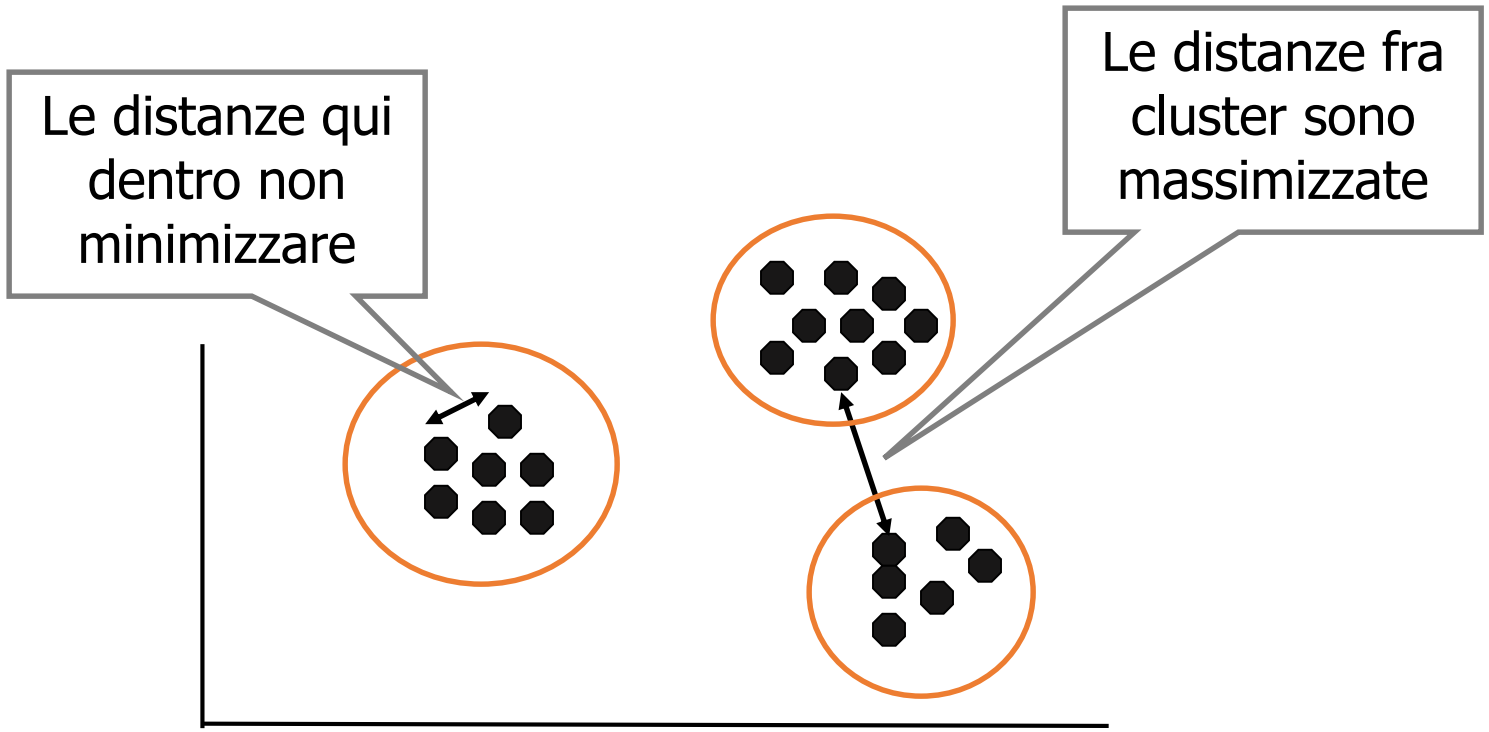
Data Mining in R
Approccio non supervisionato
ANALISI CLUSTER
1. Introduzione

Laura Grassini

Tan, Steinbach, Kumar: Introduction to data mining, 2006, Addison Wesley
http://www-users.cs.umn.edu/~kumar/dmbook/index.php

Che cosa è l'analisi cluster?

Un metodo **non supervisionato** che si propone di trovare gruppi di unità/oggetti tali che le unità di un gruppo sono più prossime fra loro che le unità di gruppi diversi



Che cosa è un cluster ?

I **gruppi o cluster** sono gruppi **esclusivi** *(ogni unità appartiene a un solo gruppo) ed **esaustivi** (tutte le unità sono collocate nei gruppi)**.**

L'analisi è **esplorativa**: sono i dati e i metodi impiegati a formare i gruppi, che vengono interpretati (nelle loro caratteristiche statistiche) a posteriori.

(*) Diversa è l'analisi «fuzzy» in cui ad ogni unità si associano degli score che misurano il grado di appartenenza dell'unità ai vari cluster

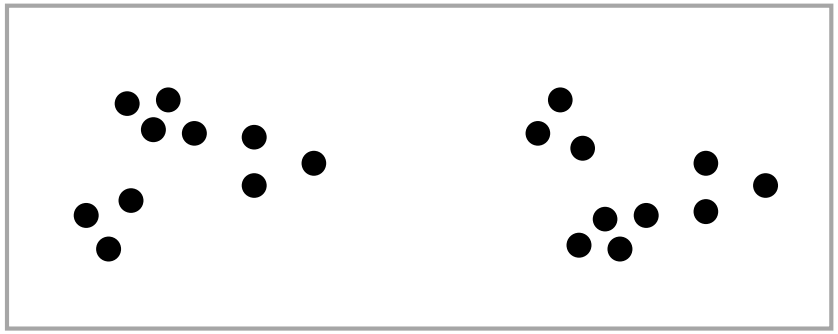
Campi di applicazioni dell'analisi cluster

- Segmentazione di mercato (clienti con comportamento simile)
- Studi geografici (aree simili rispetto a qualche caratteristica: es. precipitazioni, struttura produttiva ecc.)
- Classificazione di documenti
- Raggruppamento di titoli azionari che hanno simile andamento nel tempo (traiettorie)
- Summarization e data compression (*cluster prototypes*)

Altri obiettivi dell'analisi cluster

- Per investigare in merito alla validità di gruppi preesistenti (v. ad esempio **`iris`** data di **`R`**) o trovati in altro modo
- Per studiare la associazione fra variabili in modo «non paramaterico»: usiamo alcune variabili per individuare i gruppi mentre altre variabili sono lasciate esterne all'analisi ma vengono poi esaminate in relazione alle caratteristiche dei vari gruppi individuati. Ad esempio:
  - Individuiamo gruppi di clienti in base ai loro livelli di soddisfazione espressi in relazione agli attributi di un servizio
  - Studiamo la distribuzioni delle variabili sociodemografiche (età, genere ecc.) all'interno dei gruppi per capire se i livelli di soddisfazione si associano a qualche caratteristica dei soggetti.

I cluster non sono noti e forse non esistono ...



Quanti cluster?

6 cluster?

2 cluster?

4 cluster?

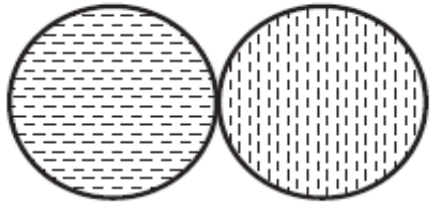
… ma quelli che troviamo ci possono orientare nella complessità dei dati

Che cosa “NON è” l’analisi cluster?

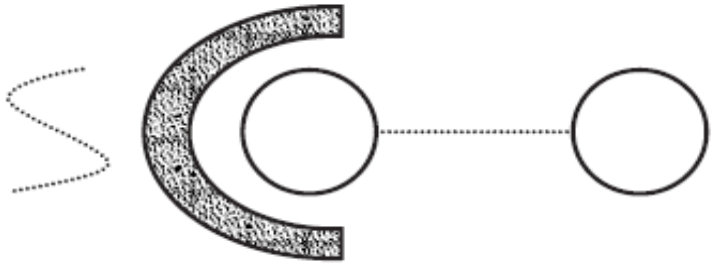
- **NON è** metodo supervisionato: le label di gruppo non sono note
- **NON è** risultato di una query: i gruppi non sono il risultato di una interrogazione esterna
- **NON è** la classificazione semplice di unità (ad es. rispetto all’iniziale del cognome)

Vediamo alcuni tipi di clusters

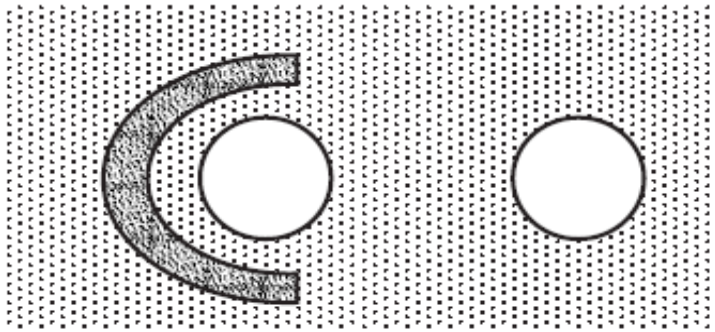
(a) Well-separated clusters. Each point is closer to all of the points in its cluster than to any point in another cluster.



(b) Center-based clusters. Each point is closer to the center of its cluster than to the center of any other cluster.



(c) Contiguity-based clusters. Each point is closer to at least one point in its cluster than to any point in another cluster.



(d) Density-based clusters. Clusters are regions of high density separated by regions of low density.

Tan, Steinbach, Kumar (2006)

Metodi gerarchici e metodi non gerarchici

Metodi gerarchici. I gruppi provengono dalla progressiva *aggregazione* (o *divisione*) di gruppi successivi, partendo da *n* unità singole (un solo gruppo comprendente tutte le unità) fino ad arrivare a un solo gruppo comprendente tutte le unità (le *n* singole unità originali).
Non è necessario specificare il numero di gruppi cercati.
La partizione dei dati viene ricavata a posteriori, esaminando i risultati.
Quando un'unità è entrata in un gruppo **non viene** da questo più rimossa.

Metodi non gerarchici. Si ricerca direttamente una partizione dell'insieme delle unità in *K* gruppi.
E' necessario specificare il numero di gruppi cercati.
Le unità possono cambiare gruppo durante il processo di clustering.