Análisis de Datos — Curso 2007/2008
Licenciatura en Biología

Técnicas de análisis multivariante para la agrupación

1. Se tienen seis ejemplares en los que se han medido dos variables: x_1 y x_2. El investigador supone que provienen de dos especies distintas, y sospecha que las tres primeras observaciones son de la misma especie. La matriz de datos es:

$$\mathbf{X} = \begin{bmatrix} 3 & 7 \\ 2 & 4 \\ 4 & 7 \\ 6 & 9 \\ 5 & 7 \\ 4 & 8 \end{bmatrix}.$$

De la siguiente salidas de SPSS se pueden obtener las distancias euclídeas entre cada par de observaciones:

Matriz de distancias

	Distancia euclídea					
	1	2	3	4	5	6
1	,000	3,162	1,000	3,606	2,000	1,414
2	3,162	,000	3,606	6,403	4,243	4,472
3	1,000	3,606	,000	2,828	1,000	1,000
4	3,606	6,403	2,828	,000	2,236	2,236
5	2,000	4,243	1,000	2,236	,000	1,414
6	1,414	4,472	1,000	2,236	1,414	,000

Se pide:

(a) ¿Cuál es la primera unión de clusters en cualquier algoritmo aglomerativo? Justifique su respuesta.

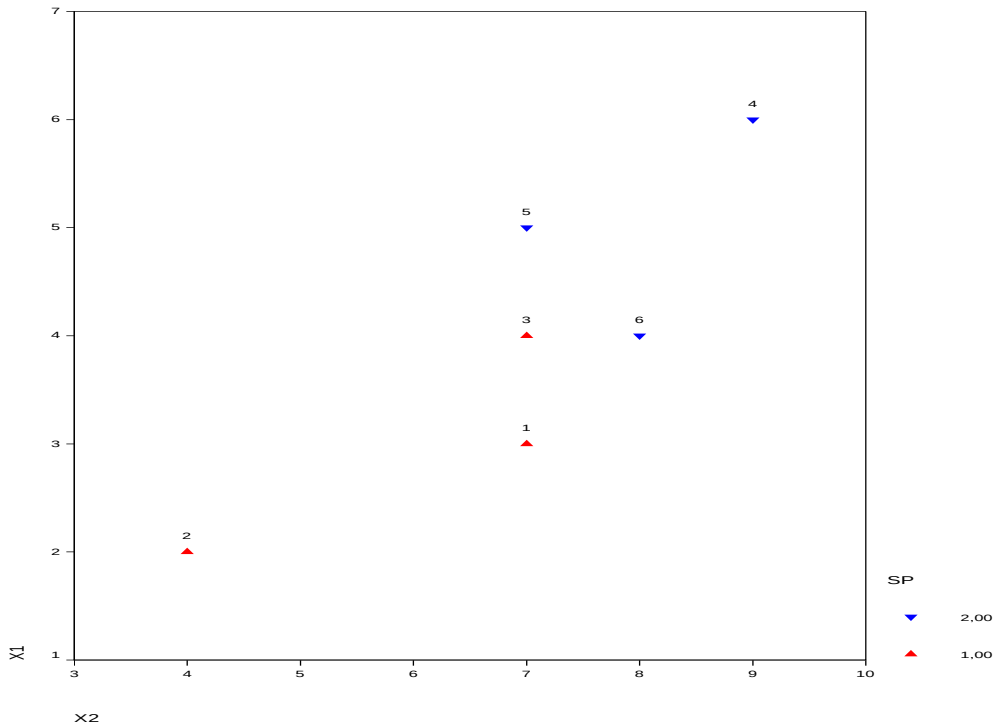
(b) En la siguiente salida de SPSS se muestra el historial de aglomeración con los datos de este ejercicio utilizando el método del vecino más próximo. ¿Coincide la primera unión del historial con la de su respuesta en (a)? ¿Y si su respuesta difiere, cree que es incorrecta?

Historial de conglomeración

Etapa	Conglomerado que se combina		Coeficientes	Etapa en la que el conglomerado aparece por primera vez		Próxima etapa
	Conglomerado 1	Conglomerado 2		Conglomerado 1	Conglomerado 2	
1	3	6	1,000	0	0	2
2	3	5	1,000	1	0	3
3	1	3	1,000	0	2	4
4	1	4	2,236	3	0	5
5	1	2	3,162	4	0	0

(c) Dibuje el dendrograma utilizando la información del historial de aglomeración del apartado (b).

(d) En la siguiente figura representamos el diagrama de dispersión de las seis observaciones (hemos representado los puntos según la hipótesis inicial del investigador). ¿Cree que, en este ejercicio, cambiar la unión inicial cambiará el resultado final? Compruebe su respuesta efectuando el análisis cluster por el método del vecino más próximo y con una unión inicial distinta de la efectuada por SPSS.



(e) (Ejercicio de ordenador) Obtenga el análisis cluster por los métodos de vecino más lejano y por agrupación de centroides. Compare los resultados con los obtenidos en el apartado

2. Con los datos del ejercicio anterior el investigador efectúa un análisis K-medias. La siguiente salida de SPSS muestra los centros iniciales de los conglomerados. Notemos que el primer

centro coincide con la observación 4, y que el segundo coincide con la observación 2.

Centros iniciales de los conglomerados

	Conglomerado	
	1	2
X1	6,00	2,00
X2	9,00	4,00

Se pide:

(a) Utilizando la matriz de distancias euclídeas del ejercicio anterior, diga a qué grupo se asigna la observación 1.

(b) La siguiente salida de SPSS muestra los centros finales de los conglomerados, diga a qué grupo se asigna cada observación.

Centros de los conglomerados finales

	Conglomerado	
	1	2
X1	4,75	2,50
X2	7,75	5,50

(c) (Ejercicio de ordenador) Cambie el orden de los datos (por ejemplo, intercambie las observaciones 1 y 2) y repita el análisis cluster de este ejercicio. Explique los resultados obtenidos.

3. Se ha realizado un estudio sencillo para establecer las similitudes y posibles relaciones entre 35 especies de animales. En cada animal se han considerado 15 variables binarias o dicotómicas, es decir, que únicamente toman los valores 0 y 1 (0 y 1 indican respectivamente la ausencia o presencia de una cierta característica). Concretamente las variables estudiadas son:

x_1 = `PELO`, x_2 = `PLUMAS`, x_3 = `HUEVOS`, x_4 = `LECHE`, x_5 = `AÉREO`, x_6 = `ACUÁTICO`, x_7 = `PREDADOR`, x_8 = `DIENTES`, x_9 = `VERTEBRAS`, x_{10} = `RESPIRACIÓN`, x_{11} = `VENENOSOS`, x_{12} = `ALETAS`, x_{13} = `COLA`, x_{14} = `DOMÉSTICO` y x_{15} = `TAMAÑO_GRANDE`.

Por ejemplo, los datos correspondientes a las especies `avispa` y `lobo` son los siguientes:

Especie	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}	x_{11}	x_{12}	x_{13}	x_{14}	x_{15}
`avispa`	1	0	1	0	1	0	0	0	0	1	1	0	0	0	0
`lobo`	1	0	0	1	0	0	1	1	1	1	0	0	1	0	1

(a) Calcúlese el coeficiente de concordancia simple y el de Jaccard entre el par de datos `avispa` y `lobo`. En este ejemplo concreto, ¿cuál es la medida de similitud más adecuada? Razónese la respuesta.

(b) Efectúese un análisis jerárquico de las especies `gorrión`, `halcón`, `león` y `ratón` basado en el método del encadenamiento simple (o vecino más próximo) a partir de la matriz de similitudes de abajo. Para el cálculo de la matriz se ha utilizado el coeficiente de

concordancia simple. Es necesario escribir claramente el historial de conglomeración y el dendograma resultante.

Matriz de distancias

Caso	medida de emparejamiento simple			
	1:gorrión	2:halcón	3:león	4:ratón
1:gorrión	1,000	,933	,467	,600
2:halcón	,933	1,000	,533	,533
3:león	,467	,533	1,000	,867
4:ratón	,600	,533	,867	1,000

Esta es una matriz de similaridades

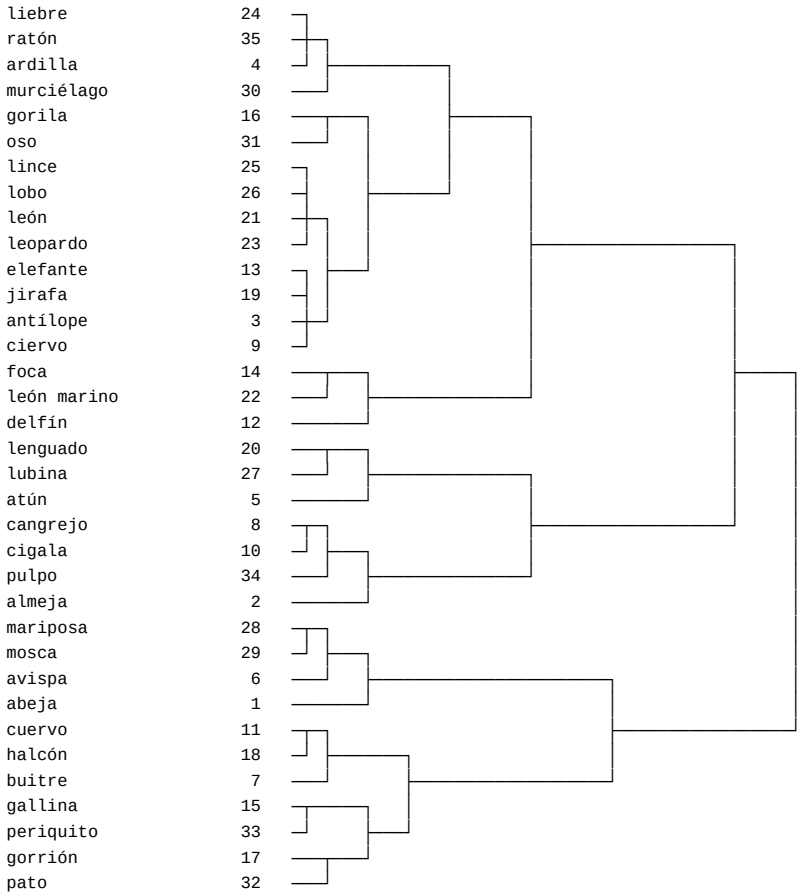
Se efectúa un análisis jerárquico basado en el método del encadenamiento completo (o vecino más lejano) para distinguir lo más posible los grupos y utilizando el coeficiente de concordancia simple. Los resultados de este análisis (historial de conglomeración e histograma) se encuentran en la siguiente página.

(c) ¿Qué significa exactamente el coeficiente 0,867 que aparece en la etapa 23 del historial de conglomeración?

(d) Coméntese el dendograma ¿Cuántos grupos se pueden distinguir? ¿Qué animales incluyen? ¿Cómo se pueden interpretar?

Historial de conglomeración

Etapa	Conglomerado que se combina		Coeficientes	Etapa en la que el conglomerado aparece por primera vez		Próxima etapa
	Conglomerado 1	Conglomerado 2		Conglomerado 1	Conglomerado 2	
1	24	35	1,000	0	0	6
2	15	33	1,000	0	0	23
3	28	29	1,000	0	0	17
4	25	26	1,000	0	0	5
5	21	25	1,000	0	4	7
6	4	24	1,000	0	1	16
7	21	23	1,000	5	0	20
8	13	19	1,000	0	0	10
9	11	18	1,000	0	0	21
10	3	13	1,000	0	8	12
11	8	10	1,000	0	0	13
12	3	9	1,000	10	0	20
13	8	34	,933	11	0	26
14	17	32	,933	0	0	23
15	16	31	,933	0	0	24
16	4	30	,933	6	0	29
17	6	28	,933	0	3	27
18	20	27	,933	0	0	22
19	14	22	,933	0	0	25
20	3	21	,933	12	7	24
21	7	11	,933	0	9	28
22	5	20	,867	0	18	31
23	15	17	,867	2	14	28
24	3	16	,867	20	15	29
25	12	14	,867	0	19	30
26	2	8	,867	0	13	31
27	1	6	,867	0	17	32
28	7	15	,800	21	23	32
29	3	4	,733	24	16	30
30	3	12	,600	29	25	33
31	2	5	,600	26	22	33
32	1	7	,467	27	28	34
33	2	3	,267	31	30	34
34	1	2	,133	32	33	0



4. Se ha registrado la duración en minutos de erupciones del geyser Old Faithful (variable X) junto con los minutos que transcurren entre una erupción y la siguiente (variable Y). Para 15 erupciones se ha llevado a cabo un análisis de agrupación jerárquico (utilizando la distancia euclídea entre observaciones estandarizadas) y un análisis de agrupación de k-medias (a partir de los datos originales no estandarizados) con los resultados que aparecen al final del enunciado. Se pide contestar a las siguientes preguntas:

(a) ¿Qué observaciones forman cada uno de los grupos que se unen en la última etapa del algoritmo?

(b) ¿En cuántos grupos sugiere el dendograma que deben dividirse los datos?

(c) Si $d(A, B)$ es la distancia entre los grupos de observaciones A y B, ¿cuál es el valor de $d(\{9, 15\}, \{11\})$?

(d) ¿Sería apropiado en este caso basar el algoritmo en el coeficiente de concordancia simple? ¿Y en el coeficiente de Jaccard?

(e) ¿Cuáles son los centros de los grupos obtenidos en el método de k-medias?

(f) ¿Son similares los resultados obtenidos mediante el método de agrupación jerárquico y el no jerárquico (k-medias)?

(g) En el resultado del método no jerárquico, ¿cuál es la observación más cercana al centro del grupo 2?

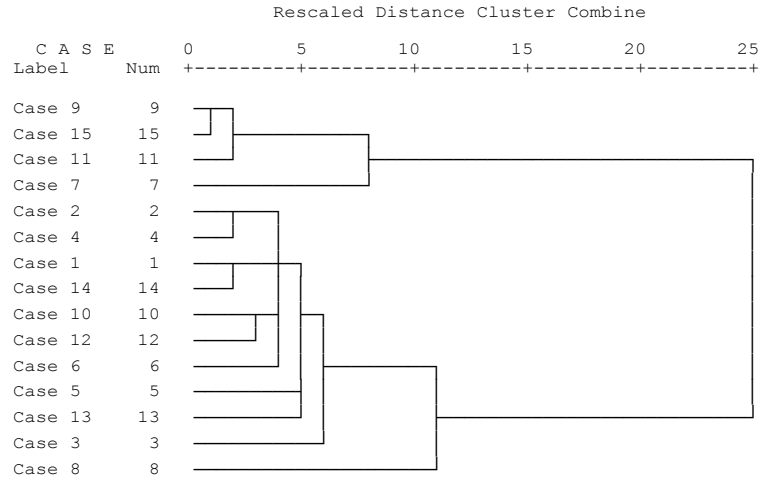
(h) ¿Cuál es el valor de la suma de cuadrados dentro de los grupos una vez aplicado el método no jerárquico?

Vinculación simple

Historial de conglomeración

Etapa	Conglomerado que se combina		Coeficientes	Etapa en la que el conglomerado aparece por primera vez		Próxima etapa
	Conglomerado 1	Conglomerado 2		Conglomerado 1	Conglomerado 2	
1	9	15	,083	0	0	2
2	9	11	,166	1	0	12
3	2	4	,188	0	0	8
4	1	14	,188	0	0	6
5	10	12	,311	0	0	6
6	1	10	,375	4	5	7
7	1	6	,375	6	0	8
8	1	2	,388	7	3	10
9	5	13	,424	0	0	10
10	1	5	,446	8	9	11
11	1	3	,505	10	0	13
12	7	9	,670	0	2	14
13	1	8	,913	11	0	14
14	1	7	2,051	13	12	0

Dendrogram using Single Linkage



Análisis de conglomerados de K medias

Pertenencia a los conglomerados

Número de caso	Conglomerado	Distancia
1	1	,236
2	1	4,009
3	1	10,001
4	1	2,007
5	1	2,107
6	1	6,000
7	2	4,771
8	1	15,010
9	2	,292
10	1	2,131
11	2	3,253
12	1	4,024
13	1	3,096
14	1	2,005
15	2	1,259

Centros de los conglomerados finales

	Conglomerado	
	1	2
Y	78,00	54,75
X	4,16	1,85

5. Se realizó un estudio sobre afecciones hepáticas que pueden aparecer a causa de la ingesta excesiva de alcohol midiendo 5 variables de interés en la analítica de un total de 345 varones. Concretamente, se consideraron las variables:

- $x_1 =$ `mcv` $\equiv$ volumen corpuscular medio.
- $x_2 =$ `alkphos` $\equiv$ fosfatasa alcalina.
- $x_3 =$ `sgpt` $\equiv$ alanina aminotransferasa.
- $x_4 =$ `sgot` $\equiv$ aspartato aminotransferasa.
- $x_5 =$ `gammagt` $\equiv$ gamma-glutamil transpeptidasa.

Además de estas 5 variables en la analítica de la sangre se consideró la variable $x_6 =$ `drinks` que mide el número de medias pintas de cerveza o equivalentes en bebidas alcohólicas consumidas al día.

Después del estudio descriptivo preliminar se pensó que los datos se podrían dividir en diferentes grupos. Por consiguiente, se utilizó el algoritmo de k-medias con $k = 2$ y $k = 3$. Contéstese a las siguientes preguntas:

(a) ¿En cuántos grupos homogéneos (dos o tres) es más adecuado separar los datos? Razónese. ¿Cómo se interpretan los grupos en promedio?

(b) ¿En qué grupo se clasificaría un nuevo dato $\mathbf{x} = [100, 70, 35, 30, 35, 3]'$ para $k = 2$?

Estadísticos

		mcv	alkphos	sgpt	sgot	gammagt	drinks
N	Válidos	345	345	345	345	345	345
	Perdidos	0	0	0	0	0	0
Media		90,159	69,8696	30,4058	24,643	38,284	3,4551
Mediana		90,000	67,0000	26,0000	23,000	25,000	3,0000
Desv. típ.		4,4481	18,34767	19,51231	10,0645	39,2546	3,33784
Mínimo		65,0	23,00	4,00	5,0	5,0	,00
Máximo		103,0	138,00	155,00	82,0	297,0	20,00
Percentiles	25	87,000	57,0000	19,0000	19,000	15,000	,5000
	75	93,000	80,0000	34,0000	27,000	46,500	6,0000
	90	96,000	95,0000	52,0000	35,000	82,800	8,0000
	95	98,000	104,1000	67,7000	44,400	115,000	10,0000
	99	100,540	123,0000	131,9000	71,780	203,000	16,0000

Estadísticos descriptivos

Número inicial de casos		N	Media	Desv. típ.	Varianza
1	mcv	307	89,971	4,3816	19,198
	alkphos	307	69,1759	18,59321	345,708
	sgpt	307	26,7980	12,58637	158,417
	sgot	307	22,860	7,4362	55,297
	gammagt	307	27,055	16,4076	269,209
	drinks	307	3,1450	3,03320	9,200
	N válido (según lista)	307			
2	mcv	38	91,684	4,7426	22,492
	alkphos	38	75,4737	15,31249	234,472
	sgpt	38	59,5526	35,34287	1249,119
	sgot	38	39,053	15,6463	244,808
	gammagt	38	129,000	51,0103	2602,054
	drinks	38	5,9605	4,49982	20,248
	N válido (según lista)	38			

Centros de los conglomerados finales

	Conglomerado		
	1	2	3
mcv	89,6	90,8	92,3
alkphos	65,24	77,74	76,16
sgpt	22,54	38,71	62,52
sgot	20,9	28,1	41,9
gammagt	18,6	51,5	150,6
drinks	2,60	4,54	6,18

Estadísticos descriptivos

Número inicial de casos		N	Suma
GRUPO 1 (K=3)	distancia al centro al cuadrado	214	94343,23
	N valido	214	
GRUPO 2 (K=3)	distancia al centro al cuadrado	106	125304,77
	N valido	106	
GRUPO 3 (K=3)	distancia al centro al cuadrado	25	117550,32
	N valido	25	