Técnicas de análisis discriminante

1. Se tienen tres ejemplares de dos especies en los que se han medido dos variables: x_1 y x_2. Las matrices de datos (por especies) son:

$$\mathbf{X}_1 = \begin{bmatrix} 3 & 7 \\ 2 & 4 \\ 4 & 7 \end{bmatrix}, \qquad \text{y} \qquad \mathbf{X}_2 = \begin{bmatrix} 6 & 9 \\ 5 & 7 \\ 4 & 8 \end{bmatrix}.$$

De las siguientes salidas de SPSS se puede obtener las medias de los grupos: $\bar{\mathbf{x}}_1$ y $\bar{\mathbf{x}}_2$, las matrices de covarianzas: $\mathbf{S}_1$ y $\mathbf{S}_2$, y la matriz de varianzas dentro de los grupos: $\mathbf{S}_w$. [1]

Estadísticos de grupo

ESPECIE		Media	Desv. típ.	N válido (según lista) No ponderados	N válido (según lista) Ponderados
1	X1	3,0000	1,00000	3	3,000
	X2	6,0000	1,73205	3	3,000
2	X1	5,0000	1,00000	3	3,000
	X2	8,0000	1,00000	3	3,000
Total	X1	4,0000	1,41421	6	6,000
	X2	7,0000	1,67332	6	6,000

Matrices de covarianza

ESPECIE		X1	X2
1	X1	1,000	1,500
	X2	1,500	3,000
2	X1	1,000	,500
	X2	,500	1,000

Matrices intra-grupo combinadas

		X1	X2
Covarianza	X1	1,000	1,000
	X2	1,000	2,000

Se pide:

(a) Obtener el vector o dirección de proyección para discriminar entre las dos especies, $\widehat{\mathbf{w}}$.

(b) Calcular las puntuaciones discriminantes para cada uno de los elementos de la muestra, $\widehat{\mathbf{w}}'\mathbf{x}$.

(c) Clasificar cada uno de los elementos de la muestra.

(d) Teniendo en cuenta los resultados de (c), calcule la tasa de error aparente.

(e) Compruebe los resultados de los apartados (c) y (d) con la siguiente salida de SPSS y diga cuál es la tasa de error estimada por validación cruzada.

[1] En este ejercicio utilizaremos las estimaciones del SPSS, es decir, no tendremos en cuenta la diferencia entre dividir por n o por $n-1$.

Resultados de la clasificación[b,c]

			Grupo de pertenencia pronosticado		
		ESPECIE	1	2	Total
Original	Recuento	1	3	0	3
		2	1	2	3
	%	1	100,0	,0	100,0
		2	33,3	66,7	100,0
Validación cruzada[a]	Recuento	1	2	1	3
		2	1	2	3
	%	1	66,7	33,3	100,0
		2	33,3	66,7	100,0

a. La validación cruzada sólo se aplica a los casos del análisis. En la validación cruzada, cada caso se clasifica mediante las funciones derivadas a partir del resto de los casos.

b. Clasificados correctamente el 83,3% de los casos agrupados originales.

c. Clasificados correctamente el 66,7% de los casos agrupados validados mediante validación cruzada.

(f) Clasifique una nueva observación $x_0 = \begin{bmatrix} 2 \\ 7 \end{bmatrix}$.

(g) Repita el apartado anterior, suponiendo que las probabilidades a priori de cada especie son: $\pi_1 = 0{,}75$, y $\pi_2 = 0{,}25$.

2. En las siguientes salidas de SPSS presentamos un análisis discriminante de los datos del fichero `turtlefm.xls` que contiene las siguientes medidas del caparazón de 48 tortugas pintadas: longitud, ancho y altura.

Estadísticos de grupo

SX		Media	Desv. típ.	N válido No ponderados
f	LONGITUD	136,04	21,249	24
	ANCHO	102,58	13,105	24
	ALTURA	52,04	8,046	24
m	LONGITUD	113,38	11,780	24
	ANCHO	88,29	7,074	24
	ALTURA	40,71	3,355	24
Total	LONGITUD	124,71	20,495	48
	ANCHO	95,44	12,676	48
	ALTURA	46,38	8,366	48

Matrices de covarianza

SX		LONGITUD	ANCHO	ALTURA
f	LONGITUD	451,520	270,975	165,955
	ANCHO	270,975	171,732	101,844
	ALTURA	165,955	101,844	64,737
m	LONGITUD	138,766	79,147	37,375
	ANCHO	79,147	50,042	21,654
	ALTURA	37,375	21,654	11,259

Matrices intra-grupo combinadas

		LONGITUD	ANCHO	ALTURA
Covarianza	LONGITUD	295,143	175,061	101,665
	ANCHO	175,061	110,887	61,749
	ALTURA	101,665	61,749	37,998

Coeficientes de las funciones canónicas discriminantes

	Función
	1
LONGITUD	-,111
ANCHO	-,056
ALTURA	,509
(Constante)	-4,446

Coeficientes no tipificados

Funciones en los centroides de los grupos

SX	Función 1
f	1,228
m	-1,228

Resultados de la clasificación[b,c]

			Grupo de pertenencia pronosticado		
		SX	f	m	Total
Original	Recuento	f	20	4	24
		m	0	24	24
	%	f	83,3	16,7	100,0
		m	,0	100,0	100,0
Validación cruzada[a]	Recuento	f	19	5	24
		m	1	23	24
	%	f	79,2	20,8	100,0
		m	4,2	95,8	100,0

a. La validación cruzada sólo se aplica a los casos del análisis. En la validación cruzada, cada caso se clasifica mediante las funciones derivadas a partir del resto de los casos.

b. Clasificados correctamente el 91,7% de los casos agrupados originales.

c. Clasificados correctamente el 87,5% de los casos agrupados validados mediante validación cruzada.

(a) Escriba la función canónica discriminante.

(b) ¿Cómo clasificaría una observación $\mathbf{x}_0' = [124{,}71,\ 95{,}44,\ 46{,}38]$? Justifique su respuesta.

(c) Calcule la tasa de error aparente y la tasa de error estimada por validación cruzada.

(d) Calcule el vector o dirección de proyección para discriminar entre los dos sexos, $\widehat{\mathbf{w}}$.

Ayuda: $\mathbf{S}_w^{-1} = \begin{bmatrix} 0{,}0718 & -0{,}0671 & -0{,}0831 \\ -0{,}0671 & 0{,}1575 & -0{,}0766 \\ -0{,}0831 & -0{,}0766 & 0{,}3729 \end{bmatrix}$.

(e) (Ejercicio de ordenador) Repita los apartados (a, b y c) con el fichero `turtlefm.xls` con las variables transformadas: `l1` = log(`longitud`), `l2` = log(`ancho`), y `l3` = log(`altura`).

3. Se han analizado las cantidades de 13 componentes en los vinos de una misma región de Italia provenientes de 3 bodegas diferentes. Concretamente, 59 botellas de la bodega 1, 71 de la bodega 2 y 48 de la 3. Supóngase que nos interesa discriminar entre la bodega 2 y la bodega 3. Realizamos un análisis discriminante lineal con los datos de las dos últimas bodegas obteniendo los siguientes resultados:

Coeficientes de las funciones canónicas discriminantes

	Función
	1
V1	-,182
V2	-,330
V3	-1,133
V4	-,034
V5	,000
V6	-,522
V7	1,330
V8	2,217
V9	,225
V10	-,369
V11	1,415
V12	,822
V13	,000
(Constante)	2,981

Coeficientes no tipificados

Funciones en los centroides de los grupos

	Función
BODEGA	1
2	2,132
3	-3,154

Funciones discriminantes canónicas no tipificadas evaluadas en las medias de los grupos

Resultados de la prueba

M de Box		438,871
F	Aprox.	4,226
	gl1	91
	gl2	32194,358
	Sig.	,000

Contrasta la hipótesis nula de que las matrices de covarianza poblacionales son iguales.

Resultados de la clasificación[b,c]

			Grupo de pertenencia pronosticado		
		BODEGA	2	3	Total
Original	Recuento	2	71	0	71
		3	0	48	48
	%	2	100,0	,0	100,0
		3	,0	100,0	100,0
Validación cruzada[a]	Recuento	2	70	1	71
		3	0	48	48
	%	2	98,6	1,4	100,0
		3	,0	100,0	100,0

a. La validación cruzada sólo se aplica a los casos del análisis. En la validación cruzada, cada caso se clasifica mediante las funciones derivadas a partir del resto de los casos.

b. Clasificados correctamente el 100,0% de los casos agrupados originales.

c. Clasificados correctamente el 99,2% de los casos agrupados validados mediante validación cruzada.

(a) Dígase cuál es la tasa de error por validación cruzada y de qué bodega o bodegas son los vinos mal clasificados (mediante este mismo método de validación cruzada).

(b) ¿Cómo se calcula el valor $2,132$, que corresponde a la puntuación discriminante canónica del centroide o media de la segunda bodega? No es necesario más que indicar teóricamente cómo ha de calcularse dicho valor.

(c) Supongamos que al analizar una botella de vino de procedencia desconocida de esa región italiana obtenemos los siguientes valores:

$$\mathbf{x}_0 = [10, 2, 2, 10, 100, 4, 2, 1, 1, 5, 1, 1, 500]'.$$

¿Cómo se clasificaría esta observación?

(d) Supongamos que conocemos que la botella anterior es del año 1995, y que en ese año la producción de la segunda bodega fue de 3 millones de litros de vino y la de la tercera bodega de 2 millones de litros de vino. ¿Cómo clasificaríamos la botella teniendo en cuenta esta nueva información?

(e) ¿Es razonable la hipótesis de igualdad de matrices de covarianzas que es necesaria para poder llevar a cabo el análisis discriminante lineal?

Supóngase que ahora nos interesase discriminar entre las tres bodegas. Para ello efectuamos un análisis discriminante entre las 3 poblaciones con los siguientes resultados:

Coeficientes de las funciones canónicas discriminantes

	Función	
	1	2
C1	,403	,872
C2	-,165	,305
C3	,369	2,346
C4	-,155	-,146
C5	,002	,000
C6	-,618	-,032
C7	1,661	-,492
C8	1,496	-1,631
C9	-,134	-,307
C10	-,355	,253
C11	,818	-1,516
C12	1,158	,051
C13	,003	,003
(Constante)	-9,231	-14,642

Coeficientes no tipificados

Funciones en los centroides de los grupos

	Función	
Bodega	1	2
1	3,422	1,692
2	,080	-2,473
3	-4,325	1,578

Funciones discriminantes canónicas no tipificadas evaluadas en las medias de los grupos

Resultados de la clasificación[b,c]

			Grupo de pertenencia pronosticado			
		PROD	1	2	3	Total
Original	Recuento	1	59	0	0	59
		2	0	71	0	71
		3	0	0	48	48
	%	1	100,0	,0	,0	100,0
		2	,0	100,0	,0	100,0
		3	,0	,0	100,0	100,0
Validación cruzada[a]	Recuento	1	59	0	0	59
		2	1	69	1	71
		3	0	0	48	48
	%	1	100,0	,0	,0	100,0
		2	1,4	97,2	1,4	100,0
		3	,0	,0	100,0	100,0

a. La validación cruzada sólo se aplica a los casos del análisis. En la validación cruzada, cada caso se clasifica mediante las funciones derivadas a partir del resto de los casos.

b. Clasificados correctamente el 100,0% de los casos agrupados originales.

c. Clasificados correctamente el 98,9% de los casos agrupados validados mediante validación cruzada.

(f) Dígase cuál es la tasa de error por validación cruzada y de qué bodega o bodegas son los vinos mal clasificados (mediante este mismo método de validación cruzada).

(g) ¿Cómo se calcula el vector $[3'422, 1'692]'$ que corresponde a la puntuación discriminante canónica del centroide o media de la primera bodega? No es necesario más que indicar teóricamente cómo ha de calcularse dicho vector.

(h) Supongamos que al analizar una botella de vino desconocida obtenemos los siguientes valores:

$$\mathbf{x}_0 = [12, 2, 2, 20, 95, 3, 2, 1, 2, 3, 1, 3, 520]'.$$

Calcúlese el vector de las puntuaciones discriminantes canónicas de esta nueva observación.

(i) ¿Cómo se clasificaría dicha observación?

4. La tasa de sedimentación de eritrocitos (ESR) mide la distancia en milímetros en que los glóbulos rojos se sedimentan en sangre no coagulada hacia el fondo de un tubo de ensayo durante 1 hora. La presencia de algunas proteínas en la sangre, asociada a algunas enfermedades, hace que los eritrocitos se unan causando aglutinaciones que se vuelven más pesadas y caen antes que una célula individual. Si al aumentar el nivel de proteína también se incrementa la ESR, ésta se puede utilizar para detectar enfermedades. En un estudio a partir de datos de 32 personas se investiga si es posible clasificar a los individuos en alguno de los dos grupos siguientes: `ESR`=0 (ESR<20 mm/h) o `ESR`=1 (ESR>20 mm/h) a partir de la cantidad de fibrinógenos (`fib`) y gammaglobulinas (`glob`) detectadas en la sangre. Para ello, se aplicó con `SPSS` la técnica de análisis discriminante con los resultados que aparecen abajo (algunos valores han sido sustituidos por letras):

(a) Calcula los valores de A, B y C en la salida de abajo.

(b) ¿En qué grupo se clasifica un individuo tal que `fib` vale 3.93 y `glob` vale 32?

(c) Lleva a cabo de nuevo la clasificación anterior pero suponiendo ahora que la probabilidad a priori de ESR=0 es el doble que la de ESR=1.

(d) Calcula la tasa de error aparente y la tasa de error por validación cruzada.

ESR		Media	Desv. típ.	N válido (según lista)	
				No ponderados	Ponderados
,00	fib	2,6504	,40566	26	26,000
	glob	35,1154	4,17925	26	26,000
1,00	fib	3,3883	1,07821	6	6,000
	glob	38,0000	5,89915	6	6,000
Total	fib	2,7888	,63707	32	32,000
	glob	35,6563	4,58335	32	32,000

Matrices intra-grupo combinadas[a]

		fib	glob
Covarianza	fib	A	-,102
	glob	-,102	20,355
Correlación	fib	1,000	-,039
	glob	-,039	B

a. La matriz de covarianza tiene 30 grados de libertad

Matrices de covarianza[a]

ESR		fib	glob
,00	fib	C	,288
	glob	,288	17,466
1,00	fib	1,163	-2,056
	glob	-2,056	34,800
Total	fib	,406	,236
	glob	,236	21,007

a. La matriz de covarianza total presenta 31 grados de libertad.

Coeficientes de las funciones canónicas discriminantes

	Función
	1
fib	1,563
glob	,105
(Constante)	-8,108

Funciones en los centroides de los grupos

ESR	Función
	1
,00	-,273
1,00	1,184

Resultados de la clasificación

		ESR	Grupo de pertenencia pronosticado		Total
			,00	1,00	
Original	Recuento	,00	20	6	26
		1,00	2	4	6
	%	,00	76,9	23,1	100,0
		1,00	33,3	66,7	100,0
Validación cruzada[a]	Recuento	,00	20	6	26
		1,00	3	3	6
	%	,00	76,9	23,1	100,0
		1,00	50,0	50,0	100,0