

# Cluster Analysis

## Metodi gerarchici

---

R. Pappadà (rpappada@units.it)

Analisi dei dati per l'impresa, a.a. 24-25

# Algoritmi gerarchici

---

I metodi gerarchici rappresentano un approccio alternativo ai metodi come il  $K$ -means e non richiedono la specificazione del numero di cluster in input



I metodi gerarchici consentono di ottenere una partizione delle unità e inoltre forniscono informazioni sulla dissimilarità tra i cluster:

- piccoli gruppi sono annidati in grandi gruppi di oggetti, e questi cluster più grandi vengono via via aggregati formando gruppi sempre più numerosi
- una partizione in gruppi disgiunti viene ottenuta “tagliando” ad un certo *livello* di aggregazione

Tra le procedure gerarchiche distinguiamo

- **metodi agglomerativi**, che procedono mediante una serie di fusioni successive delle unità  $n$  in gruppi
- **metodi divisivi**, che separano successivamente gli oggetti  $n$  in raggruppamenti più fini

Descriveremo innanzitutto il clustering *bottom-up* o agglomerativo (AHC), che è il metodo gerarchico più utilizzato.

Data una matrice di distanza/dissimilarità  $n \times n$ ,  $\Delta$ ,

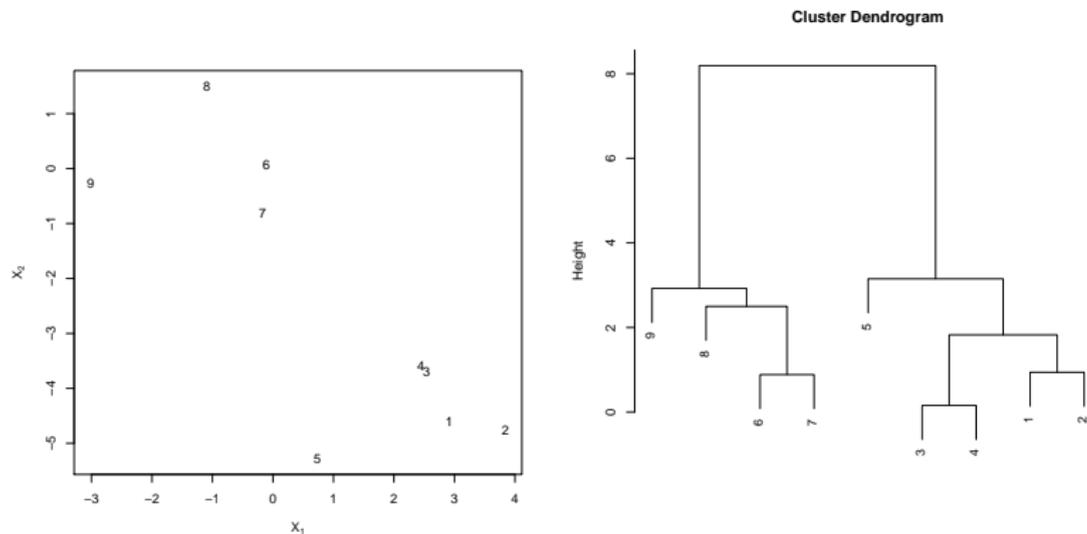
- (0) Si considerino i cluster  $C_i = \{\mathbf{x}_i\}$  per  $i \in \{1, \dots, n\}$ , formati da singole unità
- (1) si identificano i cluster con il valore più piccolo in  $\Delta$ ; si uniscono per formare un nuovo cluster e si aggiorna la matrice di dissimilarità;
- (2) Se il numero di gruppi dopo il passo (1) è uguale a 1 allora **stop**. Altrimenti si ripete (1).

## Il dendrogramma

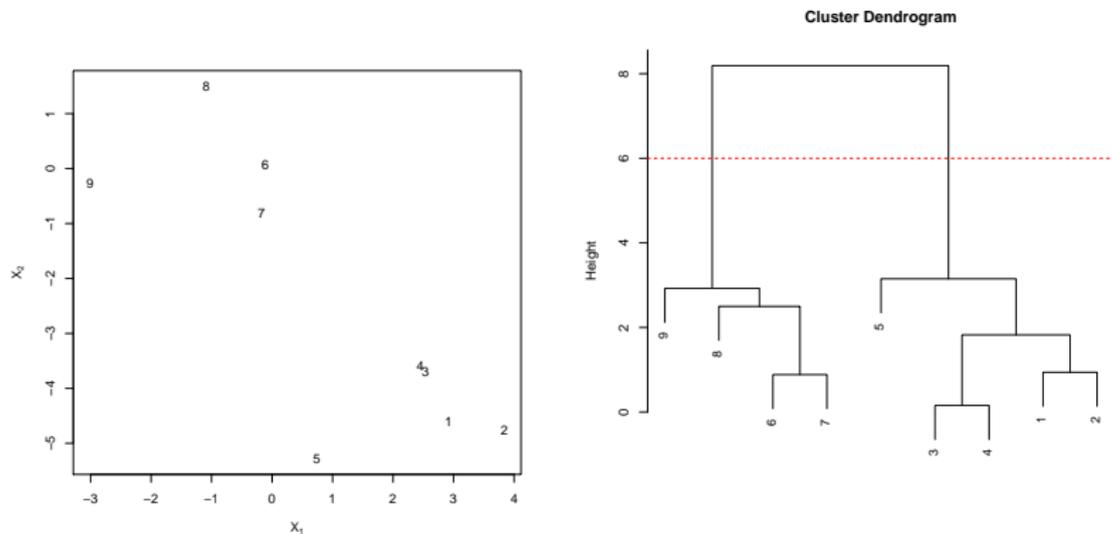
I metodi gerarchici producono una serie di partizioni annidate che sono rappresentate attraverso una struttura ad albero detta

### **dendrogramma:**

- è costruito partendo dalle foglie e unendo i gruppi fino al tronco
- partendo dal basso e procedendo verso l'alto, le foglie (unità) più simili vengono unite e formano i rami, gruppi di unità simili tra loro
- le unità che si fondono proprio alla base dell'albero sono molto simili tra loro, mentre le osservazioni che si fondono vicino alla cima dell'albero tenderanno ad essere abbastanza diverse

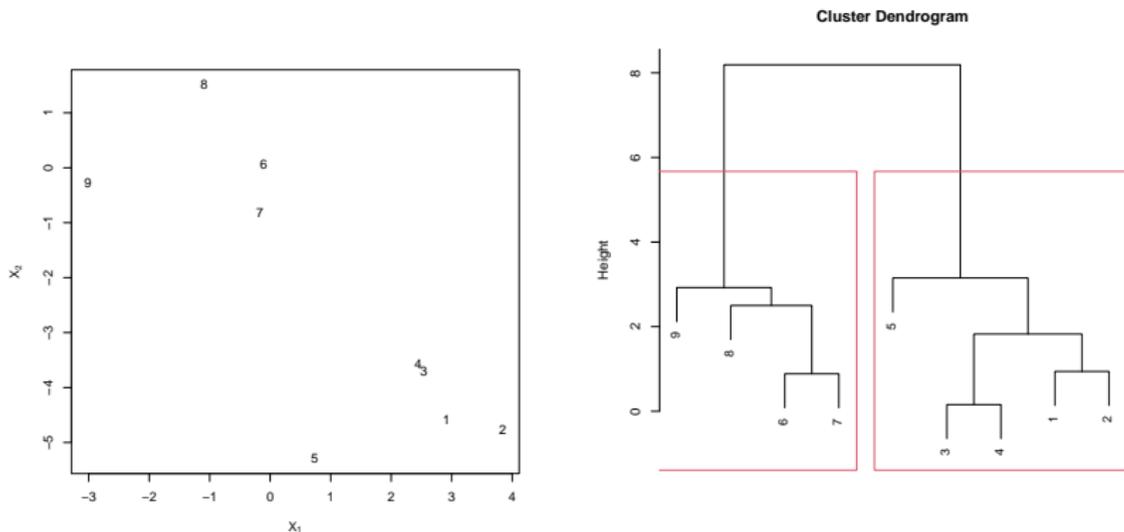


**Figura 1:** Dati (2-dim) e un possibile dendrogramma risultante da un metodo di AHC. Le unità 3 e 4 vengono raggruppate per prime, a seguire le coppie (1,2) e (6,7). L'asse  $y$  riporta l'altezza di aggregazione data dalla distanza tra i gruppi che si formano



**Figura 1:** Dati (2-dim) e un possibile dendrogramma risultante da un metodo di AHC. Le unità 3 e 4 vengono raggruppate per prime, a seguire le coppie (1,2) e (6,7). L'asse  $y$  riporta l'altezza di aggregazione data dalla distanza tra i gruppi che si formano

## Taglio del dendrogramma per ottenere 2 gruppi



La determinazione dei gruppi finali richiede il taglio dell'albero in corrispondenza di un certo livello di altezza; talvolta, è agevole individuare un opportuno taglio attraverso una analisi visiva del *salto* più evidente nel dendrogramma

## Distanza tra gruppi

Il passo (1) dell'algoritmo richiede di aggiornare la matrice di dissimilarità tenendo conto dei gruppi che si sono formati.



Occorre quindi definire una misura di dissimilarità tra  $C_l$  e  $C_m$ , per ogni  $l \neq m$ .

**Esempio** Si hanno  $n = 5$  unità e  $n(n - 1)/2$  distanze a coppie:

	1	2	3	4	5
1	0				
2	9	0			
3	3	7	0		
4	6	5	9	0	
5	11	10	2	8	0

## Distanza tra gruppi

Il passo (1) dell'algoritmo richiede di aggiornare la matrice di dissimilarità tenendo conto dei gruppi che si sono formati.



Occorre quindi definire una misura di dissimilarità tra  $C_l$  e  $C_m$ , per ogni  $l \neq m$ .

**Esempio** Si hanno  $n = 5$  unità e  $n(n - 1)/2$  distanze a coppie:

	1	2	3	4	5
1	0				
2	9	0			
3	3	7	0		
4	6	5	9	0	
5	11	10	2	8	0

$\{5\}$  e  $\{3\}$  presentano la distanza più piccola  $\rightarrow \{3,5\}$ .

## Distanza tra gruppi

Il passo (1) dell'algoritmo richiede di aggiornare la matrice di dissimilarità tenendo conto dei gruppi che si sono formati.



Occorre quindi definire una misura di dissimilarità tra  $C_l$  e  $C_m$ , per ogni  $l \neq m$ .

**Esempio** Si hanno  $n = 5$  unità e  $n(n - 1)/2$  distanze a coppie:

	1	2	3	4	5
1	0				
2	9	0			
3	3	7	0		
4	6	5	9	0	
5	11	10	2	8	0

$$d_{(3,5),1} = ? \quad d_{(3,5),2} = ? \quad d_{(3,5),4} = ?$$

## Come si misura la distanza tra i gruppi?

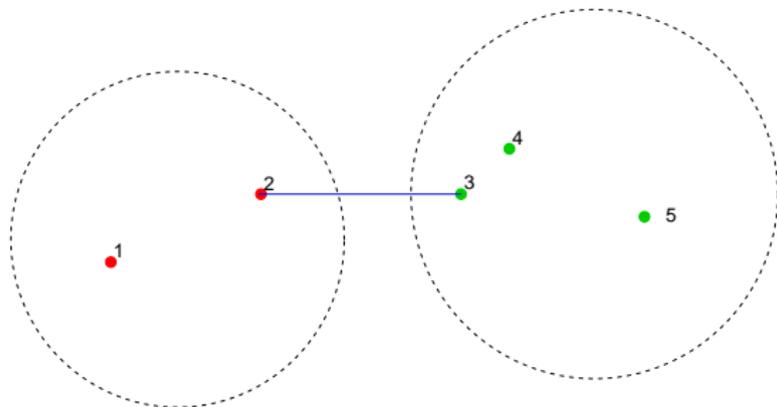
Nell'esempio precedente le unità 3 e 5 vengono unite al primo passo. Come misuriamo la distanza tra questo nuovo cluster  $\{3, 5\}$  e tutte le altre unità?



Esistono diversi modi per definire la distanza tra una singola unità e un gruppo contenente più unità, o tra due gruppi di unità. In particolare, gli algoritmi di AHC utilizzano diversi legami o *linkages*, tra cui

- legame singolo o *single linkage*
- legame completo *complete linkage*
- legame medio *average linkage*

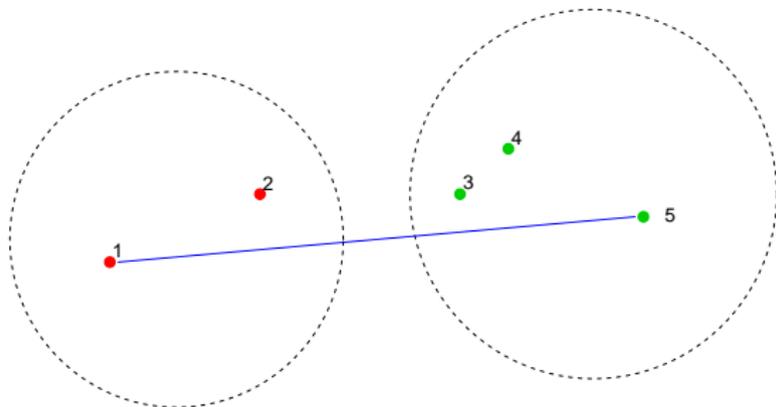
## Single linkage



dissimilarità tra i gruppi =  $\delta(\mathbf{x}_2, \mathbf{x}_3) := \delta_{2,3}$

*Metodo del vicino più vicino:* la distanza tra i gruppi corrisponde alla distanza più piccola tra tutte quelle delle coppie appartenenti ai due cluster

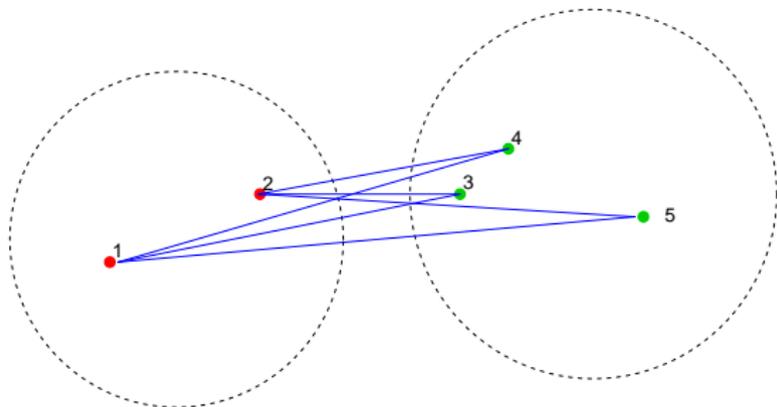
## Complete linkage



dissimilarità tra i gruppi =  $\delta(\mathbf{x}_1, \mathbf{x}_5) := \delta_{1,5}$

*Metodo del vicino più lontano:* la distanza tra i gruppi corrisponde alla distanza più grande tra tutte quelle delle coppie appartenenti ai due cluster

### Average linkage

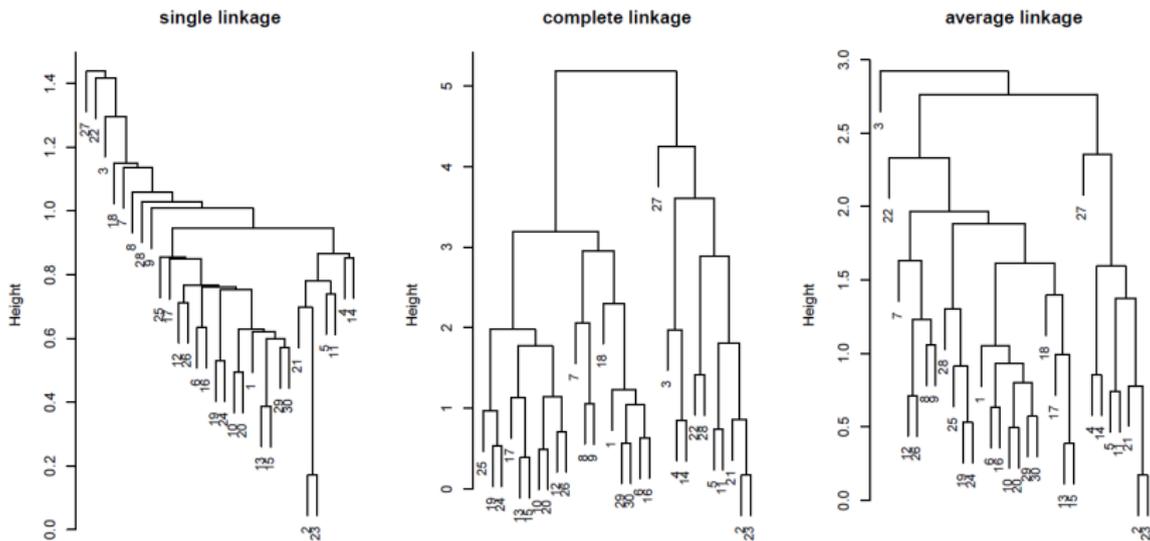


$$\text{dissimilarità tra i gruppi} = \frac{\delta_{1,3} + \delta_{1,4} + \delta_{1,5} + \delta_{2,3} + \delta_{2,4} + \delta_{2,5}}{6}$$

*Metodo della media:* la distanza tra i gruppi corrisponde alla media aritmetica di tutte le distanze tra le coppie di unità appartenenti ai due cluster

# Tipi di legame nei metodi AHC: Esempio

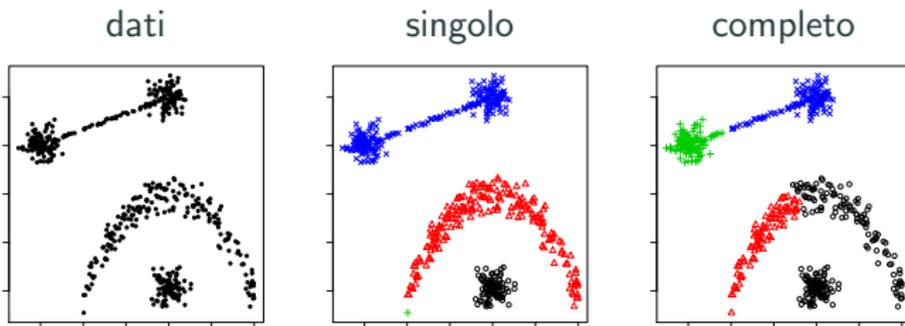
La scelta del tipo di legame può avere un forte impatto sui risultati ottenuti.



**Figura 1:** Dendrogrammi ottenuti con i diversi linkage. Average e complete linkage producono generalmente cluster più equilibrati, il single linkage determina un effetto a cascata

- Il **legame singolo** enfatizza la separazione dei cluster, vengono riconosciute le nuvole di punti allungate, ma può risultare in cluster estesi e finali in cui le singole osservazioni vengono fuse una alla volta, un problema noto come *effetto cascata*
- Il **legame completo** tende a favorire la formazione di nuovi cluster di dimensioni simili, portando a dendrogrammi di forma equilibrata anche se la struttura dei dati non lo supporta
- Il metodo del **legame medio** è intermedio tra la strategia di collegamento singolo e quella completa, tentando quindi di compensare gli svantaggi di una strategia con i vantaggi dell'altra.

## Esempio con dati simulati



Ward (1963) ha introdotto un terzo tipo di metodo, in cui l'obiettivo in ogni fase dell'algoritmo è quello di *minimizzare* una funzione obiettivo. Tale funzione è la **somma delle devianze interne ai cluster**

$$ESS_{tot,K} = ESS_1 + ESS_2 + \dots + ESS_K$$

con  $K \in \{1, \dots, n\}$  numero di cluster e

$$ESS_k = \sum_{i \in C_k} \sum_{j=1}^p (x_{ij} - \bar{x}_{kj})^2$$

dove  $\bar{x}_{kj}$  è la media del  $k$ -esimo cluster per la  $j$ -esima variabile

La tecnica di Ward si propone quindi di minimizzare la varianza delle variabili entro ciascun gruppo

- Ogni fusione genera un incremento di  $ESS_{tot,K}$ , la procedura si ferma quando tutti i gruppi sono stati fusi
- Il metodo di Ward si basa sulla distanza euclidea, funziona bene con gruppi di forma tendenzialmente sferica ed è sensibile ai valori anomali

Nota: Non esiste un metodo che risulti superiore agli altri in tutte le condizioni, molte scelte vengono fatte in base alle caratteristiche dei dati oggetto di studio

I metodi gerarchici se possono anche essere **divisivi** quando procedono per successive scissioni di gruppi

### Algoritmo

- (0) Si individua una coppia di punti nodali (punti che presentano dissimilarità massima);
- (1) Si attribuiscono le unità rimanenti ai due gruppi corrispondenti ai punti nodali, in base alla distanza minima da questi;
- (2) Si iterano i passi precedenti sino ad arrivare alla partizione in  $n$  gruppi (ciascuno composto da una unità).

I metodi divisivi

- sono più costosi in termini di tempi computazionali
- meno impiegati nelle applicazioni
- trovano applicazione soprattutto in ambito biologico