ANALISI DEI DATI PER IL MARKETING 2014

Marco Riani
mriani@unipr.it
http://www.riani.it

OBIETTIVO del MULTIDIMENSIONAL SCALING

DALLA MATRICE DI DISSIMILARITA' D RICOSTRUIRE LA MATRICE DEI DATI X

Dati di input

- I dati di input sono una o più matrici di misure di prossimità (dissomiglianza)
  - Di tipo diretto (raccogliendo i giudizi espressi da uno i più individui)
  - Calcolate in funzione di caratteri quantitativi e/o qualitativi

Caratteristiche della misura di dissomiglianza δ_{ij}

- non negatività $\delta_{ij} >= 0$ $\forall$ i,j
- identità $\delta_{ii} = 0$ $\forall$ i
- uguaglianza delle unità se $u_i = u_j$ ➔ $\delta_{ij} = 0$
- simmetria $\delta_{ij} = \delta_{ji}$

- Osservazione: se $u_i = u_j$ se e solo se $\delta_{ii} = 0$ (nullità) + disuguaglianza triangolare allora la dissimilarità diventa una distanza

Modelli di multidimensional scaling

- INPUT: una matrice di indici di prossimità n × n
- OUPUT: una matrice dei dati X di dimensioni ridotte (ad es n × 2 oppure n × 3)

Es. input sono le distanze Euclidee

$$_2d_{ij} = \sqrt{\sum_{s=1}^{p}(x_{is} - x_{js})^2}$$

$$_2d_{ij} = \sqrt{((e_i - e_j)^T X X^T (e_i - e_j))}$$

$$Q = XX^T$$

$$_2d_{ij}^2 = (e_i - e_j)^T Q (e_i - e_j)$$

$$d_{ij}^2 = e_i^T Q e_i + e_j^T Q e_j - 2e_i^T Q e_j$$

$$d_{ij}^2 = q_{ii} + q_{jj} - 2q_{ij}$$

$$q_{ii} = \sum_{i=1}^{p} x_{is}^2 \qquad q_{jj} = \sum_{j=1}^{p} x_{js}^2 \qquad -2q_{ij} = -2\sum_{i=1}^{p} x_{is}x_{js}$$

$$d_{ij}^2 = q_{ii} + q_{jj} - 2q_{ij}$$

$q_{ii} = \sum_{i=1}^{p} x_{is}^2$ $\quad q_{jj} = \sum_{j=1}^{p} x_{js}^2$ $\quad -2q_{ij} = -2\sum_{i=1}^{p} x_{is}x_{js}$

$$d_{ij}^2 = \sum_{i=1}^{p} x_{is}^2 + \sum_{j=1}^{p} x_{js}^2 - 2\sum_{i=1}^{p} x_{is}x_{js}$$

Se opero in termini di scostamenti dalla media $Q = \tilde{X}\tilde{X}^T$

- La matrice delle distanze Euclidee rimane sempre la stessa sia che si operi sulla matrice originaria oppure sulla matrice Xtilde (scostamenti dalla media)
- La somma delle righe e delle colonne della matrice Q=Xtilde*Xtilde^T è zero se si opera in termini di scostamenti dalla media

$q_{ij} = \sum_{s=1}^{p} x_{is}x_{js}$ diventa $q_{ij} = \sum_{s=1}^{p} (x_{is} - \overline{x}_s)(x_{js} - \overline{x}_s)$

$$\sum_{i=1}^{n} q_{ij} = \sum_{i=1}^{n}\sum_{s=1}^{p} (x_{is} - \overline{x}_s)(x_{js} - \overline{x}_s) = 0$$

$$\sum_{j=1}^{n} q_{ij} = \sum_{j=1}^{n}\sum_{s=1}^{p} (x_{is} - \overline{x}_s)(x_{js} - \overline{x}_s) = 0$$

Osservazione

- Senza perdita di generalità nelle slide che seguono la matrice Q indicata con X^*X^T in realtà si riferisce alla matrice

  $Q = \tilde{X}^*\tilde{X}^T$

Relazione tra l'elemento ij della matrice D^2 (quadrati distanze euclidee) e l'elemento ij della matrice $Q = XX^T$

$$d_{ij}^2 = q_{ii} + q_{jj} - 2q_{ij}$$

$$\text{Ponendo } B = tr \sum_{i=1}^{n} q_{ii}$$

$$\sum_{i=1}^{n} d_{ij}^2 = B + nq_{jj}$$

$$\sum_{j=1}^{n} d_{ij}^2 = nq_{ii} + B$$

$$\sum_{i=1}^{n} \sum_{j=1}^{n} d_{ij}^2 = 2nB$$

$$2B = \frac{\sum_{i=1}^{n} \sum_{j=1}^{n} d_{ij}^2}{n}$$

$$q_{jj} = \frac{\sum_{i=1}^{n} d_{ij}^2 - B}{n} = \frac{d_{.j}^2}{n} - \frac{B}{n}$$

$$q_{ii} = \frac{\sum_{j=1}^{n} d_{ij}^2 - B}{n} = \frac{d_{i.}^2}{n} - \frac{B}{n}$$

$$d_{ij}^2 = \frac{d_{i.}^2}{n} - \frac{B}{n} + \frac{d_{.j}^2}{n} - \frac{B}{n} - 2q_{ij}$$

$$d_{ij}^2 = \frac{d_{i.}^2}{n} + \frac{d_{.j}^2}{n} - \frac{\sum_{i=1}^{n} \sum_{j=1}^{n} d_{ij}^2}{n^2} - 2q_{ij}$$

Relazione tra l'elemento ij della matrice D^2 (quadrati distanze euclidee) e l'elemento ij della matrice $Q = XX^T$

$$d_{ij}^2 = \frac{d_{i.}^2}{n} + \frac{d_{.j}^2}{n} - \frac{\sum_{i=1}^n \sum_{j=1}^n d_{ij}^2}{n^2} - 2q_{ij}$$

$$q_{ij} = -\frac{1}{2}\left(d_{ij}^2 - \frac{d_{i.}^2}{n} - \frac{d_{.j}^2}{n} + \frac{\sum_{i=1}^n \sum_{j=1}^n d_{ij}^2}{n^2} \right)$$

$$\frac{d_{i.}^2}{n} = \text{Media della } i\text{-esima riga della matrice } D^2$$

$$\frac{d_{.j}^2}{n} = \text{Media della } j\text{-esima colonna della matrice } D^2$$

$$\frac{\sum_{i=1}^n \sum_{j=1}^n d_{ij}^2}{n^2} = \text{Media degli } n^2 \text{ elementi della matrice } D^2$$

Relazione tra l'elemento ij della matrice D^2 (quadrati distanze euclidee) e l'elemento ij della matrice $Q = XX^T$

$$d_{ij}^2 = \frac{d_{i.}^2}{n} + \frac{d_{.j}^2}{n} - \frac{\sum_{i=1}^n \sum_{j=1}^n d_{ij}^2}{n^2} - 2q_{ij}$$

$$q_{ij} = -\frac{1}{2}\left(d_{ij}^2 - \frac{d_{i.}^2}{n} - \frac{d_{.j}^2}{n} + \frac{\sum_{i=1}^n \sum_{j=1}^n d_{ij}^2}{n^2} \right)$$

In termini matriciali

$$Q = -\frac{1}{2} H D^2 H$$

$$H = I_n - 1_n \times 1_n^T / n$$

Ricostruzione della matrice X

$$Q = XX^T = -\frac{1}{2}HD^2H$$

Scomposizione spettrale della matrice Q

$$Q = XX^T = A\Lambda A^T$$

La miglior rappresentazione di rango 2 della matrice X è data da $A_{(2)}\Lambda^{0.5}_{(2)}$

$A_{(2)}$ Matrice di dimensione n x **2** che contiene (nelle colonne) i primi due autovettori della matrice Q

$\Lambda^{0.5}_{(2)}$ Matrice diagonale di dimensione 2 x **2** che contiene sulla diagonale principale le radici quadrate dei primi due autovalori della matrice Q

Scelta del numero di dimensioni

- Se la matrice di dissimilarità è una matrice di distanze allora Q ha rango n-1, tutti gli autovalori sono positivi e l'indice per valutare la bontà della soluzione in uno spazio ridotto a k dimensioni è

$$\frac{\sum_{i=1}^{k}\lambda_i}{\sum_{i=1}^{n-1}\lambda_i}$$

Scelta del numero di dimensioni

- Se la matrice di dissimilarità non è una matrice di distanze e non è semidefinita positiva l'indice per valutare la bontà della soluzione in uno spazio ridotto a k dimensioni è

$$\frac{\sum_{i=1}^{k} \lambda_i}{\sum_{i=1}^{n-1} |\lambda_i|}$$

Osservazione

- La soluzione trovata è definita a meno di una rotazione e/o traslazione

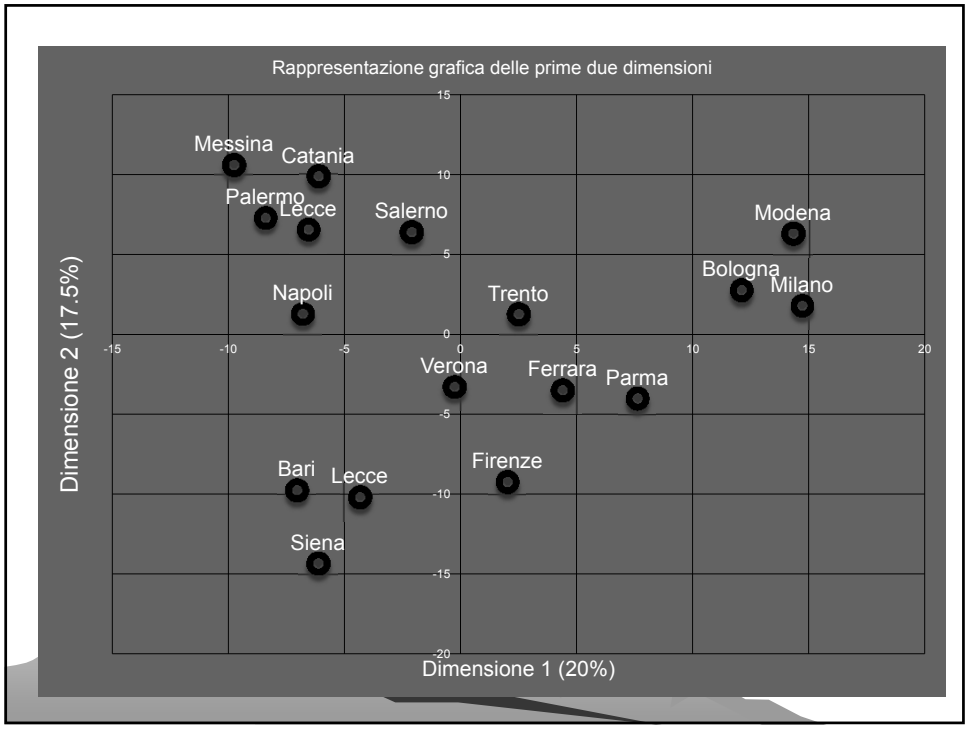
Esempio

Matrice dissimilarità derivante derivante dai questionari somministrati a 100 individui a cui è stato chiesto di valutare (sulla scala 0-30) il grado di diversità di 17 città Italiane.

	Salerno	Napoli	Firenze	Siena	Pescara	Verona	Parma	Bologna	Palermo	Messina	Roma	Catania	Bari	Lecce	Trento	Milano	Modena
Salerno	0																
Napoli	22	0															
Firenze	23	11	0														
Siena	24	10	15	0													
Pescara	26	17	19	10	0												
Verona	27	7	28	28	28	0											
Parma	26	16	20	25	20	24	0										
Bologna	23	12	24	26	27	28	22	0									
Palermo	24	21	24	28	24	29	28	27	0								
Messina	23	21	29	28	27	26	28	29	21	0							
Roma	23	22	27	20	24	26	24	23	25	23	0						
Catania	18	21	28	27	25	29	26	27	26	16	19	0					
Bari	23	19	21	8	19	18	25	25	25	25	21	26	0				
Lecce	23	18	26	26	24	29	29	25	12	24	26	26	25	0			
Trento	28	21	28	28	29	14	24	29	29	27	27	28	29	29	0		
Milano	24	15	24	28	24	28	21	12	29	29	29	27	26	28	27	0	
Modena	22	12	23	29	28	29	20	13	27	28	27	25	29	26	26	12	0

Questiti

- Applicando le tecniche di multidimensional scaling ricostruire la matrice dei dati originaria limitandosi ad analizzare la migliore approssimazione di rango 2. Fornire un giudizio sulla bontà di adattamento.
- Rappresentare graficamente la configurazione in due dimensioni commentando i risultati ottenuti.
- Inserire come titoli degli assi la bontà di adattamento delle prime due dimensioni.


Rappresentazione grafica delle prime due dimensioni
Messina
Catania
Palermo
Lecce
Salerno
Modena
Bologna
Milano
Napoli
Trento
Verona
Ferrara
Parma
Firenze
Bari
Lecce
Siena
Dimensione 2 (17.5%)
Dimensione 1 (20%)
15
10
5
0
-5
-10
-15
-20
-15
-10
-5
0
5
10
15
20