(Tarea 3) Estadística Aplicada III

David Isaac López Romero
CU: 173993

1) Tenemos la siguiente matriz

	a	b	c	d	e
a		2	1	3	4
b			2	5	3
c					
d					3
e					

Haz el análisis de cúmulos utilizando:

a) Distancia Mínima.

①

	{a}	{b}	{c}	{d}	{e}
{a}		2	1	3	4
{b}			2	5	3
{c}				3	4
{d}					3
{e}					

⇒ ②

	{a,c}	{b}	{d}	{e}
{a,c}		2	3	4
{b}			5	3
{d}				3
{e}				

③ ⇒

	{a, b, c}	{d}	{e}
{a, b, c}		3	3
{d}			3
{e}			

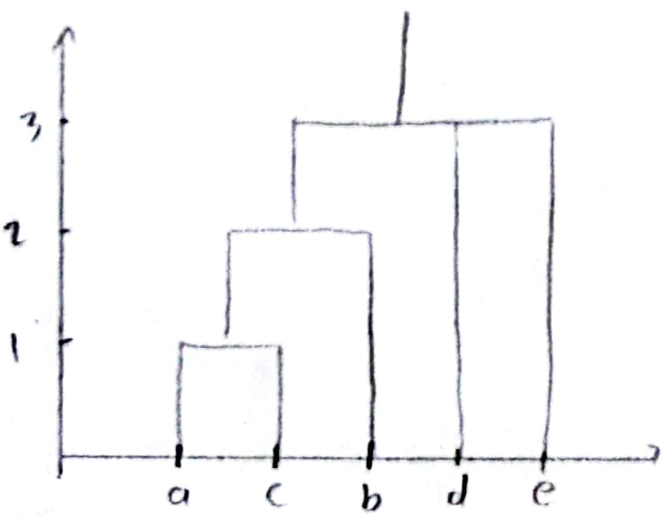
$d<1$, {a}, {b}, {c}, {d}, {e}

$1 \le d < 2$, {a, c}, {b}, {d}, {e}

$2 \le d < 3$, {a,b,c}, {d}, {e}

$3 \le d$, {a,b,c,d,e}

Dendograma



- 5 grupos: {a}, {b}, {c}, {d}, {e} ($d<1$)
- 4 grupos: {a,c}, {b}, {d}, {e} ($1 \le d < 2$)
- 3 grupos {a,b,c}, {d}, {e} ($2 \le d < 3$)
- 1 grupo: {a,b,c,d,e} ($3 \le d$)

b) Distancia máxima

①

	{a}	{b}	{c}	{d}	{e}
{a}		2	1	3	4
{b}			2	5	3
{c}				3	4
{d}					3
{e}					

$\Rightarrow$ ②

	{a,c}	{b}	{d}	{e}
{a,c}		2	3	4
{b}			5	3
{d}				3
{e}				

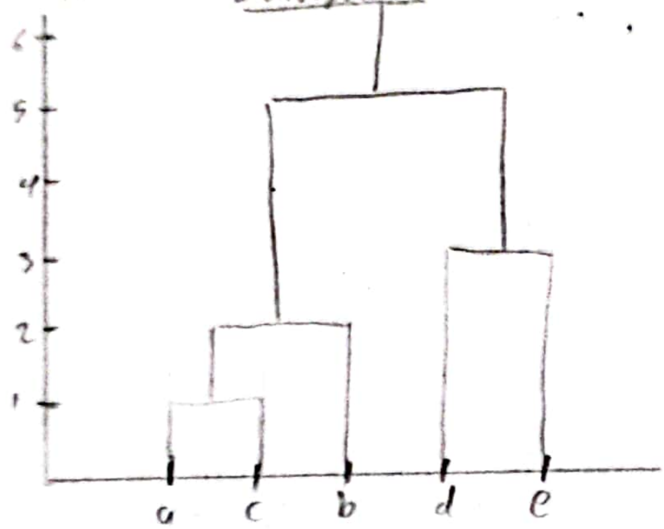
$\Rightarrow$ ③

	{a,b,c}	{d}	{e}
{a,b,c}		5	4
{d}			3
{e}			

$\Rightarrow$ ④

	{a,b,c}	{d,e}
{a,b,c}		5
{d,e}		

Dendrograma



$d<1$ {a}, {b}, {c}, {d}, {e}

$1\leq d<2$ {a,c}, {b}, {d}, {e}

$2\leq d<3$ {a,b,c}, {d}, {e}

$3\leq d<5$ {a,b,c}, {d,e}

$5\leq d$ {a,b,c,d,e}

- 5 grupos: {a}, {b}, {c}, {d}, {e} $(d<1)$
- 4 grupos: {a,c}, {b}, {d}, {e} $(1\leq d<2)$
- 3 grupos: {a,b,c}, {d}, {e} $(2\leq d<3)$
- 2 grupos: {a,b,c}, {d,e} $(3\leq d<5)$
- 1 grupo: {a,b,c,d,e} $(5\leq d)$

c) Distancia Media

①

	{a}	{b}	{c}	{d}	{e}
{a}		2	1	3	4
{b}			2	5	3
{c}				3	4
{d}					3
{e}					

$\Rightarrow$ ②

	{a,c}	{b}	{d}	{e}
{a,c}		2	3	4
{b}			5	3
{d}				3
{e}				

$$d(\{a,c\},b) = \frac{d(a,b)+d(c,b)}{2} = \frac{2+2}{2} = 2$$

$$d(\{a,c\},d) = \frac{d(a,d)+d(c,d)}{2} = \frac{3+3}{2} = 3$$

$$d(\{a,c\},e) = \frac{d(a,e)+d(c,e)}{2} = \frac{4+4}{2} = 4$$

③ $\Rightarrow$

	{a,b,c}	{d}	{e}
{a,b,c}		4	3.5
{d}			3
{e}			

$\Rightarrow$ ④

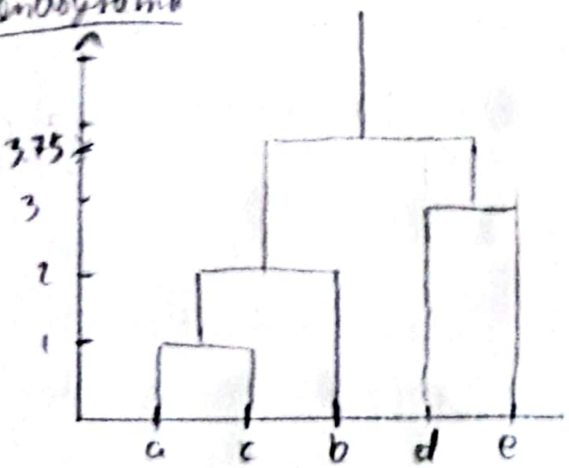
	{a,b,c}	{d,e}
{a,b,c}		3.75
{d,e}		

$$d(\{\{a,c\},b\},d) = \frac{d(\{a,c\},d)+d(b,d)}{2} = \frac{3+5}{2} = 4$$

$$d(\{\{a,c\},b\},e) = \frac{d(\{a,c\},e)+d(b,e)}{2} = \frac{4+3}{2} = 3.5$$

$$d(\{a,b,c\},\{d,e\}) = \frac{d(\{a,b,c\},d) + d(\{a,b,c\},e)}{2} = \frac{4+3.5}{2} = 3.75$$

Dendograma



$d<1$, {a}, {b}, {c}, {d}, {e}

$1\le d<2$, {a,c}, {b}, {d}, {e}

$2\le d<3$ {a,b,c} {d}, {e}

$3\le d<3.75$ {a,b,c}, {d,e}

$3.75\le d$ {a,b,c,d,e}

- 5 grupos: $\{a\}, \{b\}, \{c\}, \{d\}, \{e\}$ $(d<1)$
- 4 grupos: $\{a, c\}, \{b\}, \{d\}, \{e\}$ $(1 \leq d < 2)$
- 3 grupos: $\{a, b, c\}, \{d\}, \{e\}$ $(2 \leq d < 3)$
- 2 grupos: $\{a, b, c\}, \{d, e\}$ $(3 \leq d < 3.75)$
- 1 grupo: $\{a, b, c, d, e\}$ $(3.75 \leq d)$

2) Sea A matriz de tamaño $m \times n$. Definimos $\|A\|_2 = \max \|Ax\|_2$ s.a. $\|x\|_2 = 1$ con x vector de tamaño $n \times 1$. Demuestra que:

$\|A\|_2 = \max(\sqrt{\lambda})$ con λ eigenvalor de la matriz $A^T A$.

Dem:

Calculamos $\|A\|_2^2 = (\max \|Ax\|_2)^2$ s.a. $\|x\|_2^2 = 1$

Tenemos el problema

$$\max \|Ax\|_2^2 \quad \Leftrightarrow \quad \max \ x^T A^T A x$$
$$\text{s.a. } \|x\|_2^2 = 1 \qquad \text{s.a. } x^T x = 1$$

Consideramos al Lagrangiano

$$\mathcal{L}(x, \lambda) = x^T A^T A x - \lambda (x^T x - 1)$$

Así, vemos que

$$\frac{\partial \mathcal{L}}{\partial x}(x, \lambda) = 0 \Leftrightarrow 2A^T A x - 2\lambda x = 0 \Leftrightarrow A^T A \bar{x} = \lambda \bar{x}$$

Por lo tanto, λ es eigenvalor de $A^T A$ y $\bar{x}$ su eigenvector respectivo.

Ahora bien, si definimos a $Y = x^T A^T A x$ tenemos que (AA^T es definida positiva)

$$\mathrm{Var}(x^T A^T A x) = x^T (A^T A x) = x^T \lambda x = \lambda (x^T x) = \lambda$$

el cual es máximo para el eigenvalor máximo. Es decir,

$$\lambda = \max_{i=1,\dots,n} \{\lambda_i\} = \max \{\text{eigenvalores de } A^T A\}$$

En consecuencia,

$\|A\|_2^2 = máx(\lambda)$ con λ eigenvalor de A^TA

$\Leftrightarrow \|A\|_2 = máx(\sqrt{\lambda})$ con λ eigenvalor de A^TA

$\therefore \|A\|_2 = máx(\sqrt{\lambda})$ con λ eigenvalor de A^TA //

3) Tenemos el siguiente modelo para la probabilidad de que una persona vaya a votar por uno de los siguientes candidatos:

1) Clinton
2) Bush H
3) Perrot

$$\ln\left(\frac{P_{2i}}{P_{1i}}\right) = -0.35 + 0.88\, H_i + 0.47\, Univ_i$$

$$\ln\left(\frac{P_{3i}}{P_{1i}}\right) = -0.52 - 0.32\, H_i - 0.17\, Univ_i$$

H_i es una variable indicadora que toma el valor de 1 si es Hombre y 0 en otro caso de mujer

$Univ_i$ es una variable indicadora que toma el valor de 1 si tiene estudios universitarios y 0 en otro caso.

4) Interpreta los coeficientes

I) El momio entre Clinton y BushH aumenta $(e^{0.88}-1)\times 100\% = 141.0899\%$ por pasar de mujer a hombre.

II) El momio entre Clinton y BushH aumenta $(e^{0.47}-1)\times 100\% = 59.9997\%$ por pasar de no tener estudios universitarios a sí tenerlos

III) El momio entre Clinton y Perrot disminuye en $(e^{-0.32}-1)\times 100\% = -27.39\%$ por pasar de mujer a hombre

IV) El momio entre Clinton y Perrot disminuye en $(e^{-0.17}-1)\times 100\% = -15.633\%$ por pasar de no tener estudios universitarios a sí tenerlos.

(6)

b) Encuentra la probabilidad de que vote por cada uno de los candidatos para todos los perfiles.

I) Hombre con estudios universitarios

Clinton: $\hat{p}_1 = \dfrac{1}{1 + e^{-0.35+0.88+0.47} + e^{-0.52-0.32-0.17}} = 0.244948$

Bush H: $\hat{p}_2 = \dfrac{e^{-0.35+0.88+0.47}}{1 + e^{-0.35+0.88+0.47} + e^{-0.52-0.32-0.17}} = 0.665837$

Perrot: $\hat{p}_3 = \dfrac{e^{-0.52-0.32-0.17}}{1 + e^{-0.35+0.88+0.47} + e^{-0.52-0.32-0.17}} = 0.089215$

II) Hombre sin estudios universitarios

Clinton: $\hat{p}_1 = \dfrac{1}{1 + e^{-0.35+0.88} + e^{-0.52-0.32}} = 0.319423$

Bush H: $\hat{p}_2 = \dfrac{e^{-0.35+0.88}}{1 + e^{-0.35+0.88} + e^{-0.52-0.32}} = 0.542678$

Perrot: $\hat{p}_3 = \dfrac{e^{-0.52-0.32}}{1 + e^{-0.35+0.88} + e^{-0.52-0.32}} = 0.137898$

III) Mujer con estudios universitarios

Clinton: $\hat{p}_1 = \dfrac{1}{1 + e^{-0.35+0.47} + e^{-0.52-0.17}} = 0.380362$

Bush H: $\hat{p}_2 = \dfrac{e^{-0.35+0.47}}{1 + e^{-0.35+0.47} + e^{-0.52-0.17}} = 0.428857$

Perrot: $\hat{p}_3 = \dfrac{e^{-0.52-0.17}}{1 + e^{-0.35+0.47} + e^{-0.52-0.17}} = 0.190781$

IV) Mujer sin estudios universitarios

Clinton: $\hat{p}_1 = \dfrac{1}{1+e^{-0,35}+e^{-0,52}} = 0,434932$

Bush H: $\hat{p}_2 = \dfrac{e^{-0,35}}{1+e^{-0,35}+e^{-0,52}} = 0,306492$

Perrot: $\hat{p}_3 = \dfrac{e^{-0,52}}{1+e^{-0,35}+e^{-0,52}} = 0,258576$

4) $\begin{pmatrix} X_1 \\ X_2 \\ X_3 \end{pmatrix} \sim N_3\left(\begin{pmatrix} 1 \\ 1 \\ 2 \end{pmatrix}, \begin{pmatrix} 2 & 1 & 3 \\ 1 & 4 & 5 \\ 3 & 5 & 8 \end{pmatrix} \right)$

a) Encuentra la distribución condicional de X_3 dado que $X_1 = 1$, $X_2 = 2$

Definimos a $\underline{X_1} = \begin{pmatrix} X_1 \\ X_2 \end{pmatrix}$, $\underline{X}_2 = (X_2)$

Sabemos que $\underline{X_2} \mid \underline{X_1} \sim N(\underline{\mu}_2 + \Sigma_{21}\Sigma_{11}^{-1}(\underline{x}_1 - \underline{\mu}_1), \Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12})$

donde $\underline{\mu}_1 = \begin{pmatrix} 1 \\ 1 \end{pmatrix}$ $\quad \Sigma_{11} = \begin{pmatrix} 2 & 1 \\ 1 & 4 \end{pmatrix}$ $\quad \Sigma_{12} = \begin{pmatrix} 3 \\ 5 \end{pmatrix}$ entonces

$\underline{\mu}_2 = 2$ $\quad \Sigma_{21} = (3 \;\; 5)$ $\quad \Sigma_{22} = (8)$

$$\mu_{\underline{X}_2 \mid \underline{X}_1} = \underline{\mu}_2 + \Sigma_{21}\Sigma_{11}^{-1}(\underline{x}_1 - \underline{\mu}_1) = 2 + (3 \;\; 5)\, \frac{1}{7}\begin{pmatrix} 4 & -1 \\ -1 & 2 \end{pmatrix}\begin{pmatrix} x_1 - 1 \\ x_2 - 1 \end{pmatrix}$$

$$= 2 + \frac{1}{7}(7, 7)\begin{pmatrix} x_1 - 1 \\ x_2 - 1 \end{pmatrix} = 2 + (1, 1)\begin{pmatrix} x_1 - 1 \\ x_2 - 1 \end{pmatrix} = \cancel{2} + x_1 + x_2 \cancel{-2}$$

$$\therefore \mu_{\underline{X}_2 \mid \underline{X}_1} = x_1 + x_2 \underset{\substack{X_1 = 1 \\ X_2 = 2}}{\Longrightarrow} \mu_{\underline{X}_2 \mid \underline{X}_1} = 3$$

$$\Sigma_{\underline{X}_2|\underline{X}_1} = \Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12} = 8 - (3\ 5)\frac{1}{7}\begin{pmatrix} 4 & -1 \\ -1 & 2 \end{pmatrix}\begin{pmatrix} 3 \\ 5 \end{pmatrix}$$

$$= 8 - \frac{1}{7}(7, 7)\begin{pmatrix} 3 \\ 5 \end{pmatrix} = 8 - (1,1)\begin{pmatrix} 3 \\ 5 \end{pmatrix} = 8 - 8 = 0$$

Así, $\Sigma_{\underline{X}_2|\underline{X}_1} = 0 \quad \forall (x_1, x_2) \in S_{x_1, x_2}$

$\therefore \ X_2 | X_1 = 1, X_2 = 2 \sim N(3, 0)$

Como $\Sigma_{\underline{X}_2|\underline{X}_1} = 0$ resulta que $X_2 / X_1 = 1, X_2 = 2 \equiv 3$ (constante)

b) Encuentra la distribución de $X_1 + X_2 - X_3$

Definimos a $Y = X_1 + X_2 - X_3$ Vemos que

$$Y = X_1 + X_2 - X_3 = (1, 1, -1)\begin{pmatrix} X_1 \\ X_2 \\ X_3 \end{pmatrix} = AX, \quad A = (1, 1, -1), \quad X = \begin{pmatrix} X_1 \\ X_2 \\ X_3 \end{pmatrix}$$

entonces,

$$Y \sim N(A\underline{\mu}_X + A\Sigma_X A^T)$$

por lo que,

$$\mu_Y = A\mu_X = (1, 1, -1)\begin{pmatrix} 1 \\ 1 \\ 2 \end{pmatrix} = 0$$

$$\Sigma_Y = A\Sigma_X A^T = (1, 1, -1)\begin{pmatrix} 2 & 1 & 3 \\ 1 & 4 & 5 \\ 3 & 5 & 8 \end{pmatrix}\begin{pmatrix} 1 \\ 1 \\ -1 \end{pmatrix} = (1,1,1)\begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix} = 0$$

Así pues,

$$Y \sim N(0, 0)$$

Pero, como $\Sigma_Y = 0$, resulta que $Y = X_1 + X_2 - X_3 \equiv 0$ (constante)

5) Tenemos las siguientes variables estandarizadas:

	Componente Principal 1	Componente Principal 2
Física	0.55	−0.02
Química	0.50	0.23
Educación Física	0.02	0.02
Literatura	−0.13	0.85
Inglés	−0.15	0.84

Los eigenvalores son 3 y 1.5.

a) Interpreta los componentes Principales, obtén la correlación de las variables con los componentes principales.

Componente Principal 1: Ciencias exactas

Componente Principal 2: Lenguas

Como están estandarizadas, $\sigma^2 = 1$

Correlación $\rho(X_i, Y_j) = \frac{\sqrt{\lambda_j}\, a_{ji}}{\sigma_i} = \sqrt{\lambda_j}\, a_{ji}$

Variable	Primer componente	Segundo Componente
Física	0,952628	−0,024495
Química	0,866025	0.281691
Educación Física	0,034641	0,024495
Literatura	−0,225167	1,04103
Inglés	−0,259808	1.02879

- Las variables más correlacionadas con el componente principal son Física y Química
- Las variables más correlacionadas con el componente principal son Literatura e Inglés.

b) ¿Qué porcentaje de la varianza explican los primeros dos componentes principales?

	λ_i	$\frac{\lambda_i}{\text{Var total}}$	% λ_r	acumulado	% Acumulado
Componente 1	3	$3/5 = 0,6$	60%	3	60%
Componente 2	1.5	$1.5/5 = 0,3$	30%	4.5	30%

c) Si las calificaciones estandarizadas de un alumno son las siguientes, ubicalo en el plano cartesiano de los primeros dos componentes principales.

	Componente Principal 1
Física	0.15
Química	0.12
Educación Física	0.22
Literatura	-0.03
Inglés	-0.05

Comp 1: $(0.15)(0.55)+(0.12)(0.50)+(0.22)(0.02)+(-0.03)(-0.13)+(-0.05)(-0.15)= 0.1583$

Comp 2: $(0.15)(-0.07)+(0.12)(0.23)+(0.22)(0.02)+(-0.03)(0.85)+(-0.05)(0.04) = -0.0385$