

Data Mining in R

Approccio non supervisionato

ANALISI CLUSTER

2. Metodi gerarchici e distanze

Laura Grassini

Tan, Steinbach, Kumar: Introduction to data mining, 2006,
Addison Wesley

<http://www-users.cs.umn.edu/~kumar/dmbook/index.php>

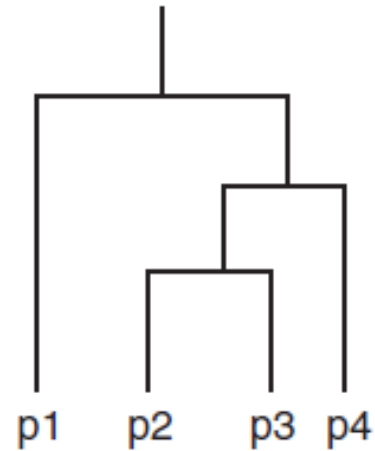
Indice

- I metodi gerarchici agglomerativi e divisivi
- La logica dei metodi di cluster gerarchici agglomerativi
- La prossimità fra unità e fra cluster
- Variabili numeriche: distanze e indici di distanza
- L'unità di misura e la scala delle variabili
- Variabili categoriche: dissimilarità
- Variabili miste: distanza di Gower

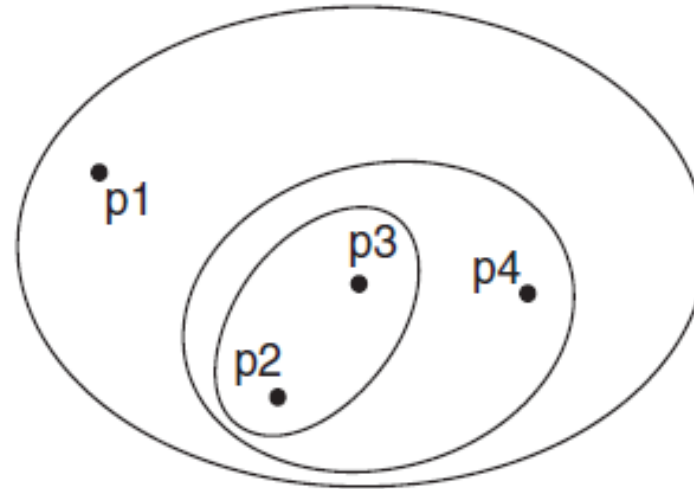
I metodi gerarchici

- **Gerarchici agglomerativi:** da n gruppi (quante sono le unità) si arriva progressivamente a raggruppare le unità più prossime in sottogruppi, e a raggruppare i sottogruppi formati in gruppi, fino ad avere un solo gruppo che contiene tutte le unità
- **Gerarchici divisivi:** da un solo gruppo, che contiene tutte le unità, si procede alla progressiva suddivisione delle unità meno prossime fino ad ottenere n gruppi ciascuno con una sola unità

Metodo gerarchico agglomerativo con 4 unità e 2 variabili



(a) Dendrogram.



(b) Nested cluster diagram.

Quando un'unità è entrata in un gruppo **non viene** da questo più rimossa.
Il metodo gerarchico **non fornisce** il numero dei gruppi. Esso deve essere scelto a posteriori esaminando i risultati.

La logica dei metodi gerarchici agglomerativi

1. Calcola la matrice di prossimità (distanze, similarità, dissimilarità) fra le unità
2. Ripeti fino a che non arrivi ad un solo cluster con tutti i dati
 3. Unisci («merge») i cluster più prossimi
 4. Aggiorna la matrice delle prossimità

La logica del raggruppamento gerarchico (1/8)

- Vediamo il funzionamento della cluster gerarchica a partire da una matrice delle distanze fra 6 unità (N.B. la matrice è **simmetrica**).


	U_1	U_2	U_3	U_4	U_5	U_6
U_1	0	3,00	0,30	1,08	1,56	1,49
U_2	3,00	0	5,02	1,76	0,57	0,91
U_3	0,30	5,02	0	1,68	2,78	2,53
U_4	1,08	1,76	1,68	0	0,33	0,15
U_5	1,56	0,57	2,78	0,33	0	0,04
U_6	1,49	0,91	2,53	0,15	0,04	0

- **PASSO 1.** Si uniscono le unità meno distanti: U_5 e U_6 formano il gruppo G_1

... la logica del raggruppamento gerarchico (2/8)

- **PASSO 2.** Si aggiorna la matrice delle distanze che ora è 5×5 e si uniscono le unità/gruppi meno distanti.
- La distanza minima si ha per le unità U_1 e U_3 che vanno a formare il gruppo G_2

	G_1	U_1	U_2	U_3	U_4
G_1	0				
U_1	1.56	0			
U_2	0.91	3.00	0		
U_3	2.78	0.30	5.02	0	
U_4	0.33	1.08	1.76	1.68	0



- **Attenzione:** queste sono distanze fra unità e **gruppo**.

... la logica del raggruppamento gerarchico (3/8)

- **PASSO 3.** Si aggiorna la matrice delle distanze che ora è 4×4 e si uniscono le unità/gruppi meno distanti.
- La distanza minima si ha per l'unità U_4 e il gruppo G_1 . L'unità U_4 si unisce al gruppo G_1 e forma il gruppo G_3

	G_1	G_2	U_2	U_4
G_1	0			
G_2	2,78	0		
U_2	0,91	5,02	0	
U_4	0,33	1,68	1,76	0

- **Attenzione:** in questa matrice ci sono **distanze fra gruppi** e **distanze fra unità e gruppi**.

... la logica del raggruppamento gerarchico (4/8)

- **PASSO 4.** Si aggiorna la matrice delle distanze che ora è 3×3 e si uniscono le unità/gruppi meno distanti.
- L'unità U_2 si unisce al gruppo G_3 e si forma il gruppo G_4

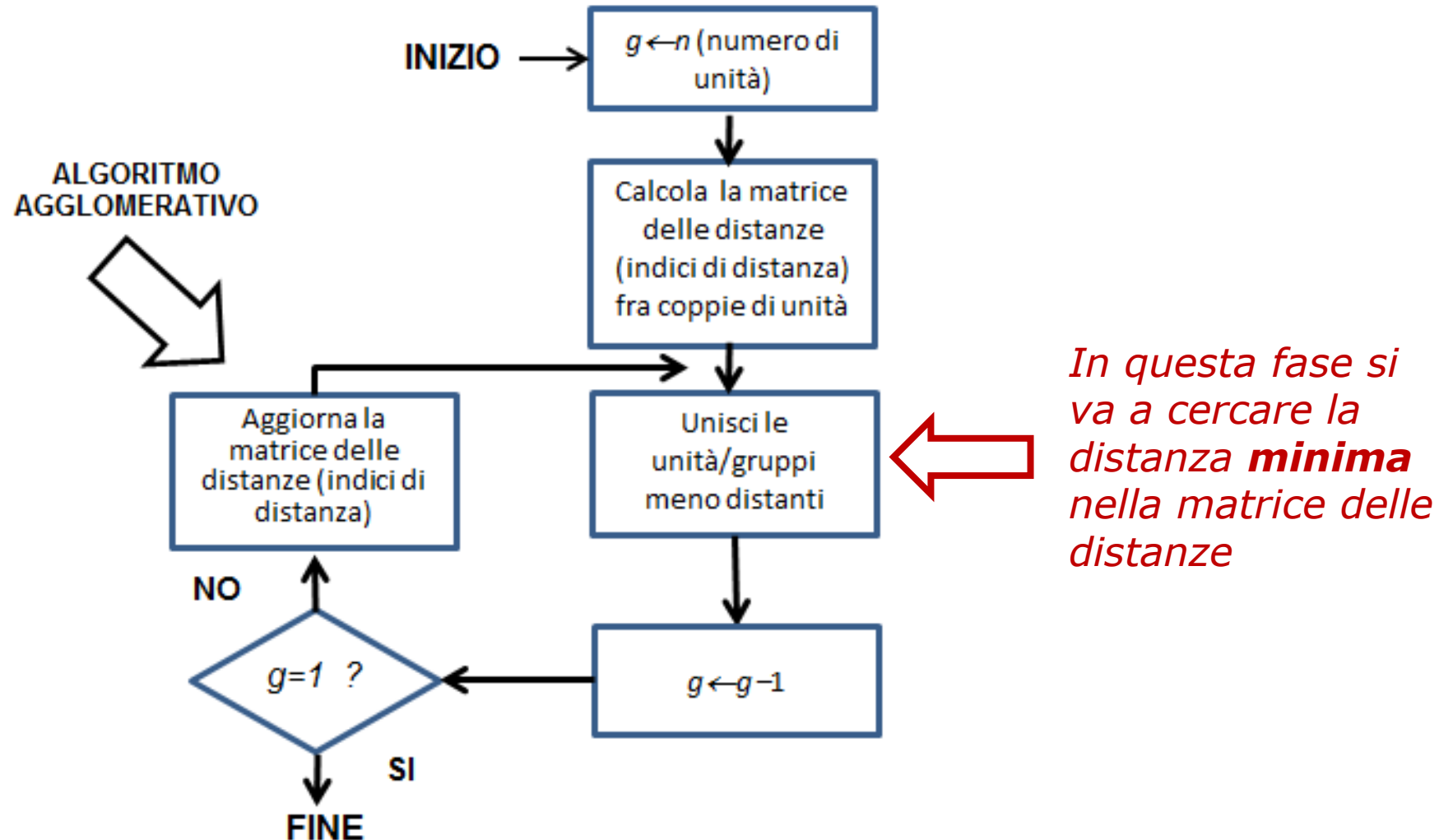
	U_2	G_3	G_2
U_2	0		
G_3	1,76	0	
G_2	5,02	2,78	0

... la logica del raggruppamento gerarchico (5/8)

- **PASSO 5.** Si aggiorna la matrice delle distanze che ora è 2×2 e si uniscono i due gruppi rimasti ottenendo un unico gruppo che contiene tutte le unità.

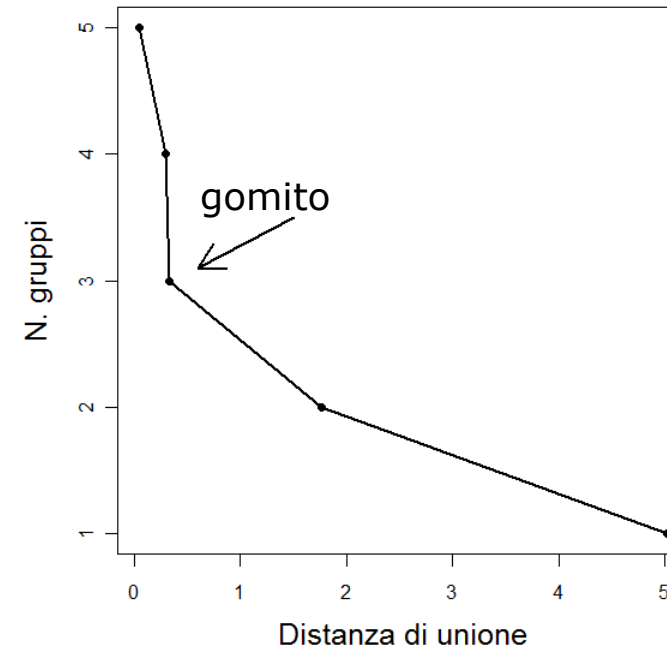
	G_4	G_2
G_4	0	
G_2	5,02	0

... la logica del raggruppamento gerarchico (6/8)



... la logica del raggruppamento gerarchico. Riepilogo dell'aggregazione: lo screeplot (7/8)

Unità/gruppi formati	N. gruppi	Distanza di unione
$U_1, U_2, U_3, U_4, U_5, U_6$	6	0,00
$G_1:\{U_5, U_6\}, U_1, U_2, U_3, U_4$	5	0,04
$G_1:\{U_5, U_6\}, G_2:\{U_1, U_3\}, U_2, U_4$	4	0,30
$G_3:\{G_1, U_4\}, G_2, U_2$	3	0,33
$G_2, G_4:\{G_3, U_2\}$	2	1,76
$G_5:\{G_2, G_4\}$	1	5,02



Lo scree plot Suggestisce il numero dei gruppi in corrispondenza del **gomito**.
In questo caso: **3** gruppi.

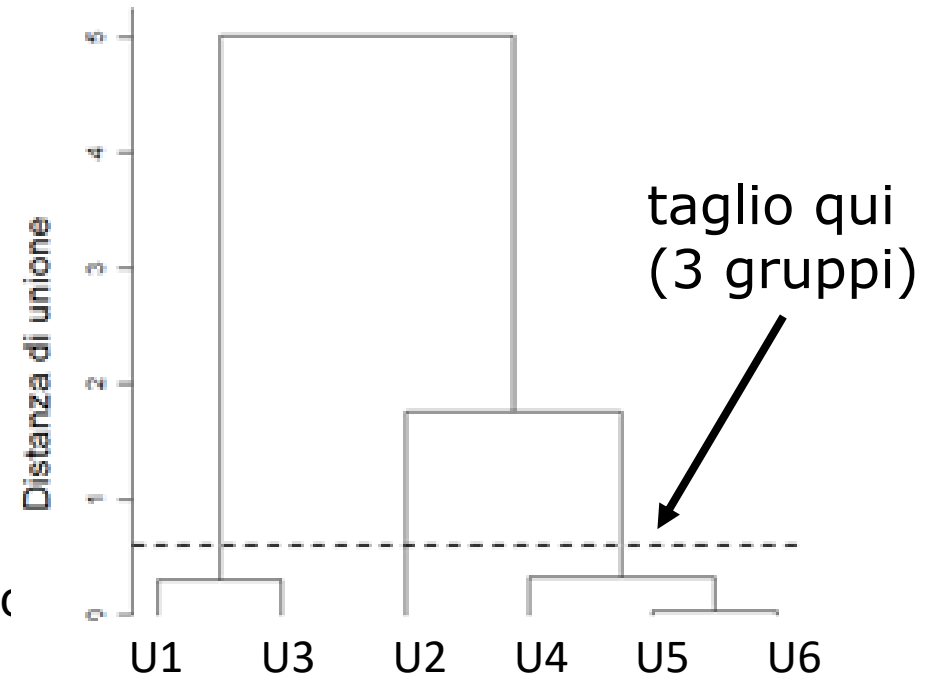
N.B. Più aggrego i gruppi e maggiore è la distanza di unione (cioè vado ad unire gruppi sempre più eterogenei fra loro e quindi formo gruppi sempre più eterogenei al loro interno)

... la logica del raggruppamento gerarchico. Dendrogramma o diagramma ad albero (8/8)

Unità/gruppi formati	N. gruppi	Distanza di unione
$U_1, U_2, U_3, U_4, U_5, U_6$	6	0,00
$G_1: \{U_5, U_6\}, U_1, U_2, U_3, U_4$	5	0,04
$G_1: \{U_5, U_6\}, G_2: \{U_1, U_3\}, U_2, U_4$	4	0,30
$G_3: \{G_1, U_4\}, G_2, U_2$	3	0,33
$G_2, G_4: \{G_3, U_2\}$	2	1,76
$G_5: \{G_2, G_4\}$	1	5,02

La cluster gerarchica non ci fornisce direttamente la composizione dei gruppi o il numero dei gruppi (cioè la **partizione**).

Si deve decidere dove fare il **pruning** (potatura)



N.B. Più taglio in alto (minore numero di gruppi) e più eterogenei al loro interno sono i gruppi che vado a formare

Metodi gerarchici: la prossimità fra coppie di unità

Per raggruppare le unità *meno diverse* è necessario stabilire una misura di **prossimità** (o distanza ecc.) fra unità sulla base delle p variabili

Se le p variabili sono **quantitative** si parla di **distanze** o di **indici di distanza** fra unità altrimenti si parla di **similarità** o **dissimilarità**

La **prossimità** (o distanza, similarità, dissimilarità) fra le unità U_i e U_j è:

$$d_{ij} = d(U_i, U_j) = d(\underline{x}_i, \underline{x}_j)$$

dove \underline{x}_i e \underline{x}_j sono le righe i e j della tradizionale matrice dei dati unità \times variabili.

Metodi gerarchici: le prossimità fra coppie di cluster

Man mano che l'agglomerazione procede, si uniscono non solo singole unità fra loro ma unità a cluster, e anche cluster fra loro.

Occorre definire che cosa si intende per prossimità fra cluster.

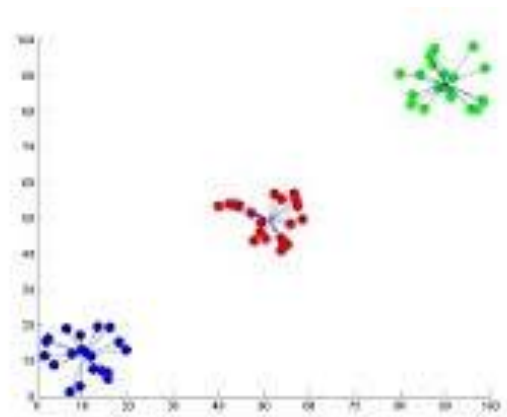
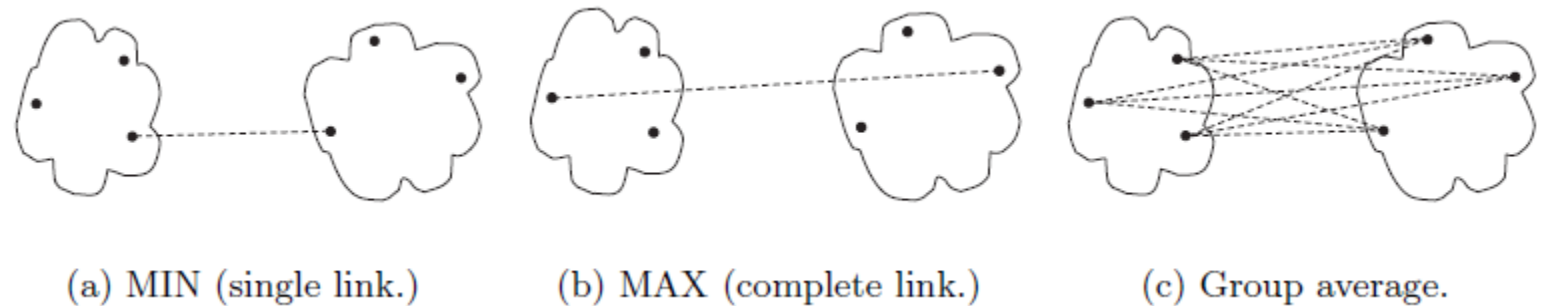
Vediamo 4 approcci

a) Single link

b) Complete link

c) Group average

d) Prototype based (prossimità rispetto a un punto «centrale»)



Metodi gerarchici: le prossimità fra cluster. Usiamo per il momento il concetto di distanza

- a) Single link: la distanza fra cluster è la **distanza minima** fra due unità una del primo cluster e una dell'altro
- b) Complete link: la distanza fra cluster è la **distanza minima** fra due unità una del primo cluster e una dell'altro
- c) Group average: la distanza fra cluster è la **media delle distanze** fra le unità del primo cluster e quelle dell'altro
- d) Ward's method: ogni cluster è rappresentato da un centroide (punto con coordinate le medie di gruppo) e la distanza fra due cluster è la **distanza fra i rispettivi centroidi** (prototipi)

Quale metodo gerarchico ?

- Non c'è un criterio generale
- I metodi gerarchici si differenziano per come si calcola la distanza fra cluster
- I metodi **single link**, **complete link**, **group average**, sono basati sulle distanze fra unità, in ogni stadio dell'aggregazione
- I metodi come il **WARD**, introducono un oggetto nuovo, il centroide, e calcolano le distanze fra questi.

I metodi, **single link** e complete link, sono i più *puri*, perché basano comunque la distanza fra cluster su distanze fra unità (le più vicine per il **single link**, le più lontane per il **complete link**)

Distanze e indici di distanza: proprietà

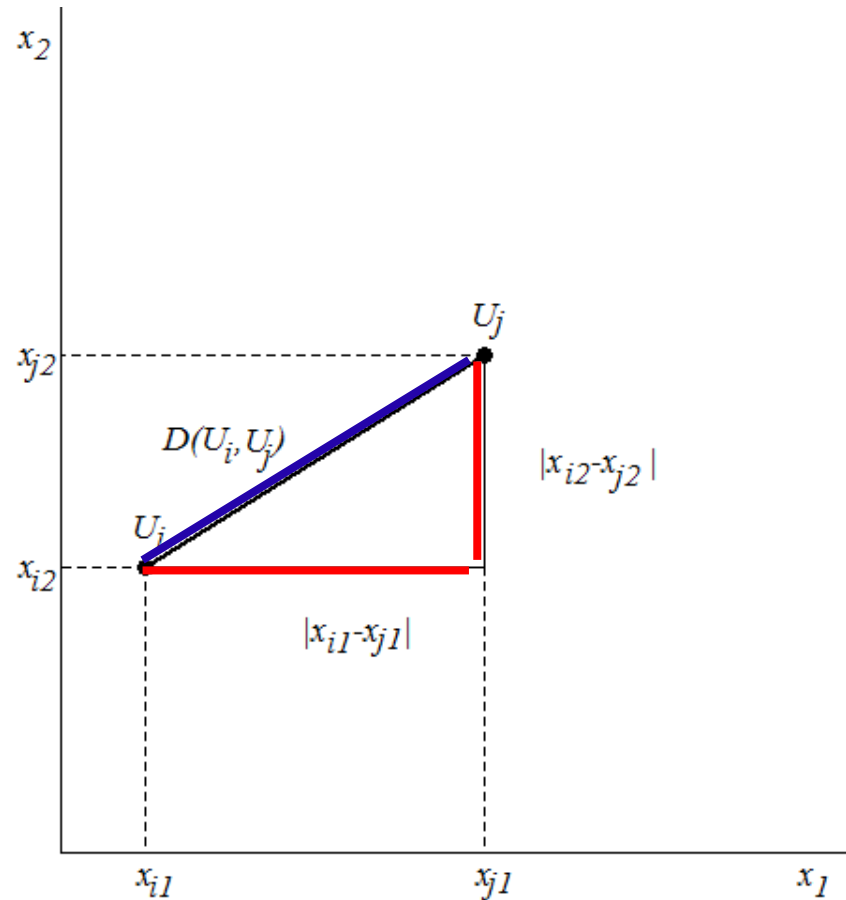
$$d_{ij} = d(U_i, U_j) = d(\underline{x}_i, \underline{x}_j)$$

1. Identità: $d_{ij} = 0$ se e solo se $\underline{x}_i = \underline{x}_j$
 2. Non negatività: $d_{ij} \geq 0$
 3. Simmetria: $d_{ij} = d_{ji}$
 4. Triangolarità: $d_{ij} \leq d_{ik} + d_{kj}$
- Indice
di
distanza
- Distanza

	U_1	U_2	...	U_j	...	U_n
U_1	0	d_{12}	...	d_{1j}	...	d_{1n}
U_2		0	...	d_{2j}	...	
...
U_i			...	d_{ij}	...	d_{in}
...				
U_n						0

La matrice delle distanze (o degli indici di distanza) fra coppie di unità è **simmetrica**

Distanza euclidea e distanza Manhattan (city block): rappresentazione grafica con $p=2$



Euclidea

$$D(U_i, U_j) = D(\underline{x}_i, \underline{x}_j) = \sqrt{\sum_{k=1}^p (x_{ik} - x_{jk})^2}$$

Manhattan

$$D_M(U_i, U_j) = D_M(\underline{x}_i, \underline{x}_j) = \sum_{k=1}^p |x_{ik} - x_{jk}|$$

Il quadrato della distanza euclidea è una distanza o è un indice di distanza ?

Distanze in R (funzione `dist()`)

“euclidean”

“maximum” massima distanza fra le componenti dei vettori delle due unità

“manhattan”

“canberra”: $\sum_{k=1}^p \frac{|x_{ik} - x_{jk}|}{|x_{ik}| + |x_{jk}|}$; questo metodo è riferito a valori non negativi come ad es. conteggi (R usa il valore assoluto al denominatore per evitare distanze negative se la formula è usata con valori negativi).

“binary”: il metodo è riferito a record binary 0/1. E’ la proporzione di “solo 1” sul numero di “almeno 1”.

“minkowski”: $(\sum_{k=1}^p (x_{ik} - x_{jk})^p)^{\frac{1}{p}}$ (con $p=2$ è la distanza euclidea)

distanza binary e canberra

```
x <- c(0, 1, 1, 1, 1)
```

```
y <- c(1, 1, 1, 0, 1)
```

```
dist(rbind(x, y), method = "binary")
```

```
dist(rbind(x, y), method = "canberra")
```

```
dist(rbind(x, y), method = "minkowski", p=2)
```

```
dist(rbind(x, y)) # euclidea è il default con dati numerici
```

Unità di misura e scala delle variabili

- le funzioni di distanza non hanno senso se applicate a variabili espresse in diversa unità di misura (es. Kg e metri)
- attenzione a quando le variabili hanno scala diversa pur essendo misurate con la stessa unità di misura (es. lunghezza del naso e altezza della persona, espresse in mm.) → diverso peso delle variabili

E' necessario procedere ad una normalizzazione (standardizzazione)

La scala delle variabili

- Su due individui abbiamo misurato lunghezza del naso (in cm.) e altezza (in cm.)

- $\underline{x}_i:(3.7,160)$ $x_j:(4.1,180)$

$$D^2(\underline{x}_i, \underline{x}_j) = (3.7 - 4.1)^2 + (160 - 180)^2 = 0.16 + 400$$

La distanza è
determinata
dall'altezza

- **Possibile soluzione:** normalizzazione/standardizzazione dei valori.
- **ATTENZIONE:** si possono generare inversioni nelle distanze fra unità

Metodi di normalizzazione delle variabili

- Standardizzazione tradizionale: $y = \frac{x - \text{media}(x)}{\text{deviazione standard}(x)}$
- Ampiezza massima unitaria: $y = \frac{x}{\text{MAX}(x)}$
- Deviazione standard unitaria: $y = \frac{x}{\text{deviazione standard}(x)}$
- Intervallo $[-1, 1]$: $y = \frac{x}{\text{MAX}(x) - \min(x)}$
- Intervallo $[0, 1]$: $y = \frac{x - \min(x)}{\text{MAX}(x) - \min(x)}$

La normalizzazione non basta: variabili correlate e misure di distanza

- Se sono presenti variabili **molto correlate**, gli aspetti descritti da tali variabili saranno *sovrarappresentati* nel calcolo della distanza e avranno quindi maggior importanza nell'analisi cluster.
- La correlazione ha anche conseguenze sulla **forma dei gruppi** (v. oltre) che diventano difficili da individuare.
- **Possibile soluzione:** applicazione dell'ACP (ricorda di valutare se dobbiamo standardizzare le variabili) e impiego delle CP standardizzate nell'analisi cluster.

Altra soluzione per variabili correlate: distanza di Mahalanobis

$$(\underline{x}_i - \underline{x}_j)' \Sigma^{-1} (\underline{x}_i - \underline{x}_j)$$

Σ è la matrice di varianza e covarianza delle p variabili

\underline{x}_i È il vettore (colonna) delle variabili misurate sull'unità i -esima

```
library(vegan)
```

```
dista<-vegdist(mtcars,method="mahalanobis")
```

Metodi gerarchici: esempio in R

```
data(mtcars)
# scelgo le variabili (prendo le quantitative)
x<-mtcars[,1:7] # attenzione: diversa unità di misura
cor(x)         # attenzione: variabili correlate
## calcolo le CP sulle variabili standardizzate
z<-prcomp(x,center = TRUE, scale. = TRUE)
dati<-z$x      ## prendo le CP che sono ortogonali
## scelgo la misura di prossimità
dista<-(dist(dati))^2    ## quadrato distanza euclidea
## scelgo l'algoritmo: media delle distanze
risultati<-hclust(dista,method='average')
```

Metodi gerarchici: esempio in R (continua)

```
risultati
summary(risultati)
plot(risultati)
### scelgo k=4 gruppi
gruppi<-cutree(risultati,k=4)
table(gruppi)    ## dimensione dei 4 gruppi
## analizzo le variabili ordinarie rispetto ai gruppi
## media di gruppo della variabile mpg
tapply(x[,1],gruppi,mean)
boxplot(x[,1]~gruppi)
```

Variabili categoriche: dissimilarità

Una misura di dissimilarità verifica le seguenti proprietà:

1. Identità: $d_{ij} = 0$ se e solo se $\underline{x}_i = \underline{x}_j$
2. Non negatività: $d_{ij} \geq 0$
3. Simmetria: $d_{ij} = d_{ji}$
4. Triangolarità: $d_{ij} \leq d_{ik} + d_{kj}$

Dissimilarità di Jaccard

$$Jaccard = \frac{n_{01} + n_{10}}{n_{01} + n_{10} + n_{11}}$$

```
y <- c(1, 1, 1, 0, 1, 0)
x <- c(0, 1, 1, 1, 1, 0)
library(vegan)
vegdist(rbind(x,y),method="jaccard")
```

x	y	
	0	1
0	1	1
1	1	3

$$Jaccard = \frac{1 + 1}{1 + 1 + 3} = \frac{2}{5} = 0.4$$

Calcolo dissimilarità Jaccard in R

```
library(vegan)
library(ade4)
dati<-data.frame(pranzo = c("casa", "casa", "bar", "bar"),
col = c("rosso","blue","verde","rosso") )
rownames(dati)<-c("Anna","Dino","Pia","Aldo")
binaria<-acm.disjonctif(dati)          # dataset binario
vegdist(binaria,metric='jaccard')
```

	Anna	Dino	Pia
Dino	0.5		
Pia	1.0	1.0	
Aldo	0.5	1.0	0.5

```
library(cluster)
daisy(dati)  ## riconosce i dati politomici
```

	Anna	Dino	Pia
Dino	0.5		
Pia	1.0	1.0	
Aldo	0.5	1.0	0.5

Variabili miste: distanza di Gower (1971)

Opera anche con variabili miste (numeriche e politomiche)

1) Si normalizzano le variabili numeriche come segue:

$$x^* = \frac{x - \min(x)}{\max(x) - \min(x)}$$

2) Si calcola la **media** della distanza Manhattan per le **w** variabili x^*

3) Si calcola un indice di dissimilarità per le variabili categoriche. **R** calcola l'indice tipo Jaccard per le **h** variabili politomiche ma si considerano gli (0,0) invece che gli (1,1) e cioè: il contributo di una variabile 0/1 è 0 se la coppia di valori è (1,1), è 1 altrimenti.

4) Si calcola la media ponderata dei due tipi di distanza, con pesi **w** e **h**.

Calcolo indice di Gower: funzione `daisy {cluster}`

La funzione **`daisy()`** :

- non richiede trasformazioni binarie
- riconosce automaticamente quali variabili sono numeriche e quali sono qualitative (*factor*)

Calcolo distanza di Gower con variabili miste

	pranzo	col	eta	eta*
Anna	casa	rosso	20	0
Dino	casa	blue	20	0
Pia	bar	verde	24	1
Aldo	bar	rosso	21	0.25

$\text{diss}(\text{Anna}, \text{Dino}) = (0.5 \times 2 + 0 \times 1) / 3 = 1 / 3 = 0.3333$
 $\text{diss}(\text{Anna}, \text{Pia}) = (1 \times 2 + 1 \times 1) / 3 = 3 / 3 = 1$
 $\text{diss}(\text{Anna}, \text{Aldo}) = (0.5 \times 2 + 0.25 \times 1) / 3 = 1.25 / 3 = 0.41667$
 $\text{diss}(\text{Dino}, \text{Pia}) = (1 \times 2 + 1 \times 1) / 3 = 3 / 3 = 1$
 $\text{diss}(\text{Dino}, \text{Aldo}) = (1 \times 2 + 0.25 \times 1) / 3 = 2.25 / 3 = 0.75$
 $\text{diss}(\text{Pia}, \text{Aldo}) = (0.5 \times 2 + 0.75 \times 1) / 3 = 1.75 / 3 = 0.58333$

Variabile «eta» normalizzata

```

library(cluster)
dati<-data.frame(pranzo = c("casa", "casa", "bar", "bar"),
col = c("rosso", "blue", "verde", "rosso"), eta=c(20,20,24,21) )
dista<-daisy(dati,metric="gower")    # calcolo la distanza di Gower

```

Le fasi di un'analisi cluster gerarchica

- **1.** scelta delle unità;
- **2.** scelta delle variabili sulle quali basare il raggruppamento, considerando anche eventuali trasformazioni dei dati;
- **3.** scelta della misura di prossimità;
- **4.** scelta del criterio (o algoritmo) gerarchico e applicazione;
- **5.** validazione dei gruppi risultanti → la vediamo più avanti
- **6.** interpretazione dei gruppi → la vediamo più avanti