CdL in Scienze Statistiche ed Economiche - Università degli Studi di Milano-Bicocca

Lezione : Approfondimenti ed esercizi

Docente: Aldo Solari

1 L'analisi dei gruppi

Example 1.1. *Distanza tra gruppi: legame completo.*

Passo ①: Inizializzare $k = n$ *e* $\underset{k\times k}{D} = \underset{n\times n}{D}$

$\underset{5\times 5}{D} = \{d_{IL}\} =$

$I\backslash L$	1	2	3	4	5
1	0				
2	9	0			
3	3	7	0		
4	6	5	9	0	
5	11	10	**2**	8	0

ITERAZIONE 1

② $\min_{I\neq L}(d_{IL}) = d_{53} = 2$

- *Le due unità (cluster) 3 e 5 vengono fuse nel cluster (35)*

③ *Aggiorno le distanze tra il nuovo cluster (35) e i rimanenti*

- $d_{(35)1} = \max\{d_{31}, d_{51}\} = \max\{3, 11\} = 11$
- $d_{(35)2} = \max\{d_{32}, d_{52}\} = \max\{7, 10\} = 10$
- $d_{(35)4} = \max\{d_{34}, d_{54}\} = \max\{9, 8\} = 9$

dove il legame completo $d_{(IL)J} = \max\{d_{IJ}, d_{LJ}\}$

$\underset{4\times 4}{D} = \{d_{IL}\} =$

$I\backslash L$	(35)	1	2	4
(35)	0			
1	11	0		
2	10	9	0	
4	9	6	**5**	0

ITERAZIONE 2

② $\min_{I\neq L}(d_{IL}) = d_{42} = 5$

- *I due cluster 2 e 4 vengono fusi nel cluster (24)*

③ *Aggiorno le distanze tra il nuovo cluster (24) e i rimanenti*

- $d_{(24)(35)} = \max\{d_{2(35)}, d_{4(35)}\} = \max\{10, 9\} = 10$
- $d_{(24)1} = \max\{d_{21}, d_{41}\} = \max\{9, 6\} = 9$

$\underset{3\times3}{D} = \{d_{IL}\} =$

$I\backslash L$	*(35)*	*(24)*	*1*
(35)	*0*		
(24)	*10*	*0*	
1	*11*	***9***	*0*

ITERAZIONE 3

② $\min_{I\neq L}(d_{IL}) = d_{1(24)} = 9$

- *I due cluster 1 e (24) vengono fusi nel cluster (124)*

③ *Aggiorno le distanze tra il nuovo cluster (124) e il rimanente*

- $d_{(124)(35)} = \max\{d_{1(35)}, d_{(24)(35)}\} = \max\{11, 10\} = 11$

$\underset{2\times2}{D} = \{d_{IL}\} =$

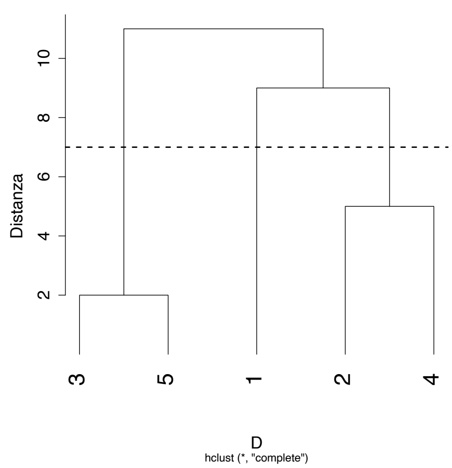
$I\backslash L$	*(35)*	*(124)*
(35)	*0*	
(124)	***11***	*0*

ITERAZIONE 4

② $\min_{I\neq L}(d_{IL}) = d_{(35)(124)} = 11$

- *I due cluster (35) e (124) vengono fusi nel cluster (12345)*

③ *STOP*



$$\cancel{y_{i,1} = \frac{1}{\sqrt{2}}(z_{i,1} + z_{i,2})} \qquad \cancel{y_{i,1} = \frac{1}{\sqrt{2}}(z_{i,1} - z_{i,2})}$$

~~La varianza spiegata é pari a:~~

$$\cancel{\frac{\lambda_1}{\lambda_1 + \lambda_2} = 0.81} \tag{6}$$

~~2. I risultati non coincidono in quanto l'analisi delle componenti principali non é invariante rispetto trasformazioni di scala pertanto quando si effettua l'analisi é necessario valutare se farla a partire da~~ $\cancel{\tilde{X}}$ ~~o da~~ $\cancel{Z}$.

~~3. Ricordando che la correlazione tra~~ $\cancel{Z_j}$ ~~e i punteggi~~ $\cancel{Y_k}$ ~~é pari a~~ $\cancel{v_{j,k}\sqrt{\lambda_k}}$
~~Segue:~~

$$\cancel{cor_{Y_1,Z_1} = \frac{\sqrt{(0.36)}}{\sqrt{2}}} \qquad \cancel{cor_{Y_1,Z_2} = \frac{\sqrt{(1.63)}}{\sqrt{2}}} \qquad \cancel{cor_{Y_2,Z_1} = -\frac{\sqrt{(0.36)}}{\sqrt{2}}}$$

Esercizio 3.
1. Data la matrice di dati si trasformino le variabili in variabili binarie e si calcolino il coefficiente di similaritá semplice e quello di Jaccard i presidenti Nixon e Johnson e tra Nixon e Kennedy.

Presidente	Luogo di Nascita	Eletto	Partito	Esperienze pregresse al congresso	Vicepresidente
Nixon	ovest	si	rep.	si	si
Kennedy	est	si	dem.	si	no
Johnson	sud	no	dem.	si	si

Soluzione:
Definendo

$$X_1 = \begin{cases} 1 \ \ se \ \ sud \\ 0 \ \ altrimenti \end{cases} \tag{7}$$

$$X_i = \begin{cases} 1 \ \ se \ \ si \\ 0 \ \ altrimenti \end{cases} \tag{8}$$

$$X_3 = \begin{cases} 1 \ \ se \ \ repubblicano \\ 0 \ \ altrimenti \end{cases} \tag{9}$$

Si ottiene

Presidente	X_1	X_2	X_3	X_4	X_5
Nixon	0	1	1	1	1
Kennedy	0	1	0	1	0
Johnson	0	0	0	1	0

I coefficienti di similaritá tra Nixon e Johnson valgono:

$$s_c = s_j = \frac{2}{5}$$

Mentre i coefficienti di similaritá semplice tra Nixon e Kennedy risultano:

$$s_c = \frac{3}{5} \qquad s_j = \frac{3}{4}$$