Capítulo 4

Medidas de Proximidad

$X_1, \ldots, X_p$ variables

$A = (a_1, \ldots, a_p)$ valores de $X_1, \ldots, X_p$ para el individuo A
$B = (b_1, \ldots, b_p)$ valores de $X_1, \ldots, X_p$ para el individuo B

$$\text{Proximidades} \begin{cases} \text{Disimilaridades: } \delta(A,B) & \rightarrow \text{ Distancias: } d(A,B) \\ \text{Similaridades: } \xi(A,B) & \rightarrow \text{ Similitudes: } s(A,B) \end{cases}$$

4.1. VARIABLES CUANTITATIVAS

Distancia: Una *distancia* es una aplicación $d : I\!R^p \times I\!R^p \rightarrow I\!R$ que verifica:

(i) $d(A,B) = 0 \Leftrightarrow A = B$,

(ii) $d(A,B) + d(C,B) \geq d(A,C)$.

La distancia mide lo "lejanos" que están dos puntos.

Teorema 4.1 Para cualquier par $A, B \in I\!R^p$, se verifica

(iii) $d(A,B) \geq 0$,

(iv) $d(A,B) = d(B,A)$.

Por contra, una medida de similitud mide "lo cercanos" que están dos puntos.

Similitud: Una *similitud* es una aplicación $s : \mathbb{R}^p \times \mathbb{R}^p \to \mathbb{R}$ que verifica:

(i) $0 \leq s(A, B) \leq 1$.

(ii) $s(A, B) = 1 \Leftrightarrow A$ y B son iguales.

(iii) $s(A, B) = s(B, A)$.

Distancia euclídea o L_2: La *distancia euclídea* entre dos puntos $A = (a_1, \ldots, a_p)$ y $B = (b_1, \ldots, b_p)$ es el módulo del vector que une A con B, es decir,

$$d_e(A, B) = |B - A| = \sqrt{(B - A)'(B - A)} = \sqrt{\sum_{k=1}^{p} (b_k - a_k)^2}\,.$$

Distancia L_1, de Manhattan o city block:

$$d_M(A, B) = \sum_{k=1}^{p} |b_k - a_k|\,.$$

Distancia de Minkowski:

$$d_m(A, B) = \left(\sum_{k=1}^{p} |b_k - a_k|^m \right)^{1/m}, \quad m \in \mathbb{N}.$$

Ejemplo 4.1 La siguiente tabla de contingencia representa la distribución conjunta de 230 personas respecto a las variables *Estado Civil* y *Sexo*.

	Mujer	Hombre	Total
Soltero	112	8	120
Casado	107	3	110
Total	219	11	230

La siguiente tabla, llamada *tabla de perfiles fila*, representa la distribución condicionada de la variable *Sexo* para cada estado civil.

	Mujer	Hombre	Total
Soltero	0.933	0.067	0.522
Casado	0.973	0.027	0.478
Total	0.952	0.048	1

Las filas de esta tabla son puntos de $\mathbb{R}^2$,

$$R_1 = (0.9332, 0.0667), \quad R_2 = (0.9726, 0.0273).$$

Calcular la distancia euclídea entre los puntos R_1 y R_2, escribiendo la cantidad que aporta cada coordenada a la distancia:

$d_e(R_1, R_2) = ...$

Cada coordenada aporta la misma cantidad de distancia, a pesar de que intuitivamente parece que las primeras coordenadas están más cerca. La distancia ji-cuadrado permite ponderar cada coordenada, para tener en cuenta su magnitud.

Distancia chi-cuadrado: La *distancia ji-cuadrado* entre los puntos $A = (a_1, \ldots, a_p)$ y $B = (b_1, \ldots, b_p)$, con pesos $\omega = (\omega_1, \ldots, \omega_p)$ es

$$d_{\mathcal{X}}(A, B) = \sum_{k=1}^{p} \omega_k (b_k - a_k)^2 .$$

Ejemplo 4.2 Si en el ejemplo anterior tomamos como pesos los inversos de la última fila de la tabla, es decir,

$$\omega_1 = \frac{1}{0{,}952}, \quad \omega_2 = \frac{1}{0{,}048}\,,$$

obtenemos

$d_{\mathcal{X}}(R_1, R_2) = ...$

Observa que ahora es la segunda coordenada la que aporta mayor cantidad a la distancia.

Ninguna de las distancias anteriores tiene en cuenta la correlación entre las variables $X_1, \ldots, X_p$. La distancia de Mahalanobis tiene en cuenta las varianzas y la correlación que existe entre ellas.

Distancia de Mahalanobis: Se miden p variables $X_1, \ldots, X_p$ a n individuos, y se obtiene la matriz de varianzas-covarianzas muestral S_X. Sean $A = (a_1, \ldots, a_p)$ y $B = (b_1, \ldots, b_p)$ los valores de las variables para dos individuos A y B. La *distancia de Mahalanobis* entre $A = (a_1, \ldots, a_p)$ y $B = (b_1, \ldots, b_p)$ es

$$d_M(A, B) = (B - A)' S_X^{-1} (B - A)\,.$$

Obsérvese que esta distancia solo se puede calcular si se dispone de un conjunto de n mediciones de $X_1, \ldots, X_p$.

Ejemplo 4.3 Consideremos los datos económicos (en millones de dólares) de las 10 corporaciones industriales estadounidenses más importantes:

Compañía	Ventas	Beneficios	Bienes
General Motors	126.974	4.224	173.297
Ford	96.933	3.835	160.893
Exxon	86.656	3.510	83.219
IBM	63.438	3.758	77.734
General Electric	55.264	3.939	128.344
Mobil	50.976	1.809	39.08
Philip Morris	39.069	2.946	38.528
Chrisler	36.156	0.359	51.038
Du Pont	35.209	2.480	34.715
Texaco	32.416	2.413	25.636

Calcular la distancia de Mahalanobis entre Ford y Exxon.

La matriz de varianzas-covarianzas es

$$S_X = \begin{pmatrix} 1000{,}509 & 23{,}018 & 1360{,}644 \\ 23{,}018 & 1{,}430 & 41{,}090 \\ 1360{,}644 & 41{,}090 & 2980{,}49 \end{pmatrix}.$$

Por tanto, la distancia de Mahalanobis entre Ford y Exxon es

$$(\ldots \qquad) \times$$

$$\begin{pmatrix} 1000{,}509 & 23{,}018 & 1360{,}644 \\ 23{,}018 & 1{,}430 & 41{,}090 \\ 1360{,}644 & 41{,}090 & 2980{,}49 \end{pmatrix}^{-1} \begin{pmatrix} \\ \vdots \\ \\ \end{pmatrix}$$

Similitud a partir de distancia:

A partir de una distancia se puede calcular una similitud:

$$s(A,B) = \frac{1}{1 + d(A,B)}.$$

4.2. VARIABLES BINARIAS

Una valor binario (0 ó 1) representa la presencia (1) o ausencia (0) de determinada característica. Ahora suponemos que $X_1, \ldots, X_p$ son variables binarias, con lo cual

$$A = (a_1 \ldots, a_p), \quad B = (b_1 \ldots, b_p), \quad a_k, b_k \in \{0, 1\}, \quad k = 1, \ldots, p.$$

Medidas de distancia

Distancia euclídea:

$$d_e(A, B) = \sqrt{\sum_{k=1}^{p} (b_k - a_k)^2}\,.$$

Se verifica

$$(b_k - a_k)^2 = \begin{cases} \ldots, & si\ b_k = a_k\,, \\ \ldots, & si\ b_k \neq a_k\,. \end{cases}$$

La distancia euclídea al cuadrado es el número de componentes de A y B que no coinciden.

Representamos los datos en una tabla de contingencia

		B		
		1	0	Total
A	1	a	b	$a+b$
	0	c	d	$c+d$
	Total	$a+c$	$b+d$	$t = a+b+c+d$

Según la tabla, la distancia euclídea es ... . Las medidas de distancia que aparecen en SPSS son:

Distancia euclídea	...
Distancia euclídea al cuadrado	...
Diferencia de tamaño	$\frac{(b-c)^2}{t^2}$
Diferencia de configuración	$\frac{bc}{t^2}$
Varianza	$\frac{b+c}{4t}$
Lance y Williams	$\frac{b+c}{2a+b+c}$

Medidas de similitud:

Coeficiente	Descripción
$\frac{a+d}{t}$	Igual peso a coincidencias 1-1 y 0-0.
$\frac{2(a+d)}{2(a+d)+b+c}$	Doble peso a las coincidencias.
$\frac{a+d}{a+d+2(b+c)}$	Doble peso a las no coincidencias.
$\frac{a}{t}$	No se tienen en cuenta las coincidencias 0-0 en el numerador.
$\frac{a}{a+b+c}$	No se tienen en cuenta las coincidencias 0-0.
$\frac{2a}{2a+b+c}$	No se tienen encuenta las coincidencias 0-0, y se da doble peso a las coincidencias 1-1.
$\frac{a}{a+2(b+c)}$	No se tienen encuenta las coincidencias 0-0, y se da doble peso a las no coincidencias.
$\frac{a}{b+c}$	Cociente de coincidencias y no coincidencias, con exclusión de coincidencias 0-0.

El primer coeficiente que aparece en la tabla se conoce como el *coeficiente de comparación simple*.

Ejemplo 4.4 Queremos medir la similaridad entre dos individuos según su actitud positiva (1) o negativa (0) hacia la compra de 10 artículos

	1	2	3	4	5	6	7	8	9	10
Indiv. 1	1	0	1	1	0	0	0	1	1	0
Indiv. 2	1	0	0	1	1	0	0	1	0	0

Construimos la tabla de contingencia de ambos individuos

		Indiv. 2	
		1 0	Total
Indiv. 1	1		
	0		
	Total		

El coeficiente de comparación simple vale ... .

4.3. VARIABLES NOMINALES O CATEGÓRICAS

Ejemplo 4.5 Medimos las variables X_1= "tipo de whisky", X_2= "tipo de botella" y X_3= "región de fabricación" a dos marcas de whisky escocés A y B. La variable X_1 toma valores m="puro de malta", b="mezclado" (en Inglés blended), y c="cereales diferentes de la cebada", y X_2 toma valores s="standard", cc="cilíndrica corta", cl="cilíndrica larga" y c="cuadrada". Por último, la procedencia del whisky (X_3) puede ser h="Highlands", l="lowlands" y wi="western islands". Supongamos que los valores obtenidos son

$$A = (m, s, h),\ B = (m, cc, wi)\,.$$

Una similaridad entre los whiskies A y B sería

$$\xi(A, B) = ... \quad .$$

Ahora suponemos que $X_1, \ldots, X_p$ son nominales, de manera que estas variables pueden tomar valores de un conjunto de categorías. Es decir, ahora deseamos comparar la similaridad de $A = (a_1, \ldots, a_p)$ y $B = (b_1, \ldots, b_p)$, donde a_k y b_k son ciertas categorías de la variable X_k, $k = 1, \ldots, p$.

Una similaridad sencilla entre A y B es la proporción de coincidencias en las coordenadas de A y B: para la coordenada k-ésima, definimos

$$\xi_k(A, B) = 1 \text{ si } a_k = b_k, \quad \text{y} \quad \xi_k(A, B) = 0 \text{ si } a_k \neq b_k,$$

entonces la similaridad final entre A y B es

$$\xi(A, B) = \frac{1}{p} \sum_{k=1}^{p} \xi_k(A, B)\,.$$

En lugar de asignar el valor 0 a $\xi_k(A, B)$ si las coordenadas a_k y b_k no coinciden, se puede asignar un valor entre 0 y 1 en función del grado de semejanza entre a_k y b_k.

Ejemplo 4.6 En el Ejemplo 4.5, supongamos que la similitud entre las categorías de la variable X_2 se puede cuantificar mediante los coeficientes siguientes

	s	cc	cl	c
s	1	0.5	0.5	0
cc	0.5	1	0.3	0
cl	0.5	0.3	1	0
c	0	0	0	1

En este caso, para

$$A = (m, s, h) \text{ y } B = (m, cc, wi)$$

tenemos

$$\xi_1(A,B) = ..., \quad \xi_2(A,B) = ... \quad \text{y} \quad \xi_3(A,B) = ...,$$

con lo cual la similaridad entre A y B es

$$\xi(A,B) = ... \qquad .$$

4.4. VARIABLES ORDINALES

Supongamos ahora que las m categorías $c_1, c_2, \ldots, c_m$ de una de las variables (X_k) están ordenadas de forma natural. Entonces se construyen $m-1$ variables dummy, con los siguientes valores

Categoría de X_k	I_1	I_2	$\cdots$	I_{m-1}
c_1	0	0	$\cdots$	0
c_2	1	0	$\cdots$	0
c_3	1	1	$\cdots$	0
$\vdots$	$\vdots$	$\vdots$		$\vdots$
c_m	1	1	$\cdots$	1

Para los sujetos $A = (a_1, \ldots, a_p)$ y $B = (b_1, \ldots, b_p)$, los elementos k-ésimos son ciertas categorías de la variable X_k; por ejemplo $a_k = c_i$ y $b_k = c_j$. Se construye una tabla de contingencia de ambas categorías para las variables I_1, I_2, ..., I_{m-1},

		c_j		
		1	0	Total
c_i	1	a	b	$a+b$
	0	c	d	$c+d$
	Total	$a+c$	$b+d$	$t = a+b+c+d$

Así, el coeficiente de comparación simple nos proporciona un coeficiente de similaridad para la variable X_k, llamado $\xi_k(A, B)$. Promediando para todas las variables, obtenemos la similaridad entre A y B.

Ejemplo 4.7 Supongamos que en el ejemplo anterior, la variable X_2 es "altura de la botella", con valores p="pequeña", s="standard", l="larga" y el="extra larga". Los whiskies A y B toman los valores

$$A = (m, s, h),\ B = (m, p, wi)\,.$$

Para la variable X_2 con cuatro categorías, construimos tres variables dummy de la forma

Categoría	I_1	I_2	I_3
p			
s			
l			
el			

Así, la tabla de contingencia sería

		s		
		1	0	Total
p	1			
	0			
	Total			

El coeficiente de la segunda componente es $\xi_2(A, B) = ...$. La similaridad entre A y B es

$$\xi(A, B) = ... \qquad .$$

Disimilaridades a partir de similaridades

Disimilaridades se pueden obtener a partir de similaridades de varias formas, entre ellas las siguientes

$$\delta(A,B) = 1 - \xi(A,B)\,,$$
$$\delta(A,B) = c - \xi(A,B), \text{ para alguna constante } c\,,$$
$$\delta(A,B) = \sqrt{2(1-\xi(A,B))}\,.$$