93	0.00								
Informática	5.00	6.86	6.50	0.00							
Economía	8.00	8.14	8.21	4.79	0.00						
Geografía	4.29	7.00	7.64	7.71	5.93	0.00					
Historia	8.07	8.14	8.71	8.57	5.86	3.86	0.00				
Matemática	3.64	7.14	4.43	1.43	3.57	7.07	9.07	0.00			
Medicina	8.21	2.50	2.93	6.36	8.43	7.86	8.43	6.29	0.00		
Física	2.71	5.21	4.57	4.21	8.36	7.29	8.64	2.21	5.07	0.00	
Sicología	9.36	5.57	7.29	7.21	6.86	8.29	7.64	8.71	3.79	8.64	0.0

5.3 Variables Binarias

Las variables dicotómicas tienen dos posibles resultados o estados('+', '-').

$$x_{if} = \begin{cases} 1 & \text{si } A \\ 0 & \text{si } \bar{A} \end{cases}$$

5.3.1 Tabla de Contingencia

objeto j

	1	0	
1	a	b	$a+b$
0	c	d	$c+d$
	$a+c$	$b+d$	p

a número de variables en que ambos objetos valen **+**

b número de variables en que el j-ésimo vale **+**

c número de variables en que el i-ésimo vale **-**

d número de variables en que ambos objetos valen **-**

p $= a + b + c + d$

Los valores **a, b, c, d** son combinados en un coeficiente que describe cuan cercanos están los objetos i y j de acuerdo a esas variables binarias.

Existen dos tipos de variables binarias

simétricas Los dos estados('+','-') tienen el mismo peso. Ejemplo: hombre—mujer

asimétricas Sus resultados nos son igualmente importantes. Ejemplo: presencia—ausencia de un atributo relativamente raro

Para enfrentar el caso de variables binarias simétricas se utilizan Medidas de Similaridad Invariantes, cuyo resultados no cambian aunque alguna de las variables sean codificadas de forma diferente.

5.3.2 Simple Matching Coefficient

$$d(i,j) = \frac{b+c}{p}$$

En el caso de variables binarias asimétricas se utiliza un coeficiente no invariante.

5.3.3 Coeficiente de Jaccard

$$d(i,j) = \frac{b+c}{a+b+c}$$

5.4 Ejemplo

Datos Originales

	V1	V2	V3	V4	V5	V6	V7	V8	V9	V10
aa_1	1	1	1	1	1	1	1	1	1	1
aa_2	1	0	1	0	0	0	1	0	1	0
aa_3	0	0	1	1	0	1	1	1	0	0
aa_4	0	1	1	1	1	0	1	0	0	1
aa_5	1	1	1	1	0	1	0	1	1	1
aa_6	0	1	0	0	0	1	1	0	1	1
aa_7	0	1	1	1	0	1	1	0	1	0
aa_8	0	1	1	1	1	1	1	0	0	0
aa_9	1	1	1	0	1	0	1	0	0	0
aa_10	0	0	0	0	1	1	1	1	1	0
aa_11	0	0	0	0	1	1	1	1	0	0
aa_12	1	0	0	0	0	1	0	0	1	0
aa_13	0	1	0	1	0	1	0	0	1	0
aa_14	0	1	1	1	1	1	1	0	1	1
aa_15	0	1	0	0	0	1	1	1	0	1
aa_16	0	0	1	1	0	1	1	1	0	0
aa_17	1	1	1	1	1	1	1	1	1	1

Matriz de Disimilaridades

	aa_1	aa_2	aa_3	aa_4	aa_5	aa_6	aa_7	aa_8	aa_9	aa_10
aa_1	0.0000									
aa_2	0.6000	0.0000								
aa_3	0.5000	0.7143	0.0000							
aa_4	0.4000	0.7500	0.6250	0.0000						
aa_5	0.2000	0.6667	0.5556	0.6000	0.0000					
aa_6	0.5000	0.7143	0.7500	0.6250	0.5556	0.0000				
aa_7	0.4000	0.5714	0.4286	0.5000	0.4444	0.4286	0.0000			
aa_8	0.4000	0.7500	0.4286	0.2857	0.6000	0.6250	0.2857	0.0000		
aa_9	0.5000	0.5000	0.7500	0.4286	0.7000	0.7500	0.6250	0.4286	0.0000	
aa_10	0.5000	0.7143	0.5714	0.7778	0.7000	0.5714	0.6250	0.6250	0.7500	0.0000
aa_11	0.6000	0.8571	0.5000	0.7500	0.8000	0.7143	0.7500	0.5714	0.7143	0.2000
aa_12	0.7000	0.6000	0.8571	1.0000	0.6250	0.6667	0.7143	0.8750	0.8571	0.6667
aa_13	0.6000	0.8571	0.7143	0.7500	0.5000	0.5000	0.3333	0.5714	0.8750	0.7143
aa_14	0.2000	0.6667	0.5556	0.2500	0.4000	0.3750	0.2500	0.2500	0.5556	0.5556
aa_15	0.5000	0.8750	0.5714	0.6250	0.5556	0.3333	0.6250	0.6250	0.7500	0.5714
aa_16	0.5000	0.7143	0.0000	0.6250	0.5556	0.7500	0.4286	0.4286	0.7500	0.5714
aa_17	0.0000	0.6000	0.5000	0.4000	0.2000	0.5000	0.4000	0.4000	0.5000	0.5000

5.5 Variables Nominales

Las variables nominales toman mas de dos estados

$$x_{if} = \begin{cases} 1 & \text{si } A_1 \\ 2 & \text{si } A_2 \\ \vdots & \text{si } \vdots \\ M & \text{si } A_M \end{cases}$$

Ejemplos: Nacionalidad, Estado Conyugal, colores de ojos

5.5.1 Simple Matching

$$d(i, j) = \frac{p - u}{p}$$

u número de coincidencias (i,j pertenecen al mismo estado)

5.6 Variables Ordinales

Los M estados tienen un orden y un sentido sequencial.
A veces se obtienen de discretizar variables cuantitativas dividiendo un eje continuo en un número finito de clases.

$$x_{if} = \begin{cases} 1 & \text{detesta} \\ 2 & \text{disgusta} \\ 3 & \text{indiferente} \\ 4 & \text{gusta} \\ 5 & \text{encanta} \end{cases}$$

- Tratar los rangos como continuas
- Tratarlas como "continuas" en el intervalo $[0, 1]$. Reemplazando los rangos r_{if} del objeto $i - esimo$ en la $f - esima$ variable por

$$z_{if} = \frac{r_{if} - 1}{M_f - 1}$$

donde M_f es el mayor valor de la variable f

$\Rightarrow$ Calcular disimilaridades utilizando la Distancia Euclidea

5.7 Variables Cuantitativas

Sea x_{ik} el valor de la $k-esima$ variable cuantitativa tomado por el objeto $i-esimo$. $(i=1,\dots n)$ y $(k=1,\dots,p)$

Una familia de medidas de disimilaridad, indexada por el parametro λ

5.7.1 Metrica de Minkowski

$$d(i,j)=\left(\sum_{k=1}^{p} w_k^{\lambda}|x_{ik}-x_{jk}|^{\lambda}\right)^{1/\lambda}$$

donde $\{w_k\}$ con $k=1,\dots,p$ son pesos no-negativos asociados con las p variables.

5.7.2 Metrica de Manhattan (city block)

Caso particular de Minkowski con $\lambda = 1$.

$$d(i,j) = \sum_{k=1}^{p} w_k |x_{ik} - x_{jk}|$$

5.7.3 Distancia Euclídea

Caso particular de Minkowski con $\lambda = 2$

$$d(i,j) = \sqrt{\sum_{k=1}^{p} w_k^2 (x_{ik} - x_{jk})^2}$$

5.7.4 Metrica de Canberra

Tiene incorporada una tipificación

$$d(i,j) = \sum_{k=1}^{p} \frac{|x_{ik} - x_{jk}|}{(|x_{ik}| + |x_{jk}|)}$$

definiendo $d(i,j) = 0$ si $x_{ik} = 0 = x_{jk}$.

Su sensibilidad en valores cercanos a 0 la hace una posible generalización de las binarias.