

Noi, gli altri e l'analisi statistica dei dati



La crescita delle capacità dei computer

La disponibilità di tali mezzi a costi contenuti

L'efficienza ed efficacia del software per utilizzarli

Hanno reso strumenti sofisticati alla portata di tutti.

Chi sfreccia veloce sul mare dei dati non può che raggiungere risultati facili e immediati.

Per arrivare a mete prestigiose è necessaria la pensosa pazienza dei sub.



Strutture di dati



Matrice dei dati: organizzazione righe-colonne (nxm) collocata al centro di molte tecniche di analisi multivariata

n numero di entità
 m numero di variabili

$X =$

Matrice rettangolare

	V_1	V_2	V_j	V_m
U_1	x_{11}	x_{12}	x_{1j}	x_{1m}
U_2	x_{21}	x_{22}	x_{2j}	x_{2m}
U_i	x_{i1}	x_{i2}	x_{ij}	x_{im}
U_n	x_{n1}	x_{n2}	x_{in}	x_{nm}



Matrice delle dissimilarità o distanze di ordine (nxn)

Spesso la matrice delle dissimilarità o delle affinità è la prima elaborazione della matrice dei dati per poi servire da base per ulteriori analisi.

Altre volte è il punto di partenza delle analisi

$D =$

Matrice quadrata

	U_1	U_2	U_i	U_n
U_1	0	d_{12}	d_{1i}	d_{1n}
U_2	d_{12}	0	d_{2i}	d_{2n}
U_i	d_{1i}	d_{2i}	0	d_{in}
U_n	d_{1n}	d_{2n}	d_{in}	0

La matrice dei dati

Il risultato di molti percorsi di ricerca è l'individuazione della matrice dei dati, X , che costituirà poi l'oggetto dell'analisi.

Si tratterà di una matrice rettangolare di dimensioni $(n \times m)$, in cui le n righe saranno costituite dalle unità oggetto di indagine, e le m colonne rappresenteranno le variabili che sono state rilevate per ciascuna unità.

Vettori riga per le unità →

Vettori colonna per le variabili ↓

Unità	Indicatori					
	X_1	X_2		X_j		X_m
I_1	x_{11}	x_{12}		x_{1j}		x_{1m}
I_2	x_{21}	x_{22}		x_{2j}		x_{2m}
.....						
I_i	x_{i1}	x_{i2}		x_{ij}		x_{im}
.....						
I_n	x_{n1}	x_{n2}		x_{nj}		x_{nm}

Con x_{ij} possiamo indicare quindi il valore che il j -esimo indicatore di base assume nell'unità i -esima

Esempio

Analisi di un modello di crescita

DATA APPENDIX

Country	y_{1980}	y_{2000}	n	s_n
Argentina	11747	17494	1.4	17.8
Australia	17245	38079	1.8	24.7
Austria	11263	34566	0.4	26.1
Burundi	1073	1158	1.9	5.0
Belgium	35065	35625	0.3	24.0
Benin	2140	2248	2.5	6.5
Burkina Faso	1281	1898	1.9	8.4
Bangladesh	1964	2855	2.5	10.0
Bolivia	3975	4854	2.4	10.1
Brazil	4644	10901	2.7	20.6
Canada	17247	39512	1.7	21.9
Switzerland	21154	39476	0.8	27.8
Chile	7103	15097	2.1	16.0
Cote d'Ivoire	3686	3464	3.7	8.2
Cameroon	3158	4051	2.4	6.9
Congo	1191	4394	2.6	23.0
Colombia	4840	8634	2.8	11.5
Costa Rica	7158	8753	3.5	14.2
Denmark	16786	40343	0.5	23.4
Dominican Rep.	3354	8472	2.9	12.4
Algeria	4729	9247	3.0	17.9
Ecuador	4113	5650	3.0	20.1
Egypt	3006	6813	2.5	7.0
Spain	7222	26556	0.8	24.5
Ethiopia	1026	1294	2.5	4.4
Finland	11850	34318	0.6	26.4
France	12903	34931	0.8	24.7
United Kingdom	14819	35009	0.3	18.3
Ghana	3273	2297	2.8	10.1
Greece	6178	21747	0.7	25.8
Guatemala	5090	7656	2.7	8.0
Hong Kong	4946	36058	2.6	25.8
Honduras	3214	3678	2.2	12.3
Indonesia	1354	5811	2.3	12.2
India	1521	4928	2.3	11.6
Ireland	9029	38195	1.1	17.9
Israel	10285	28838	2.8	28.2
Italy	10328	31764	0.4	24.9
Jamaica	4343	5249	1.5	19.1
Jordan	3911	6940	4.7	13.1
Japan	7386	35982	0.9	31.0
Kenya	1764	2277	3.4	11.1
Korea	2685	19552	2.3	27.3

Variabili

Unità

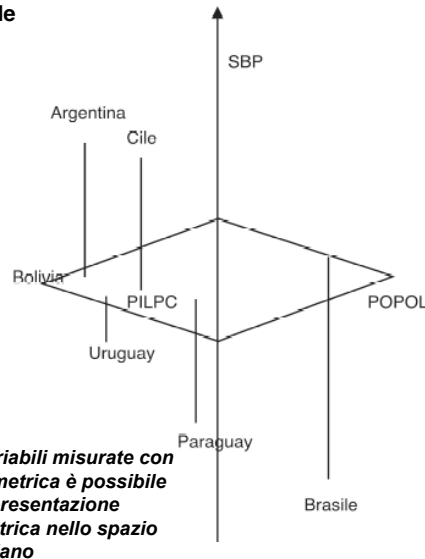
Rappresentazione geometrica dei dati

Ogni riga della matrice dei dati corrisponde ad un punto nello spazio R^m

$$(x_{i1}, x_{i2}, \dots, x_{im})$$

Sono incluse le rilevazioni sulle variabili relative alla i-esima unità. Di tali vettori se ne usano n (l'ampiezza della rilevazione). L'insieme delle rilevazioni costituisce la matrice dei dati.

	Pop.	Pil.pc	Sbp	
$X =$	34.6	8.121	2.165	Argentina
	155.9	4.335	-3.200	Brasile
	4.8	1.854	-1.675	Paraguay
	3.2	5.563	-0.761	Uruguay
	8.1	8.640	-0.222	Bolivia
	14.2	4.739	1.374	Cile



Per variabili misurate con scala metrica è possibile la rappresentazione geometrica nello spazio cartesiano

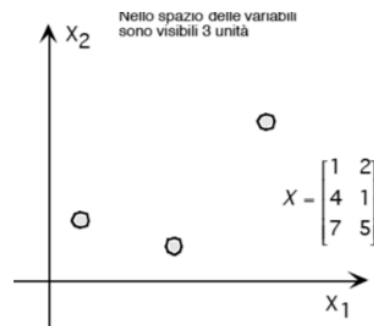
Spazio delle variabili

In alternativa si potrebbe puntare di più sulle relazioni tra le unità ed in questo caso la matrice dei dati è considerata un aggregato di n righe.

$$X = \begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ u_n \end{bmatrix} \quad u_i = [u_{i1}, u_{i2}, \dots, u_{in}]$$

Ogni u_i è il vettore riga formato dalle rilevazioni effettuate su tutte le m variabili rispetto alla i-ma unità.

Si dice che in queste analisi ci si muove nello spazio delle variabili, sottoinsieme di R^m .

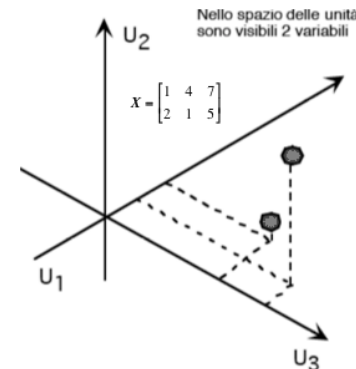


Spazio delle unità

Le analisi statistiche della matrice dei dati sono talvolta dirette alle relazioni tra le variabili ed in questo caso la si considera un aggregato di m colonne

$$X = [x_1, x_2, \dots, x_m]$$

$$x_i = \begin{bmatrix} x_{1i} \\ x_{2i} \\ \vdots \\ x_{ni} \end{bmatrix}$$



Ogni x_i è un vettore colonna formato dalle osservazioni per la i-esima variabile su tutte le n unità coinvolte nella rilevazione.

Si dice che in questa analisi ci si muove nello spazio delle unità che è un sottoinsieme di R^n .

Qui, ogni colonna corrisponde ad un punto.

Obiettivo: semplificare!

La disposizione dei punti dallo spazio (dalle variabili o delle unità) è la struttura che deve essere individuata e spiegata.

L'analisi multivariata può fare riferimento sia alle unità che alle variabili ed in base questa scelta si richiamano certe tecniche e non altre.

L'intento rimane comunque lo stesso: semplificare!



Riduzione del numero di unità: gli n distinti soggetti sono ricondotti a pochi casi tipici le cui caratteristiche rappresentano gli aspetti salienti nel data set



Riduzione del numero di variabili: gli m distinti indicatori sono combinati in pochi altri fattori o variabili latenti che descrivono sinteticamente, al meglio possibile, il contenuto informativo del data set.

Cofronto delle unità

Dalla matrice dei dati si passa alla matrice delle affinità e da questa alla matrice delle dissimilarità o distanze (è possibile pure il passaggio diretto)

$$\begin{array}{c}
 \mathbf{X} = \begin{array}{c|cccc} & V_1 & V_2 & V_j & V_m \\ \hline U_1 & x_{11} & x_{12} & x_{1j} & x_{1m} \\ U_2 & x_{21} & x_{22} & x_{2j} & x_{2m} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ U_i & x_{i1} & x_{i2} & x_{ij} & x_{im} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ U_n & x_{n1} & x_{n2} & x_{in} & x_{nm} \end{array} \longrightarrow \begin{array}{c} \mathbf{D} = \begin{array}{c|cccc} & U_1 & U_2 & U_i & U_n \\ \hline U_1 & 0 & d_{12} & d_{1i} & d_{1n} \\ U_2 & d_{12} & 0 & d_{2i} & d_{2n} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ U_i & d_{1i} & d_{2i} & 0 & d_{in} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ U_n & d_{1n} & d_{2n} & d_{in} & 0 \end{array} \end{array}
 \end{array}$$

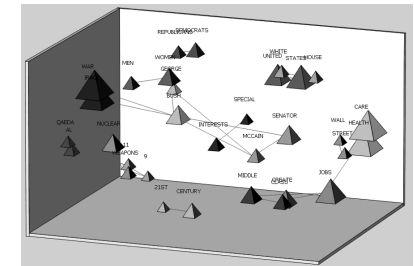
La prima struttura riporta analiticamente tutti i dati, la seconda ne condensa alcune caratteristiche.

La matrice delle dissimilarità può sia essere costruita in base alle variabili di una matrice di dati, ma anche essere ottenuta dal confronto diretto delle unità rispetto a caratteristiche immateriali o indirette.

Matrice delle distanze/dissimilarità

Una matrice quadrata (n x n) in cui ogni elemento rappresenta una misura di come differiscono tra di loro le due unità a confronto.

$$\mathbf{D} = \begin{array}{c|cccc} & U_1 & U_2 & U_i & U_n \\ \hline U_1 & 0 & d_{12} & d_{1i} & d_{1n} \\ U_2 & d_{12} & 0 & d_{2i} & d_{2n} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ U_i & d_{1i} & d_{2i} & 0 & d_{in} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ U_n & d_{1n} & d_{2n} & d_{in} & 0 \end{array}$$



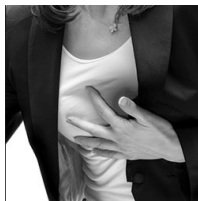
Lo scopo di molte tecniche di analisi multivariata è di ottenere una configurazione geometrica di punti.

Questa dovrebbe essere di aiuto a comprendere come i soggetti valutano gli aspetti sottoposti rispetto a poche dimensioni latenti

Esempio: heart disease

N=303, m=14

1. Eta'
2. Genere
3. Tipo di dolore toracico (1:angina tipica; 2:angina atipica; 3:non anginoso; 4:Asintomatico)
4. Pressione del sangue a riposo (in mm Hg alla accettazione in ospedale)
5. Livello di colesterolo (mg/dl)
6. Glicemia >120 mg (dicotomica)
7. Elettrocardiogramma a riposo (0:normale, 1:con anomalie dell'onda ST-T*, 2:ipertrofia probabile o definita ventricolare sinistra secondo i criteri di Estes)
8. Numero massimo di battiti
9. Comparsa di dolore anginoso sotto stress (categorica)
10. Diminuzione della distanza ST sotto stress rispetto a quella a riposo
11. Pendenza del segmento ST sotto stress (1: ascendente; 2: piatto; 3: Discendente)
12. Numero di vasi principali evidenziati dalla fluoroscopia (da zero a tre)
13. Thal (3: normale; 6: difetto irreversibile; 7: difetto reversibile)
14. Diagnosi di cardiopatia ischemica (rispetto all'angiografia): (0: volume > 50%; 1: volume < 50%); i valori da 1 a 4 si riferiscono al numero di vasi principali risultati occlusi all'analisi angiografica.



Esempio: heart disease/2

Varianza e norme delle variabili di Cleveland con indicazione della scala di misurazione

Variabili	σ_i^2	$\ H\ _{\infty}$	$\ H\ _1$	$\ H\ _2$
1 - numerale	81,8977	77	16199	952,77
2 - dicotomica	0,2195	1	201	14,177
3 - a intervalli	0,9310	4	938	56,903
4 - numerale	31,5173	200	39113	2290
5 - numerale	$2737 \cdot 10^3$	564	73463	4355,6
6 - dicotomica	0,1242	1	43	6,5574
7 - categorica	0,9899	2	296	24,249
8 - numerale	526,3153	202	44431	2608,2
9 - categorica	0,2207	1	97	9,8489
10 - numerale	1,3597	6,2	313,5	27,082
11 - categorica	0,3822	3	476	29,597
12 - numerale?	0,8817	3	201	19,925
13 - categorica	3,7583	7	1405	88,085
14 - numerica	1,5241	4	281	26,777

Ogni tipologia di variabile necessita di una specifica misura di affinità o di dissimilarità.

Alle variabili deve poi essere assegnato un peso: in blocco o singolarmente

Il tema dell'affinità

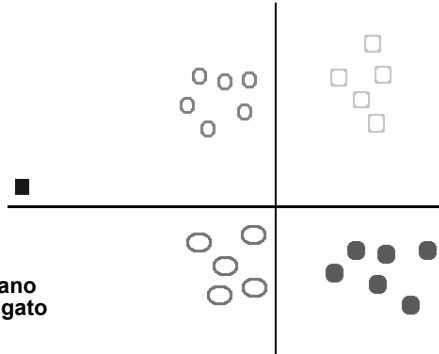
La percezione della similarità rappresenta forse l'asse portante del nostro pensiero.

Una nozione di similarità ha un ruolo fondamentale in contesti quali la diagnosi delle malattie, economia, psicologia, marketing, teoria delle decisioni, sociologia e scienza della politica,

Il tema si affronta valutando su di un piano il modo in cui dei soggetti organizzano degli item al fine di collocare vicini quelli somiglianti e lontani quelli diversi.

Il grafico aiuta a visualizzare le percezioni

La rappresentazione geometrica su piano cartesiano è un passaggio quasi obbligato per lo studio dell'affinità, prossimità, similarità tra item distinti.



La misura dell'affinità

Il problema del confronto nasce quando si considerano almeno due unità rispetto ad una caratteristica suscettibile di almeno due valori.

Il caso minimale è appunto $n=2$ unità, $m=1$ variabile con 2 modalità

$$S(x) = \{x_1, x_2\}$$

Il confronto dei due soggetti i e j rispetto alla variabile X potrà configurare le seguenti situazioni

		unità j	
		x_1	x_2
unità i	x_1	=	$\neq$
	x_2	$\neq$	=

La situazione di indagine è tale che per ogni rilevazione potrà verificarsi una ed una sola delle celle previste nella suddetta tabella tetracorica

La misura dell'affinità/2

L'affinità (o prossimità o contiguità oppure somiglianza) tra due unità è la percezione di qualche loro tratto palese o latente che porta a collocarle in un'unica categoria piuttosto che in categorie diverse.

Se i due stati possibili della X sono due località o due periodi allora le entità potrebbero essere giudicate affini se sono coeve oppure occupano lo stesso sito (contiguità temporale o spaziale).

Se le due modalità indicano due aspetti che siano parti, componenti, organi, etc. allora l'affinità potrebbe derivare da qualcosa che le unità hanno in comune o su cui hanno un comune effetto

Se invece X_1 è la presenza di una caratteristica X_2 la sua assenza, l'affinità sarà legata alla condivisione di quella presenza e/o dell'assenza.

All'aumentare del numero di modalità ed all'aumentare delle relazioni che si possono instaurare tra le modalità stesse (tenuto conto della loro scala di misurazione), l'idea di affinità diventa più articolata e la sua misura sempre più ricca di possibilità.

Rilevazione dell'affinità

Ecco tre diversi modi per rilevare l'affinità

◆ **RATING.** Un modo immediato di acquisire un valore per l'affinità e di richiedere ai soggetti di assegnare un voto: (1-10) (18-30) (0-100) in cui l'estremo inferiore implica zero somiglianza e l'estremo superiore la massima somiglianza.

◆ **PROBABILITA' DI CONFUSIONE.** L'affinità si può rilevare ipotizzando la probabilità di confusione tra gli item. Se la probabilità di confusione è nulla allora gli item sono estremamente dissomiglianti. Se è certa la confusione allora gli item sono identici.

◆ **SORTING.** Gli item a confronto sono disposti in gruppo da un certo numero di soggetti. L'affinità tra due item è data dalla proporzione in cui gli item sono stati collocati nello stesso gruppo.

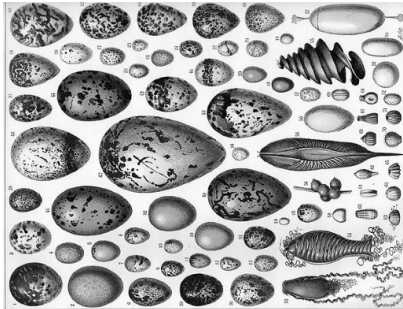
"anything which, by an act of faith, can be considered a similarity"
(Shepard)

Rilevazione dell'affinità/2

Occorre dedicare molta cura alla definizione della metrica: non è facile e ci vuole inventiva oltre che tecnica.

I benefici di una metrica ben fatta sono però altamente remunerativi

Il grado di affinità tra due uova è dato dal confronto tra i volumi d'acqua spostati.



Se il numero di item da comparare è elevato diventa difficile ottenere giudizi attendibili.

$$\binom{n}{2} = \frac{n(n-1)}{2}$$

Spesso, ai "giudici" sono presentati set diversi di coppie.

Se possibile occorre basarsi su elementi neutri.

Nel caso delle figure, ogni affinità riferita alla forma dell'oggetto può dare buoni risultati.

Esempio: codice Morse

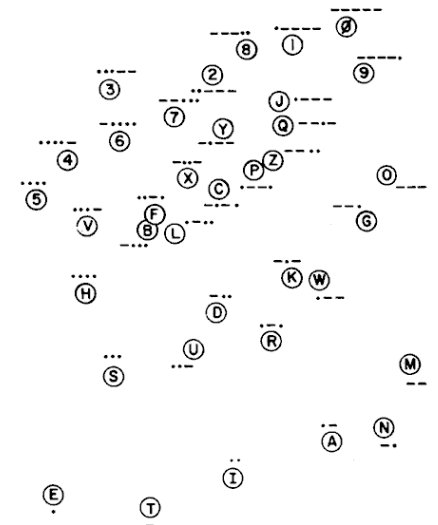
Il codice Morse consiste brevi segnali composti da punti e linee per rappresentare 10 cifre e 26 lettere.

Per valutare la loro confondibilità i segnali sono stati tradotti in suoni (brevi e lunghi) e sottoposti a circa 600 soggetti che non conoscevano il Morse.

I segnali sono stati presentati a coppie 2 volte (prima uno e poi un altro e viceversa)

L'affinità è rilevata con il numero di mancate corrispondenze.

Nel grafico c'è il risultato di una delle procedure di scaling



Rilevazione dell'affinità/3

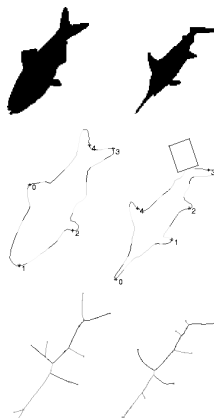
La definizione della metrica è la fase cruciale in una vastissima gamma di ricerche.

In molte occasioni è necessario definire una metrica ad hoc



Che tipo di metrica adottare per valutare la somiglianza-dissomiglianza ne confronto di due pesci?

- La conformazione del bordo
- La lisca
- La combinazione di genere, peso, colore
- Lo sforzo comunque misurato di trasformarne uno in un altro.



Esempio di Gordon (1990)

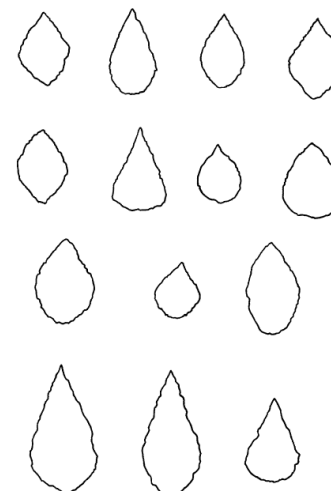
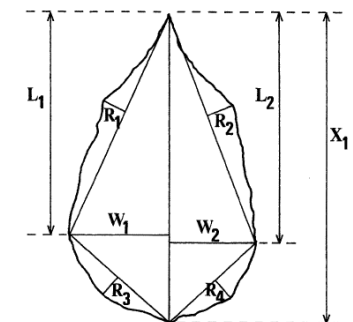


Figure 1. Outlines of 14 leaf-shaped flint arrowheads recovered during archaeological investigations in northeast Scotland.



Definition of variables used to describe the flint arrowheads; see also Figure 2

X_1 = maximum length of arrowhead.

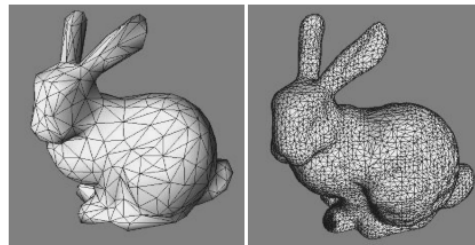
$X_2 = (W_1 + W_2)/X_1$, ratio of maximum width to maximum length.

$X_3 = (L_1 + L_2)/X_1$, location of maximum width.

$X_4 = (R_1 + R_2)/X_1$, convexity of upper part of arrowhead.

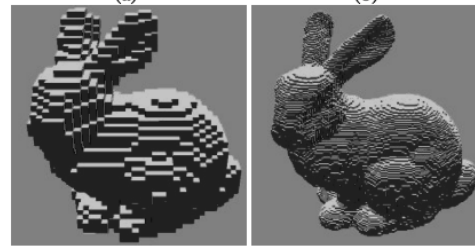
$X_5 = (R_3 + R_4)/X_1$, convexity of lower part of arrowhead.

Affinità e dettagli



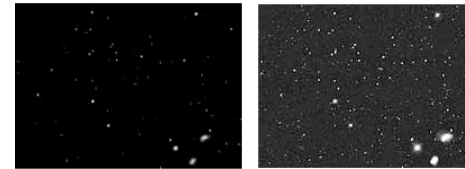
il riconoscimento delle forme e la loro modellazione sono riferite al numero di dettagli che si riescono a considerare.

Maggiori sono i dettagli maggiore è la accuratezza della ricostruzione.



Alcuni dettagli possono essere tralasciati come l'artista che nel dipingere un quadro ignora molte cose che sono invece presenti in una foto.

Analogia del telescopio

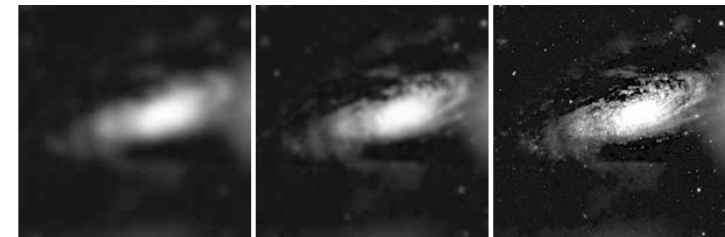


bassa sensibilità

alta sensibilità

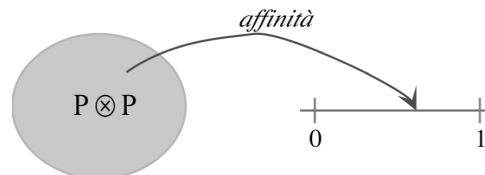
La sensibilità è il livello minimo di affinità che l'indice riesce a distinguere rispetto al rumore di fondo

Maggiore è la risoluzione cioè più numerose sono le variabili, tante più solo differenze che la misura è grado di percepire tra le entità a confronto.



Confronti e affinità

Per ragioni di semplicità preferiamo pensare all'affinità come ad una misura normalizzata compresa nell'intervallo unitario.



Dove $P = \{u_1, u_2, \dots, u_n\}$ è l'insieme delle "n" entità considerate nell'indagine.

Ad ogni elemento del prodotto cartesiano di insiemi $P \otimes P$ è associato un numero reale a_{ij} che esprime il grado di affinità, di prossimità, di contiguità comunque percepito tra le entità a confronto.

L'affinità o similarità è un numero reale non negativo a_{ij} , variamente ottenuto, in grado di quantificare in modo univoco le differenze che separa due entità, se differenti.

Un esempio

Fitch and Margoliash (Science, 1967) hanno determinato la dissimilarità tra specie in base al numero di posizioni nella molecola della proteina cytochrome-c in cui le due specie avevano aminoacidi diversi.

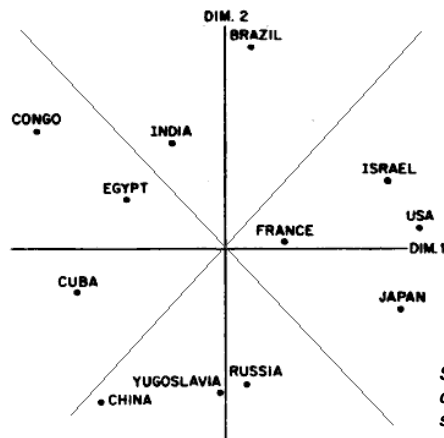
La matrice qui riportata ne include una parte

	Man	Monkey	Horse	Pig	Pigeon	Tuna	Mould	Fungus
Man	0	1	17	13	16	31	63	66
Monkey	1	0	16	12	15	32	62	65
Horse	17	16	0	5	16	27	64	68
Pig	13	12	5	0	13	25	64	67
Pigeon	16	15	16	13	0	27	59	66
Tuna	31	32	27	25	27	0	72	69
Mould	63	62	64	64	59	72	0	61
Fungus	66	65	68	67	66	69	61	0

Da notare la non negatività dei valori, la diagonale nulla e la simmetria

Altro esempio

Un gruppo di soggetti (18) ha classificato n=12 Paesi in base ad un differenziale di 9 livelli dove 1= estremamente diversi e 9=identici



Si possono notare alcuni cluster di Paesi che sono percepiti simili. Gli assi non sono solo un oggetto geometrico, ma sono variabili latenti dietro la percezione dei soggetti che li hanno classificati

STIMOLI	DIM. 1	DIM. 2
BRAZIL	0.15	1.22
CONGO	-1.15	0.71
CUBA	-0.90	-0.29
EGYPT	-0.60	0.29
FRANCE	0.36	0.02
INDIA	-0.33	0.64
ISRAEL	0.96	0.40
JAPAN	1.04	-0.39
CHINA (MAINLAND)	-0.76	-0.96
RUSSIA	0.12	-0.85
U.S.A.	1.14	0.12
YUGOSLAVIA	-0.03	-0.90

Requisiti per le misure di affinità

Indistinguibilità degli identici: Se $u_i = u_j \Rightarrow a_{ij} = 1$

Questo implica che $a_{ij} = a_{ji} = 1$ cioè se si confrontano una entità con se stessa questo deve risultare in un valore unitario (o massimo) dell'affinità.

Distinguibilità dei diversi: Se $u_i \neq u_j \Rightarrow a_{ij} < 1$

Il valore uno deve essere riservato all'identità tra i due soggetti a confronto. Ogni altra comparazione deve dar luogo ad un grado o misura dell'affinità diverso da quello massimo. Non sempre questo requisito può essere rispettato

Monotonicità: se $a_{i,j} < a_{r,s}$ allora le entità i e j sono meno affini di quanto non lo siano le unità r ed s

Questa proprietà garantisce la possibilità di ordinare in sequenza diversi casi di affinità riscontrati

$$0 \leq a_{i_1 i_2} \leq a_{i_2 i_3} \leq \dots \leq a_{i_{n-1} i_n} \leq 1$$

Requisiti per le misure di affinità/2

Simmetria: $a_{ij} = a_{ji} \forall i, j$

Talvolta questa condizione deve essere abbandonata dato che è incongrua con certi aspetti intuitivi dell'affinità: il viaggio da una località A ad un'altra località B non è necessariamente lo stesso che da B ad A dato che i due lati della stessa strada potrebbero avere condizioni di usura diverse.

In alcune soluzioni chimiche l'ordine di miscelazione degli ingredienti implica la formazione di composti differenti



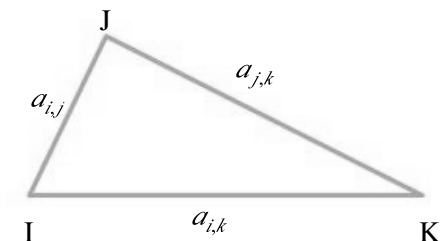
Requisiti per le misure di affinità/3

L'affinità che si può rilevare dal confronto di due unità non può superare quella riscontrabile in una triade di entità comparate due a due:

$$a_{ij} \leq a_{ik} + a_{jk} \forall i, j, k$$

Se i valori delle variabili fossero univocamente rappresentabili come proiezioni su degli assi ortogonali allora ogni unità sarebbe un punto.

La disuguaglianza triangolare assicura che tali punti in triade formino dei triangoli: scaleni, isoscele, equilateri.



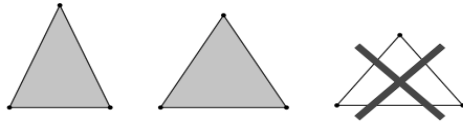
Sebbene la disuguaglianza triangolare sia considerata fondamentale, altri autori la considerano una condizione valida in generale, ma che per un numero ristretto di comparazioni può essere violata.

Requisiti per le misure di affinità/4



Ultrametricità: $a_{ij} \leq \max\{a_{ik}, a_{jk}\} \quad \forall i, j, k$

In questo caso le triadi di unità, se viste come punti dello spazio, possono formare solo un triangolo isoscele.



Se la condizione di ultrametricità è verificata, allora anche la disuguaglianza triangolare è verificata in quanto entrambe rientrano nello schema:

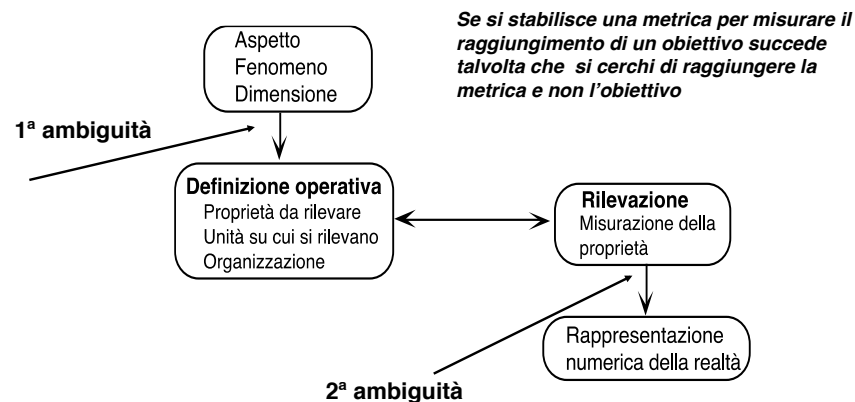
$$a_{ij} \leq \left[a_{ik}^\alpha + a_{jk}^\alpha \right]^{\frac{1}{\alpha}} \quad \forall i, j, k$$

Se $\alpha=1$ si ottiene la disuguaglianza triangolare, ma se α va all'infinito si ottiene la disuguaglianza ultrametrica.

Inoltre, è possibile dimostrare che se la ultrametricità è vera per α allora è anche vera $\beta < \alpha$; ne consegue che la disuguaglianza ultrametrica è una condizione più stringente della triangolare, in quanto la implica e non ne è implicata.

Tecniche di misurazione

il concetto di misurazione è uno dei più controversi tanto che il dibattito è sempre aperto.



Affinità e scale di misurazione

La funzione adottata per misurare l'affinità e le proprietà che la funzione stessa possiede dipende dalla scala di misurazione delle variabili.

Noi ipotizziamo che per ogni unità possano essere presenti variabili del tipo



Binarie simmetriche e asimmetriche



Politome sconnesse



Politome ordinate



Gradatorie



Scale a rapporti o intervallari



In inglese: scale=bilancia

La definizione operativa

E' l'insieme di regole con cui classificare un concetto, determinarne la misura o, in generale, per agganciarlo alla realtà osservabile.

cioè impone la definizione operativa solo con variabili di cui sia possibile seguire con facilità il meccanismo di conversione di una proprietà delle unità in una categoria o valore del dominio.

ESEMPIO

il concetto di interconnessione tra due centri abitati, diciamo "A" e "B", è misurato con la semisomma degli automezzi che si sono spostati da "A" a "B" e quelli che da "B" sono andati ad "A".

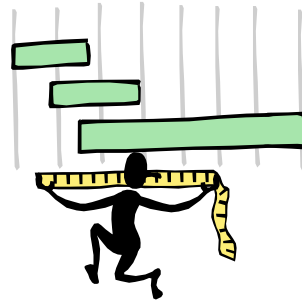


Classificazione e misurazione

L'acquisizione dei dati può avvenire classificando in categorie distinte la proprietà di cui l'unità è portatrice oppure misurandola in base ad una determinata unità di misura.

con la **CLASSIFICAZIONE** si identificano l'unità (e le modalità numeriche del dominio sono equivalenti ad ogni altro insieme di simboli);

con la **MISURAZIONE** si quantifica una proprietà posseduta ed i numeri sono utilizzati in quanto tali.



Classificazione e misurazione/2

La classificazione e la misurazione possono scaturire da due procedure di assegnazione dei valori: enumerazione delle unità rispetto alla proprietà posseduta

oppure comparazione della proprietà studiata rispetto ad un ventaglio di possibilità che, identico per tutte le unità, non dipende né dal numero.

		ASSEGNAZIONE VALORI	
		Enumerazione	Comparazione
RILEVAZIONE DEL DATO	Classificazione	Nominazioni	Scala nominale
	Misurazione	Graduatorie	Scale ordinali semplici Scale ordinali graduate Scale intervallari Scale proporzionali

Scale di misurazione

Molte analisi multidimensionali si collegano all'idea di dissomiglianza o distanza tra una coppia di entità.

Pertanto la scelta della scala di misurazione è essenziale e condiziona pesantemente i risultati.

La scala è una utile caratterizzazione della variabile ed è legata al contesto dell'indagine.

Gli stessi dati possono essere interpretati su scale diverse secondo che interessi solo distinguere le unità oppure graduarne gli scarti.



Scale proporzionali o a rapporti

Ad un incremento relativo nella misura, corrisponde un incremento relativo in ciò che si misura

Ad esempio, per la lunghezza di un segmento, la misura di due centimetri è -senza incertezze- il doppio di uno con lunghezza di un centimetro.



Per queste scale esiste un elemento minimo che si può far coincidere univocamente con l'assenza completa di ciò che si misura.

Nelle scale proporzionali rientrano volume, altezza, area, inclinazione di un piano, resistenza alla tensione dei materiali, prezzo di un bene, durata delle componenti di un sistema di controllo.

La caratteristica distintiva di queste scale è l'invarianza dei rapporti se le misure sono moltiplicate per una costante:

$$Y_1 = bX_1; \quad Y_2 = bX_2 \Rightarrow \frac{Y_1}{Y_2} = \frac{bX_1}{bX_2} = \frac{X_1}{X_2} \quad \text{se } b \neq 0$$

Scale ad intervalli

Nelle scale intervallari si valuta ciò che succede al fenomeno ponendolo, in relazione con un movimento a scansione prefissata lungo un'asta graduata.

Le tacche sono separate -al livello minimo- da una unità convenzionale che può essere dilatata o contratta senza interferire con ciò che si misura o sulla sua interpretazione.

L'origine agisce solo come riferimento e può essere cambiata a piacere. In questo tipo di scala un incremento assoluto tra due misurazioni ha lo stesso significato qualunque sia il livello da cui si calcola l'incremento.



$$30\text{ C} \rightarrow \frac{9 \cdot 30}{5} + 32 = 86\text{ F}; \quad 40\text{ C} \rightarrow \frac{9 \cdot 40}{5} + 32 = 104\text{ F}; \quad 20\text{ C} \rightarrow \frac{9 \cdot 20}{5} + 32 = 68\text{ F};$$

le differenze tra le temperature non sono cambiate: $40-30=30-20$ come $104-86=86-68$ e non sono cambiati i rapporti tra gli scarti delle due diverse scale: $(40-30)/(30-20)=(104-86)/(86-68)$.

Ciò che si modifica è lo scarto tra le misure della stessa scala che passa da 10 a 18 perché è stato moltiplicato per il fattore 9/5.

Variabili nominali o politòme

Le modalità di queste variabili esprimono categorie, qualità, status: le $\{X_i\}$ in "hanno la sola funzione di etichettare le unità per formarne un elenco o per raggrupparle in classi omogenee".

ESEMPI:

La variabile "Regione" si manifesta con le usuali 20 modalità $S=\{\text{Calabria, Sicilia, ..., Val d'Aosta, Piemonte}\}$.

Un'impresa può ricadere nel settore $\{\text{agricoltura, industria, altre attività}\}$.

Le differenze possono essere accertate, ma non ordinate né misurate: si possono scambiare di posto senza che ciò influisca sulla validità dei dati così raccolti

Operazioni ammesse

Level	Examples	Numerical Operations	Descriptive Statistics
Interval	100-Point Job Performance Ratings Assigned by Supervisors: 0% = Worst Performers 100% = Best Performers Temperature Type Attitude Scales: Low Temperature = Bad Attitude High Temperature = Good Attitude	Common Arithmetic Operations	• Mean • Median • Variance • Standard Deviation
Ratio	Amount Purchased Salesperson Sales Volume Likelihood of performing some act: • 0% = No Likelihood to • 100% = Certainty Number of stores visited Time spent viewing a particular web page Number of web pages viewed	All Arithmetic Operations	• Mean • Median • Variance • Standard Deviation

Uso dei numeri

la codifica delle modalità porta ad usare dei numeri. Questo però non significa che siano lecite delle operazioni aritmetiche:

i ruoli di una squadra di calcio sono indicati con dei numeri, ma non si può dire che l'ala sinistra ("11") sia maggiore dello stopper ("5") o che l'unità di misura "1" dei calciatori sia il portiere;

ESEMPI

il numero civico delle abitazioni:



Non ha alcun significato operativo la eventuale progressione delle modalità;

Variabili nominali o politome

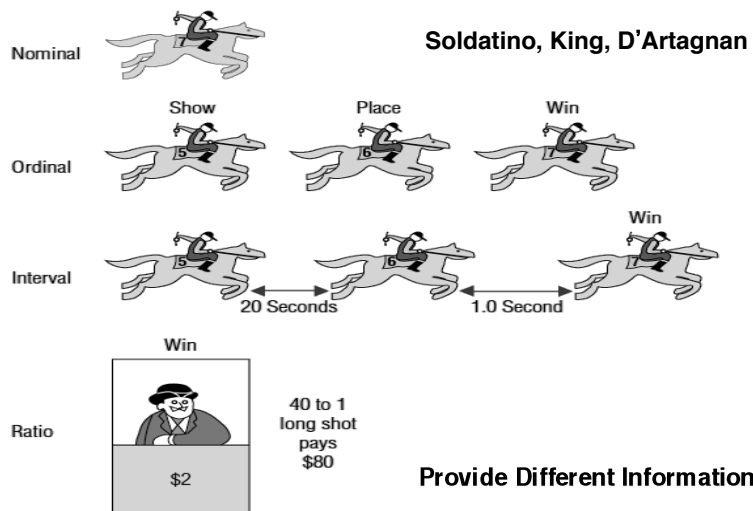
Il livello di misura della variabile è tale che, date due qualsiasi modalità: x_r , x_s , è possibile affermare solo che:

$$x_r = x_s \quad \text{oppure} \quad x_r \neq x_s$$

La scelta di una sola tra le bevande incluse nelle modalità è una politomia



Differenza di informazioni



Operazioni ammesse/2

Level	Examples	Numerical Operations	Descriptive Statistics
Nominal	Yes - No Female - Male Buy - Did Not Buy Postal Code: ____	Counting	• Frequencies • Mode
Ordinal	Rankings Choose from the Following: • Dissatisfied • Satisfied • Very Satisfied • Delighted Indicate Your Level of Education: • HS Diploma • Some College • Bachelor's Degree • Graduate Degree	Counting and Ordering	• Frequencies • Mode • Median • Range

Ordinamenti

Il termine "scala" ha senso se tra le modalità di "S" sono possibili degli ordinamenti.

- 1) $X_i < X_j$ oppure $X_i > X_j$ per ogni $i \neq j$
- 2) $X_i < X_j \Rightarrow X_i \neq X_j$
- 3) $X_i < X_j$ e $X_j < X_k \Rightarrow X_i < X_k$ per ogni $i < j < k$

Maggiore è il contenuto di "fenomeno" maggiore è la modalità che la rappresenta; esiste perciò una disposizione delle modalità che non può essere alterata senza che ne risulti modificata la rilevazione.

Il dominio si esprime con interi consecutivi:

$$S = \{a, a+1, a+2, \dots, a+k-1\}$$

Variabili ordinali

i ranghi sono dei voti che esprimono la stima della proprietà posseduta: ogni unità è confrontata con una linea di valutazione che incasella l'unità in una data categoria di valore a prescindere da quello che succede alle altre unità.

Spesso, le modalità di una variabile ordinale esprimono soglie di vicinanza ad un ideale che fungerebbe da “metro” o “campione” di misurazione del concetto.

ESEMPLI:

a) Voti di un giudice: $S=\{0, 1, 2, \dots, 10\}$;

b) Ammontare di punti da ripartire: $\{0 -100\}$;

c) Voti grafici: $\{++, +0, 0+, 00, -0, 0-, - -\}$;

d) Quantificatore verbale: $\{ \text{pianura, collina, montagna} \}$

Invarianza rispetto a trasformazioni monotone

$$f(X_i) < f(X_j) \text{ se } X_i < X_j$$

Differenziale semantico

Per attenuare le ambiguità derivanti delle scale ordinali si possono usare delle scale bipolari in cui sono inserite solo le valutazioni più opposte dell'aspetto indagato collocando tra di esse, ad opportune interdistanze, una serie di riquadri

Chi risponde dovrà poi individuare il punto più prossimo al suo giudizio ovvero indicare quale descrizione numerica o verbale si adatti al proprio sentire.

ESEMPIO:

“Come giudicate l'operato dei rappresentanti degli studenti nel Senato accademico?”



Numero di modalità e posizione

Non esiste un numero ottimale di livelli: $k=7\pm 2$ o $k=6$ sono considerati uno standard nelle ricerche di mercato (Kinnear e Taylor, 1979, p. 30, Malhotra 1996, p. 298).

3 o 4 gradini comportano risultati confusi per l'accorpamento di giudizi eterogenei; d'altra più di sei è utile solo per acquisire variazioni di quantità molto piccole di cui non sempre si ha bisogno.

Anche la disposizione deve essere equilibrata:

ESEMPIO:

Quale delle tre seguenti moltiplicazioni

P1. $9 * 7 * 8 * 6 * 5 * 4 * 3$

P2. $3 * 4 * 5 * 6 * 7 * 8 * 9$

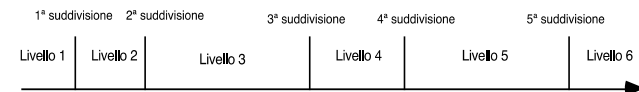
P3. $7 * 3 * 8 * 4 * 6 * 5 * 9$

*Effetto
posizione*

darà il risultato più alto?

Continuo percettivo

L'intensità con cui si avverte una sensazione varia in una successione continua di stati: al tessuto uniforme del concetto si sovrappone una griglia più o meno regolare



una unità che sia X_i in una rilevazione ed X_j in una successiva con $X_i < X_j$ sarà passata per tutti gli stati intermedi tra X_i ed X_j .

Le suddivisioni non sono però oggettive: osservatori diversi scelgono divisioni diverse ovvero lo stesso punto di separazione ha senso diverso.

N.B. Talvolta la proprietà studiata ha natura discontinua: si modifica con una scansione non frazionabile per un numero finito di stati che sono i soli a poter essere osservati.

Rilevazione diretta dei ranghi

Un certo insieme di n oggetti o situazioni sono ordinate secondo il grado con cui presentano una certa caratteristica X .

Supponiamo ...

Che la caratteristica sia un *mix* di immaterialità graduabile, ma non misurabile.

Che le valutazioni siano espresse con i voti $\{1, 2, \dots, n\}$ così ottenendo la permutazione $\{s_1, s_2, \dots, s_n\}$

Ripetiamo la rilevazione per una Y rilevata allo stesso modo e che produce la permutazione: $\{r_1, r_2, \dots, r_n\}$



*Condizione di ansia e stress
Prima e dopo una separazione*

Gli indici di dipendenza tra ranghi cercano di quantificare l'intensità del legame tra i due insiemi di giudizi

Esempio: giudizi degli esperti



Ad un esperto è stato chiesto di pronunciarsi sulla posizione che le 20 squadre di un campionato di calcio occuperanno alla fine: $\{s_1, s_2, \dots, s_{20}\}$.

Alla fine della stagione i giudizi sono comparati con le posizioni reali: $\{r_1, r_2, \dots, r_{20}\}$.

Per semplificare il calcolo possiamo disporre le due serie di posizioni secondo l'ordine crescente della prima

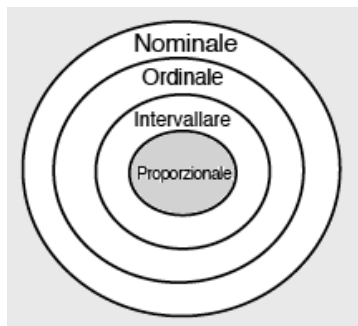
Squadra	A	B	C	D	E	F	G	H	I	L	M	N	O	P	Q	R	S	T	U	V	Totale
Prima	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	210
Dopo	9	2	4	7	5	1	3	8	6	11	13	10	14	18	15	12	16	20	17	19	210

L'esperto ha dato un buon giudizio sebbene sembri più in grado di indovinare le squadre che avranno una cattiva stagione rispetto a quelle che l'avranno buona

Gerarchia tra scale di misurazione

Se una variabile è su scala proporzionale, con un processo di arrotondamenti è possibile riportarla su scala intervallare; questa a sua volta instaura un ordinamento che è anche utilizzabile per valutare la similarità delle categorie a quella di riferimento (scala nominale).

Fra le scale esiste perciò una gerarchia:



Variabili binarie o dicotome

La variabile binaria è l'espressione di una dicotomia tra due possibili stati in cui può trovarsi l'unità.

Di solito si rileva la presenza o l'assenza di una proprietà e talvolta si fa riferimento alla condizione bianco/nero, oppure ON/OFF di un circuito logico

	u_j			u_j			u_j			u_j		
u_i	1	x_1	x_2	2	x_1	x_2		x_1	x_2	m	x_1	x_2
	x_1	1	0	x_1	0	0	x_1	0	0	x_1	0	1
	x_2	0	0	x_2	0	1	x_2	1	0	x_2	0	0

Ognuna di queste situazioni contribuisce all'affinità in modo autonomo e separato ed è subito evidente una sorta di gerarchia di importanza:

- a) contemporanea presenza: le due unità sono più simili tra di loro perché hanno in comune un aspetto.
- b) contemporanea assenza: le due unità sono più simili tra di loro perché sono entrambe privi di una caratteristica.
- c) e d) le due unità sono meno simili perché hanno un comportamento diverso rispetto alla presenza/assenza, di una proprietà.

Variabili binarie o dicotome/2

La valutazione dell'affinità tra soggetti passa di solito per più variabili binarie perché sarebbe troppo riduttivo affrontare una comparazione solo in base allo stato attivo/passivo di un'unica caratteristica.

Per ogni variabile si può produrre una delle combinazioni: 00, 11, 01, 10 e l'esito del confronto risulterà dalla aggregazione dei singoli confronti parziali

Per ottenere la massima flessibilità di giudizio sull'affinità che scaturisce dai confronti, conviene associare una tabella di valutazione che specifica il contributo che il confronto stesso dà all'affinità complessiva

Tabella di confronto

k	x_1	x_2
x_1	=	≠
x_2	≠	=

Tabella di giudizio

k	x_1	x_2	
x_1	a_k	b_k	$a_k + b_k$
x_2	c_k	d_k	$c_k + d_k$
	$a_k + c_k$	$b_k + d_k$	T_k

Dove a_k, b_k, c_k, d_k sono valori non negativi ed eventualmente frazionari che quantificano il peso da dare al singolo confronto

Esempio

Si intende valutare l'affinità di un data set di comuni rispetto ad alcune variabili binarie dando ad ognuna un peso diverso.

Variabile	Peso Variabile	Peso delle binarie	Peso specifico binarie
X1=Litoraneità	45	30	13.5
X2=Ufficio turistico	15	30	4.5
X3=Alberghi 5 stelle	40	30	12.0

100

X_1				X_2				X_3			
	x_1	x_2			x_1	x_2			x_1	x_2	
x_1	45	0	45	x_1	7.5	0	7.5	x_1	25	0	25
x_2	0	0	0	x_2	0	7.5	7.5	x_2	0	10	10
	45	0	45		7.5	7.5	15		25	5	35

La tabella si riferisce alla presenza/assenza della caratteristica indicata nella variabile e tiene conto che nella analisi di riferimento, l'affinità derivata dalle variabili binarie ha un peso pari al 30% dell'affinità totale riscontrabile tra le unità.

Variabili binarie o dicotome/3

Il confronto complessivo produce la tabella:

$$\begin{matrix} \sum_{k=1}^m a_k & \sum_{k=1}^m b_k \\ \sum_{k=1}^m c_k & \sum_{k=1}^m d_k \end{matrix}$$

Il totale $T_k = a_k + b_k + c_k + d_k$ costituisce il peso complessivo che la variabile k-esima ottiene nella comparazione di due soggetti ed il suo contributo al formarsi della valutazione della loro affinità.

$$\text{Ponendo } a = \sum_{k=1}^m a_k; \quad b = \sum_{k=1}^m b_k; \quad c = \sum_{k=1}^m c_k; \quad d = \sum_{k=1}^m d_k$$

La tabella riassuntiva dei confronti sarà:

$$u_i \begin{matrix} & u_j \\ \begin{matrix} a & b \\ c & d \end{matrix} \end{matrix}$$

Esempio/2

I pesi delle variabili sono così ripartiti in base a considerazioni a priori.

X_1				X_2				X_3			
	x_1	x_2			x_1	x_2			x_1	x_2	
x_1	45	0	45	x_1	7.5	0	7.5	x_1	25	0	25
x_2	0	0	0	x_2	0	7.5	7.5	x_2	0	10	10
	45	0	45		7.5	7.5	15		25	5	35

Per la X_2 e la X_3 si è ritenuto che l'assenza della caratteristica generasse un grado di affinità positivo, ma inferiore a quello della comune presenza per la X_3 e di pari livello per la X_2 .

Invece, l'assenza di litoraneità nella X_1 non è stata considerata elemento per una affinità positiva.

Londra e Staiti (RC) non sono comuni litoranei, ma sapere questo non aggiunge molto alla percezione della loro vicinanza.

Comune	X1	X2	X3
Roccasecca	0	1	0
Cajaniello	0	1	1
Capri	1	1	1
Sorrento	1	0	1

Il computo dell’affinità passa per la valutazione della compresenza e della assenza congiunta ovvero della non presenza in una o entrambe le unità delle caratteristiche misurate dalle tre variabili

I comuni più affini sono Capri e Sorrento dato che la somma delle affinità binarie raggiunge il massimo 70 ed i comuni meno affini sono Roccasecca e Sorrento che non hanno alcuna affinità rispetto agli aspetti considerati nell'esempio.

Esempio/3

X1	Roccasecca	Cajaniello	Capri	Sorrento
Roccasecca		d1=1	b1=1	b1=1
Cajaniello	d1=1		b1=1	b1=1
Capri	b1=1	b1=1		a1=1
Sorrento	b1=1	b1=1	a1=1	
X2	Roccasecca	Cajaniello	Capri	Sorrento
Roccasecca		a2=1	a2=1	c2=1
Cajaniello	a2=1		a2=1	a2=1
Capri	a2=1	a2=1		c2=1
Sorrento	c2=1	a2=1	c2=1	
X3	Roccasecca	Cajaniello	Capri	Sorrento
Roccasecca		b3=1	b3=1	b3=1
Cajaniello	b3=1		a3=1	a3=1
Capri	b3=1	a3=1		a3=1
Sorrento	b3=1	a3=1	a3=1	
X1,X2,X3	Roccasecca	Cajaniello	Capri	Sorrento
Roccasecca		45+7.5+0	0+7.5+0	0+0+0
Cajaniello	45+7.5+0		0+7.5+25	0+0+25
Capri	0+7.5+0	0+7.5+25		45+0+25
Sorrento	0+0+0	0+0+25	45+0+25	
X1,X2,X3	Roccasecca	Cajaniello	Capri	Sorrento
Roccasecca		52.5	7.5	0.0
Cajaniello	52.5		32.5	25.0
Capri	7.5	32.5		70.0
Sorrento	0.0	25.0	70.0	

Coefficienti di affinità/binarie

In questo contesto assume grande rilevanza il problema di come considerare la assenza congiunta.



Se il fatto di non possedere un attributo è irrilevante ai fini della somiglianza allora la cella “d” non deve entrare nella misura.

Se invece la dicotomia è fra due stati complementari aventi uguale rilevanza allora “a” e “d” entrano nello stesso modo nell’indice. Nel primo caso si parla di varabili binarie asimmetriche e nel secondo le variabili sono dette binarie simmetriche.



Il leone ed il coniglio sono privi di ali, ma questo non li rende più affini.

Esempi di coefficienti per binarie

Coefficiente	Tipo	Formula
Jaccard	T	$\frac{a}{a + b + c}$
Andenberg	T	$\frac{a}{a + 2(b + c)}$
Czekanowski	T	$\frac{2a}{2a + b + c}$
Ochiai	T	$\frac{a}{\sqrt{(a + b)(a + c)}}$
Sokal – Sneath	S	$\frac{a + d}{a + 0.5b + 0.5c + d}$
Hamann	S	$\frac{(a + d) - (b + c)}{a + b + c + d}$
Rogers – Tanimoto	S	$\frac{a + d}{a + 2b + 2c + d}$
Simple Matching	S	$\frac{a + d}{a + b + c + d}$
Russell – Rao	S	$\frac{a}{a + b + c + d}$

Se la compresenza o la coassenza nell’attributo sono tanto peculiari nel generare molta somiglianza tra le due unità allora “a” e/o “d” devono entrare nel coefficiente con peso doppio o comunque maggiore degli altri.

Nella tabella sono distinti gli indici che non includono il conteggio della coassenza (T) con quelli che la includono (S).

In caso di valori nella forma 0/0 si può porre, per convenzione, il coefficiente pari a zero.

I coefficienti qui inseriti variano tra zero ed uno e generano una matrice delle distanze euclidea.

Esempio

Entità	x1	x2	x3	x4	x5	x6	x7	x8	x9	x10
A	1	1	0	1	0	1	0	0	1	1
B	0	0	0	0	0	1	1	1	1	1
C	1	0	1	0	1	0	1	0	1	0
D	1	0	1	1	0	0	1	0	1	0
E	0	1	0	1	0	1	0	1	0	1
F	1	1	1	1	0	0	0	0	0	0
G	1	0	0	1	0	0	0	0	1	0
H	0	0	1	1	1	0	0	1	0	0

Jaccard

	A	B	C	D	E	F	G	H
A	0.0000	0.7906	0.8819	0.7906	0.6547	0.7906	0.7071	0.9428
B	0.7906	0.0000	0.8660	0.8660	0.7559	1.0000	0.9258	0.9354
C	0.8819	0.8660	0.0000	0.5774	1.0000	0.7559	0.8165	0.8452
D	0.7906	0.8660	0.5774	0.0000	0.9428	0.7559	0.6325	0.8452
E	0.6547	0.7559	1.0000	0.9428	0.0000	0.8660	0.9258	0.8452
F	0.7906	1.0000	0.7559	0.7559	0.8660	0.0000	0.8165	0.7071
G	0.7071	0.9258	0.8165	0.6325	0.9258	0.8165	0.0000	0.9129
H	0.9428	0.9354	0.8452	0.8452	0.8452	0.7071	0.9129	0.0000

Simple matching

	A	B	C	D	E	F	G	H
A	0.0000	0.7071	0.8367	0.7071	0.5477	0.7071	0.5477	0.8944
B	0.7071	0.0000	0.7746	0.7746	0.6325	1.0000	0.7746	0.8367
C	0.8367	0.7746	0.0000	0.4472	1.0000	0.6325	0.6325	0.7071
D	0.7071	0.7746	0.4472	0.0000	0.8944	0.6325	0.4472	0.7071
E	0.5477	0.6325	1.0000	0.8944	0.0000	0.7746	0.7746	0.7071
F	0.7071	1.0000	0.6325	0.6325	0.7746	0.0000	0.6325	0.5477
G	0.5477	0.7746	0.6325	0.4472	0.7746	0.6325	0.0000	0.7071
H	0.8944	0.8367	0.7071	0.7071	0.7071	0.5477	0.7071	0.0000

Altri esempi

Matrici di associazione: possono essere simmetriche e asimmetriche

	Actor	Recipient					
		A	S	N	D	W	T
Grooming rates of capuchin monkeys (Perry 1996)	A	-	5.8	3.5	2.1	2.3	0.04
	S	41.6	-	28.6	18.1	9.0	7.4
	N	10.3	25.5	-	9.6	9.9	4.3
	D	23.3	9.3	10.5	-	13.4	6.9
	W	21.2	15.2	14.6	25.1	-	10.4
	T	2.5	2.9	3.7	3.6	5.3	-

GRI -0.24
VAX 0.02 0.08
KRI 0.02 -0.04 -0.19
MYR -0.27 0.44 -0.03 -0.11
WOW 0.22 0.11 0.32 -0.10 0.10
HOB -0.04 0.11 -0.17 -0.13 -0.08 -0.12
WBE 0.15 0.07 -0.08 0.08 -0.08 0.23 0.13
HOR -0.08 0.21 -0.14 -0.23 0.18 0.12 0.11 0.26
AJA -0.24 0.23 -0.04 -0.16 -0.01 -0.16 0.07 0.25 0.32
PIK -0.11 0.35 -0.07 0.04 0.02 -0.05 0.09 0.60 0.21 0.27
ANV -0.05 -0.23 -0.39 -0.39 -0.21 -0.13 -0.41 0.11 0.11 0.02 -0.06
VEE 0.14 0.02 0.15 -0.11 -0.08 0.00 -0.09 -0.05 0.06 0.01 -0.17 -0.17
LAT GRI VAX KRI MYR WOW HOB WBE HOR AJA PIK ANV

Genetic relatedness among
bottlenose dolphins (Krutzen et
al. 2003)

Affinità/dissimilarità in ambiente R

stats, vegan, ade4, cluster, FD, arules.

Although the literature provides similarity as well as distance measures, in R all similarity measures are converted to distances to compute a square matrix of class "dist" in which the diagonal (distance between each object and itself) is 0 and can be ignored.

The conversion formula varies with the package used, and this may not be without consequences:

In stats, FD and vegan, the conversion from similarities S to dissimilarities D is $D = 1 - S$.

In ade4, it is computed as $D = (1 - S)^{0.5}$. This allows some indices to become Euclidean

Distance matrices computed by other packages that are not Euclidean can often be made Euclidean by computing $D \leftarrow \sqrt{D}$.

In cluster, all available measures are distances, so no conversion has to be made.

Dalla affinità alla dissimilarità

Ad ogni indice di affinità/prossimità è associato un indice di dissimilarità o dissomiglianza o distanza.

Se a_{ij} è simmetrico e non negativo allora la dissomiglianza d_{ij} dovrà avere la stessa proprietà; inoltre, deve diminuire quando la prima aumenta e viceversa.

Per indici di affinità normalizzati (0, 1) la trasformazione in una dissimilarità è semplice

$$\begin{aligned}
 1. \quad d_{ij} &= 1 - a_{ij}^\alpha & 2. \quad d_{ij} &= (1 - a_{ij})^\alpha \quad \alpha > 0 \\
 3. \quad d_{ij} &= \frac{\ln(1 + a_{ij}) - \ln(2)}{-\ln(2)} & 4. \quad d_{ij} &= \frac{\alpha - a_{ij}}{\alpha + a_{ij}} \\
 5. \quad d_{ij} &= \frac{e^{(-a_{ij})} - 1}{e - 1}
 \end{aligned}$$

Ogni funzione f definita nell'intervallo (0,1), tale che f(0)=1 e f(1)=0 e che in tale intervallo abbia derivata negativa è idonea a trasformare un indice di affinità in un indice di dissomiglianza.

Esempio: SPECTF heart data

The dataset describes diagnosing of cardiac Single Proton Emission Computed Tomography (SPECT) images.

Each of the patients is classified into two categories: normal and abnormal.

The database of 267 SPECT image sets (patients) was processed to extract features that summarize the original SPECT images.

Package	Package stats				
	1	2	3	4	5
Package ade4: simple matching	1	0.0000	0.1667	0.7778	1
	2	0.1667	0.0000	0.6667	1
	3	0.7778	0.6667	0.0000	1
	4	1.0000	1.0000	1.0000	0
	5	0.7143	0.7500	0.7500	1
Package vegan, jaccard	1	0.0000	0.4366	0.4513	0.2975
	2	0.4366	0.0000	0.2997	0.5395
	3	0.4513	0.2997	0.0000	0.3633
	4	0.2975	0.5395	0.3633	0.0000
	5	0.4951	0.3659	0.4436	0.6109

Politome sconnesse o multistato

Il dominio è formato da modalità che distinguono gli aspetti, le categorie, gli attributi, stati che le unità possiedono in vario modo senza che tra le modalità possa essere stabilito un ordinamento univoco in termini quantitativi.

Il livello di misura della variabile è tale che, date due qualsiasi modalità: x_r , x_s , è possibile solo affermare che:

$$x_r = x_s \quad \text{oppure} \quad x_r \neq x_s$$

Le modalità hanno qui la sola funzione di etichettare le unità per formarne una lista o per raggrupparle in categorie omogenee. L'ordinamento alfabetico con cui sono spesso presentate le modalità semplifica l'esposizione, ma non stabilisce una gerarchia.

Le differenze tra le unità possono essere accertate, ma non ordinate né misurate.

Esempio

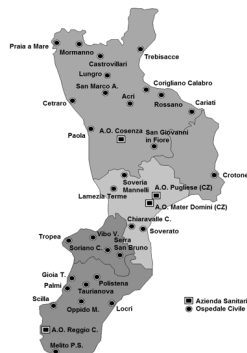
Giudizi personali sul confronto delle cinque province calabresi rispetto alla attività economica prevalente:

$S(X) = \{\text{Commercio, Turismo, Amministrazione pubblica, Industria manu-fatturiera, Agricoltura}\}$

L'idea è di sottoporre poi a delle unità (un campione oppure un censimento di soggetti in posizioni strategiche) una domanda del tipo "Quale considera sia l'attività prevalente della sua provincia?"

L'incrocio dei dati produrrà poi valori in base alla tabella delle compresenze.

	Commercio	Turismo	Amm. Pubblica	Industria Manif.	Agricoltura
Commercio	1	0.7	0.8	0.4	0.1
Turismo		1	0.6	0.3	0.5
Amm. Pubblica			1	0.4	0.2
Industria Manif.				1	0.1
Agricoltura					1



Politome sconnesse o multistato/2

Supponiamo che la variabile X sia rilevata con il dominio.

$$S(X) = \{x_1, x_2, \dots, x_m\}$$

Per ciascun confronto tra due unità può verificarsi una qualsiasi delle combinazioni di modalità

$$\{(x_r, x_s), r, s = 1, 2, \dots, n\}$$

Per ciascuna combinazione di modalità occorre esprimere un giudizio di somiglianza: a_{ij}

	x_1	x_2	...	x_j	...	x_m
x_1	1	a_{12}	...	a_{1j}	...	a_{1m}
x_2	a_{21}	1	...	a_{2j}	...	a_{2m}
...	...	...	...	...	...	...
x_i	a_{i1}	a_{i2}	...	1	...	a_{im}
...	...	...	...	...	...	...
x_m	a_{m1}	a_{m2}	...	a_{mj}	...	1

Le entrate a_{ij} esprimono la valutazione della affinità tra due unità in cui si verificano le due modalità della coppia considerata.

Non è necessario che verifichino la disuguaglianza triangolare, ma è preferibile

Altro esempio

Ad un gruppo di 10 studenti scelti accuratamente per posizione curricolare, estrazione sociale, formazione secondaria, etc. è stato chiesto di esprimere con un voto da 1 a 6 un gruppo di quattro insegnanti:

0= diametralmente opposti e 6= del tutto equivalenti

Confronto	Stud_01	Stud_02	Stud_03	Stud_04	Stud_05	Stud_06	Stud_07	Stud_08	Stud_09	Stud_10	MEDIA	MEDIANA	Medio6
T1,T2	5	6	4	6	5	5	6	4	6	5	5.2	5.0	0.9
T1,T3	4	2	3	4	3	2	2	3	3	4	3.0	3.0	0.5
T1,T4	2	0	1	1	2	1	2	2	1	3	1.6	1.5	0.3
T2,T3	6	5	5	6	4	3	4	5	4	5	4.7	5.0	0.8
T2,T4	1	1	2	3	1	1	2	1	1	0	1.4	1.0	0.2
T3,T4	3	2	4	5	2	3	4	4	3	4	3.4	3.5	0.6

La sintesi dei giudizi può avvenire con la media aritmetica, con la mediana, il voto minimo, il voto massimo, etc.

Qui si è scelta la media aritmetica, rapportata a 6 per avere numeri tra zero ed uno. A questo punto la matrice dei giudizi è pronta per comparare due qualsiasi soggetti

	T1	T2	T3	T4
T1	1	0.87	0.50	0.27
T2		1	0.78	0.23
T3			1	0.57
T4				1

Coefficiente di Beijnen

Un coefficiente che risponde al requisito di disuguaglianza triangolare ed anche a quella di euclideanità della corrispondente matrice è stato proposto da Beijnen (1973).

$$a_{ij} = \sqrt{\frac{\sum_{r=1}^m \alpha(x_{ri}, x_{rj})}{m}} \quad \text{con} \quad \alpha(x_{ri}, x_{rj}) = \begin{cases} \frac{L_r}{p} & \text{se } x_{ri} = x_{rj} \\ 0 & \text{altrimenti} \end{cases}$$

dove L_r è il numero di modalità previsto per la variabile multistato r -esima.

$$p = \sum_{r=1}^m L_r$$

Il coefficiente proposto somma una frazione pari al numero di modalità della variabile politoma rispetto a tutte le modalità delle diverse politomie presenti nel data set.

L'affinità è tanto maggiore quanto maggiore è il numero di stati nella politomia

Scomposizione della politomia in variabili binarie

La rilevazione della variabile politoma è suddivisa in vari confronti binari (tante quanto sono le modalità del suo dominio).

$$S(X) = \{x_1, x_2, \dots, x_m\}$$

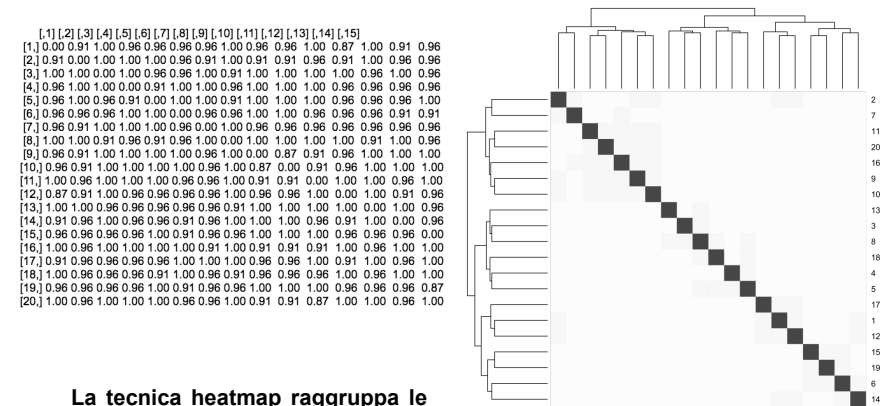
l'accertamento della modalità presentata da una qualsiasi unità si realizza valutando una sequenza di valori binari (ad esempio zero ed uno)

	X_1	X_2	...	X_k	...	X_m
Presente	1	1	...	1	...	1
Assente	0	0	...	0	...	0

Il confronto di due unità generiche: i e j passa per m variabili binarie di tipo asimmetrico (perché manca la complementarità)

Esempio

Consideriamo dei dati simulati su $n=15$ unità e su $k=3$ politomie con 3, 4, 5 diversi stati.



La tecnica heatmap raggruppa le unità con maggiore prossimità

Politomia in variabili binarie

In sintesi, per ogni indicatore di stato del dominio S si presenta una delle quattro configurazioni: 00, 01, 10, 11 che poi vanno adeguatamente sommate come si è fatto per il confronto su variabili binarie autonome.

$$u_i \begin{array}{c|cc} & u_j & & \\ \hline X_1 & 1 & 0 \\ \hline 1 & a & b \\ \hline 0 & c & d \end{array} \quad u_j \begin{array}{c|cc} & u_j & & \\ \hline X_2 & 1 & 0 \\ \hline 1 & a & b \\ \hline 0 & c & d \end{array} \quad \dots \quad u_j \begin{array}{c|cc} & u_j & & \\ \hline X_k & 1 & 0 \\ \hline 1 & a & b \\ \hline 0 & c & d \end{array}$$

Dove $X_1, X_2, \dots, X_k$ non sono variabili vere e proprie, ma pseudo-variabili derivate dalle modalità della politomia.

A queste tabelle, si può applicare uno degli indici di somiglianza tra variabili binarie, adeguatamente ponderato.

Esempio

Consideriamo dei dati simulati su n=15 unità e su k=3 politomie a diversi stati

unità	X1	X2	X3	Range	Xa1	Xa2	Xa3	Xb1	Xb2	Xb3	Xb4	Xc1	Xc2	Xc3	Xc4	Xc5
u_1	2	3	1	X1=1:3	0	1	0	0	0	1	0	1	0	0	0	0
u_2	3	3	1	X2=1:4	0	0	1	0	0	1	0	1	0	0	0	0
u_3	1	2	2	X3=1:5	1	0	0	0	1	0	0	0	0	0	0	0
u_4	2	1	4		0	1	0	1	0	0	0	0	0	0	1	0
u_5	2	1	2		0	1	0	1	0	0	0	0	0	0	0	0
u_6	1	3	5		1	0	0	0	0	1	0	0	0	0	0	1
u_7	3	3	3		0	0	1	0	0	1	0	0	0	1	0	0
u_8	1	1	2		1	0	0	1	0	0	0	0	0	0	0	0
u_9	3	4	1		0	0	1	0	0	0	1	1	0	0	0	0
u_10	3	4	1		0	0	1	0	0	0	1	1	0	0	0	0
u_11	3	4	5		0	0	1	0	0	0	1	0	0	0	0	1
u_12	2	3	1		0	1	0	0	0	1	0	1	0	0	0	0
u_13	1	1	3		1	0	0	1	0	0	0	0	0	1	0	0
u_14	2	3	5		0	1	0	0	0	1	0	0	0	0	0	1
u_15	1	3	4		1	0	0	0	0	1	0	0	0	0	1	0

Le tre politomie sono diventate 15 dicotomie e la similarità tra i soggetti pu essere calcolata con uno dei coefficienti di tipo T ad esempio lo Jaccard o lo Ochiai.

Affinità per i ranghi

Riguarda le variabili riportate in scala quantitativa ordinale.

- Perché non esiste una vera misura, ma solo un punteggio o valutazione
- Perché le misurazioni su sono imprecise o viziate da errore
- Perché sono presenti dei valori remoti

Le modalità sono poste in corrispondenza con dei numeri naturali (ranghi)

Per ogni unità si osserva una coppia di modalità che si trasforma poi in una coppia di ranghi

$$(X_p, Y_i) \longrightarrow (\pi_p, \sigma_i)$$

E' possibile che i ranghi siano rilevati direttamente come risposte ad un quantificatore verbale

Esempio/continua

Con il frazionamento in binarie e coefficiente di Jaccard

	[,1]	[,2]	[,3]	[,4]	[,5]
[1,]	0.0000	0.7071	1.0000	0.8944	0.8944
[2,]	0.7071	0.0000	1.0000	1.0000	1.0000
[3,]	1.0000	1.0000	0.0000	1.0000	0.8944
[4,]	0.8944	1.0000	1.0000	0.0000	0.7071
[5,]	0.8944	1.0000	0.8944	0.7071	0.0000

Coefficiente di Beijnen

	[,1]	[,2]	[,3]	[,4]	[,5]
[1,]	0.00	0.5	1.00	0.75	0.75
[2,]	0.50	0.0	1.00	1.00	1.00
[3,]	1.00	1.0	0.00	1.00	0.75
[4,]	0.75	1.0	1.00	0.00	0.50
[5,]	0.75	1.0	0.75	0.50	0.00

La dicotomizzazione delle politomie genera molte variabili binarie e potrebbe ingenerare somiglianze anche nei confronti di unità in cui queste siano assenti.

Inoltre, nel confronto di unità rispetto distanze multiscala (che studieremo più avanti) potrebbe esagerare oltre misura il ruolo delle variabili binarie.

Le situazioni che coinvolgono i ranghi sono di due tipi



Politome ordinate.

La o le variabili hanno come dominio un insieme di numeri naturali ovvero quantificatori verbali descritti con dei numeri.

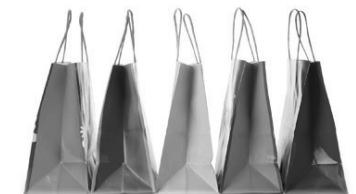
Su ogni unità si rileva una singola modalità



Graduatorie.

Ciò che si rileva sulle unità è una graduatoria che considera un gruppo prefissato di oggetti.

La graduatoria è una singola variabile a se, anche se si presenta con diverse modalità in forma di ranghi



Applicazione politome ordinate

Response Pattern	No of Cases
2 2 2 2 2	97
3 3 2 3 3	70
3 2 2 2 2	49
3 3 3 3 3	45
3 3 2 2 2	45
3 2 2 3 3	40
3 3 2 3 2	32
3 3 2 3 2	31
2 2 1 2 2	25
1 1 1 1 1	23
3 2 2 3 2	20
2 2 1 1 1	18
3 3 2 2 3	18
2 3 2 2 2	17
3 2 2 3 2	16
3 3 3 3 3	16
3 3 2 3 2	15
3 2 3 3 3	15
2 1 2 2 2	14
2 2 2 3 2	13

agree strongly (1), agree (2), disagree (3), disagree strongly (4) per 6 domande sulla partecipazione alla vita politica sottoposte a 1554 soggetti.

NOSAY	People like me have no say in what the government does
VOTING	Voting is the only way that people like me can have any say about how the government runs things
COMPLEX	Sometimes politics and government seem so complicated that a person like me cannot really understand what is going on
NOCARE	I don't think that public officials care much about what people like me think
TOUCH	Generally speaking, those we elect to Congress in Washington lose touch with the people pretty quickly
INTEREST	Parties are only interested in people's votes but not in their opinions description

Le combinazioni di risposte possibili sono $4^6=4096$, ma molte sono ripetute ed altre non si sono realizzate.

La matrice delle affinità potrà avere dimensioni 1554×1554 se si misura l'affinità tra i soggetti, oppure 6×6 se si misura l'affinità tra le domande.

Affinità/dissimilarità per politome ordinate

Gli indici che si possono utilizzare si basano sugli scarti tra modalità

$$d_{i,j} = \left(\frac{1}{m} \right) \left[\sum_{i=1}^m \left(\frac{|\pi_i - \sigma_i|}{d_i} \right)^\alpha \right]^{\frac{1}{\alpha}} \quad \text{dove} \quad \begin{cases} \pi_i & \text{rango } i\text{-esimo di } A \\ \sigma_i & \text{rango } i\text{-esimo di } B \\ d_i & \text{scarto massimo possibile in } i\text{-esima posizione} \end{cases} \quad \text{Arcavata distance}$$

Il coefficiente varia tra zero (le due entità coincidono esattamente su tutte le variabili di questa tipologia) ed uno (per ogni variabile si riscontra lo scarto massimo ammesso per quella variabile).

L'equivalente misura della affinità si ottiene considerando il complemento ad uno del coefficiente di dissimilarità

$$a_{i,j} = 1 - \left(\frac{1}{m} \right) \left[\sum_{h=1}^m \left(\frac{|\pi_h - \sigma_h|}{d_h} \right)^\alpha \right]^{\frac{1}{\alpha}}$$

Applicazione/2

Consideriamo due soggetti che abbiano dato le seguenti risposte su $m=6$ domande politome ordinate con 4 modalità potenziali

3	2	2	3	2	1	A
1	2	3	3	3	3	B

Come si può misurare l'affinità/vicinanza?

Qui non sembra logico riscalare le risposte in una graduatoria unica del tipo (1,2,...,n) dato che si tratta di domande diverse.

Si può utilizzare il valore numerico del rango e valutare la quantità:

$$d_{i,j} = \left(\frac{1}{6} \right) \sum_{i=1}^6 \left(\frac{|\pi_i - \sigma_i|}{d_i} \right) \quad \text{dove} \quad \alpha = 1, m = 6$$

$$\frac{|3-1|}{4-1} + \frac{|2-2|}{4-1} + \frac{|2-3|}{4-1} + \frac{|3-3|}{4-1} + \frac{|2-3|}{4-1} + \frac{|1-3|}{4-1} = \frac{2}{3} + \frac{0}{3} + \frac{1}{3} + \frac{0}{3} + \frac{1}{3} + \frac{2}{3} = \frac{6}{3} = 2 \Rightarrow \frac{2}{6} = \frac{1}{3}; 1 - \frac{1}{3} = \frac{2}{3} \quad \leftarrow \text{Distanza}$$

Affinità per variabili politome ordinate

Per ogni unità si osserva un valore ordinale (o rango) di una variabile con scala di misurazione che consente di stabilire delle priorità univoche tra le modalità osservate.

I valori riportati sono dei numeri interi che variano in un intervallo limitato e valori diversi indicano stati diversi in cui uno precede l'altro.

Detersivi. Graduatoria a somma costante



Attributo	Giudice1	Giudice2	Giudice3
Fragranza	19	23	27
Confezione	14	16	18
Prezzo	31	24	28
Detergenza	26	17	19
Schiumosità	9	11	6
Consistenza	1	9	2
	100	100	100

Gli attributi sono delle unità e i giudici sono delle variabili

Affinità politome ordinate/2

Gli scarti tra valori successivi NON sono significativi e quindi dovrebbero essere sempre ricondotti ad un intervallo di numeri naturali consecutivi

	Graduatorie attributi								
	Gi1	Gi2	G3	G1	G2	G3	A1	A2	A3
Fragranza	19	23	27	4	5	5	1	2	3
Confezione	14	16	18	3	3	3	1	2	3
Prezzo	31	24	28	6	6	6	3	1	2
Detergenza	26	17	19	5	4	4	3	1	2
Schiumosità	9	11	6	2	2	2	2	3	1
Consistenza	1	9	2	1	1	1	1	3	2
	100	100	100						

Graduatorie giudici

L'indagine qui prospettata si configura come un confronto di graduatorie (di giudici o di attributi).

Esempio

Un gruppo di clienti di una banca classificato per reddito e per importo del prestito. Convertiamo i valori osservati in ranghi.

Cliente	Reddito (x)	Rango (x)	Prestito (y)	Rango (y)
A	25 600	6	8 600	5
B	17 800	9	8 800	4
C	167 200	1	500	8
D	44 200	3	6 600	6
E	36 400	4	10 500	3
F	27 400	5	74 400	1
G	83 600	2	0	9
H	18 600	8	6 300	7
I	24 500	7	12 100	2



E' evidente la perdita di informazione. Lo scarto tra i ranghi in X per i clienti H ed I è $9-7=3$ e sarebbe questo per qualunque coppia di valori compresi tra 18'600 e 24'500.

In breve, conoscere i ranghi poco ci dice sui valori originari

Situazione di studio

Un insieme fisso di n oggetti è graduato da un giudice rispetto a due specifici attributi ovvero due giudici graduan gli n oggetti rispetto ad un solo attributo.

Lo stesso scenario vale per lo studio di variabili metriche poi trasformate in ranghi

Ci troviamo di fronte due permutazioni dei primi n numeri naturali

$$\begin{array}{ccccccc}
 1 & 2 & \dots & i & \dots & n-1 & n \\
 \sigma_1 & \sigma_2 & \dots & \sigma_i & \dots & \sigma_{n-1} & \sigma_n & X_1 \\
 \pi_1 & \pi_2 & \dots & \pi_i & \dots & \pi_{n-1} & \pi_n & X_2
 \end{array}$$

Sono possibili le parità (ex aequo)

Convenience Store	Distance from CAM (m)	Rank distance	Price of 50cl bottle (€)	Rank price
1	50	10	1.80	2
2	175	9	1.20	3.5
3	270	8	2.00	1
4	375	7	1.00	6
5	425	6	1.00	6
6	580	5	1.20	3.5
7	710	4	0.80	9
8	790	3	0.60	10
9	890	2	1.00	6
10	980	1	0.85	8

Un indice di correlazione tra ranghi riassume l'intensità e la direzione del legame tra le due graduatorie.

Altro esempio

Ad un campione di soggetti è sottoposto un elenco di situazioni o item in cui debbono indicare il grado di problematicità ovvero assegnare un livello di priorità:

K.	Aumentare l'occupazione	R.	Assistenza agli anziani
A.	Dare speranze ai giovani	C.	Tutelare l'infanzia
J.	Ridurre la criminalità	L.	Costruire case a basso costo
L.	Ridurre l'orario di lavoro	W.	Dare un salario ai disoccupati
B.	Migliorare la sanità	F.	Investire in ricerca scientifica
Z.	Incrementare la solidarietà	H.	Dare dignità alla condizione umana
D.	Eliminare la povertà	T.	Migliorare il sistema carcerario

Ogni soggetto esprime una gerarchia (per comodità, crescente) delle posizioni che gli item occupano sulla propria scala di priorità.

Sono possibili le parità di posizione.

C'è un potenziale di $14!=87'178'291'200$ possibili configurazioni

U1) 10 8 1 2 14 6 12 7 4 9 3 5 11 13
 U2) 5 4 14 11 7 2 13 10 9 3 8 6 12 1
 U3) 4 12 9 8 3 7 10 5 13 6 1 2 11 14

Requisiti degli indici

1. *Normalizzazione*: $-1 \leq r(\sigma, \pi) \leq 1$
con $r(\sigma, \sigma) = r(\pi, \pi) = 1$, $r(\sigma, \sigma^*) = r(\pi, \pi^*) = -1$
dove $\pi^* = n+1-\pi$, $\sigma^* = n+1-\sigma$
2. *Simmetria*: $r(\sigma, \pi) = r(\pi, \sigma)$
3. *Antisimmetria dopo l'antiteticità*: $r(\sigma, \pi^*) = r(\sigma^*, \pi) = -r(\sigma, \pi)$
4. *Invarianza a destra*: $r(\sigma, \pi) = r[\sigma(\theta), \pi(\theta)] \quad \forall \sigma, \pi, \theta$
5. *Valore atteso nullo in caso di indipendenza*
 $E_{\sigma, \pi \in S_n} [r(\sigma, \pi)] = 0$, dove S_n è l'insieme delle $n!$ permutazioni

I desiderata in questo elenco non sono tutti ritenuti cogenti. Ad esempio gli indici top-rank sono asimmetrici in quanto pesano di più i confronti per i ranghi iniziali della graduatoria (web-search)

Indici ammissibili

<i>Spearman</i> (r_1)	$\left(\frac{3}{n^3 - n} \right) \left\{ \sum_{i=1}^n \pi_i^* - \sigma_i ^2 - \sum_{i=1}^n \pi_i - \sigma_i ^2 \right\}$	1906
<i>Gini</i> (r_2)	$\left(\frac{1}{n^2/2} \right) \left\{ \sum_{i=1}^n \pi_i^* - \sigma_i - \sum_{i=1}^n \pi_i - \sigma_i \right\}$	1914
<i>Kendall</i> (r_3)	$\left(\frac{1}{n^2 - n} \right) \sum_{i=1}^n \text{sign} \left\{ (\sigma_i - \sigma_j) \text{sign}(\pi_i - \pi_j) \right\}$	1938
<i>Blomqvist</i>	$\left(\frac{1}{n} \right) \sum_{i=1}^n \text{sign} \left\{ [\sigma_i - \text{median}(\sigma)] [\pi_i - \text{median}(\pi)] \right\}$	1950
<i>Gideon - Hollister</i>	$\left(\frac{1}{n/2} \right) \left\{ \max_{1 \leq i \leq n} \sum_{j=1}^n [I(\sigma_j \leq n - \pi_i)] - \max_{1 \leq i \leq n} \sum_{j=1}^n [I(\sigma_j > \pi_i)] \right\}$	1987



Robustezza: l'indice rimane stabile se i dati da cui derivano i ranghi cambiano poco.



Sensitività: l'indice è in grado di differenziare permutazioni diverse sia pure somiglianti. r_1, r_2, r_3 hanno un range di valori ridotto risultando poco informativi. Gli ultimi due indici sono troppo "robusti" per essere utili in molti contesti.

Importanza dei requisiti



La presenza dei due limiti consente di valutare l'indice rispetto alle situazioni estreme a -1 e 1 assegnate a situazioni estreme.



E' poco utile un indice che cambia valore secondo l'ordine di considerazione delle permutazioni

$$r(\sigma, \pi) = r(\pi, \sigma)$$



E' poco utile un indice che cambia valore oltre che segno se il confronto avviene con la permutazione antitetica

$$r(\sigma, \pi^*) = r(\sigma^*, \pi) = -r(\sigma, \pi)$$



Il valore dell'indice non deve cambiare se cambia l'ordine di considerazione delle varie coppie di ranghi

$$(\sigma_i, \pi_i), i = 1, 2, \dots, n$$



Se le $n!$ permutazioni sono equiprobabili l'indice deve avere aspettativa zero altrimenti i test risultano viziati

ρ (rho) di Spearman

La misura forse più popolare della dipendenza tra i ranghi è la seguente

$$r_1 = \frac{\sum_{i=1}^n \left(\pi_i - \frac{n+1}{2} \right) \left(\sigma_i - \frac{n+1}{2} \right)}{\sqrt{\sum_{i=1}^n \left(\pi_i - \frac{n+1}{2} \right)^2 \sum_{i=1}^n \left(\sigma_i - \frac{n+1}{2} \right)^2}}$$

Caso delle n coppie di valori senza posizioni di parità.

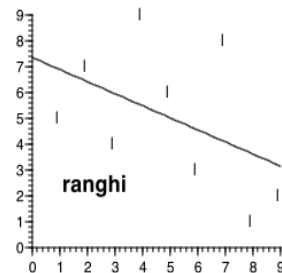
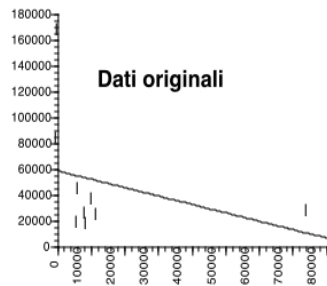
La definizione di r_1 è la stessa del coefficiente di correlazione. Comunque il particolare tipo di dati coinvolti consente delle semplificazioni. Ad esempio:

$$r_s = 1 - \frac{6 \sum_{i=1}^n (r_i - s_i)^2}{n(n^2 - 1)}$$

Esempio

Cliente	Reddito	r_i	Prestito	s_i	$(r_i - s_i)^2$
A	25600	6	8600	5	1
B	17800	9	8800	4	25
C	167200	1	500	8	49
D	44200	3	6600	6	9
E	36400	4	10500	3	1
F	27400	5	74400	1	16
G	83600	2	0	9	49
H	18600	8	6300	7	1
I	24500	7	12100	2	25
					176

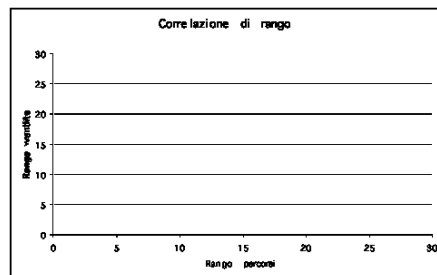
$$r_s = 1 - \frac{6 \cdot 176}{9 \cdot (81 - 1)} = -0.4667$$



Esempio

Unità	X Percorsi	Y Vendite	Rank(X)	Rank(Y)
A	121.5	373	21	25
B	151.5	314	25	21
C	146.2	301	24	20
D	106.7	263	16	17
E	98.9	204	11	9
F	95.1	176	9	7
G	90.1	138	4	1
H	115.5	329	19	23
I	71.7	225	1	11
J	111.7	300	18	19
K	93.6	164	7	5
L	109.6	284	17	18
M	105.3	252	15	16
N	125.0	400	22	26
O	91.7	239	6	15
P	88.7	161	3	4
Q	101.9	226	13	12
R	162.3	322	26	22
S	96.4	185	10	8
T	90.7	143	5	2
U	100.0	212	12	10
V	102.6	232	14	13
X	94.5	171	8	6
Y	88.6	143	2	2
W	119.4	358	20	24
Z	142.9	232	23	13

Venditori porta-a-porta per vendite e km percorsi

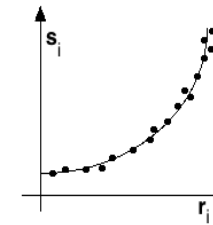
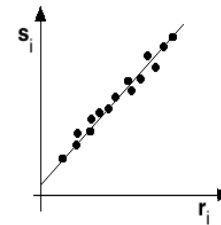


rho= 0.850084703
gdl= 25
tc= 7.907679188
p-Value 2.90161E-08

La correlazione è elevata sebbene si notino diversi disturbi

Considerazione sul rho di Spearman

- Per costruzione l'indice rho varia tra -1 ed 1
- Misura la dipendenza monotonica tra le due variabili. Assume il valore massimo (minimo) quando gli ordinamenti sono perfettamente concordi (discordi)
- $r_s = \pm 1$ se tra la Y e X esiste una relazione monotona, crescente o decrescente, senza vincolo di linearità
- Rho mette tutte le osservazioni sullo stesso piano (considera solo l'ordine) e perciò annulla l'effetto dei valori remoti
- L'indice rimane invariato se una o entrambe le variabili subiscono una trasformazione monotona, ad esempio $Y' = \text{Log}(Y)$ e/o $X' = \exp(X)$



tau di Kendall

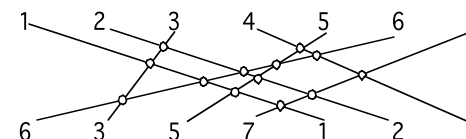
E' una misura alternativa di dipendenza tra ranghi

$$\tau = 1 - \frac{4C}{n(n-1)} \quad \text{con } -1 \leq \tau \leq 1$$

"C" è il numero minimo di scambi necessari per trasformare una graduatoria nell'altra. Gli estremi sono interpretabili come nel rho di Spearman

ESEMPIO

Calcolo con il metodo di Holmes (1920)



Le linee che congiungono i ranghi nelle due graduatorie si incrociano C volte

$$\tau = 1 - \frac{4(13)}{7(6)} = 1 - 1.2381 = -0.2381$$

Presenza di parità

Se i ranghi graduano misure soggette a errori è utile dare rango eguale a valori molto prossimi.

Anche in caso di giudizi succede di non riuscire a stabilire una preferenza tra due aspetti egualmente percepiti.

Le parità o ex aequo richiedono l'uso di pseudo-ranghi nelle posizioni coinvolte e modifiche nelle formule di calcolo

e	σ	s/r.m.	σ/max
1	1	1	1
2	2	3	4
3	2	3	3
4	2	3	2
5	3	5	5
6	4	6	6
7	5	7	7
8	6	8.5	9
9	6	8.5	8
10	7	11.5	13
11	7	11.5	12
12	7	11.5	11
13	7	11.5	10
14	8	14	14
15	9	15	15

Esempi:

1) Rango medio

2) Sub-graduatoria che rende massima o minima la correlazione di rango. Media tra le due

3) Bootstrap di sub-permutazioni

Il problema è nella determinazione una tantum del massimo possibile in caso di ex aequo

Formula di rho in caso di parità

$$r_1 = \frac{(n^3 - 3) - 6 \sum_{i=1}^n d_i^2 - \frac{1}{2} \left\{ \sum_{j=1}^{n_x} [(t_j^x)^3 - (t_j^x)] + \sum_{j=1}^{n_y} [(t_j^y)^3 - (t_j^y)] \right\}}{\sqrt{\left[(n^3 - 3) - \sum_{j=1}^{n_x} [(t_j^x)^3 - (t_j^x)] \right] \left[(n^3 - 3) - \sum_{j=1}^{n_y} [(t_j^y)^3 - (t_j^y)] \right]}}$$

dove

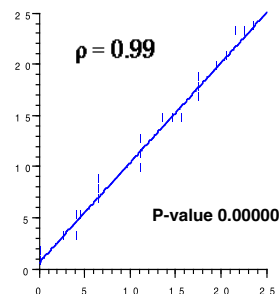
$$\begin{cases} n_x = \text{numero di gruppi di } X \text{ con parità} \\ t_j^x = \text{numero di valori uguali per la } j\text{-esima parità in } X \\ n_y = \text{numero di gruppi di } Y \text{ con parità} \\ t_j^y = \text{numero di valori uguali per la } j\text{-esima parità in } Y \end{cases}$$

Esempio

E	E-or	rE	B	B-or	rB	(rE) ²	(rB) ²	rE*rB
10.3	9.2	0.5	-1.3	-1.8	1	0.25	1	0.5
9.9	9.2	0.5	-1.1	-1.6	2	0.25	4	1
10.3	9.3	3.0	1.3	-1.5	3.5	9	12.25	10.5
9.7	9.4	4.5	0.5	-1.5	3.5	20.25	12.25	15.75
9.6	9.4	4.5	1.8	-1.3	5.5	20.25	30.25	24.75
9.5	9.5	5.0	-0.4	-1.3	5.5	25	30.25	27.5
10.8	9.6	7.0	-1.6	-1.1	7	49	49	49
9.7	9.6	7.0	-1.8	-0.8	8	49	64	56
9.4	9.6	7.0	1.0	-0.7	9	49	81	63
10.1	9.7	11.5	-1.5	-0.5	10	132.25	100	115
9.4	9.7	11.5	-0.8	-0.4	11.5	132.25	132.25	132.25
10.4	9.7	11.5	-1.3	-0.4	11.5	132.25	132.25	132.25
9.6	9.7	11.5	-0.7	-0.2	13	132.25	169	149.5
9.7	9.8	14.0	-0.4	0.5	15	196	225	210
10.6	9.9	15.0	0.5	0.5	15	225	225	225
9.3	10.1	16.0	1.3	0.5	15	256	225	240
10.5	10.3	18.0	1.1	0.7	17	324	289	306
10.7	10.3	18.0	0.9	0.8	18	324	324	324
9.2	10.3	18.0	0.5	0.9	19	324	361	342
9.7	10.4	20.0	0.8	1.0	20	400	400	400
9.2	10.5	21.0	-1.5	1.1	21	441	441	441
9.6	10.6	22.0	-0.2	1.3	23.5	484	552.25	517
9.8	10.7	23.0	0.7	1.3	23.5	529	552.25	540.5
10.3	10.8	24.0	-0.5	1.8	24	576	576	576
						4830	4988	4898.5

Accertamento di una relazione d'ordine tra il tasso di interesse effettivo "E" dei BOT trimestrali e l'indice di borsa "B"

$$r_s = \frac{\frac{4898.5}{24} - \frac{(25)^2}{4}}{\sqrt{\left[\frac{4830}{24} - \frac{(25)^2}{4} \right] \left[\frac{4988}{24} - \frac{(25)^2}{4} \right]}}$$



Esempio

Distanza da un punto inquinante e concentrazione dell'agente nell'aria

	Distanza (X)	Concen. (Y)	ranghi(X)	ranghi (Y)	d(x,y)
	0	510	1	12	121
	50	380	2	9	49
Il rango medio sottostima la variabilità nei ranghi.	300	450	3.5	10	42.25
	300	480	3.5	11	56.25
Induce a ritenere che sia presente più correlazione di quanto non ve ne sia in realtà	800	300	5	7.5	6.25
	900	300	6	7.5	2.25
	1000	170	7	6	1
	1500	94	9	3.5	30.25
	1500	94	9	3.5	30.25
	1500	108	9	5	16
	2000	45	11	1	100
	5000	89	12	2	100
					554.5

$$\rho_s = \frac{1725 - 3327 - 0.5 \left\{ (8-2) + (27-3) \right\} + \left\{ (8-2) + (8-2) \right\}}{\sqrt{[1725 - 30][1725 - 12]}} = -0.95$$

Formula di tau in caso di parità

$$\tau_b = \frac{S}{\sqrt{\left[\binom{n}{2} - \sum_{j=1}^{n_x} \binom{r_j^x}{2} \right] \left[\binom{n}{2} - \sum_{j=1}^{n_y} \binom{r_j^y}{2} \right]}}$$

N.B. $\binom{n}{2} = \frac{n*(n-1)}{2}$

dove $\begin{cases} n_x = \text{numero di gruppi di } X \text{ con parità} \\ r_j^x = \text{numero di valori uguali per la } j\text{-esima parità in } X \\ n_y = \text{numero di gruppi di } Y \text{ con parità} \\ r_j^y = \text{numero di valori uguali per la } j\text{-esima parità in } Y \\ s = \text{numero minimo di interscambi che trasforma } X \text{ in } Y \end{cases}$

$$S = \sum_{i=1}^{n-1} \sum_{j=1}^i \text{sgn}(r_i - r_j) \text{sgn}(s_i - s_j) \quad \text{dove} \quad \text{sgn}(x) = \begin{cases} 1 & \text{se } x > 0 \\ 0 & \text{se } x = 0 \\ -1 & \text{se } x < 0 \end{cases}$$

Applicazione

Una selezione di n=50 giudici ha disposto secondo l'ordine di preferenza 15 versioni di uno stesso prodotto



Non sono presenti parità, ma è raro con tanti gradi di giudizio

Giudice	P1	P2	P3	P4	P5	P6	P7	P8	P9	P10	P11	P12	P13	P14	P15
J_01	10	6	9	7	8	4	5	2	1	3	15	11	14	13	12
J_02	10	7	9	8	6	15	13	11	12	14	1	3	4	5	2
J_03	1	5	4	3	2	6	7	9	8	10	15	13	12	11	14
J_04	10	8	9	6	7	4	5	1	2	3	12	14	13	11	15
J_05	7	8	10	9	6	15	12	13	11	14	4	1	2	5	3
J_06	9	6	10	7	8	1	3	4	2	5	15	11	12	13	14
J_07	8	7	9	6	10	1	2	4	5	3	13	11	14	12	15
J_08	14	15	13	12	11	7	6	9	8	10	5	2	1	3	4
J_09	8	6	7	10	9	5	4	3	1	2	15	14	11	12	13
J_10	9	8	6	7	10	13	11	15	12	14	5	4	3	1	2
J_11	7	6	8	10	9	2	5	3	1	4	11	13	12	15	14
J_12	9	7	10	8	6	1	2	5	4	3	13	15	11	12	14
J_13	4	3	2	5	1	6	8	9	10	7	14	15	12	11	13
J_14	10	9	8	6	7	4	2	5	1	3	15	12	13	14	11
J_15	3	1	2	4	5	9	6	8	7	10	11	12	14	13	15
J_16	8	10	6	9	7	2	5	1	4	3	15	13	11	14	12
J_17	10	6	7	9	8	1	4	2	3	5	11	15	13	12	14
J_18	10	7	6	9	8	15	12	11	14	13	3	4	1	2	5
J_19	9	8	6	7	10	12	11	13	14	15	5	1	4	3	2
J_20	2	1	5	3	4	7	6	8	9	10	15	11	12	14	13

In questi casi non si considerano 15 distinte variabili, ma una sola variabile: la graduatoria fissata dal giudice (che si articola in quindici valori non scindibili)

Calcolo del tau-b di Kendall

Una delle graduatorie è disposta in ordine ascendente (con eventuali parità). L'altra segue per abbinamento.

Per ogni rango del secondo si contano quanti, tra quelli alla sua destra, ne sono superiori

Il totale di questi conteggi darà il valore di "S" nel numeratore del tau.

ESEMPIO

	A	B	C	D	E	F	G	H	I	J	10
	1	2	4.5	4.5	4.5	4.5	8	8	8	8	10
	1	2.5	2.5	4.5	4.5	6.5	6.5	8	9.5	9.5	
1	1	1	1	1	1	1	1	1	1	1	9
	0	1	1	1	1	1	1	1	1	1	7
		0	0	0	1	1	1	1	1	1	4
			0	0	1	1	1	1	1	1	4
				0	1	1	1	1	1	1	4
					0	1	1	1	1	1	3
						0	1	1	1	1	1
							0	0	1	1	1
								0	1	1	1
									0	1	1
										0	0
											33

$$\tau_b = \frac{33}{\sqrt{[45 - (6 + 3)][45 - (1 + 1 + 1 + 1)]}} = \frac{33}{\sqrt{41 * 36}} = 0.859$$

Applicazione/2

Riportiamo una parte della matrice delle dissimilarità

	[.1]	[.2]	[.3]	[.4]	[.5]	[.6]	[.7]
[1.]	0.0000	0.9097	0.9634	0.9825	0.9520	0.9479	0.9572
[2.]	0.9097	0.0000	0.9769	0.9640	0.9654	0.9214	0.9267
[3.]	0.9634	0.9769	0.0000	0.9681	0.9996	0.9807	0.9835
[4.]	0.9825	0.9640	0.9681	0.0000	0.9934	0.9566	0.9582
[5.]	0.9520	0.9654	0.9996	0.9934	0.0000	0.9190	0.9316
[6.]	0.9479	0.9214	0.9807	0.9566	0.9190	0.0000	0.9309
[7.]	0.9572	0.9267	0.9835	0.9582	0.9316	0.9309	0.0000

Gli autovalori sono negativi tranne il primo (quello massimo in valore assoluto) che è positivo.

La traccia della matrice di dissimilarità ovviamente è nulla

[1]	47.363	-0.706	-0.742	-0.759	-0.782	-0.785	-0.813	-0.818	-0.829	-0.843	-0.858
[12]	-0.860	-0.863	-0.868	-0.884	-0.890	-0.896	-0.905	-0.911	-0.912	-0.917	-0.922
[23]	-0.926	-0.931	-0.935	-0.938	-0.942	-0.951	-0.956	-0.959	-0.968	-0.970	-0.978
[34]	-0.988	-0.994	-1.000	-1.008	-1.021	-1.024	-1.056	-1.070	-1.084	-1.089	-1.105
[45]	-1.147	-1.204	-1.256	-1.348	-1.366	-1.389					

Conversione per le correlazioni

Se la misura di affinità è ottenuta come coefficiente di associazione quale ad esempio il coefficiente di correlazione di rango, la conversione richiede una riflessione in più in quanto la presenza del segno specifica la direzione in cui si muove una entità al variare dell'altra.

La trasformazione più ovvia per portare i valori in zero/uno sarebbe:

$$6. \delta_{ij} = \left[\frac{1 - a_{ij}}{2} \right]^\alpha \quad \alpha > 0$$

Ha il difetto logico di far corrispondere la dissomiglianza massima alla dissociazione (cioè unità con modalità opposte) e non alla mancanza di affinità: $a_{ij}=0$.

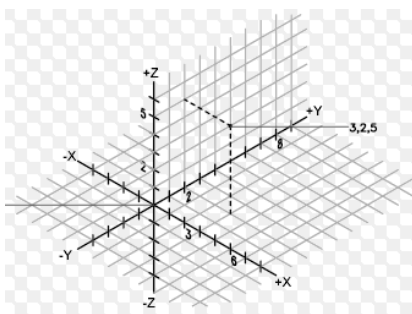
Se pensiamo alla associazione negativa come ad una forma più dettagliata di affinità che oltre a dare la misura dell'intensità del legame tra le due unità è in grado di specificare il grado di opposizione, la (6.) torna ad essere intellegibile.

Affinità per variabili metriche

Il concetto di affinità diventa vicinanza, contiguità, prossimità quando è riferito a variabili quantitative per il significato metrico e quindi di distanza che hanno le modalità.

Ogni unità può essere rappresentata come un punto in un sistema cartesiano di assi (assi formanti angoli retti) ed ogni idea di somiglianza tra di esse potrà essere espressa come vicinanza.

$$\begin{array}{c|ccc} & X_1 & \cdots & X_m \\ \hline u_1 & x_{11} & \cdots & x_{1m} \\ u_2 & x_{21} & \cdots & x_{2m} \\ \vdots & \vdots & & \vdots \\ u_i & x_{i1} & \ddots & x_{im} \\ \vdots & \vdots & & \vdots \\ u_n & x_{n1} & & x_{nm} \end{array}$$



L'insieme delle distanze forma una matrice simmetrica e nonnegativa avente degli zeri sulla diagonale

Conversione Affinità/Dissimilarità

Per trasformare le correlazioni in dissimilarità o distanze esistono varie formule

$$1) \quad d_{ij} = \sqrt{1 - \frac{r_{ij} + 1}{2}}, \quad \forall i, j$$

Le distanze hanno valore zero se la r_{ij} è +1.

Hanno valore massimo 1 se r_{ij} è -1.

$$2) \quad d_{ij} = \frac{1 - r_{ij}}{2}, \quad \forall i, j$$

L'incertezza è sul significato da dare a $r_{ij}=0.5$

L'uso del coefficiente in valore assoluto o Del quadrato elimina le incertezze

$$3) \quad d_{ij} = \sqrt{1 - |r_{ij}|}, \quad \forall i, j$$

$$d_{i,j} = \sqrt{1 - |r_{i,j}|}$$

$$d_{i,j} = \sqrt{1 - (r_{i,j})^2}$$

Affinità per variabili metriche/2

I numeri sono veri e propri numeri adoperati per registrare l'esito di conteggi (variabili discrete) o di rapporti di misurazione (variabili continue).

In entrambi i casi le modalità del dominio si presentano come una successione più o meno fitta di valori

$$\mathcal{S}(x) = \{x_1, x_2, \dots, x_p, \dots\}$$

I cui valori delle continue sono stati arrotondati ed eventualmente moltiplicati per una potenza del dieci. Il dominio può essere finito o infinito, anche se, per ogni data applicazione è possibile proporre dei ragionevoli limiti estremi.

$$X_{4 \times 2} = \begin{bmatrix} 42 & 4 \\ 52 & 5 \\ 48 & 8 \\ 58 & 3 \end{bmatrix}$$

Vettore delle medie (centroide) $\bar{x} = \begin{bmatrix} 50 \\ 4 \end{bmatrix}$

Matrice di varianze-covarianze e di correlazione $S_n = \begin{bmatrix} 34 & -1.5 \\ -1.5 & 0.5 \end{bmatrix}$, $R = \begin{bmatrix} 1 & -0.36 \\ -0.36 & 1 \end{bmatrix}$

Distanze tra entità

Ipotizziamo, per semplicità, che le unità siano descritte da un vettore di valori X

$d(X_i, X_j) = 0$ se e solo se $X_i = X_j$;	Identità
$d(X_i, X_j) > 0$ se $X_i \neq X_j$;	Positività
$d(X_i, X_j) = d(X_j, X_i)$;	Simmetria
$d(X_i, X_k) + d(X_k, X_j) \geq d(X_i, X_j)$;	Disuguaglianza triangolare

La quarta proprietà porta alla distinzione tra dissimilarità e distanza nel senso che alle prime non è richiesta la verifica della disuguaglianza triangolare.

L'insieme delle distanze forma una matrice simmetrica e nonnegativa avente degli zeri sulla diagonale.

Se per ognuna delle permutazioni degli n indici presi a tre a tre si verifica la quarta proprietà, la matrice si dice euclidea e gode di particolari proprietà

Una distanza elementare

Verifichiamo le potenzialità come metrica della seguente funzione

$$d(x_i, x_j) = \begin{cases} 1 & \text{se } i \neq j \\ 0 & \text{se } i = j \end{cases} \quad \text{Un punto non può essere più vicino ad un altro di quanto non lo sia a se stesso.}$$

E' immediato accertare che possiede le caratteristiche della identità (dato che è zero solo se $i=j$), della positività (dato che è maggiore e uguale a zero) e della simmetria (dato che è 1 se $i \neq j$ e 1 se $j \neq i$).

Per la disuguaglianza triangolare dobbiamo considerare tre punti: i, j, k.

Se sono tutti diversi, allora $d(x_i, x_k) + d(x_k, x_j) \geq d(x_i, x_j) \Leftrightarrow 1 + 1 \geq 1$

$$x_i = x_j \neq x_k \Rightarrow d(x_i, x_k) + d(x_k, x_j) \geq d(x_i, x_j) \Leftrightarrow 1 + 1 \geq 0$$

$$x_i = x_k \neq x_j \Rightarrow d(x_i, x_k) + d(x_k, x_j) \geq d(x_i, x_j) \Leftrightarrow 0 + 1 \geq 1$$

Se due o tutti sono uguali, allora

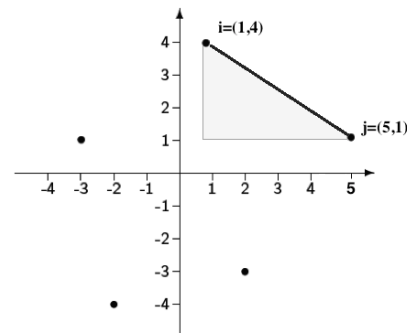
$$x_i \neq x_j = x_k \Rightarrow d(x_i, x_k) + d(x_k, x_j) \geq d(x_i, x_j) \Leftrightarrow 1 + 0 \geq 1$$

$$x_i = x_j = x_k \Rightarrow d(x_i, x_k) + d(x_k, x_j) \geq d(x_i, x_j) \Leftrightarrow 0 + 0 \geq 0$$

Significato

Due punti che sono entrambi vicini ad un altro debbono pure essere in qualche modo vicini

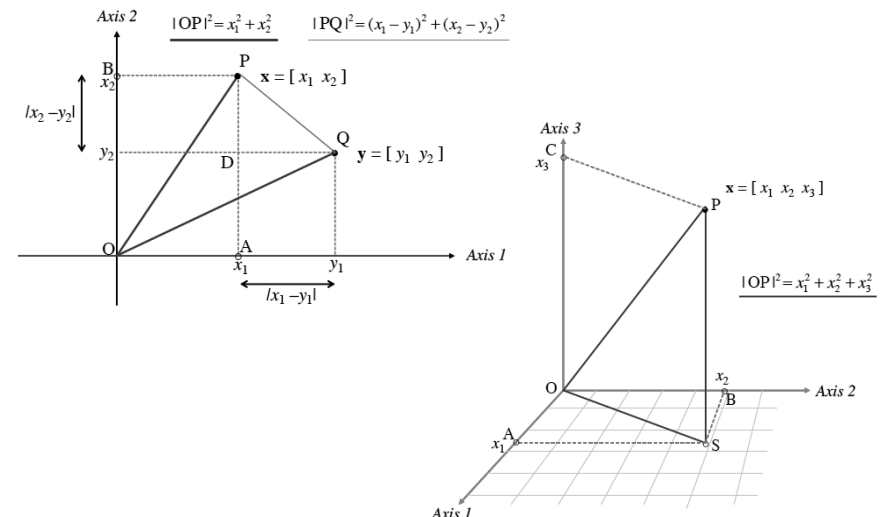
$$\begin{aligned} d(x_i, x_j) &= \sqrt{(1-5)^2 + (4-1)^2} \\ &= \sqrt{16+9} = \sqrt{25} = 5 \end{aligned}$$



Il percorso diretto è sempre più breve o al massimo uguale che con una svolta intermedia.

Questa nozione di distanza ingloba l'idea di invarianza rispetto alle traslazioni ed alle rotazioni degli assi

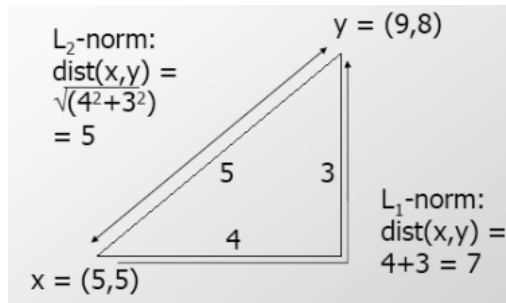
Teorema di Pitagora e distanza 2D e 3D



Metriche di Minkowski

$$d_{ij} = \left[\sum_{r=1}^m |x_{ri} - x_{rj}|^p \right]^{1/p}$$

$p=1$ (city block o Manhattan)
 $p \rightarrow \infty$ $p=2$ (Euclidea)
 (Tchebycheff o Max)



Se d_{ij} è una distanza allora lo sono anche:

$$d_{ij}^+ = \alpha d_{ij} \quad \text{con } \alpha > 0$$

$$d_{ij}^+ = \log(1 + d_{ij})$$

$$d_{ij}^+ = (d_{ij})^\beta \quad \text{con } 0 < \beta \leq 1$$

Una caratteristica delle metriche di Minkowski è che il loro valore numerico aumenta all'aumentare del numero di variabili.

Esempio

point	x	y
p1	0	2
p2	2	0
p3	3	1
p4	5	1

L1	p1	p2	p3	p4
p1	0	4	4	6
p2	4	0	2	4
p3	4	2	0	2
p4	6	4	2	0

L2	p1	p2	p3	p4
p1	0	2.828	3.162	5.099
p2	2.828	0	1.414	3.162
p3	3.162	1.414	0	2
p4	5.099	3.162	2	0

L _∞	p1	p2	p3	p4
p1	0	2	3	5
p2	2	0	1	3
p3	3	1	0	2
p4	5	3	2	0

La scelta tra metrica euclidea e city-block implica la decisione se le variabili sono da considerarsi separate oppure interagiscono

La city-block porta ad un giudizio sulla distanza in cui le variabili agiscono in modo indipendente per sommarsi nel giudizio di prossimità. Ad esempio valutando una persona per peso ed altezza la somiglianza tra persone potrebbe avvenire giudicandole separatamente per i due aspetti.

Con la euclidea il giudizio deriva da un mix tra due aspetti collegati.

Metriche di Minkowski/2

Le metriche di Minkowski verificano la disuguaglianza triangolare e ciò deriva proprio dalla disuguaglianza di Minkowski

$$\left[\sum_{r=1}^m |x_{ri} + x_{rj}|^p \right]^{1/p} \leq \left[\sum_{r=1}^m |x_{ri}|^p \right]^{1/p} + \left[\sum_{r=1}^m |x_{rj}|^p \right]^{1/p}$$

Poniamo $y_{ri} = x_{ri} - x_{rk}$, $k \neq i$, $k \neq j$ Ne consegue:

$$\begin{aligned} \left[\sum_{r=1}^m |y_{ri} - y_{rj}|^p \right]^{1/p} &\leq \left[\sum_{r=1}^m |y_{ri} + y_{rj}|^p \right]^{1/p} \leq \left[\sum_{r=1}^m |y_{ri}|^p \right]^{1/p} + \left[\sum_{r=1}^m |y_{rj}|^p \right]^{1/p} \\ \left[\sum_{r=1}^m |y_{ri} - y_{rj}|^p \right]^{1/p} &\leq \left[\sum_{r=1}^m |y_{ri} + y_{rj}|^p \right]^{1/p} \leq \left[\sum_{r=1}^m |x_{ri} - x_{rk}|^p \right]^{1/p} + \left[\sum_{r=1}^m |x_{rj} - x_{rk}|^p \right]^{1/p} \\ \left[\sum_{r=1}^m |x_{ri} - x_{rk} - (x_{rj} - x_{rk})|^p \right]^{1/p} &\leq \left[\sum_{r=1}^m |y_{ri} + y_{rj}|^p \right]^{1/p} \leq d_{ik}^p + d_{jk}^p \\ \left[\sum_{r=1}^m |x_{ri} - x_{rj}|^p \right]^{1/p} &\leq d_{ik}^p + d_{jk}^p \Rightarrow d_{ij}^p \leq d_{ik}^p + d_{jk}^p \end{aligned}$$

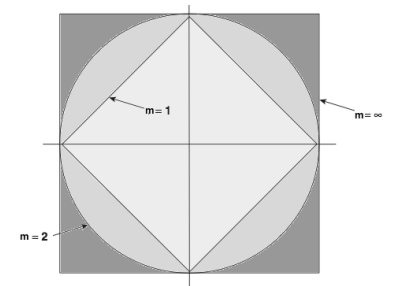
Metriche di Minkowski/3

Le metriche di Minkowski verificano la disuguaglianza triangolare e ciò deriva proprio dalla disuguaglianza di Minkowski

$$\lim_{n \rightarrow \infty} \left[\sum_{r=1}^m |x_{ri} - x_{rj}|^n \right]^{1/n} = \max_{1 \leq r \leq m} |x_{ri} - x_{rj}|$$

$$\lim_{n \rightarrow -\infty} \left[\sum_{r=1}^m |x_{ri} - x_{rj}|^n \right]^{1/n} = \min_{1 \leq r \leq m} |x_{ri} - x_{rj}|$$

In generale, non si può dire che le metriche di Minkowski siano monotone rispetto al loro indice

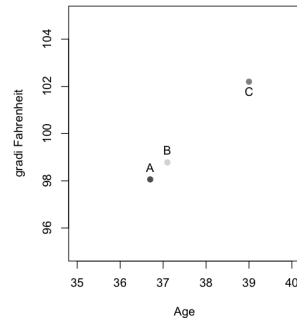
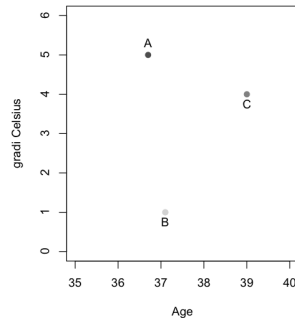


Sono molto sensibili ai valori remoti. Se questi fossero presenti sarebbe opportuno adoperare metodi robusti per la determinazione di queste distanze.

Dipendenza dall'unità di misura

Sog.	Temp_C	Temp_F	Age	TC*	TF*	A*
A	36.70	98.06	5	-0.7324	-0.7324	0.8006
B	37.10	98.78	1	-0.4069	-0.4069	-1.1209
C	39.00	102.2	4	1.1393	1.1393	0.3203
	37.600	99.680	3.333	0.000	0.000	0.000
	1.229	2.212	2.082	1.000	1.000	1.000

Le metriche di Minkowski dipendono dall'unità di misura delle variabili



Con i gradi Celsius il soggetto A e C sono più vicini rispetto al soggetto B. Con i gradi Fahrenheit risulta separata la C.

Ponderazione delle variabili

Si intende la trasformazione dei valori di una variabile (di solito con la moltiplicazione per un fattore) al fine di contrarre o espandere l'impatto che la variabile stessa ha sulle unità.

La procedura è controversa:



Da un lato c'è chi ritiene di disporre di informazioni sulla importanza relativa delle diverse variabili ed intende farne uso.

Un geologo che prospezta un terreno può decidere che gli strati abbiano rilevanza inversa rispetto alla profondità. Chi analizza un farmaco può disporre in ordine decrescente di valutazione le componenti



Se le variabili sono state inserite per una qualche ragione allora le ragioni sono tutte egualmente valide e non è corretto dare un peso diverso alle variabili.

Attenzione alla ponderazione implicita! Variabili molto correlate in realtà fanno pesare molto di più quello che hanno in comune

Dietro ogni normalizzazione/standardizzazione c'è sempre una perdita di informazioni

Metriche di Minkowski ponderate

$$d_{ij} = \left[\sum_{r=1}^m |y_{ri} - y_{rj}|^p \right]^{1/p} ; y_{ri} = w_r^{1/p} x_{ri}, y_{rj} = w_r^{1/p} x_{rj}; w_r \geq 0, \sum_{r=1}^m w_r = 1$$

Il modo più semplice di determinare i pesi è di scegliere una quantità da collocare al denominatore delle variabili.

La standardizzazione costringe le nuove variabili ad avere varianza uno.

La unitarizzazione costringe le nuove variabili a variare nell'intervallo [0,1]

La divisione per la media porta al coefficiente di variazione.

$$\text{Standardizzazione: } \frac{x_{r,j}}{\sigma(x_j)}$$

$$\text{Unitarizzazione: } \frac{x_{r,j}}{\max(x_j) - \min(x_j)}$$

$$\text{Rapporto alla media: } \frac{x_{r,j}}{\mu(x_j)}$$

$$\text{Scarto medio assoluto: } \frac{nx_{r,j}}{\sum_{r=1}^n |x_{r,j} - Me(x_j)|}$$

Metriche di Minkowski ponderate/2

La scelta della trasformazione (e quindi della ponderazione ovvero del tipo di vincolo a cui le variabili debbono sottostare) può favorire alcune variabili a danno di altre.

ESEMPIO: In un data set vi sono n entità per le quali la variabile X può assumere valore uno oppure zero. In partiolare vi sono "a" entità che hanno valore 1.

Misuriamo la distanza con la metrica city-block.

$$0 \leq |x_i - x_j| \leq 1$$

Prima della standardizzazione, lo scarto tra due entità era

$$\text{Dopo la standardizzazione è } 0 \leq |x_i^* - x_j^*| \leq \frac{n}{\sqrt{a(n-a)}}$$

Se a=1 oppure a=(n-1) il valore assunto sarà $\sqrt{n-1}$ e sarà pari a 2 se a=1/(2n) cosicché una variabile con pochi valori 1 può ottenere una rilevanza sproporzionata rispetto al ruolo atteso.

Metriche relative

Rendono invariante la distanza rispetto a trasformazioni moltiplicative o di scala, ma non necessariamente rispetto a trasformazioni additive

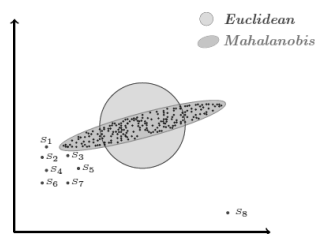
Equalizzano l'impatto delle diverse variabili che ora contribuiscono alla distanza tra entità senza far valere la loro specifica unità di misura

$$\text{Bray - Curtis : } \frac{\sum_{r=1}^m |x_{ri} - x_{rj}|}{\sum_{r=1}^m |x_{ri}| + \sum_{r=1}^m |x_{rj}|}; \quad \text{Canberra : } \sum_{r=1}^m \left(\frac{|x_{ri} - x_{rj}|}{|x_{ri}| + |x_{rj}|} \right)$$

$$\text{Soergel : } \frac{\sum_{r=1}^m |x_{ri} - x_{rj}|}{\sum_{r=1}^m \max(|x_{ri}|, |x_{rj}|)}; \quad \text{Ware - Hedges : } \frac{1}{m} \sum_{k=1}^m \left[1 - \frac{\min(|x_{ki}|, |x_{kj}|)}{\max(|x_{ki}|, |x_{kj}|)} \right]$$

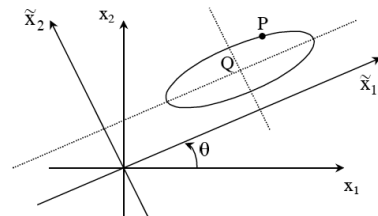
Distanza di Mahalanobis/2

La distanza di Mahalanobis è la distanza euclidea ponderata ovvero equivale ad una combinazione lineare degli scarti tra le variabili



Con variabili correlate la distanza si misura in effetti:

- 1) Ruotando gli assi di un angolo legato alla correlazione tra le variabili
- 2) Calcolando la distanza euclidea nel piano degli assi ruotati



In questo modo si può verificare che la distanza di Mahalanobis è in effetti una metrica.

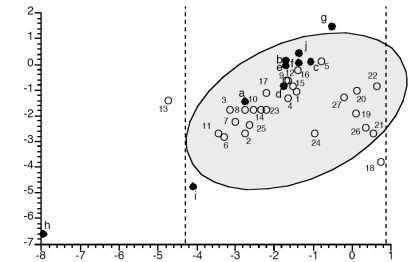
Non è più così se si ignora la radice quadrata.

Distanza di Mahalanobis

Le metriche di Minkowski ignorano i legami di correlazione tra le variabili.

Un modo per includere le relazioni lineari nella misura della distanza quantitative è la metrica:

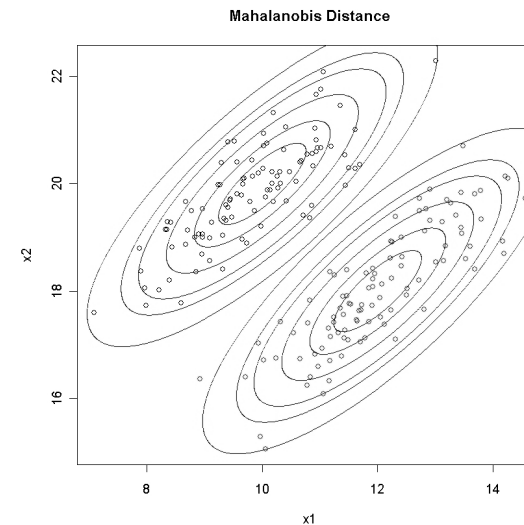
$$d_{i,j} = \sqrt{(x_i - \mu)^t \Omega^{-1} (x_j - \mu)}$$



Per m=2 variabili si ha

$$d(x, y) = [(x_1 - y_1)^2 a_{11}^2 + 2(x_1 - y_1)(x_2 - y_2) a_{12} + (x_2 - y_2)^2 a_{22}]^{1/2}$$

La metrica di Mahalanobis è molto utile se si dispone di dati sufficienti per stimare in modo attendibile i suoi parametri.



La distanza di Mahalanobis tiene conto non solo degli scarti tra valori, ma anche della correlazione tra le variabili.

Perché si possa calcolare le correlazioni non debbono essere perfette.

In questo caso ci sarebbe dipendenza lineare tra le colonne della matrice dei dati ed il determinante della matrice di varianze-covarianze sarebbe zero.

Se le variabili metriche del data set fossero incorrelate, l'uso della distanza di Mahalanobis equivarrebbe all'impiego della euclidea per variabili standardizzate.

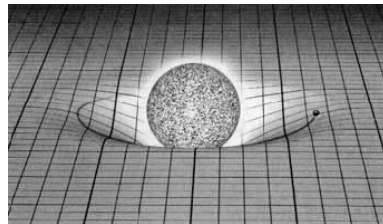
Presupposto euclideo

Le nostre analisi sono sempre riferite allo spazio euclideo multidimensionale privo di curvaturei.

Non è l'unico possibile. Se ci trovassimo su di una sfera la distanza più breve non sarebbe lungo una linea retta e le rette parallele non potrebbero esistere. Dovremmo ragionare nell'ambito della Geometria ellittica di Riemann.



E' una delle piu' strane, ma piu' vicina delle altre alla geometria del mondo reale



Einstein realized this, and a lot of his relativity work was the development of this different geometry. Much of relativity followed easily once he got the geometry right. N. Vasconcelos, UCSD

Conversione per le distanze



Si possono anche trasformare le distanze in modo da portarle nell'intervallo unitario

$$d_{i,j}^* = \frac{d_{i,j} - \min\{d_{i,j}\}}{\max\{d_{i,j}\} - \min\{d_{i,j}\}}; \quad d_{i,j}^* = \sqrt{1 - \frac{d_{i,j}}{\max\{d_{i,j}\}}};$$

$$d_{i,j}^* = e^{-d_{i,j}}; \quad d_{i,j}^* = \frac{2e^{-d_{i,j}}}{1 + e^{-d_{i,j}}}; \quad d_{i,j}^* = \frac{d_{i,j}}{1 + d_{i,j}}$$

che non solo convertono l'affinità in dissomiglianza, ma sono in grado di trasformare la matrice delle affinità/prossimità in matrici euclidee che hanno un ruolo importante nell'analisi multivariata.

Normalizzazione delle distanze

Se una misura che varia tra zero ed infinito (come la varianza) non è ritenuta soddisfacente per il confronto tra distanze si possono percorrere due strade



Scegliere metriche che sono costruite per variare in un intervallo predefinito con zero che significa assenza di distanza ed uno (o altro limite) che implichi lo sostamento massimo ottenibile.

$$\text{Ware - Hedges : } \sum_{r=1}^m \left[1 - \frac{\min(x_{ri}, x_{rj})}{\min(x_{ri}, x_{rj})} \right];$$

$$\text{Angolare : } 1 - \cos^2(\theta_{ij}); \quad \theta_{ij} = \text{angolo fra } x_i \text{ e } x_j$$

Si ottiene una soglia massima rispetto alla quale confrontare il risultato ottenuto.

Ha il difetto di non dipendere dal numero delle variabili per cui il raggiungimento degli estremi può essere pilotato scegliendo le variabili opportune.

Esempi di data set

	Region	Product	Subsidiary	Store	Sales	Inventory	Returns
322	United States	Slipper	New York	24	\$202,768	\$1,069,20	\$6,655
323	United States	Sport Sho	New York	23	\$22,190	\$141,143	\$647
324	United States	Women's C	New York	13	\$173,842	\$450,625	\$4,236
325	United States	Women's D	New York	17	\$283,972	\$1,125,59	\$16,287
326	United States	Boot	Seattle	12	\$70,750	\$225,639	\$2,009
327	United States	Men's Cas	Seattle	6	\$66,642	\$112,638	\$2,295
328	United States	Men's Dre	Seattle	10	\$132,387	\$274,310	\$3,161
329	United States	Sandal	Seattle	10	\$6,696	\$18,768	\$181
330	United States	Slipper	Seattle	10	\$155,255	\$553,502	\$6,277
331	United States	Sport Sho	Seattle	29	\$19,340	\$105,240	\$1,021
332	United States	Women's C	Seattle	2	\$17,262	\$26,510	\$790
333	United States	Women's D	Seattle	21	\$132,436	\$392,527	\$4,037
334	Western Europe	West	Copenhagen	2	\$1,663	\$4,667	\$129
335	Western Europe	Men's Cas	Copenhagen	1	\$38,462	\$92,472	\$167
336	Western Europe	Men's Dre	Copenhagen	2	\$13,416	\$36,275	\$666
337	Western Europe	Slipper	Copenhagen	4	\$16,536	\$59,153	\$294
338	Western Europe	Sport Sho	Copenhagen	1	\$1,527	\$17,643	\$21
339	Western Europe	Sport Sho	Copenhagen	10	\$191,922	\$327,742	\$4,594

Questo è pure un data set, per quanto derivato da elaborazioni su un altro data set

Variable	N	Mean	Std Dev	Minimum	Maximum
Age	20	46.7000000	12.8107685	15.0000000	63.0000000
Height	20	66.9500000	4.1482400	61.0000000	75.0000000
Weight	20	174.6500000	36.3119163	102.0000000	240.0000000
Pulse	20	75.1000000	7.9795792	65.0000000	100.0000000
FastGluc	20	299.1000000	125.6146572	152.0000000	568.0000000
PostGluc	20	355.1000000	125.5153167	206.0000000	625.0000000

Restaurant	Type	Price (\$)	Score
Bertucci's	Italian	16	77
Black Angus Steakhouse	Seafood/Steakhouse	24	79
Bonefish Grill	Seafood/Steakhouse	26	85
Bravo! Cucina Italiana	Italian	18	84
Buca di Beppo	Italian	17	81
Bugaboo Creek Steak House	Seafood/Steakhouse	18	77
Carrabba's Italian Grill	Italian	23	86
Charlie Brown's Steakhouse	Seafood/Steakhouse	17	75
Il Fornaio	Italian	28	83
Joe's Crab Shack	Seafood/Steakhouse	15	71
Johnny Carino's Italian	Italian	17	81
Lone Star Steakhouse & Saloon	Seafood/Steakhouse	17	76
LongHorn Steakhouse	Seafood/Steakhouse	19	81
Maggiano's Little Italy	Italian	22	83
McGrath's Fish House	Seafood/Steakhouse	16	81
Olive Garden	Italian	19	81
Outback Steakhouse	Seafood/Steakhouse	20	80
Red Lobster	Seafood/Steakhouse	18	78
Romano's Macaroni Grill	Italian	18	82
The Old Spaghetti Factory	Italian	12	79
Uno Chicago Grill	Italian	16	76

Distanze multiscala

La matrice dei dati potrebbe contenere variabili misurate su scala diversa.
Ad esempio

m_1	Metriche (rapporti o intervalli)	$d_{i,j}^1$
m_2	Ordinali	$d_{i,j}^2$
m_3	Politome	$d_{i,j}^3$
m_4	Binarie simmetriche	$d_{i,j}^4$
m_5	Binarie asimmetriche	$d_{i,j}^5$

La misura sintetica di distanza può essere definita con una combinazione lineare delle distanze

Poiché l'ordine di grandezza può essere diverso si debbono usare distanze normalizzate (cioè comprese tra zero ed uno).

$$\text{distanza di Gower } d_{i,j}^g = \sum_{k=1}^5 \gamma_k d_{i,j}^k; \quad \gamma_k \geq 0, \quad \sum_{k=1}^5 \gamma_k = 1$$

Appendix

We want to show that if both $d_1(x, y)$ and $d_2(x, y)$ satisfy the triangle inequality, then the same is true for $d(x, y) = \sqrt{d_1^2(x, y) + d_2^2(x, y)}$. Let $a_i = d_1(x, z)$, $b_i = d_1(x, y)$, and $c_i = d_1(y, z)$ so that we have $a_i \leq b_i + c_i$, for $i = 1, 2$. Also, let $a = d(x, z) = \sqrt{a_1^2 + a_2^2}$, $b = d(x, y) = \sqrt{b_1^2 + b_2^2}$, and $c = d(y, z) = \sqrt{c_1^2 + c_2^2}$. In order to show that $a \leq b + c$ holds, it suffices to show that $a_1^2 + a_2^2 = a^2 \leq b^2 + c^2 + 2bc = b_1^2 + b_2^2 + c_1^2 + c_2^2 + 2bc$. This holds if

$$(b_1^2 + c_1^2 + 2b_1c_1) + (b_2^2 + c_2^2 + 2b_2c_2) \leq b_1^2 + b_2^2 + c_1^2 + c_2^2 + 2bc,$$

i.e.,

$$b_1c_1 + b_2c_2 \leq bc,$$

i.e.,

$$\begin{aligned} b_1^2c_1^2 + b_2^2c_2^2 + 2b_1c_1b_2c_2 &\leq b^2c^2 \\ &= (b_1^2 + b_2^2)(c_1^2 + c_2^2) \\ &= b_1^2c_1^2 + b_2^2c_2^2 + b_1^2c_2^2 + b_2^2c_1^2, \end{aligned}$$

i.e.,

$$0 \leq (b_1c_2 - b_2c_1)^2,$$

which is clearly true and this completes the proof that $d(x, y)$ satisfies the triangle inequality.

Teorema

Scelta dei pesi



Pesi uguali

Accettabile se si ignora se nessuna tipologia è più rilevante delle altre



Pesi come frazione di variabili

Numero di variabili di un tipo sul totale delle variabili. Corretto se ogni tipologia da un proprio contributo che rimane lo stesso all'interno della tipologia.



Pesi per eguagliare la media (o la deviazione standard)

I pesi sono determinati in modo che la media (o la deviazione standard) delle distanze sia la stessa per ciascuna tipologia

Applicazione

Ecological traits and phylogeographic structure for all 27 alpine plant species



Binarie asimmetriche: 3,4,5

Politome ordinate: 1,2,6,7

Species	Landolt's indicators					Succession
	Moisture	Soil ecology	Nitrogen	Light	Temperature	
<i>Androsace obtusifolia</i> Alt.	2	1	1	1	1	2
<i>Arabis alpina</i> L.	2	3	1	1	2	2
<i>Campanula barbata</i> L.	2	1	1	1	2	2
<i>Carex firma</i> Myrd.	1	3	1	2	1	3
<i>Carex sempervirens</i> Vill.	1	2	1	1	1	3
<i>Cerastium uniflorum</i> Clairv.	2	1	1	2	1	2
<i>Cirsium spinosissimum</i> (L.) Scop.	3	2	2	1	2	2
<i>Dryas octopetala</i> L.	1	3	1	2	1	3
<i>Gentiana nivalis</i> L.	2	2	1	1	1	2
<i>Geum montanum</i> L.	2	1	1	1	2	2
<i>Geum reptans</i> L.	2	1	1	2	1	1
<i>Gypsophila repens</i> L.	2	3	1	2	2	2
<i>Hedysarum hedysaroides</i> (L.) Schinz & Thell. s.l.	2	3	2	1	2	2
<i>Hornungia alpina</i> (L.) Appel s.l.	3	3	1	2	1	1
<i>Hypochaeris uniflora</i> Vill.	2	1	1	1	2	2
<i>Juncus trifidus</i> L.	1	1	1	2	1	3
<i>Ligusticum mutellinoides</i> (Cr.) Vill.	2	2	1	2	1	3
<i>Loiseleuria procumbens</i> (L.) Desv.	1	1	1	2	1	3
<i>Luzula alpinopilosa</i> (Chaix) Brestr.	3	1	1	2	1	1
<i>Peucedanum ostruthium</i> (L.) W.D. Koch	2	2	2	1	2	2
<i>Phyteuma betonicifolium</i> Vill. s.l.	2	1	1	1	2	2
<i>Phyteuma hemisphaericum</i> L.	1	1	1	1	1	2
<i>Ranunculus alpestris</i> L. s.l.	3	3	1	1	1	1
<i>Rhododendron ferrugineum</i> L.	2	1	1	1	2	1
<i>Saxifraga stellaris</i> L.	3	2	1	2	2	1
<i>Sesleria caerulea</i> (L.) Ard.	1	3	1	1	2	3
<i>Trifolium alpinum</i> L.	1	1	1	1	2	2

Applicazione/2

	[,1]	[,2]	[,3]	[,4]	[,5]
[1,]	0.000	0.500	0.000	0.625	0.375
[2,]	0.500	0.000	0.500	0.375	0.625
[3,]	0.000	0.500	0.000	0.625	0.375
[4,]	0.625	0.375	0.625	0.000	0.250
[5,]	0.375	0.625	0.375	0.250	0.000

Binarie asimmetriche
(matrice parziale)

Russell-Rao

	[,1]	[,2]	[,3]	[,4]	[,5]
[1,]	0.0000	0.6667	0.3333	0.6667	0.6667
[2,]	0.6667	0.0000	0.6667	0.6667	0.6667
[3,]	0.3333	0.6667	0.0000	0.6667	0.6667
[4,]	0.6667	0.6667	0.6667	0.0000	0.3333
[5,]	0.6667	0.6667	0.6667	0.3333	0.0000

Politome ordinate
(matrice parziale)

Media di scarti relativi

	[,1]	[,2]	[,3]	[,4]	[,5]
[1,]	0.0000	0.5952	0.1905	0.6488	0.5417
[2,]	0.5952	0.0000	0.5952	0.5417	0.6488
[3,]	0.1905	0.5952	0.0000	0.6488	0.5417
[4,]	0.6488	0.5417	0.6488	0.0000	0.2976
[5,]	0.5417	0.6488	0.5417	0.2976	0.0000

Matrice ponderata con pesi 3/7 e 4/7
in base al numero delle variabili

Applicazione/3

Ipo<-c(1,2,9);Ime<-3:8

Species	Host	Country	Av_Len	Max_Len	Min_Len	Av_Wid	Max_Wid	Min_Wid	Class
I1	Cat	Liberia	64	72	58	41	47	38	uterobilateralis
I2	Cat	Liberia	71	77	65	42	43	40	uterobilateralis
I3	Mongoose	Liberia	70	76	65	42	47	40	uterobilateralis
I4	Mongoose	Gabon	72	78	67	41	44	37	uterobilateralis
I5	Albinos_rat	Liberia	63	70	53	39	43	35	uterobilateralis
I6	Man	Liberia	70	78	62	44	46	39	uterobilateralis
I7	Man	Guinea	72	95	62	44	52	39	uterobilateralis
I8	Man	Congo	66	75	60	44	48	39	uterobilateralis
I9	Cat	Cameroon	93	105	74	48	52	45	africanus
I10	Cat	Cameroon	90	99	85	48	54	46	africanus
I11	Cat	Cameroon	94	103	80	49	55	43	africanus
I12	Civet_cat	Cameroon	89	120	72	50	68	44	africanus
I13	Man	Cameroon	88	112	70	52	62	47	africanus
I14	Cat	Cameroon	98	117	84	57	70	45	westernman
I15	Cat	Cameroon	94	112	71	53	61	47	westernman
I16	Cat	Cameroon	94	110	85	58	75	45	westernman
I17	Man	Gabon	100	112	88	58	67	53	westernman
I18	Man	Cameroon	98	113	84	64	76	51	westernman
I19	Cat	Cameroon	84	93	65	48	55	38	Euparagonimus
I20	Cat	Cameroon	85	95	75	49	55	40	Euparagonimus
I21	Cat_doga	Ivory_Coast	83	102	63	46	54	36	Euparagonimus
I22	Man	Cameroon	85	105	70	49	55	43	Euparagonimus

	[,1]	[,2]	[,3]	[,4]
[1,]	0.000	0.143	0.571	1.000
[2,]	0.143	0.000	0.571	1.000
[3,]	0.571	0.571	0.000	0.571
[4,]	1.000	1.000	0.571	0.000

	[,1]	[,2]	[,3]	[,4]
[1,]	0.000	0.040	0.030	0.041
[2,]	0.040	0.000	0.009	0.013
[3,]	0.030	0.009	0.000	0.019
[4,]	0.041	0.013	0.019	0.000