



Analisi dei
dati

A. Iodice

Clustering:
classificazione
automatica

Clustering
gerarchico

Clustering non
gerarchico

Analisi dei dati

Cluster Analysis

Alfonso Iodice D'Enza
iodicede@unina.it

Università degli studi di Cassino



A. Iodice

Analisi dei dati



Outline

Analisi dei
dati

A. Iodice

Clustering:
classificazione
automatica

Clustering
gerarchico

Clustering non
gerarchico

1 Clustering: classificazione automatica



Outline

Analisi dei
dati

A. Iodice

Clustering:
classificazione
automatica

Clustering
gerarchico

Clustering non
gerarchico

1 Clustering: classificazione automatica

2 Clustering gerarchico



Outline

Analisi dei
dati

A. Iodice

Clustering:
classificazione
automatica

Clustering
gerarchico

Clustering non
gerarchico

1 Clustering: classificazione automatica

2 Clustering gerarchico

3 Clustering non gerarchico



Clustering

Analisi dei
dati

A. Iodice

Clustering:
classificazione
automatica

Clustering
gerarchico

Clustering non
gerarchico

Le tecniche di clustering consistono in procedure automatiche per raggruppare gli oggetti a disposizione in classi composte da record omogenei.

Esempi

- Clustering dei profili di comportamento all'acquisto per identificare comportamenti di nicchia
- Raggruppare insieme i geni che presentano caratteristiche comuni



Clustering

Analisi dei
dati

A. Iodice

Clustering:
classificazione
automatica

Clustering
gerarchico

Clustering non
gerarchico

Obbiettivo delle tecniche di clustering

Le tecniche di clustering puntano a raggruppare gli le unità statistiche considerate (oggetti, records) in gruppi (**cluster**). L'obbiettivo è creare gruppi massimamente omogenei al loro interno e massimamente eterogenei tra loro.



Clustering: misure di dissimilarità

Analisi dei dati

A. Iodice

Clustering:
classificazione
automatica

Clustering
gerarchico

Clustering non
gerarchico

Dati due oggetti x e y e un indice $d()$ che ne misura della lontananza logica, possono valere le seguenti proprietà:

Caratteristiche dell'indice di misura

- **separabilità**

$$d(x, y) = 0 \Rightarrow x = y$$

- **simmetria**

$$d(x, y) = d(y, x)$$

- **disuguaglianza triangolare** si considerino tre oggetti x , y e z

$$d(x, y) \leq d(x, z) + d(z, y)$$

Dissimilarità e distanza

- **indice di dissimilarità**: indice caratterizzato da **separabilità** e **simmetria**
- **indice di distanza (metrica)**: indice caratterizzato da **separabilità** e **simmetria** per il quale risulta verificata la **disuguaglianza triangolare**.



Clustering: misure di similarità

Analisi dei
dati

A. Iodice

Clustering:
classificazione
automatica

Clustering
gerarchico

Clustering non
gerarchico

Date due osservazioni $\mathbf{x} = x_1, x_2, \dots, x_p$ e $\mathbf{y} = y_1, y_2, \dots, y_p$ descritte da p variabili quantitative

Distanze

- **distanza euclidea**

$$d(\mathbf{x}, \mathbf{y}) = \sqrt{\sum_i (x_i - y_i)^2}$$

- **distanza city-block**

$$d(\mathbf{x}, \mathbf{y}) = \sum_i |x_i - y_i|$$



Clustering: misure di similarità

Analisi dei
dati

A. Iodice

Clustering:
classificazione
automatica

Clustering
gerarchico

Clustering non
gerarchico

Distanze

- le precedenti sono casi particolari della **distanza di Minkowski**

$$d(\mathbf{x}, \mathbf{y}) = \left(\sum_i |x_i - y_i|^q \right)^{1/q}$$

- in particolare per $q = 1$ si ottiene la distanza city-block; per $q = 2$ si ottiene la distanza euclidea.

Dissimilarità e dati qualitativi

Nel caso di **variabili categoriche** il grado di similarità/dissimilarità che caratterizza le coppie di osservazioni si misura rispetto al numero di modalità comuni.

A. Iodice

Analisi dei dati



Distanze rispetto ad una soglia

Analisi dei dati

A. Iodice

Clustering:
classificazione
automatica

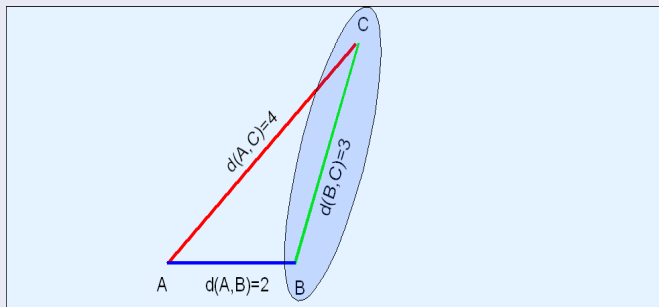
Clustering
gerarchico

Clustering non
gerarchico

Paradosso e distanza metrica

Per determinare i cluster si determina una distanza **soglia** tale che dati due punti A e B , se la distanza tra loro è tale che $d(A, B) > \text{soglia}$ allora A e B appartengono a gruppi diversi, viceversa saranno classificati nello stesso gruppo.

- **problema:** utilizzando una distanza metrica si va incontro al paradosso descritto nel seguente esempio: si considerino tre punti A , B e C , se la distanza che li separa è minore o uguale a **3** allora i punti vengono assegnati allo stesso gruppo



A. Iodice

Analisi dei dati



Distanze rispetto ad una soglia

Analisi dei dati

A. Iodice

Clustering:
classificazione
automatica

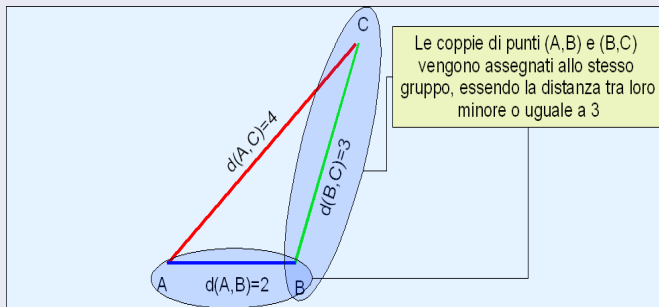
Clustering
gerarchico

Clustering non
gerarchico

Paradosso e distanza metrica

Per determinare i cluster si determina una distanza **soglia** tale che dati due punti A e B , se la distanza tra loro è tale che $d(A, B) > \text{soglia}$ allora A e B appartengono a gruppi diversi, viceversa saranno classificati nello stesso gruppo.

- **problema:** utilizzando una distanza metrica si va incontro al paradosso descritto nel seguente esemio: si considerino tre punti A , B e C , se la distanza che li separa è minore o uguale a **3** allora i punti vengono assegnati allo stesso gruppo



A. Iodice

Analisi dei dati



Distanze rispetto ad una soglia

Analisi dei dati

A. Iodice

Clustering:
classificazione
automatica

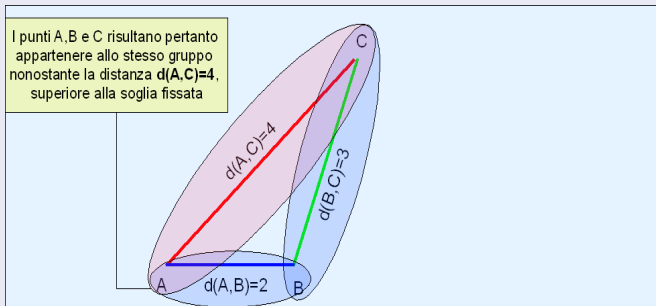
Clustering
gerarchico

Clustering non
gerarchico

Paradosso e distanza metrica

Per determinare i cluster si determina una distanza **soglia** tale che dati due punti A e B , se la distanza tra loro è tale che $d(A, B) > \text{soglia}$ allora A e B appartengono a gruppi diversi, viceversa saranno classificati nello stesso gruppo.

- **problema:** utilizzando una distanza metrica si va incontro al paradosso descritto nel seguente esemio: si considerino tre punti A , B e C , se la distanza che li separa è minore o uguale a **3** allora i punti vengono assegnati allo stesso gruppo



A. Iodice

Analisi dei dati



Distanze rispetto ad una soglia

Analisi dei
dati

A. Iodice

Clustering:
classificazione
automatica

Clustering
gerarchico

Clustering non
gerarchico

Passaggio alle ultrametriche

Per ovviare al problema si passa alle distanze **ultrametriche**: le distanze ultrametriche che caratterizzano ciascuna terna di punti sono date dal **triangolo isoscele** la cui **base** è data dalla distanza dei punti più vicini tra loro. Il **lato** del triangolo isoscele è rappresentato da una delle altre due distanze. In particolare si ha:

- **ultrametrica superiore minima** se il **lato** del triangolo isoscele corrisponde alla maggiore delle altre due distanze
- **ultrametrica inferiore massima** se il **lato** del triangolo isoscele corrisponde alla minore delle altre due distanze

A. Iodice

Analisi dei dati



Distanze rispetto ad una soglia

Analisi dei dati

A. Iodice

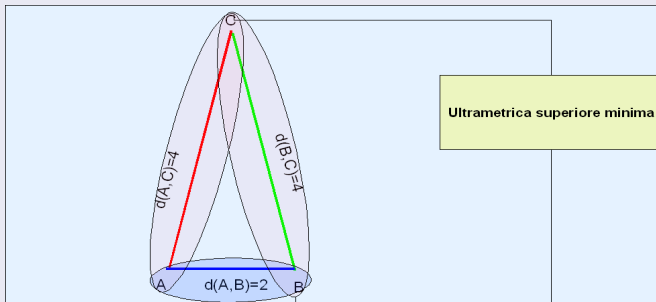
Clustering:
classificazione
automatica

Clustering
gerarchico

Clustering non
gerarchico

Passaggio alle ultrametrichi

- **ultrametrica superiore minima** se il lato del triangolo isoscele corrisponde alla maggiore delle altre due distanze
- **ultrametrica inferiore massima** se il lato del triangolo isoscele corrisponde alla minore delle altre due distanze



A. Iodice

Analisi dei dati



Distanze rispetto ad una soglia

Analisi dei dati

A. Iodice

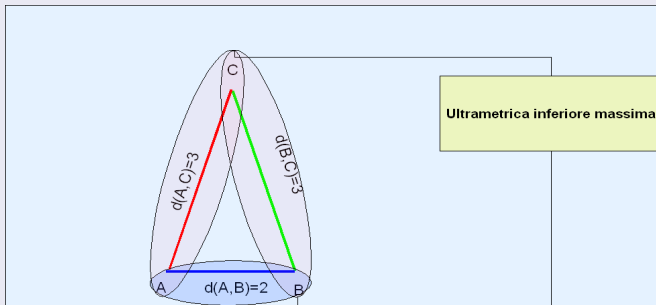
Clustering:
classificazione
automatica

Clustering
gerarchico

Clustering non
gerarchico

Passaggio alle ultrametrichi

- **ultrametrica superiore minima** se il lato del triangolo isoscele corrisponde alla maggiore delle altre due distanze
- **ultrametrica inferiore massima** se il lato del triangolo isoscele corrisponde alla minore delle altre due distanze



A. Iodice

Analisi dei dati



Trasformazione delle variabili

Analisi dei
dati

A. Iodice

Clustering:
classificazione
automatica

Clustering
gerarchico

Clustering non
gerarchico

Normalizzare. standardizzare

Per rendere direttamente confrontabili le variabili si possono utilizzare le seguenti trasformazioni

- **normalizzazione min-max**

$$X^* = \frac{X - \min(X)}{\text{range}(X)}$$

- **standardizzazione (z-score)**

$$X^* = \frac{X - \mu_X}{\sigma_X}$$



Tipologie di clustering

Analisi dei
dati

A. Iodice

Clustering:
classificazione
automatica

Clustering
gerarchico

Clustering non
gerarchico

gerarchie e partizioni

- **clustering gerarchico** determina una **gerarchia**, ovvero una struttura di cluster in cui l'insieme di unità viene bipartita ricorsivamente in corrispondenza di diversi livelli di aggregazione
- **clustering non gerarchico**: l'insieme delle unità viene partizionato in k cluster omogenei disgiunti



Partizioni e gerarchie

Analisi dei dati

A. Iodice

Clustering:
classificazione
automatica

Clustering
gerarchico

Clustering non
gerarchico

Partizioni

Si consideri un insieme O ed una sua partizione

$$P(O) = \{A, B, C, D, E\}.$$

- Due elementi A e B sono disgiunti oppure coincidono.
 $A \cap B = \emptyset$ se $A \neq B$,
 $A, B \in P(O)$.

- se $P(O) = \{A, B, C, D, E\}$, allora
 $A \cup B \cup C \cup D \cup E = O$.

Gerarchie

Si consideri un insieme O . La gerarchia $H(O)$ è un insieme di classi tali che

- Tutti gli oggetti $o_j \in O$ appartengono ad $H(O)$

$$o_j \in O \rightarrow o_j \in H(O)$$

- la gerarchia $H(O)$ contiene anche la classe contenente tutti gli oggetti considerati

$$O \in H(O)$$

- Due oggetti o_i e $o_j \in H(O)$ o sono disgiunti oppure uno dei due contiene l'altro

$$o_i \cap o_j = \emptyset$$

oppure risulta una tra

$$o_i \subset o_j \text{ e } o_j \subset o_i$$

A. Iodice

Analisi dei dati



Clustering gerarchico

Analisi dei
dati

A. Iodice

Clustering:
classificazione
automatica

Clustering
gerarchico

Clustering non
gerarchico

- **clustering gerarchico: algoritmo agglomerativo.** La soluzione si ottiene a partire dalle singole unità statistiche, ad ogni iterazione si aggregano le unità statistiche più *vicine*; la procedura termina quando tutte le unità risultano aggregate in un unico cluster.
- **clustering gerarchico: algoritmo divisivo.** In questo caso tutte le unità sono in una stessa classe e, ad ogni iterazione successiva, l'unità più dissimile dalle altre viene assegnata ad un nuovo cluster.



Scelta del criterio di aggregazione

Analisi dei
dati

A. Iodice

Clustering:
classificazione
automatica

Clustering
gerarchico

Clustering non
gerarchico

- **metodo del legame semplice.** La distanza tra due classi A e B viene calcolata considerando la distanza minima tra un elemento di A ad uno di B (**nearest-neighbour**)
- **metodo del legame completo.** La distanza tra due classi A e B viene calcolata considerando la distanza massima tra un elemento di A ad uno di B (**farthest-neighbour**)
- **metodo del legame medio.** La distanza tra due classi A e B viene calcolata considerando la distanza media tra gli elementi di A e di B



Scelta del criterio di aggregazione

Analisi dei
dati

A. Iodice

Clustering:
classificazione
automatica

Clustering
gerarchico

Clustering non
gerarchico

- **metodo dei centroidi**. La distanza tra due classi A e B viene calcolata considerando la distanza tra i centroidi (elementi medi) di A e di B .
- **metodo di Ward**: tale metodo parte da gruppi contenenti un solo oggetto; ad ogni passo aggrega gli oggetti che determinano il **minimo decremento di inerzia**
 - L'obiettivo di una cluster analysis è massimizzare l'inerzia **tra** i gruppi, ovvero minimizzare l'inerzia **interna** ai gruppi.
 - Il metodo di Ward aggrega di volta in volta la coppia di oggetti che minimizza la perdita di inerzia **tra** i gruppi.



Esempio di classificazione gerarchico

Analisi dei dati

A. Iodice

Clustering:
classificazione
automatica

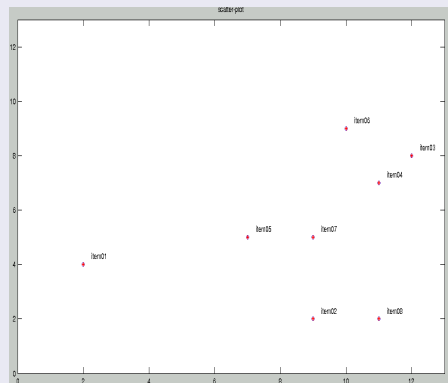
Clustering
gerarchico

Clustering non
gerarchico

esempio dati bivariati

	X	Y
item01	2	4
item02	9	2
item03	12	8
item04	11	7
item05	7	5
item06	9	9
item07	10	5
item08	11	2

scatter-plot



A. Iodice

Analisi dei dati



Esempio classificazione gerarchica

Analisi dei dati

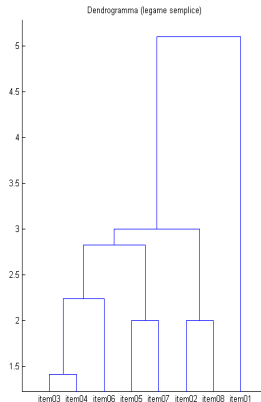
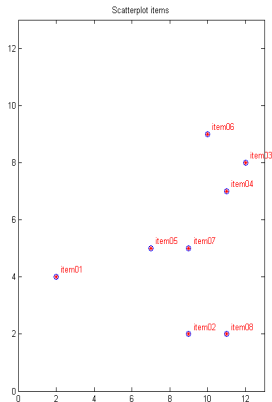
A. Iodice

Clustering:
classificazione
automatica

Clustering
gerarchico

Clustering non
gerarchico

legame semplice



A. Iodice

Analisi dei dati



Esempio classificazione gerarchica

Analisi dei dati

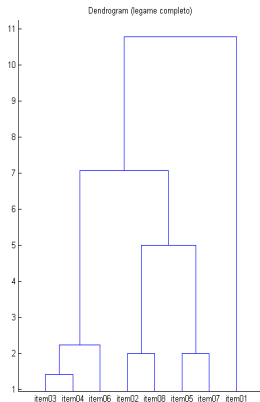
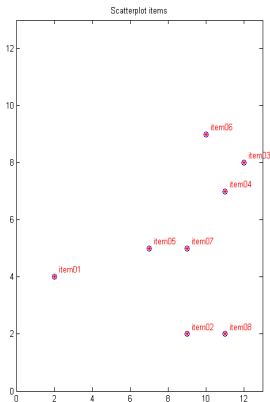
A. Iodice

Clustering:
classificazione
automatica

Clustering
gerarchico

Clustering non
gerarchico

legame completo



A. Iodice

Analisi dei dati



Esempio classificazione gerarchica

Analisi dei dati

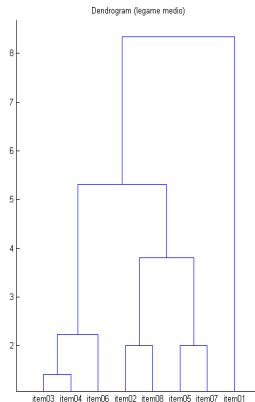
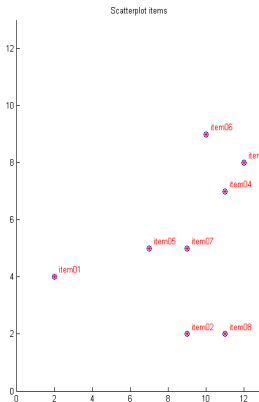
A. Iodice

Clustering:
classificazione
automatica

Clustering
gerarchico

Clustering non
gerarchico

legame medio



A. Iodice

Analisi dei dati



Qualità della soluzione

Analisi dei dati

A. Iodice

Clustering:
classificazione
automatica

Clustering
gerarchico

Clustering non
gerarchico

La qualità della gerarchia ottenuta può essere valutata confrontando le distanze tra le unità statistiche originarie e le distanze ultrametriche. Sia

- d_{ij} : la distanza tra l'individuo i e l'individuo j .
- d_{ij}^* : la distanza ultrametrica tra l'individuo i e l'individuo j .
- n : numero di unità statistiche considerate.

Indice di deformazione del
valore assoluto

$$\sum_{i=1}^n \sum_{j=i+1}^n \frac{|d_{ij} - d_{ij}^*|}{(n(n-2))/2}$$

Indice di deformazione del
quadrato degli scarti

$$\sum_{i=1}^n \sum_{j=i+1}^n \frac{(d_{ij} - d_{ij}^*)^2}{(n^2 - n)/2}$$

coefficiente di correlazione

$$\frac{\sum_{i=1}^n \sum_{j=i+1}^n (d_{ij} - \bar{d})(d_{ij}^* - d^*)}{\sum_{i=1}^n \sum_{j=i+1}^n (d_{ij} - \bar{d})^2 \sum_{i=1}^n \sum_{j=i+1}^n (d_{ij}^* - d^*)^2}$$



Qualità della soluzione

Analisi dei dati

A. Iodice

Clustering:
classificazione
automatica

Clustering
gerarchico

Clustering non
gerarchico

esempio dati bivariati

	X	Y
item01	2	4
item02	9	2
item03	12	8
item04	11	7
item05	7	5
item06	9	9
item07	10	5
item08	11	2

matrice di distanza

	i1	i2	i3	i4	i5	i6	i7	i8
i1	0							
i2	7.28	0						
i3	10.77	6.71	0					
i4	9.49	5.39	1.41	0				
i5	5.10	3.61	5.83	4.47	0			
i6	8.60	7.00	3.16	2.83	4.47	0		
i7	8.06	3.16	3.61	2.24	3.00	4.12	0	
i8	9.22	2.00	6.08	5.00	5.00	7.28	3.16	0

A. Iodice

Analisi dei dati



Qualità della soluzione

Analisi dei dati

A. Iodice

Clustering:
classificazione
automatica

Clustering
gerarchico

Clustering non
gerarchico

matrice di distanza

	i1	i2	i3	i4	i5	i6	i7	i8
i1	0							
i2	7.28	0						
i3	10.77	6.71	0					
i4	9.49	5.39	1.41	0				
i5	5.10	3.61	5.83	4.47	0			
i6	8.60	7.00	3.16	2.83	4.47	0		
i7	8.06	3.16	3.61	2.24	3.00	4.12	0	
i8	9.22	2.00	6.08	5.00	5.00	7.28	3.16	0

matrice delle ultrametriche superiori minime

	i1	i2	i3	i4	i5	i6	i7	i8
i1	0							
i2	10.7	0						
i3	10.7	7.2	0					
i4	10.7	7.2	1.41	0				
i5	10.7	5.00	7.2	7.2	0			
i6	10.7	7.2	3.16	3.16	7.2	0		
i7	10.7	5.00	7.2	7.2	3.00	7.2	0	
i8	10.7	2.00	7.2	7.2	3.00	7.2	5.00	0

A. Iodice

Analisi dei dati





Esempio con data set di dimensioni maggiori

Analisi dei dati

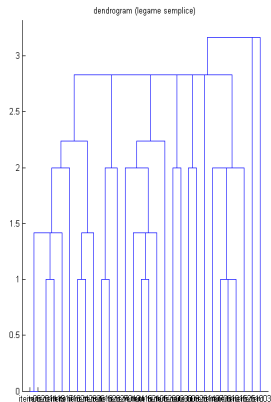
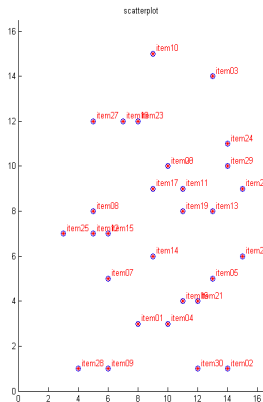
A. Iodice

Clustering:
classificazione
automatica

Clustering
gerarchico

Clustering non
gerarchico

legame semplice



A. Iodice

Analisi dei dati



Esempio con data set di dimensioni maggiori

Analisi dei dati

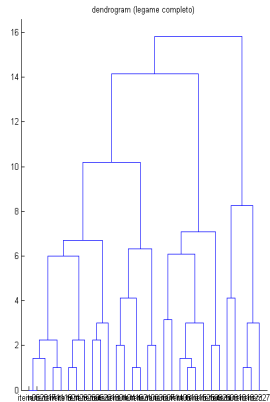
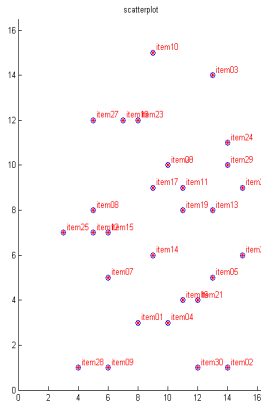
A. Iodice

Clustering:
classificazione
automatica

Clustering
gerarchico

Clustering non
gerarchico

legame completo



A. Iodice

Analisi dei dati



Esempio con data set di dimensioni maggiori

Analisi dei dati

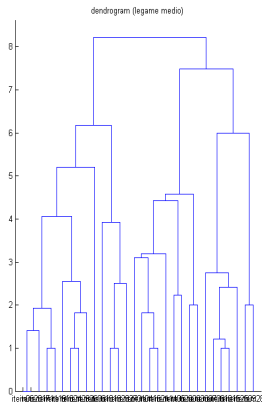
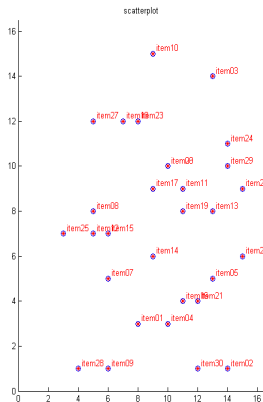
A. Iodice

Clustering:
classificazione
automatica

Clustering
gerarchico

Clustering non
gerarchico

legame medio



A. Iodice

Analisi dei dati



Esempio data set strutturato

Analisi dei dati

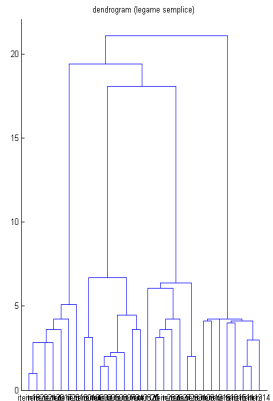
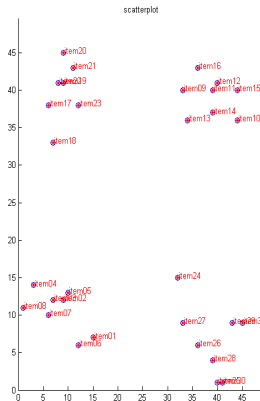
A. Iodice

Clustering:
classificazione
automatica

Clustering
gerarchico

Clustering non
gerarchico

legame semplice



A. Iodice

Analisi dei dati



Esempio data set strutturato

Analisi dei dati

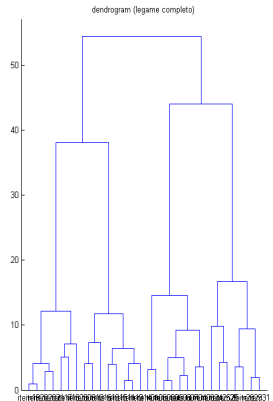
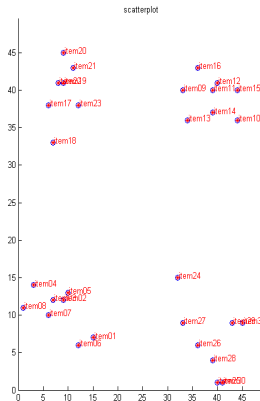
A. Iodice

Clustering:
classificazione
automatica

Clustering
gerarchico

Clustering non
gerarchico

legame completo



A. Iodice

Analisi dei dati



Esempio con data set strutturato

Analisi dei dati

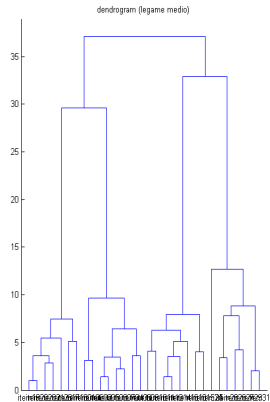
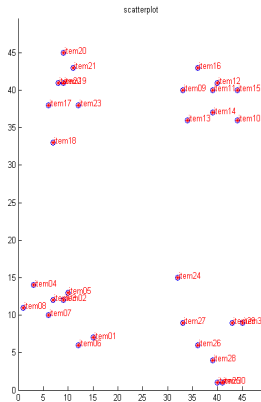
A. Iodice

Clustering:
classificazione
automatica

Clustering
gerarchico

Clustering non
gerarchico

legame medio



A. Iodice

Analisi dei dati



Clustering non gerarchico: centri mobili

Analisi dei
dati

A. Iodice

Clustering:
classificazione
automatica

Clustering
gerarchico

Clustering non
gerarchico

centri mobili

- 1 scegliere il parametro k numero di classi
- 2 scegliere k **centri** c_1, \dots, c_k in modo casuale
- 3 assegnare ciascuna osservazione al centro più vicino, ottenendo C_1, \dots, C_k classi
- 4 aggiornare i centri $c_1 = \bar{C}_1, \dots, c_k = \bar{C}_k$
- 5 ripetere gli step 3 e 4 fino a convergenza



Clustering non gerarchico: nubi dinamiche

Analisi dei
dati

A. Iodice

Clustering:
classificazione
automatica

Clustering
gerarchico

Clustering non
gerarchico

nubi dinamiche

- 1 scegliere il parametro k numero di classi
- 2 scegliere k **nuclei** C_1^j, \dots, C_k^j : ciascun nucleo è composto da j elementi
- 3 assegnare ciascuna osservazione al nucleo più vicino, ottenendo C_1, \dots, C_k classi
- 4 aggiornare i nuclei prendendo gli elementi di ciascuna classe più vicini tra loro
- 5 ripetere gli step 3 e 4 fino a convergenza



Clustering non gerarchico: K-means

Analisi dei
dati

A. Iodice

Clustering:
classificazione
automatica

Clustering
gerarchico

Clustering non
gerarchico

K-means

- 1 scegliere il parametro k numero di classi
- 2 scegliere k **nuclei** c_1, \dots, c_k : ciascun nucleo è composto da più elementi
- 3 assegnare ciascuna osservazione al nucleo più vicino
- 4 aggiornare **contestualmente** il nucleo associato alla classe cui l'osservazione è stata assegnata
- 5 ripetere gli step 3 e 4 fino a convergenza



Procedure non gerarchiche

Analisi dei dati

A. Iodice

Clustering:
classificazione
automatica

Clustering
gerarchico

Clustering non
gerarchico

Centri mobili

- Ogni centro corrisponde ad un singolo punto.
- Ciascun oggetto viene assegnato al centro più vicino. Poi si ricalcolano i centri.

Nubi dinamiche

- Ogni centro corrisponde ad un nucleo di punti.
- Ciascun oggetto viene assegnato al nucleo più vicino. Poi si ricalcolano i nuclei.

K -means

- Ogni centro corrisponde ad un nucleo di punti.
- Ciascun oggetto viene assegnato al nucleo più vicino. Ad ogni assegnazione si ricalcola il nucleo corrispondente.

Stabilità dei risultati

Sebbene il K -medie (e le sue numerose varianti) determinino soluzioni migliori rispetto al metodo dei centri mobili, il problema della dipendenza della soluzione ottenuta dalla scelta iniziale dei centri (o dei nuclei) permane.



Le forme forti

Analisi dei
dati

A. Iodice

Clustering:
classificazione
automatica

Clustering
gerarchico

Clustering non
gerarchico

Individuazione delle forme forti

Gli oggetti che in diverse applicazioni successive di una procedura di classificazione automatica vengono collocate sempre nello stesso gruppo si definiscono **forme forti**

- Effettuare s volte la classificazione automatica
- Gli oggetti che in ognuna delle s ripetizioni vengono assegnate ad uno stesso gruppo sono una forma forte
- Gli oggetti che in ognuna delle s ripetizioni vengono assegnate a gruppi diversi sono una forma debole



Esempio di classificazione non gerarchico

Analisi dei dati

A. Iodice

Clustering:
classificazione
automatica

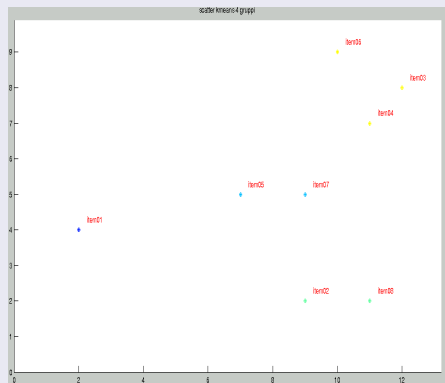
Clustering
gerarchico

Clustering non
gerarchico

esempio dati bivariati

	X	Y
item01	2	4
item02	9	2
item03	12	8
item04	11	7
item05	7	5
item06	9	9
item07	10	5
item08	11	2

scatter-plot



A. Iodice

Analisi dei dati



Esempio con data set strutturato

Analisi dei dati

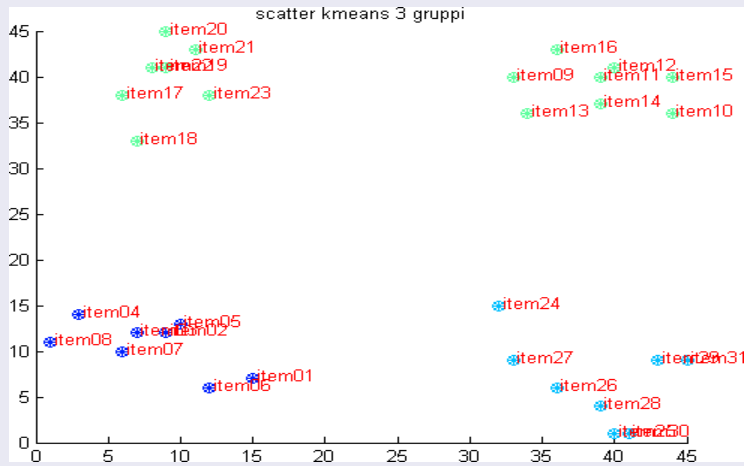
A. Iodice

Clustering:
classificazione
automatica

Clustering
gerarchico

Clustering non
gerarchico

scelta del numero di classi: $k = 3$



A. Iodice

Analisi dei dati



Esempio con data set strutturato

Analisi dei dati

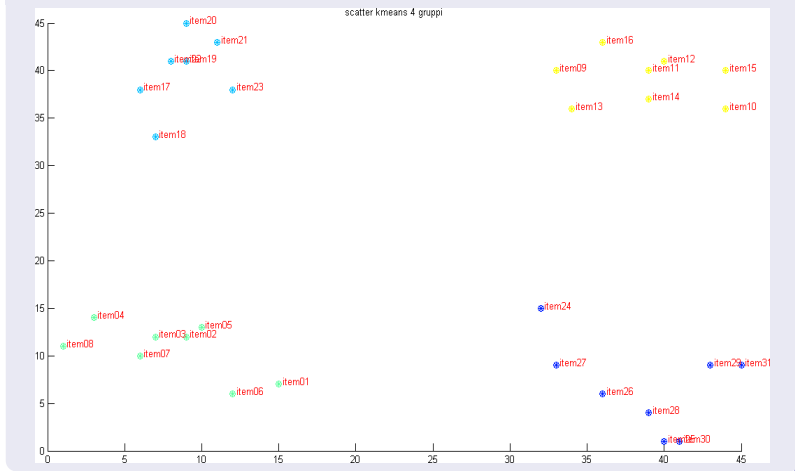
A. Iodice

Clustering:
classificazione
automatica

Clustering
gerarchico

Clustering non
gerarchico

scelta del numero di classi: $k = 4$



A. Iodice

Analisi dei dati



Scelta del metodo

Analisi dei dati

A. Iodice

Clustering:
classificazione
automatica

Clustering
gerarchico

Clustering non
gerarchico

procedure non gerarchiche

- La **soluzione** consiste in una partizione dei dati, di facile interpretazione
- il **costo computazionale** per ottenere una soluzione è $n \times k$ (n oggetti, k classi)
- la scelta **iniziale** (e casuale) dei centri determina l'esito della soluzione
- necessità di conoscere a priori k il numero di classi, altrimenti sono necessarie prove ripetute su diversi k alla ricerca di quello ottimale

procedure gerarchiche

- La **soluzione** consiste in una gerarchia indicizzata:
 - lettura verticale: *come* si formano i gruppi
 - lettura orizzontale: ad ogni livello della gerarchia gli oggetti vengono assegnati ai diversi gruppi
- il **costo computazionale** per ottenere una soluzione è $n(n-1)/2$, diventa proibitivo in caso di elevata numerosità
- le aggregazioni **iniziali** sono incluse nelle successive e condizionano l'esito della soluzione

A. Iodice

Analisi dei dati