

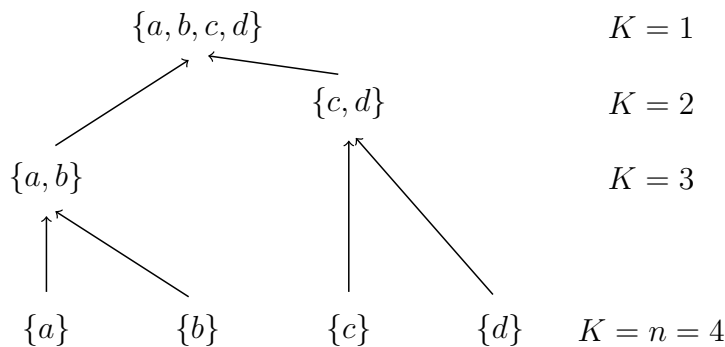
## Lezione : Metodi gerarchici

Docente: Aldo Solari

### 1 Metodi gerarchici

Nei *metodi gerarchici* si individua una sequenza di partizioni nidificate: la partizione in  $K + 1$  gruppi si ottiene dalla partizione in  $K$  gruppi facendo di due degli elementi di questa un elemento di quella (AGNES), o viceversa (DIANA)

- Algoritmo Scissorio (DIANA, DIvisive ANAlysis)
- Algoritmo Agglomerativo (AGNES, AGGlomerative NESting)

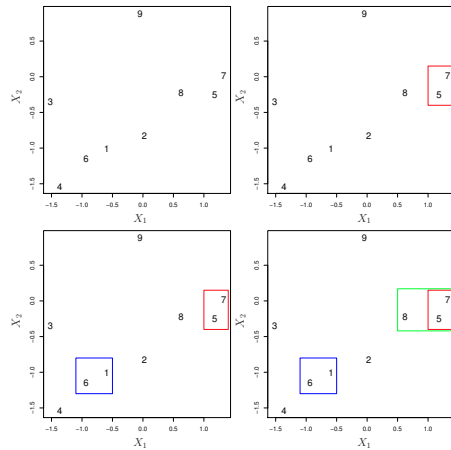


#### 1.1 Algoritmo agglomerativo

1. Si parte dalla partizione in  $n$  gruppi, ciascuno singoletto; Inizializzare  $k = n$
2. Determinare quale coppia di gruppi sia quella 'migliore' da unire, tra le  $\binom{k}{2} = \frac{k(k-1)}{2}$  coppie di gruppi possibili;
3. Fondere la 'migliore' coppia di gruppi in un unico gruppo; impostare  $k = k - 1$  e andare al passo ② se  $k > 1$ , altrimenti STOP

Per questo algoritmo sono previste  $n - 1$  iterazioni di ② e ③ prima dell'arresto

**Example 1.1.**  $n = 9$  unità statistiche



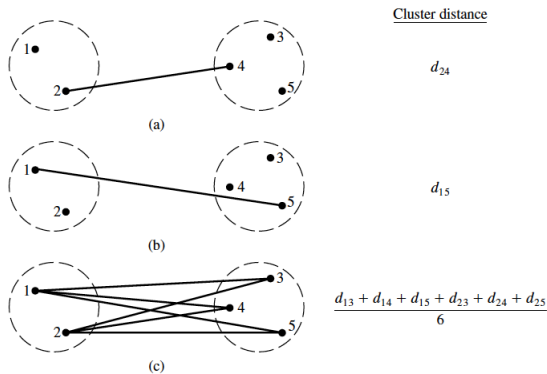
- Partizione*
- 
- 1, 2, 3, 4, 5, 6, 7, 8, 9
  - (5,7), 1, 2, 3, 4, 6, 8, 9
  - (5,7), (1,6), 2, 3, 4, 8, 9
  - (5,7,8), (1,6), 2, 3, 4, 9
  - ⋮
  - (1,2,3,4,5,6,7,8,9)

**1.2 Distanza/dissimilarità tra gruppi**

Dobbiamo precisare come si determina al passo ② la ‘migliore’ coppia di gruppi da fondere in un unico gruppo. Se abbiamo  $k$  gruppi con matrice delle distanze/dissimilarità  $D_{k \times k}$ , basta determinare quale sia la coppia di gruppi con minore distanza/dissimilarità (se più di una coppia, si sceglie una)

1. Inizializzare  $k = n$  e  $D_{k \times k} = D_{n \times n}$ ;
2. Determinare in  $D_{k \times k}$  quale coppia di gruppi ha distanza minima
3. Fondere la coppia di gruppi con distanza minima in un unico gruppo; impostare  $k = k - 1$  e aggiornare  $D_{k \times k}$  calcolando la distanza del nuovo gruppo con i rimanenti; andare al passo ② se  $k > 1$ , altrimenti STOP

Distanza tra due gruppi  $G_I$  e  $G_L$ :



- Legame singolo (*single linkage*)

$$d(G_I, G_L) = \min\{d(u_i, u_l), u_i \in G_I, u_l \in G_L\}$$

- Legame completo (*complete linkage*)

$$d(G_I, G_L) = \max\{d(u_i, u_l), u_i \in G_I, u_l \in G_L\}$$

- Legame medio (*average linkage*)

$$d(G_I, G_L) = \frac{1}{n_{G_I} n_{G_L}} \sum_{u_i \in G_I} \sum_{u_l \in G_L} d(u_i, u_l)$$

dove  $n_{G_I}$  e  $n_{G_L}$  sono le numerosità dei gruppi  $G_I$  e  $G_L$

**Example 1.2.** Distanza tra gruppi: legame singolo. Passo ①: Inizializzare  $k = n$  e  $D = \underset{k \times k}{D} = \underset{n \times n}{D}$

$$D = \{d_{IL}\}_{5 \times 5} = \begin{array}{c|ccccc} I \setminus L & 1 & 2 & 3 & 4 & 5 \\ \hline 1 & 0 & & & & \\ 2 & 9 & 0 & & & \\ 3 & 3 & 7 & 0 & & \\ 4 & 6 & 5 & 9 & 0 & \\ 5 & 11 & 10 & 2 & 8 & 0 \end{array}$$

ITERAZIONE 1

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

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

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

- $d_{(35)1} = \min\{d_{31}, d_{51}\} = \min\{3, 11\} = 3$

- $d_{(35)2} = \min\{d_{32}, d_{52}\} = \min\{7, 10\} = 7$

- $d_{(35)4} = \min\{d_{34}, d_{54}\} = \min\{9, 8\} = 8$

dove con il legame singolo  $d_{(IL)J} = \min\{d_{IJ}, d_{LJ}\}$

$$D = \{d_{IL}\}_{4 \times 4} = \begin{array}{c|cccc} I \setminus L & (35) & 1 & 2 & 4 \\ \hline (35) & 0 & & & \\ 1 & 3 & 0 & & \\ 2 & 7 & 9 & 0 & \\ 4 & 8 & 6 & 5 & 0 \end{array}$$

ITERAZIONE 2

②  $\min_{I \neq L}(d_{IL}) = d_{1(35)} = 3$

- I due cluster 1 e (35) vengono fusi nel cluster (135)

③ Aggiorno le distanze tra il nuovo cluster (135) e i rimanenti

- $d_{(135)2} = \min\{d_{(35)2}, d_{12}\} = \min\{7, 9\} = 7$

- $d_{(135)4} = \min\{d_{(35)2}, d_{14}\} = \min\{8, 6\} = 6$

$$D = \{d_{IL}\}_{3 \times 3} = \begin{array}{c|cc} I \setminus L & (135) & 2 & 4 \\ \hline (135) & 0 & & \\ 2 & 7 & 0 & \\ 4 & 6 & 5 & 0 \end{array}$$

ITERAZIONE 3

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

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

③ Aggiorno le distanze tra il nuovo cluster (24) e il rimanente

- $d_{(135)(24)} = \min\{d_{(135)2}, d_{(135)4}\} = \min\{7, 6\} = 6$

$$D = \{d_{IL}\}_{2 \times 2} = \begin{array}{c|c} I \setminus L & (135) & (24) \\ \hline (135) & 0 & \\ (24) & 6 & 0 \end{array}$$

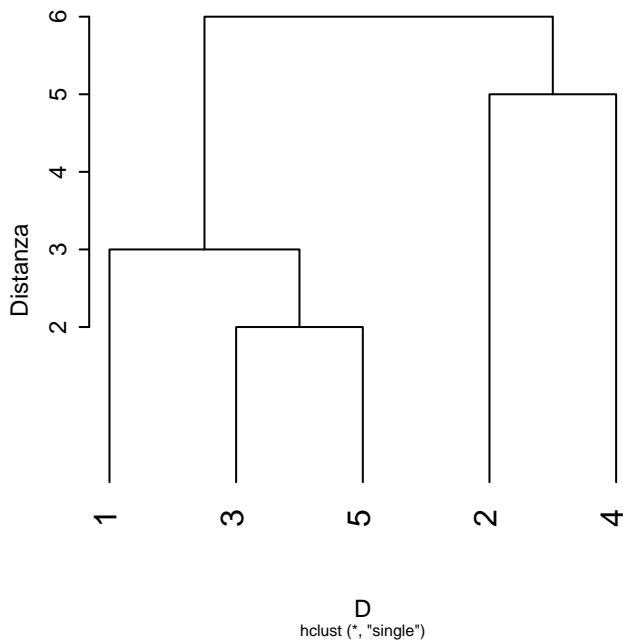
ITERAZIONE 4

②  $\min_{I \neq L}(d_{IL}) = d_{(135)(24)} = 6$

- I due cluster (135) e (24) vengono fusi nel cluster (12345)

③ STOP

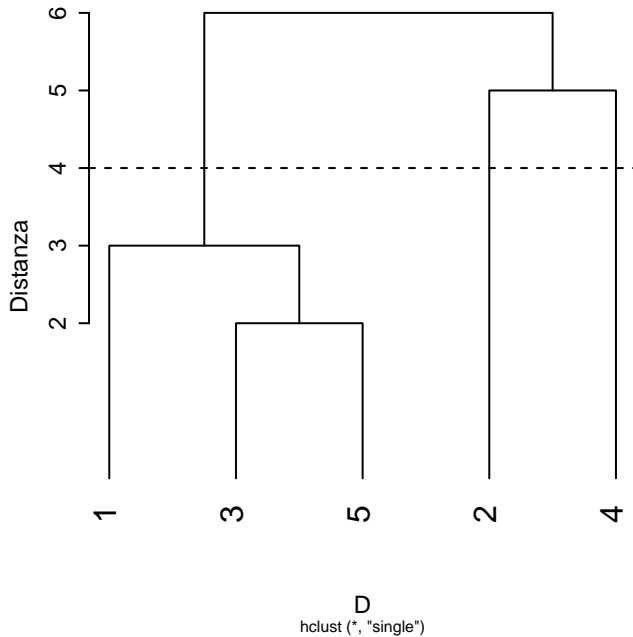
### 1.3 Il dendrogramma



- La successione di partizioni individuate può essere rappresentata con il dendrogramma
- Nell'esempio abbiamo  $n = 5$  unità statistiche, indicate con le cifre da 1 a 5
- Le unità 3 e 5 sono unite tra di loro da una linea spezzata a forma di U rovesciata, che indica che vengono messe nello stesso gruppo, e si ottiene la partizione  $\{(3, 5), 1, 2, 4\}$
- Procedendo verso l'alto, la successiva unione tra gruppi è tra 1 e (3, 5), quindi al livello successivo si ottiene la partizione  $\{(1, 3, 5), 2, 4\}$ .
- Andando su ancora di un livello, vengono uniti i gruppi 2 e 4, formando la partizione  $\{(1, 3, 5), (2, 4)\}$ .
- Procedendo ulteriormente si arriva alla partizione formata da un unico elemento  $\{(1, 2, 3, 4, 5)\}$ .
- Si noti che le unità sono rappresentate in un ordine scelto in modo che i rami dell'albero non si incrocino nel disegno (ovviamente non c'è un unico ordine siffatto)
- Le altezze a cui sono disegnati i segmenti che uniscono le unità viene disegnato all'altezza corrispondente alla distanza tra essi
  - 3 e 5 hanno distanza 2
  - (3,5) e 1 hanno distanza 3
  - 2 e 4 hanno distanza 5

- (1,3,5) e (2,4) hanno distanza 6

Fissata una distanza  $c > 0$ , disegnando una linea orizzontale ad altezza  $c$  si taglia il dendrogramma e si ottiene il numero di gruppi, corrispondente al numero di aste intersecate dalla linea orizzontale. Nell'esempio, per  $c = 4$  (linea tratteggiata), risultano formati i tre gruppi (1, 3, 5), 2 e 4.



**Example 1.3.** *Distanza tra gruppi: legame singolo. Passo ①: Inizializzare  $k = n$  e  $D = D_{k \times k} = D_{n \times n}$*

$$D_{5 \times 5} = \{d_{IL}\} = \begin{array}{c|ccccc} I \backslash L & 1 & 2 & 3 & 4 & 5 \\ \hline 1 & 0 & & & & \\ 2 & 9 & 0 & & & \\ 3 & 3 & 7 & 0 & & \\ 4 & 6 & 5 & 9 & 0 & \\ 5 & 11 & 10 & 2 & 8 & 0 \end{array}$$

*ITERAZIONE 1*

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

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

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

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

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

$$D = \{d_{IL}\}_{4 \times 4} = \begin{array}{c|ccc} I \setminus L & (35) & 1 & 2 & 4 \\ \hline (35) & 0 & & & \\ 1 & 11 & 0 & & \\ 2 & 10 & 9 & 0 & \\ 4 & 9 & 6 & 5 & 0 \end{array}$$

ITERAZIONE 2

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

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

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

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

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

$$D = \{d_{IL}\}_{3 \times 3} = \begin{array}{c|cc} I \setminus L & (35) & (24) & 1 \\ \hline (35) & 0 & & \\ (24) & 10 & 0 & \\ 1 & 11 & 9 & 0 \end{array}$$

ITERAZIONE 3

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

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

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

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

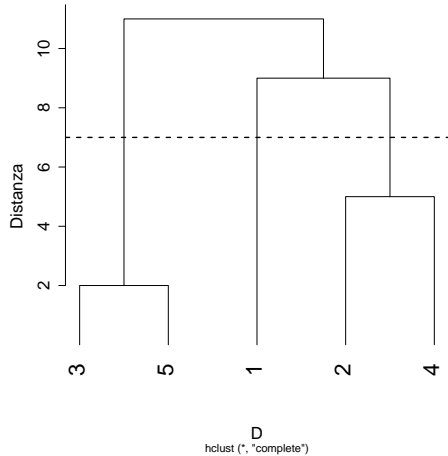
$$D = \{d_{IL}\}_{2 \times 2} = \begin{array}{c|cc} I \setminus L & (35) & (124) \\ \hline (35) & 0 & \\ (124) & 11 & 0 \end{array}$$

ITERAZIONE 4

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

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

③ STOP



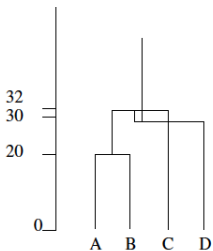
## 1.4 Proprietà

### 1.4.1 Interpretazione del taglio

In termini di distanza/dissimilarità tra unità statistiche, tagliare il dendrogramma ad altezza  $c > 0$

- Interpretazione del taglio per il legame singolo: per ogni  $u_i$  in un cluster (non singoletto), c'è almeno un'altra unità  $u_l$  tale per cui  $d(u_i, u_l) < c$
- Interpretazione del taglio per il legame completo: per ogni  $u_i$  in un cluster (non singoletto), tutte le altre unità  $u_l$  sono tali per cui  $d(u_i, u_l) < c$
- Interpretazione del taglio per il legame medio: nessuna

### 1.4.2 Dendrogramma con inversioni

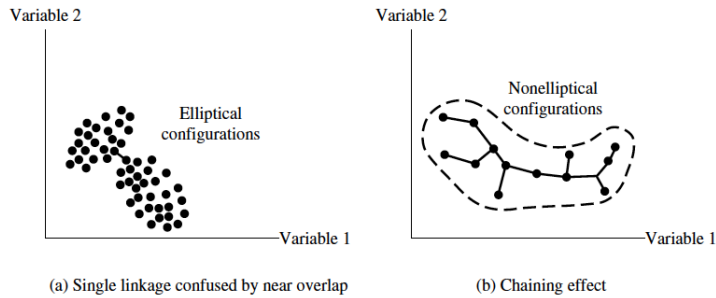


Il metodo del legame singolo, completo, medio non producono mai un dendrogramma con inversioni, ovvero la distanza/dissimilarità tra cluster non decresce mai nell'iterazione successiva dell'algoritmo

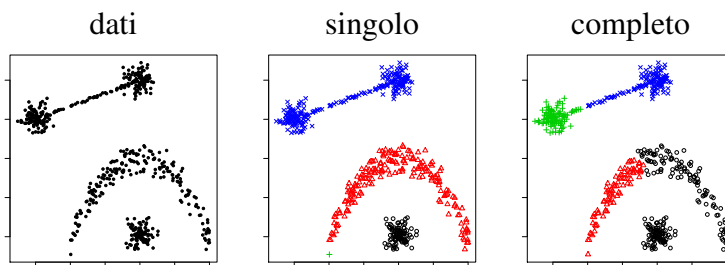
### 1.4.3 Peculiarità dei legami

- Una peculiarità del legame singolo è l'effetto catena (*chaining*). da un lato consente di collegare gruppi di forma particolare, come in Figura (b) dall'altro rischia di legare osservazioni che non appartengono a uno stesso gruppo, come in Figura (a)





- Il metodo del legame completo, d'altra parte, tende a individuare gruppi molto compatti al loro interno ma di forma circolare (ipersferica, in generale) quindi si rischia di perdere gruppi di forma irregolare.



#### 1.4.4 Invarianza rispetto a trasformazioni monotone

Si consideri una trasformazione monotona crescente  $f$

$$f(x) \leq f(y) \quad \text{se } x \leq y$$

Cosa succede se consideriamo  $f(d_{ij})$  invece di  $d_{ij}$  come elementi della matrice di distanze/dissimilarità? Ad esempio se considero  $f(d_{ij}) = d_{ij}^2$ ? I risultati con il legame medio cambiano, mentre con il legame singolo o completo non cambiano.

## 2 Metodo del legame del centroide

Distanza/dissimilarità tra due gruppi  $G_I$  e  $G_L$

$$d(G_I, G_L) = d_2(\bar{x}_I, \bar{x}_L)$$

dove

$$\bar{x}_I = \begin{bmatrix} \frac{1}{n_I} \sum_{i:u_i \in G_I} x_{i1} \\ \dots \\ \frac{1}{n_I} \sum_{i:u_i \in G_I} x_{ip} \end{bmatrix}$$

è il vettore delle medie del gruppo  $G_I$  e  $n_I$  è la numerosità del gruppo  $G_I$

- Input: la matrice di dati  $X$  (utilizzabile solo se tutte le variabili sono quantitative)

- Può produrre inversioni
- Non invariate rispetto a trasformazioni monotone

### 3 Confronto

| Legame    | Inversione | Trasformazioni monotone | Interpr. taglio | Peculiarità       |
|-----------|------------|-------------------------|-----------------|-------------------|
| Singolo   | No         | Invariante              | Si              | <i>chaining</i>   |
| Completo  | No         | Invariante              | Si              | forme sferiche    |
| Medio     | No         | Non invariante          | No              |                   |
| Centroide | Si         | Non invariante          | No              | solo quantitative |