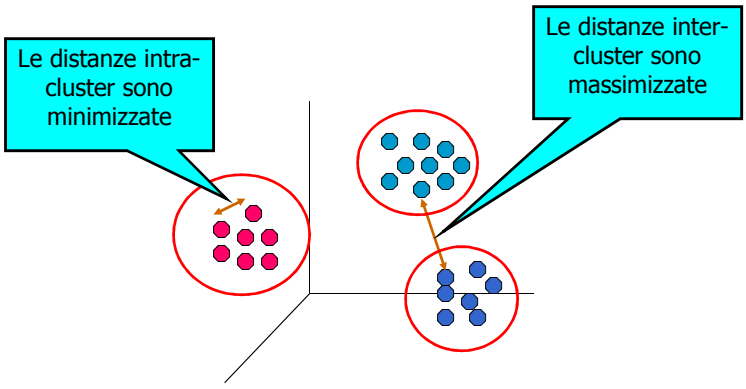
Clustering

Prof. Matteo Golfarelli
Alma Mater Studiorum - Università di Bologna

Cosa è la Clustering analysis

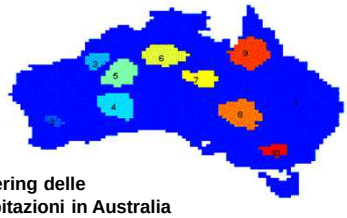
- Ricerca di gruppi di oggetti tali che gli oggetti appartenenti a un gruppo siano “simili” tra loro e differenti dagli oggetti negli altri gruppi



Applicazioni delle analisi dei cluster

- **Comprendere**
  - ✓ Gruppi di documenti correlati per favorire la navigazione, gruppi di geni e proteine che hanno funzionalità simili, gruppi di azioni che hanno fluttuazioni simili
- **Riassumere**
  - ✓ Ridurre la dimensione di data set grandi

	Discovered Clusters	*Industry Group*
1	Applied-Matl-DOWN,Bay-Network-Down,3-COM-DOWN, Cabletron-Sys-DOWN,CISCO-DOWN,HP-DOWN, DSC-Comm-DOWN,INTEL-DOWN,LSI-Logic-DOWN, Micron-Tech-DOWN,Texas-Inst-Down,Tellabs-Inc-Down, Natl-Semiconduct-DOWN,Oracl-DOWN,SGI-DOWN, Sun-DOWN	Technology1-DOWN
2	Apple-Comp-DOWN,Autodesk-DOWN,DEC-DOWN, ADV-Micro-Device-DOWN,Andrew-Corp-DOWN, Computer-Assoc-DOWN,Circuit-City-DOWN, Compaq-DOWN, EMC-Corp-DOWN, Gen-Inst-DOWN, Motorola-DOWN,Microsoft-DOWN,Scientific-Atl-DOWN	Technology2-DOWN
3	Fannie-Mae-DOWN,Fed-Home-Loan-DOWN, MBNA-Corp-DOWN,Morgan-Stanley-DOWN	Financial-DOWN
4	Baker-Hughes-UP,Dresser-Inds-UP,Halliburton-HLD-UP, Louisiana-Land-UP,Phillips-Petro-UP,Unocal-UP, Schlumberger-UP	Oil-UP



Clustering delle precipitazioni in Australia

Cosa non è la Clustering analysis

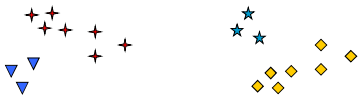
- Classificazione supervisionata
  - ✓ Parte dalla conoscenza delle etichette di classe
- Segmentazione
  - ✓ Suddividere gli studenti alfabeticamente in base al cognome
- Risultati di una query
  - ✓ Il raggruppamento si origina in base a indicazioni esterne

La nozione di cluster può essere ambigua

Quanti cluster?

6 Cluster

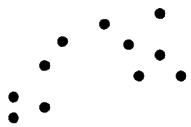
2 Cluster

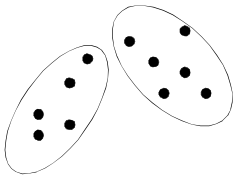


4 Cluster

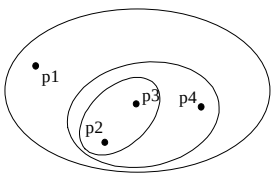
Tipi di clustering

- Un clustering è un insieme di cluster. Una distinzione importante è tra:
  - ✓ Clustering partizionante: una divisione degli oggetti in sottoinsiemi (cluster) non sovrapposti. Ogni oggetto appartiene esattamente a un cluster.

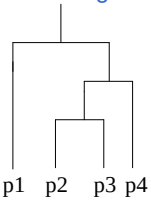




  - ✓ Clustering gerarchico: un insieme di cluster annidati organizzati come un albero gerarchico



Clustering gerarchico tradizionale



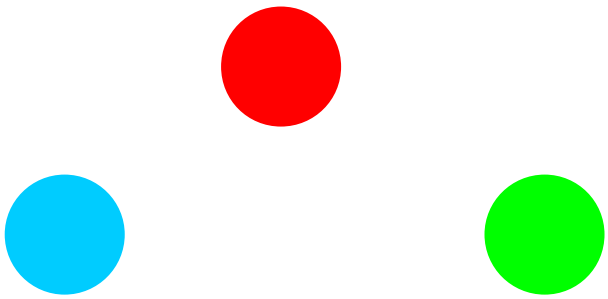
Dendrogramma

Altre distinzioni tra insiemi di cluster

- Esclusivo vs non esclusivo
  - ✓ In un clustering non esclusivo, i punti possono appartenere a più cluster.
  - ✓ Utile per rappresentare punti di confine o più tipi di classi.
- Fuzzy vs non-fuzzy
  - ✓ In un fuzzy clustering un punto appartiene a tutti i cluster con un peso tra 0 e 1.
  - ✓ La somma dei pesi per ciascun punto deve essere 1.
  - ✓ I clustering probabilistici hanno caratteristiche similari.
- Parziale vs completo
  - ✓ In un clustering parziale alcuni punti potrebbero non appartenere a nessuno dei cluster.
- Eterogeneo vs omogeneo
  - ✓ In un cluster eterogeneo i cluster possono avere dimensioni, forme e densità molto diverse.

Tipi di cluster: Well-Separated

- Well-Separated Cluster:
  - ✓ Un cluster è un insieme di punti tali che qualsiasi punto nel cluster è più vicino (più simile a) ogni altro punto del cluster rispetto a ogni altro punto che non appartenga al cluster.



3 well-separated cluster

Tipi di cluster: Center-Based

- Center-based
  - ✓ Un cluster è un insieme di punti tali che un punto nel cluster è più vicino (o più simile a) al “centro”del cluster, piuttosto che al centro di ogni altro
  - ✓ Il centro di un cluster è chiamato centroide, la media di tutti i punti che appartengono al cluster, oppure medioide, il punto più “representativo” del cluster

4 center-based cluster

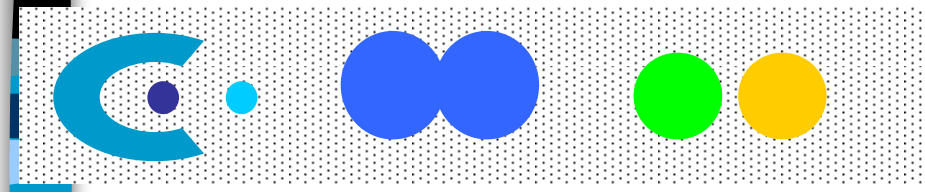
Tipi di cluster: Contiguity-Based

- Cluster contigui (Nearest neighbor o Transitive)
  - ✓ Un cluster è un insieme di punti tali che un punto nel cluster è più vicino (o più simile) ad almeno uno dei punti del cluster rispetto a ogni punto che non appartenga al cluster.

8 contiguous cluster

Tipi di cluster: Density-Based

- Density-based
  - ✓ Un cluster è una regione densa di punti, che è separata da regioni a bassa densità, dalle altre regioni a elevata densità.
  - ✓ Utilizzata quando i cluster hanno forma irregolare o "attorcigliata", oppure in presenza di rumore o di outliers

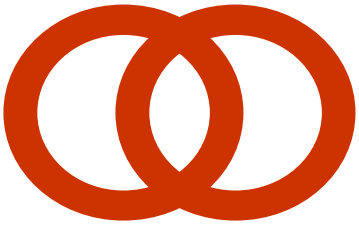


6 density-based cluster

Tipi di cluster: Cluster concettuali

- Cluster con proprietà condivise o in cui la proprietà condivisa deriva dall'intero insieme di punti (rappresenta un particolare concetto)
  - ✓ Sono necessarie tecniche sofisticate in grado di esprimere il concetto sotteso

.



2 cerchi sovrapposti

K-means Clustering

- Si tratta di una tecnica di clustering partizionante
- Ogni cluster è associato a un centroide
- Ogni punto è assegnato al cluster con il cui centroide è più vicino
- Il numero di cluster, K, deve essere specificato

1: Select K points as the initial centroids.
2: **repeat**
3: Form K clusters by assigning all points to the closest centroid.
4: Recompute the centroid of each cluster.
5: **until** The centroids don't change

K-means Clustering – Dettagli

- L'insieme iniziale di centroidi è normalmente scelto casualmente
  - ✓ I cluster prodotti variano ad ogni esecuzione
- Il centroide è (tipicamente) la media dei punti del cluster.
- La 'prossimità' può essere misurata dalla distanza euclidea, cosine similarity, correlazione, ecc.
- L'algoritmo dei K-means converge per le più comuni misure di similarità e la convergenza si verifica nelle prime iterazioni
  - ✓ L'algoritmo può convergere a soluzioni sub-ottime
  - ✓ Spesso la condizione di stop è rilassata e diventa 'continua fino a che un numero ridotto di punti passa da un cluster a un altro'
- La complessità dell'algoritmo è O(n * K * I * d)
  - ✓ n = numero di punti, K = numero di cluster, I = numero di iterazioni, d = numero di attributi

Valutazione della bontà dei cluster K-means

- La misura più comunemente utilizzata è lo scarto quadratico medio (SSE - Sum of Squared Error)
  - ✓ Per ogni punto l'errore è la distanza dal centroide del cluster a cui esso è assegnato.

$$SSE = \sum_{i=1}^{K} \sum_{x \in C_i} dist^2(m_i, x)$$

  - ✓ x è un punto appartenente al cluster C_i e m_i è il rappresentante del cluster C_i
    - è possibile dimostrare che il centroide che minimizza SSE quando si utilizza come misura di prossimità la distanza euclidea è la media dei punti del cluster.

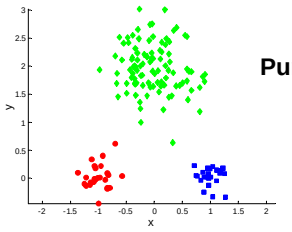
$$m_i = \sum_{x \in C_i} x$$

  - ✓ Ovviamente il valore di SSE si riduce incrementando il numero dei cluster K
    - Un buon clustering con K ridotto può avere un valore di SSE più basso di un cattivo clustering con K più elevato

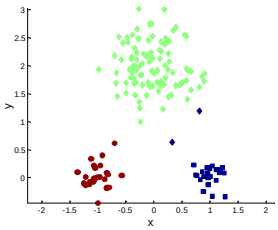
Convergenza e ottimalità

- C'è soltanto un numero finito di modi di partizionare n record in k gruppi. Quindi c'è soltanto un numero finito di possibili configurazioni in cui tutti i centri sono centroidi dei punti che possiedono.
- Se la configurazione cambia in una iterazione, deve avere migliorato la distorsione. Quindi ogni volta che la configurazione cambia, deve portare in uno stato mai visitato prima
  - ✓ Il riassegnamento dei record ai centroidi è fatto sulla base delle distanze minori
  - ✓ Il calcolo dei nuovi centroidi minimizza il valore di SSE per il cluster
- Quindi l'algoritmo deve arrestarsi per non disponibilità di ulteriori configurazioni da visitare
- Non è detto tuttavia che la configurazione finale sia quella che in assoluto presenta il minimo valore di SSE come evidenziato nella seguente slide
  - ✓ Spostare un centroide della soluzione sul lato destro comporta sempre un aumento di SSE, ma la configurazione sul lato sinistro presenta un SSE minore

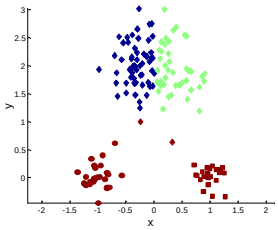
Convergenza e ottimalità



Punti e cluster naturali

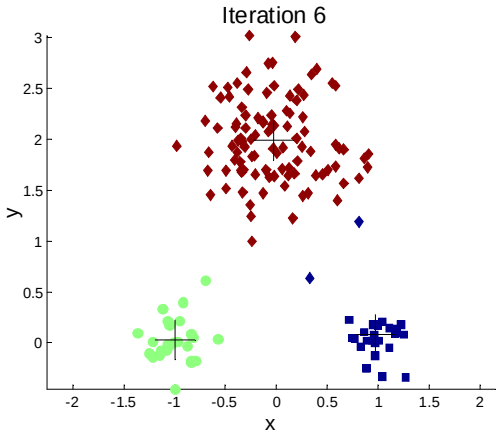


Clustering ottimale

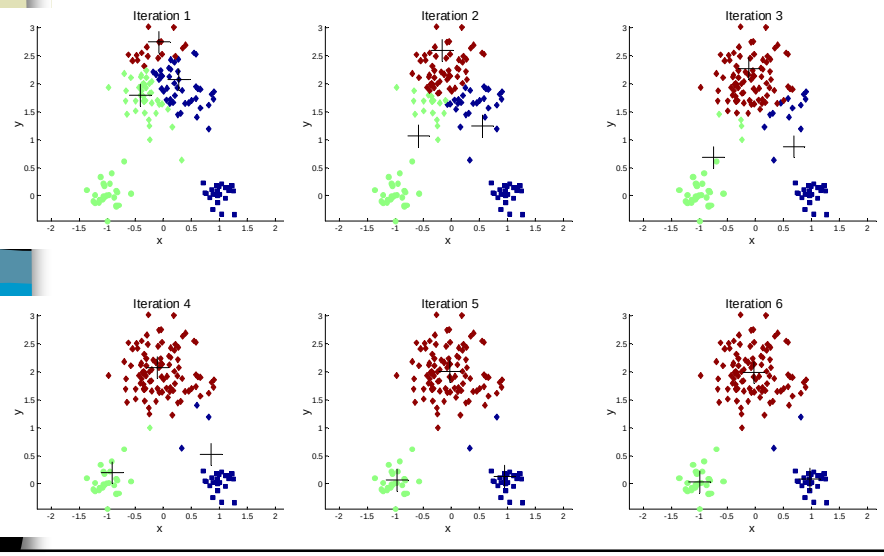


Clustering sub-ottimale

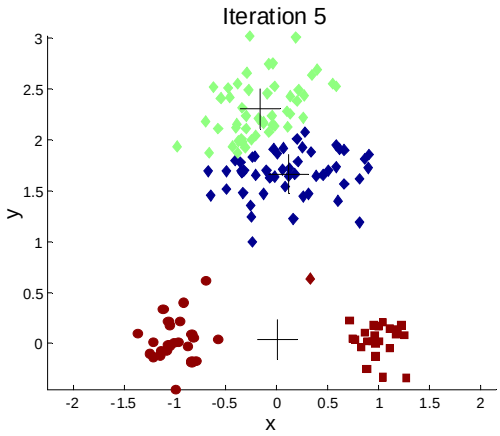
Importanza della scelta dei centroidi di partenza



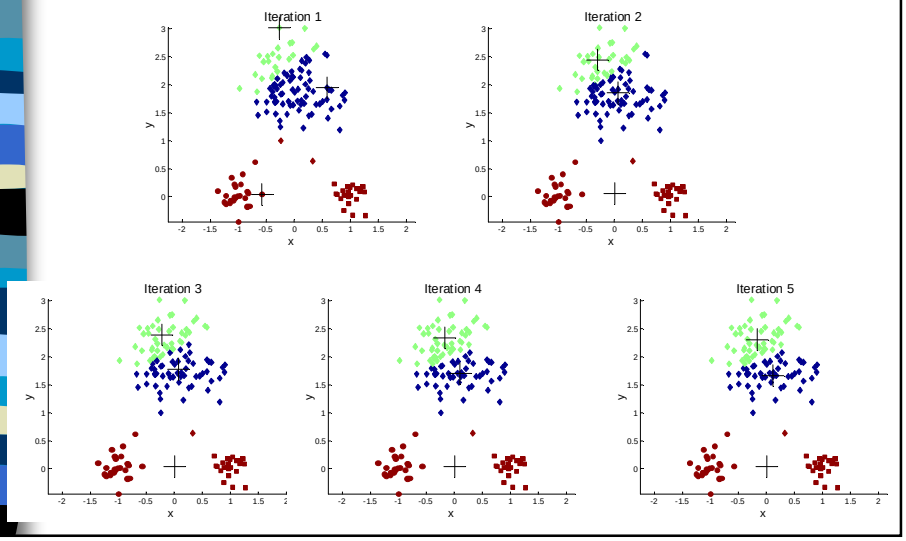
Importanza della scelta dei centroidi di partenza



Importanza della scelta dei centroidi di partenza



Importanza della scelta dei centroidi di partenza



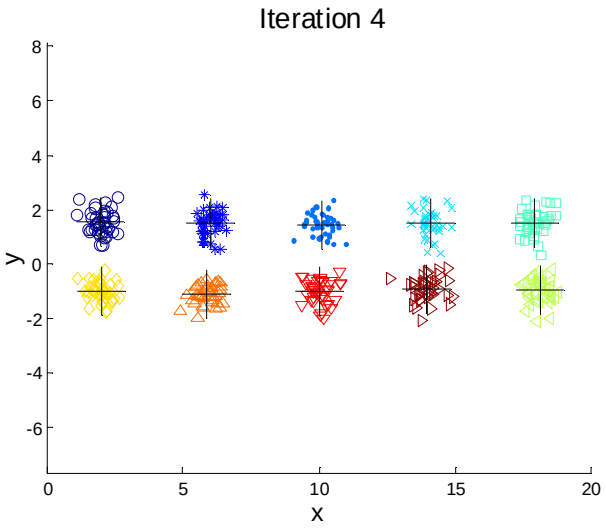
Problema della selezione dei centroidi iniziali

- Se ci sono K cluster reali la probabilità di scegliere un centroide da ogni cluster è molto limitata
  - ✓ Se i cluster hanno la stessa cardinalità n:

$$P = \frac{\#\text{ modi di scegliere un centroide per cluster}}{\#\text{ modi di scegliere un centroide}} = \frac{K!n^K}{(Kn)^K} = \frac{K!}{K^K}$$

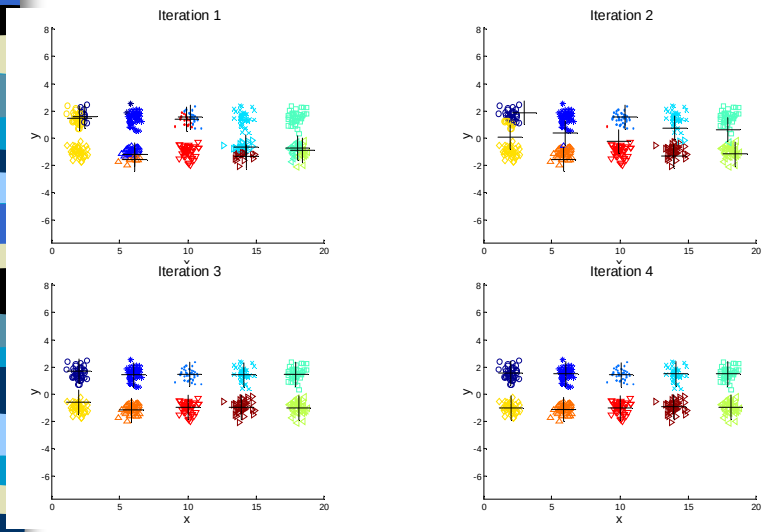
  - ✓ K = 10, la probabilità è $10!/10^{10} = 0.00036$
  - ✓ Alcune volte i centroidi si riposizioneranno correttamente altre no…

Esempio con 10 Cluster



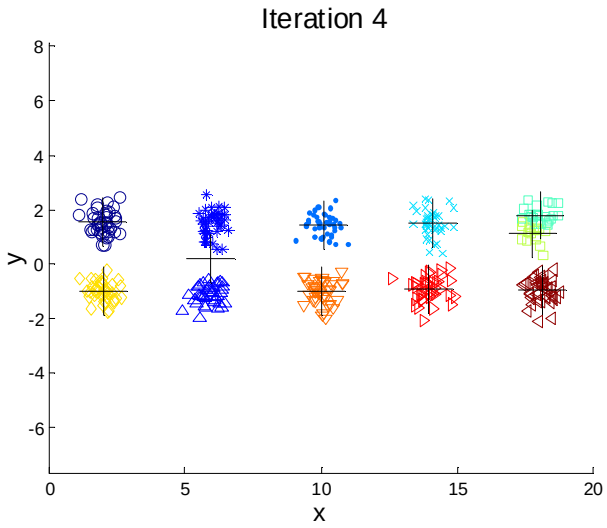
Partendo con cluster con 2 centroidi e cluster con 0 centroidi

Esempio con 10 Cluster



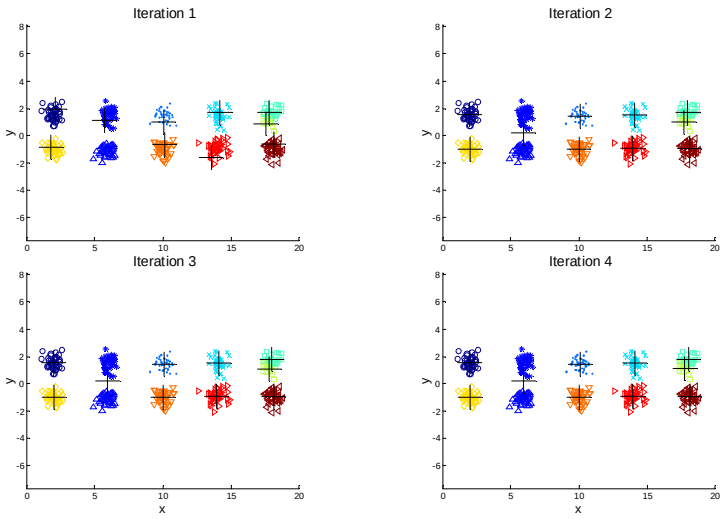
Partendo con cluster con 2 centroidi e cluster con 0 centroidi

Esempio con 10 Cluster



Partendo con coppie di cluster con 3 centroidi
e coppie di cluster con 1 centroide

Esempio con 10 Cluster



Partendo con coppie di cluster con 3 centroidi
e coppie di cluster con 1 centroide

Soluzione ai problemi indotti dalla scelta dei centroidi iniziali

- Esegui più volte l'algoritmo con diversi centroidi di partenza
  - ✓ Può aiutare, ma la probabilità non è dalla nostra parte!
- Esegui un campionamento dei punti e utilizza una tecnica di clustering gerarchico per individuare k centroidi iniziali
- Seleziona più di k centroidi iniziali e quindi seleziona tra questi quelli da utilizzare
  - ✓ Il criterio di selzione è quello di mantenere quelli maggiormente "separati"
- Utilizza tecniche di post-processing per eliminare i cluster erronemante individuati
- Bisecting K-means
  - ✓ Meno suscettibile al problema

Gestione dei Cluster vuoti

- L'algoritmo K-means può determinare cluster vuoti qualora, durante la fase di assegnamento, ad un centroide non venga assegnato nessun elemento.
  - ✓ Questa situazione può determinare un SSE elevato poichè uno dei cluster non viene "utilizzato"
- Sono possibili diverse strategie per individuare un centroide alternativo
  - ✓ Scegliere il punto che maggiormente contribuisce al valore di SSE
  - ✓ Scegliere un elemento del cluster con il maggior SSE. Normalmente ciò determina lo split del cluster in due cluster che includono gli elementi più vicini.

Gestione degli outlier

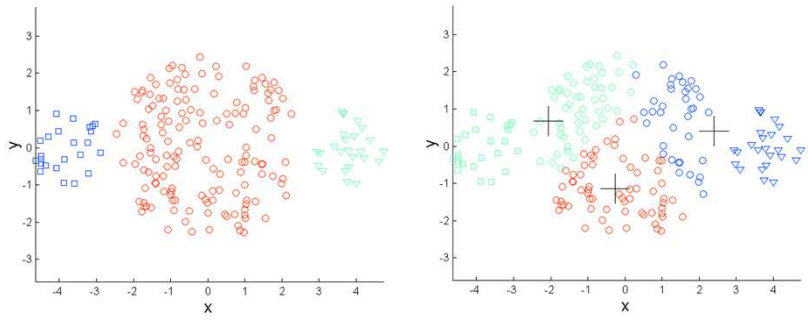
- La bontà del clustering può essere negativamente influenzata dalla presenza di outlier che tendono a ”spostare” il centroide dei cluster al fine di ridurre l’aumento dell’SSE determinato indotto dall’outlier
  - ✓ Dato che SSE è un quadrato di una distanza, i punti molto lontani incidono pesantemente sul suo valore
- Gli outlier se identificati possono essere eliminati in fase di preprocessing
  - ✓ Il concetto di outlier dipende dal dominio di applicazione
  - ✓ Studieremo opportune tecniche per la loro definizione

K-means: Limitazioni

- L’algoritmo k-means non raggiunge buoni risultati quando i cluster naturali hanno:
  - ✓ Diverse dimensioni
  - ✓ Diversa densità
  - ✓ Forma non globulare
  - ✓ I dati contengono outlier

Limitazioni di k-means: differenti dimensioni

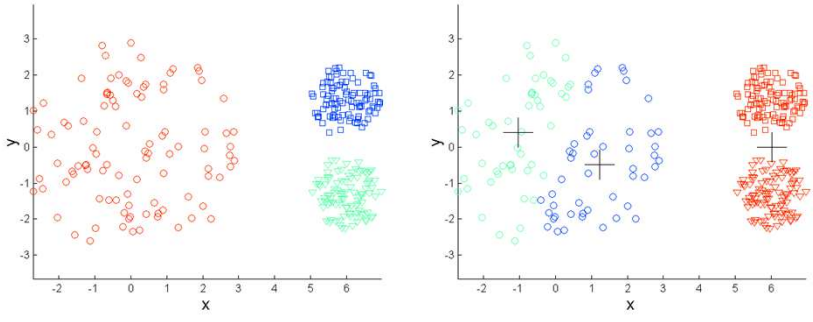
- Il valore di SSE porta a identificare i centroidi in modo da avere cluster delle stesse dimensioni se i cluster non sono well-separated



Punti orignali **K-means (3 Cluster)**

Limitazioni di k-means: differenti densità

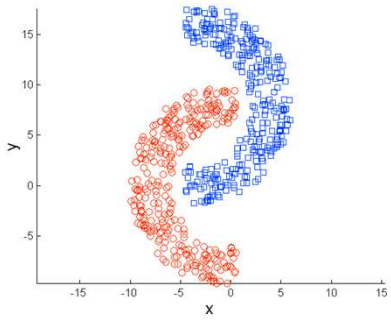
- Cluster più densi comportano distanze intra-cluster minori, quindi le zone meno dense richiedono più mediani per minimizzare il valore totale di SSE



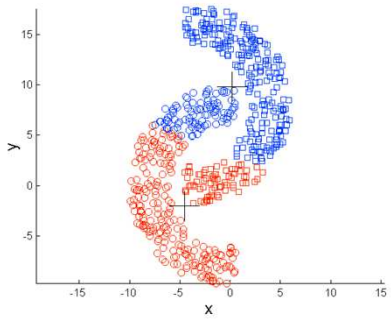
Punti originali **K-means (3 Cluster)**

Limitazioni di k-means: forma non globulare

- SSE si basa su una distanza euclidea che non tiene conto della forma degli oggetti



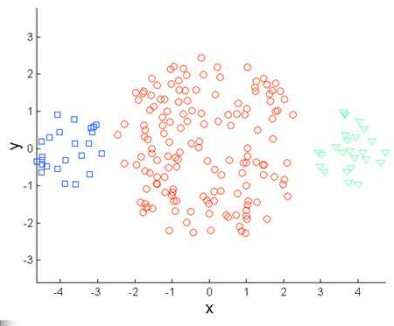
Punti originali



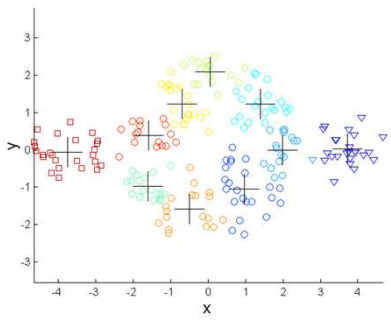
K-means (2 Cluster)

K-means: possibili soluzioni

- Una possibile soluzione è quella di utilizzare un valore di k più elevato individuando così porzioni di cluster.
- La definizione dei cluster “naturali” richiede poi una tecnica per mettere assieme i cluster individuati

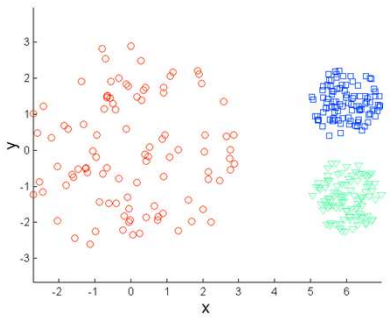


Punti originali

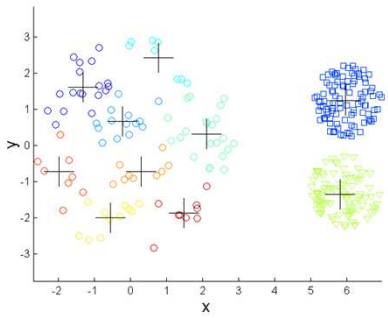


K-means Clusters

K-means: possibili soluzioni

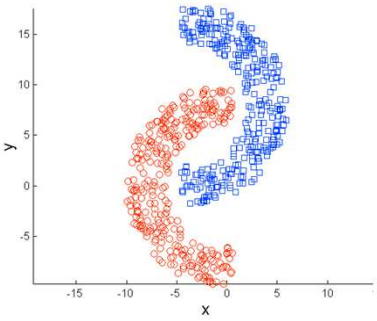


Punti originali

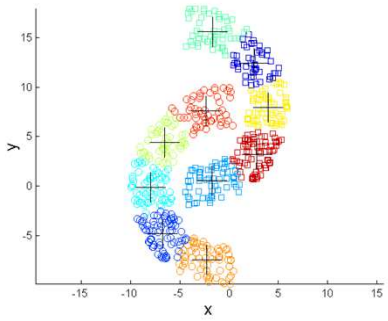


K-means Cluster

K-means: possibili soluzioni



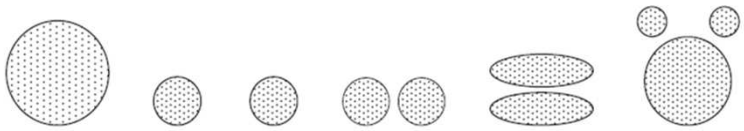
Punti originali



K-means Cluster

Esercizio

- Indicare la suddivisione in cluster e la posizione approssimata dei centroidi scelta dall'algoritmo k-means assumendo che:
  - ✓ I punti siano equamente distriubiti
  - ✓ La funzione distanza sia SSE
  - ✓ Il valore di K è indicato sotto le figure

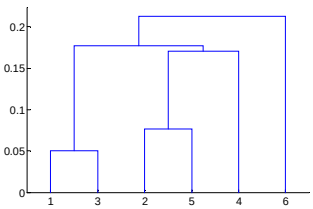


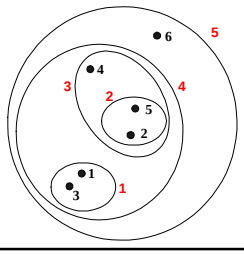
K=2 K=3 K=3 K=2 K=3

- Se ci possono essere più soluzioni quali sono ottimi globali?

Clustering gerarchico

- Produce un insieme di cluster organizzati come un albero gerarchico
- Può essere visualizzato come un dendrogramma
  - ✓ Un diagramma ad albero che mostra la sequenza di fusioni tra cluster
- A valori diversi sulle ordinate possono corrispondere clustering composti da un numero diverso di elementi





Clustering gerarchico: pro e contro

☺ Non richiede di definire a priori il numero dei cluster
- ✓ Il numero desiderato di cluster può essere ottenuto 'tagliando' il dendrogramma al livello opportuno

☺ Può identificare una tassonomia (classificazione gerarchica) di concetti
- ✓ Gli elementi più simili saranno fusi prima di elementi meno simili

☹ Una volta che una decisione è presa (fusione) non può più essere annullata

☹ In molte configurazioni è sensibile a rumore e outlier

☹ Manca una funzione di ottimizzazione globale

Clustering gerarchico

- Esistono due approcci per la creazione di un clustering gerarchico
  - ✓ Agglomerativo:
    - Si parte con cluster formati da elementi singoli
    - A ogni passo, fondi i due cluster più 'vicini' fino a che non rimane un solo cluster (o k cluster)
  - ✓ Divisivo:
    - Si parte con un unico cluster che include tutti gli elementi
    - A ogni passo, separa il cluster più 'lontano' fino a che i cluster contengono un solo elemento (oppure sono stati creati k cluster)
- Normalmente gli algoritmi gerarchici utilizzano una matrice di similarità o delle distanze (proximity matrix)

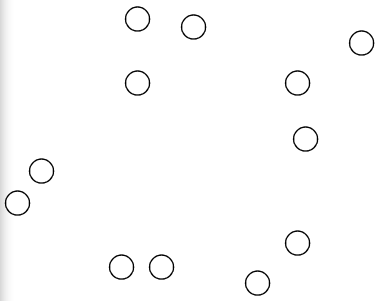
»

Approcci agglomerativi

- E' la tecnica di clustering gerarchico più comune
- L'algoritmo di base è molto semplice
  1. Crea un cluster per ogni elemento
  2. Calcola la proximity matrix
  3. **Repeat**
  4. Fondi i due cluster più vicini
  5. Aggiorna la proximity matrix
  6. **Until** rimane un solo cluster
- L'operazione fondamentale è il calcolo della similarità/vicinanza tra due cluster
  - ✓ Gli approcci per calcolare la distanza tra i cluster distinguono i diversi approcci.

Situazione iniziale

- La situazione iniziale prevede cluster formati da singoli elementi e una proximity matrix

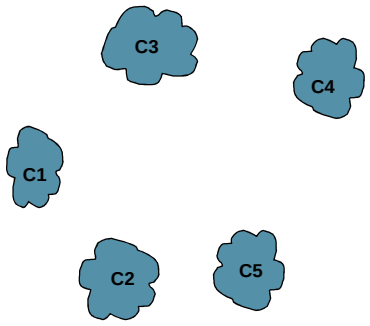


	p1	p2	p3	p4	p5	. . .
p1						
p2						
p3						
p4						
p5						
.						
.						
.						

Proximity Matrix

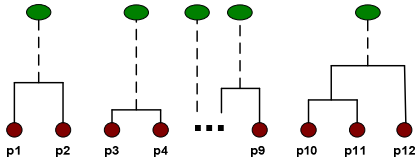
Situazione intermedia

- Dopo alcune iterazioni si saranno formati diversi cluster



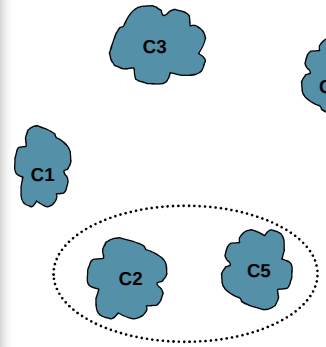
	C1	C2	C3	C4	C5
C1					
C2					
C3					
C4					
C5					

Proximity Matrix



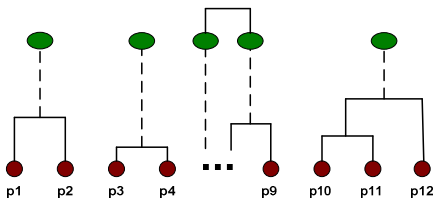
Situazione intermedia

- E' necessario fondere i due cluster più vicini (C2 e C5) e aggiornare la proximity matrix.
  - ✓ Le uniche celle interessate sono quelle che coinvolgono C2 e C5

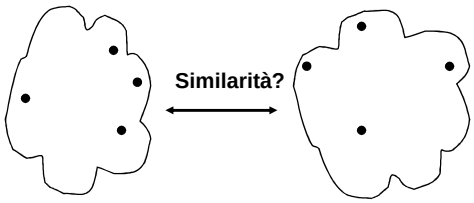


	C1	C2	C3	C4	C5
C1					
C2					
C3					
C4					
C5					

Proximity Matrix

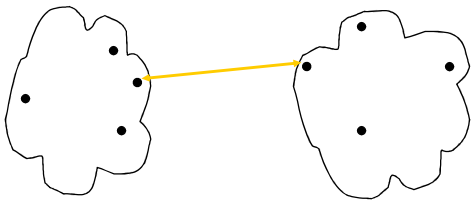


Come definire la similarità inter-cluster



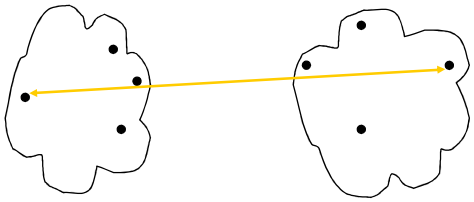
- MIN o Single link
- MAX o Complete link
- Group Average
- Distanza tra i centroidi
- Altri metodi guidati da una funzione obiettivo
  - ✓ Il Ward Method utilizza lo scarto quadratico medio

Come definire la similarità inter-cluster



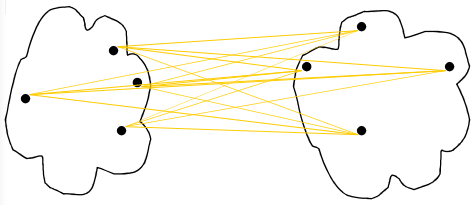
- MIN o Single link: è la minima distanza tra due elementi dei cluster
- MAX o Complete link
- Group Average
- Distanza tra i centroidi
- Altri metodi guidati da una funzione obiettivo
  - ✓ Il Ward Method utilizza lo scarto quadratico medio

Come definire la similarità inter-cluster



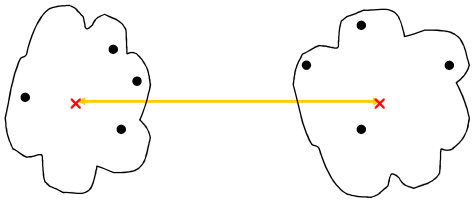
- MIN o Single link
- MAX o Complete link: è la massima distanza tra due punti del cluster
- Group Average
- Distanza tra i centroidi
- Altri metodi guidati da una funzione obiettivo
  - ✓ Il Ward Method utilizza lo scarto quadratico medio

Come definire la similarità inter-cluster



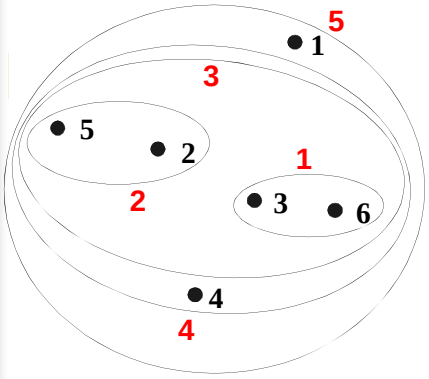
- MIN o Single link
- MAX o Complete link
- Group Average: è la media delle distanze tra tutti i punti dei cluster
- Distanza tra i centroidi
- Altri metodi guidati da una funzione obiettivo
  - ✓ Il Ward Method utilizza lo scarto quadratico medio

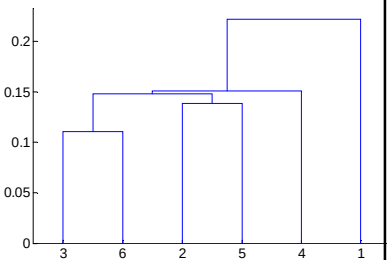
Come definire la similarità inter-cluster



- MIN o Single link
- MAX o Complete link
- Group Average
- Distanza tra i centroidi
- Altri metodi guidati da una funzione obiettivo
  - ✓ Il Ward Method utilizza lo scarto quadratico medio

Clustering gerarchico: MIN



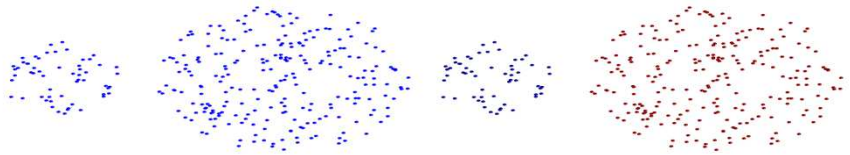


Dist({3,6}, {2,5})=min(dist({2,3}), dist({2,6}), dist({5,3}), dist({5,6})) = min(0.15,0.25,0.28,0.39)=0.15

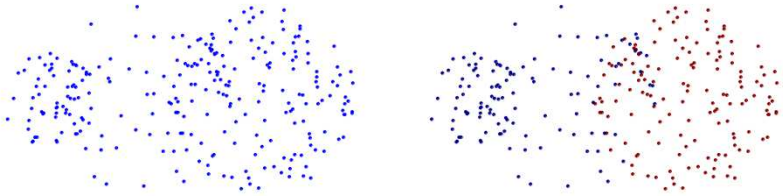
	p1	p2	p3	p4	p5	p6
p1	0.00	0.24	0.22	0.37	0.34	0.23
p2	0.24	0.00	0.15	0.20	0.14	0.25
p3	0.22	0.15	0.00	0.15	0.28	0.11
p4	0.37	0.20	0.15	0.00	0.29	0.22
p5	0.34	0.14	0.28	0.29	0.00	0.39
p6	0.23	0.25	0.11	0.22	0.39	0.00

MIN: pro e contro

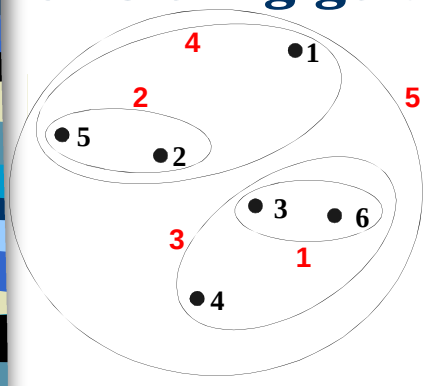
- Permette di gestire anche cluster non sferici

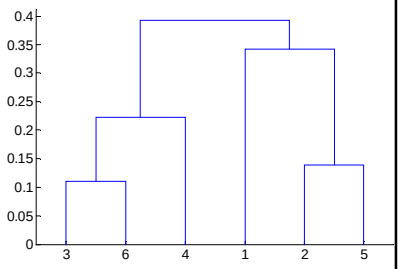


- E' soggetto a outlier e rumori



Clustering gerarchico: MAX





Dopo il passo 2…

Dist({3,6}, {4})=max(dist({3,4}), dist({6,4}))
= max(0.15,0.22)=0.22

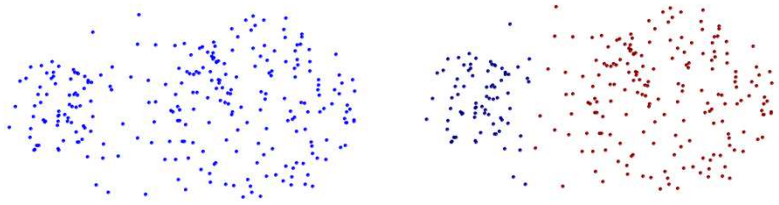
Dist({3,6}, {2,5})=max(dist({3,2}),
dist({3,5}), dist({6,2}) dist({6,5})) =
max(0.15,0.25,0.28,0.39)=**0.39**

Dist({3,6}, {1})=max(dist({3,1}), dist({6,1}))
= max(0.22,0.23)=0.23

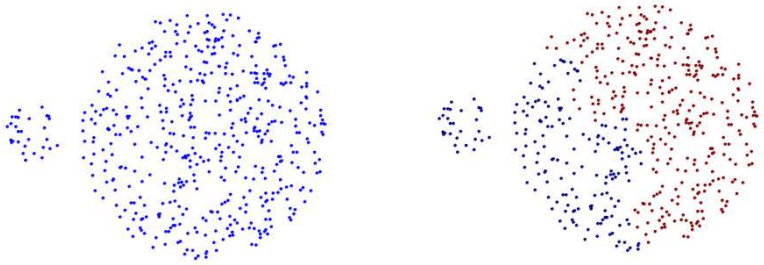
	p1	p2	p3	p4	p5	p6
p1	0.00	0.24	0.22	0.37	0.34	0.23
p2	0.24	0.00	0.15	0.20	0.14	0.25
p3	0.22	0.15	0.00	0.15	0.28	0.11
p4	0.37	0.20	0.15	0.00	0.29	0.22
p5	0.34	0.14	0.28	0.29	0.00	0.39
p6	0.23	0.25	0.11	0.22	0.39	0.00

MAX: pro e contro

- Meno suscettibile al rumore



- Tende a separare cluster grandi
- Privilegia cluster globulari



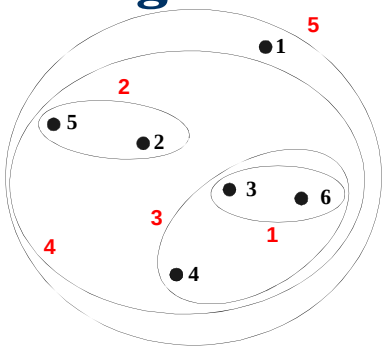
Clustering gerarchico: Group Average

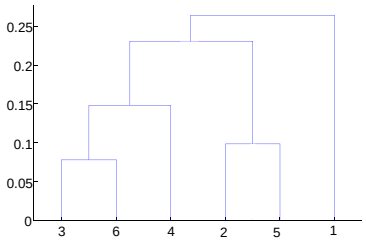
- La similarità tra due cluster è la media tra la similarità delle coppie di punti dei cluster

$$\text{proximity}(\text{Cluster}_i, \text{Cluster}_j) = \frac{\sum_{\substack{p_i \in \text{Cluster}_i \\ p_j \in \text{Cluster}_j}} \text{proximity}(p_i, p_j)}{|\text{Cluster}_i| * |\text{Cluster}_j|}$$

- Rappresenta un compromesso tra MIN e MAX
  - ✓ PRO: meno suscettibile al rumore
  - ✓ CONTRO: orientato a cluster sferici

Clustering gerarchico: Group Average





Dist({3,6,4}, {1}) = (0.22+0.37+0.23)/(3×1) = 0.28

Dist({2,5}, {1}) = (0.2357+0.3421)/(2×1) = 0.2889

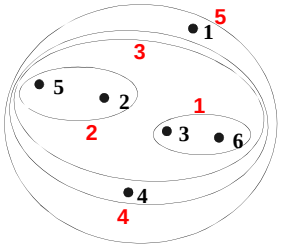
Dist({3,6,4}, {2,5}) =(0.15+0.28+0.25+0.39+0.20+0.29)/(3×2) = **0.26**

	p1	p2	p3	p4	p5	p6
p1	0.00	0.24	0.22	0.37	0.34	0.23
p2	0.24	0.00	0.15	0.20	0.14	0.25
p3	0.22	0.15	0.00	0.15	0.28	0.11
p4	0.37	0.20	0.15	0.00	0.29	0.22
p5	0.34	0.14	0.28	0.29	0.00	0.39
p6	0.23	0.25	0.11	0.22	0.39	0.00

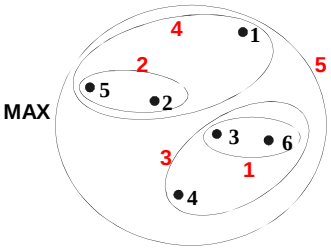
Clustering gerarchico: Ward's Method

- La similarità tra due cluster è data dall'aumento di errore quadratico quando due cluster sono fusi
  - ✓ Richiede di calcolare il centroide dei cluster per il calcolo di SSE
  - ✓ Simile al group average se per calcolare la prossimità tra due punti si usa la loro distanza al quadrato
- PRO: Meno suscettibile a outlier e rumore
- CONTRO: Orientato ai cluster sferici
- Utilizza la stessa funzione obiettivo dell'algoritmo K-means
  - ✓ Può essere utilizzata per inizializzare k-means: poichè permette di identificare il corretto valore di k e una suddivisione iniziale dei punti
    - Si ricorda che non potendo annullare le scelte fatte le soluzioni trovate dai metodi gerarchici sono spesso sub-ottime.

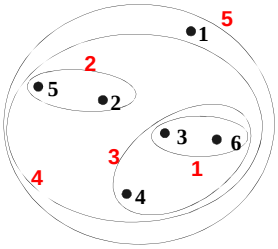
Clustering gerarchici a confronto



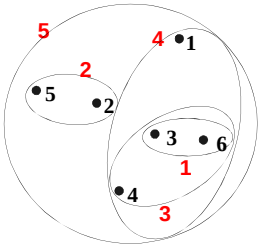
MIN



MAX



Group Average



Ward's Method

Clustering gerarchico: complessità

- Spazio: $O(N^2)$ è lo spazio occupato dalla matrice di prossimità quando il numero di punti è N
- Tempo: $O(N^3)$
  - ✓ Sono necessari N passi per costruire il dendrogramma. Ad ogni passo la matrice di prossimità deve essere aggiornata e scorsa
  - ✓ Proibitivo per dataset di grandi dimensioni

Esercizio

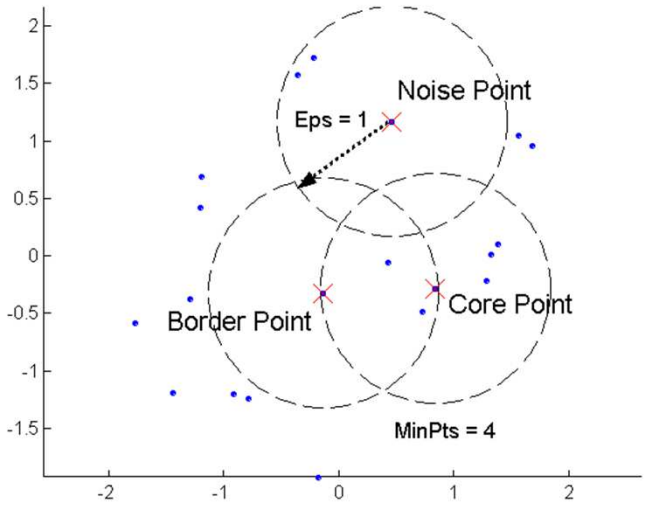
- Data la seguente matrice di similarità rappresentare i dendrogrammi derivanti dai clustering gerarchici ottenuti mediante
  - ✓ Single link
  - ✓ Complete link

simil	p1	p2	p3	p4	p5
p1	1.00	0.10	0.41	0.55	0.35
p2	0.10	1.00	0.64	0.47	0.98
p3	0.41	0.64	1.0	0.44	0.85
p4	0.55	0.47	0.44	1.00	0.76
p5	0.35	0.98	0.85	0.76	1.00

DBSCAN

- DBSCAN è un approccio basato sulla densità
  - ✓ Densità = numero di punti all'interno di un raggio Eps specificato
  - ✓ Core point sono i punti la cui densità è superiore a una soglia MinPts
    - Questi punti sono interni a un cluster
  - ✓ Border point hanno una densità minore di MinPts, ma nelle loro vicinanze (ossia a distanza < Eps) è presente un core point
  - ✓ Noise point tutti i punti che non sono Core point e Border point

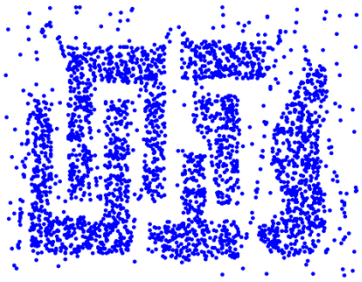
DBSCAN: Core, Border e Noise Point



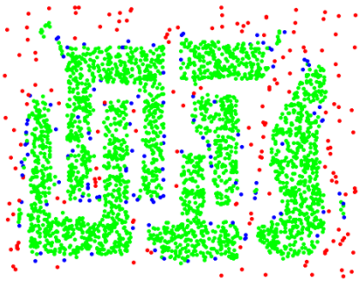
Algoritmo DBSCAN

1. // Input:Dataset **D**, MinPts, Eps
2. // Insieme dei cluster **C**
3. Classifica i punti in D come core, border o noise
4. Elimina tutti i punti di tipo noise
5. Assegna al cluster c_i i punti core che abbiano distanza < di Eps da almeno uno degli altri punti assegnato al cluster
6. Assegna i punti border a uno dei cluster a cui sono associati i corrispondenti punti core

DBSCAN: Core, Border and Noise Points



Original Points

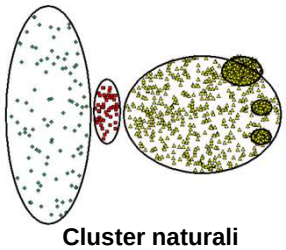


Point types: core, border and noise

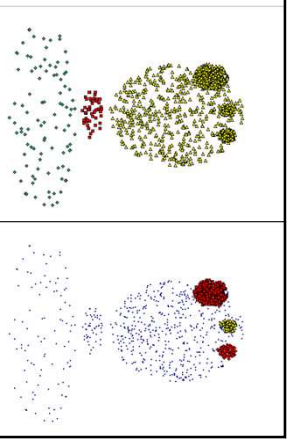
Eps = 10, MinPts = 4

DBSCAN: pro e contro

- Pro
  - ✓ Resistente al rumore
  - ✓ Può generare cluster con forme e dimensioni differenti
- Contro
  - ✓ Dati con elevata dimensionalità
    - Rende difficile definire efficacemente il concetto di densità a causa dell'elevata sparsità
  - ✓ Dataset con densità variabili



Cluster naturali

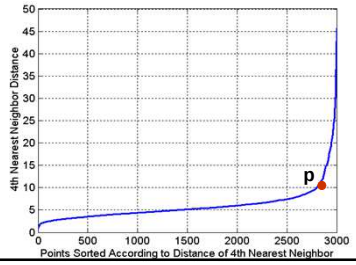


MinPts = 4
Eps=9.92

MinPts = 4
Eps = 9.75

DBSCAN: scelta di EPS e MinPts

- L'idea di base è che per i core point i k-esimi nearest neighbor siano circa alla stessa distanza e piuttosto vicini
- I noise point avranno il k-esimo nearest neighbor più lontano
- Visualizziamo i punti ordinati in base alla distanza del loro k-esimo vicino. Il punto p in cui si verifica un repentino cambio della distanza misurata segnala la separazione tra core point e noise point
  - ✓Il valore di Eps è dato dall'ordinata di p
  - ✓Il valore di MinPts è dato da k
  - ✓Il risultato dipende dal valore di k, ma l'andamento della curva rimane similare per valori sensati di k
  - ✓Un valore di k normalmente utilizzato per dataset bidimensionali è 4



Validità dei Cluster

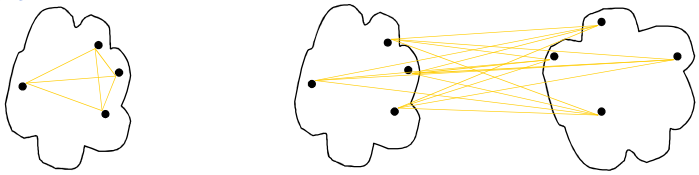
- Per le tecniche di classificazione supervisionata esistono più misure per valutare la bontà dei risultati basate sul confronto tra le label note per il test set e quelle calcolate dall'algoritmo
  - ✓ Accuracy, precision, recall
- Le motivazioni per la valutazione di un clustering
  1. Valutare, senza l'utilizzo di informazioni esterne, come il risultato del clustering modella i dati
  2. Determinare che si sia determinato il "corretto" numero di cluster
  3. Verificare la clustering tendency di un insieme di dati, ossia identificare la presenza di strutture non-randomiche
  4. Valutare, utilizzando informazioni esterne (etichette di classe), come il risultato del clustering modella i dati
  5. Comparare le caratteristiche di due insiemi di cluster per valutare quale è il migliore
  6. Comparare le caratteristiche di due algoritmi di clustering per valutare quale è il migliore
- I punti 1,2,3 non richiedono informazioni esterne
- I punti 5 e 6 possono essere basati sia su informazioni interne, sia esterne

Misure di validità

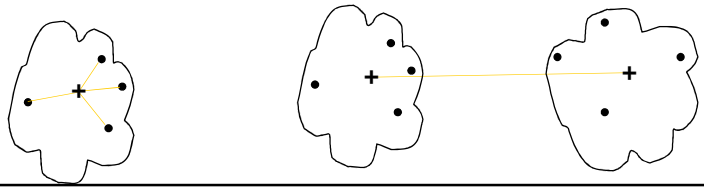
- I quantificatori numerici utilizzati per valutare i diversi aspetti legati alla validità dei cluster sono classificati in:
  - ✓ Misure esterne o supervisionate: calcolano in che misura le label dei cluster corrispondono alle label delle classi
    - Entropia
  - ✓ Misure interne o non supervisionate: misurano la bontà di un clustering *senza* utilizzare informazioni esterne
    - Somma al quadrato degli errori (SSE)
  - ✓ Misure relative: utilizzate per comparare due diversi clustering o cluster
    - Possono basarsi sia su misure interne, sia su misure esterne.

Misure interne: Coesione e Separazione

- Coesione e separazione possono essere calcolati sia per rappresentazioni basate su grafi…
  - ✓La coesione è la somma dei pesi degli archi tra i nodi appartenenti a un cluster
  - ✓La separazione è la somma dei pesi degli archi tra i nodi appartenenti a cluster distinti



- … sia per rappresentazioni basate su prototipi
  - ✓La coesione è la somma dei pesi degli archi tra i nodi appartenenti a un cluster e il relativo centroide
  - ✓La separazione è la somma dei pesi degli archi tra i centroidi



Misure interne: Coesione e Separazione

- Coesione e separazione possono essere calcolate sia per rappresentazioni basate su grafi…
  - ✓La coesione è la somma dei pesi degli archi tra i nodi appartenenti a un cluster
  - ✓La separazione è la somma dei pesi degli archi tra i nodi appartenenti a cluster distinti

$$cohession(C_i) = \sum_{\mathbf{x} \in C_i, \mathbf{y} \in C_i} proximity(\mathbf{x}, \mathbf{y})$$

$$separation(C_i, C_j) = \sum_{\mathbf{x} \in C_i} \sum_{\mathbf{y} \in C_j} proximity(\mathbf{x}, \mathbf{y})$$

- … sia per rappresentazioni basate su prototipi
  - ✓La coesione è la somma dei pesi degli archi tra i nodi appartenenti a un cluster e il relativo centroide
  - ✓La separazione è la somma dei pesi degli archi tra i centroidi

$$cohession(C_i) = \sum_{\mathbf{x} \in C_i} proximity(\mathbf{x}, \mathbf{c_i})$$

$$separation(C_i, C_j) = proximity(\mathbf{c_i}, \mathbf{c_j})$$

$$separation(C_i) = proximity(\mathbf{c_i}, \mathbf{c})$$

  - ✓La separazione tra due protipi e tra un prototipo e il centroide dell'intero dataset sono correlati

Misure interne: Coesione e Separazione

- Le formule precedenti vanno poi generalizzate per considerare tutti i cluster che compongono il clustering

$$validity\ measure = \sum_{i=1}^{K} w_i \cdot validity(C_i)$$

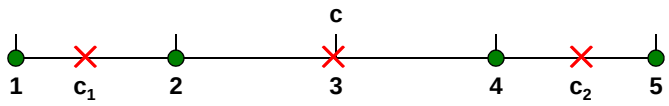
- Diverse sono le misure di prossimità utilizzabili. Se si utilizza SSE, in una rappresentazione basata su centroidi, le formule precedenti diventano:
  - ✓ SSB= Sum of Squared Between group

$$SSE = \sum_i cohession(C_i) = \sum_i \sum_{\mathbf{x} \in C_i} (\mathbf{x} - \mathbf{c_i})^2$$

$$SSB = \sum_i separation(C_i) = |C_i| (\mathbf{c_i}, \mathbf{c})^2$$

- E' possibile dimostrare che SSE+SSB=costante. Quindi minimizzare la coesione corrisponde a massimizzare la separazione

Misure interne: Coesione e Separazione



K=1 cluster:

$$SSE = (1-3)^2 + (2-3)^2 + (4-3)^2 + (5-3)^2 = 10$$
$$SSB = 4 \times (3-3)^2 = 0$$
$$Totale = 10 + 0 = 10$$

K=2 clusters:

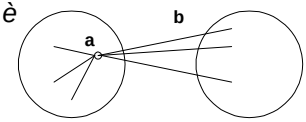
$$SSE = (1-1.5)^2 + (2-1.5)^2 + (4-4.5)^2 + (5-4.5)^2 = 1$$
$$SSB = 2 \times (3-1.5)^2 + 2 \times (4.5-3)^2 = 9$$
$$Totale = 1 + 9 = 10$$

Misure interne: silhouette

- Combina la misura di coesione e separazione
- Dato un punto *i* appartenente al cluster *C*

$$a_i = \operatorname{avg}_{j \in C}(\text{dist}(i,j)) \qquad b_i = \min_{C' \neq C}(\operatorname{avg}_{j \in C'}(\text{dist}(i,j))$$

- Il coefficiente di silhouette per il punto *i* è



$$s_i = (b_i - a_i) / \max(a_i, b_i)$$

  - ✓ Varia tra -1 and 1.
  - ✓ E' auspicabile che il coefficiente sia quanto più possibile vicino a 1 il che implica a_i piccolo (cluster coesi) e b_i grande (cluster ben separati)

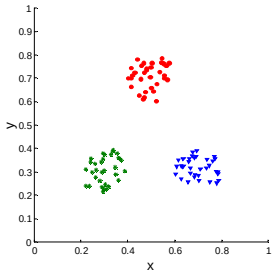
- Il coefficiente può essere mediato su tutti i punti per calcolare la silhouette dell'intero clustering

Misurare la validità per mezzo della correlazione

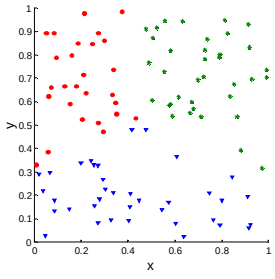
- Si utilizzano due matrici
  - ✓ Proximity Matrix
    - Matrice delle distanze tra gli elementi
  - ✓ “Incidence” Matrix
    - Una riga e una colonna per ogni elemento
    - La cella è posta a 1 se la coppia di punti corrispondenti appartiene allo stesso cluster
    - La cella è posta a 0 se la coppia di punti corrispondenti appartiene a cluster diversi
- Si calcola la correlazione tra le due matrici
- Una correlazione elevata indica che punti che appartengono allo stesso cluster sono vicini
- Non rappresenta una buona misura per cluster non sferici (ottenuti con algoritmi density based o con misure di contiguità)
  - ✓ In questo caso le distanze tra i punti non sono correlate con la loro appartenenza allo stesso cluster

Misurare la validità per mezzo della correlazione

- Correlazione tra matrice di incidenza e matrice di prossimità per il risultato dell’algoritmo k-means sui seguenti data set.
  - ✓ La correlazione è negativa perchè a distanze piccole nella matrice di prossimità corrispondono valori grandi (1) nella matrice di incidenza
  - ✓ Ovviamente, se si fosse usata la matrice delle distanze al posto della matrice di simlarità la correlazione sarebbe stata positiva



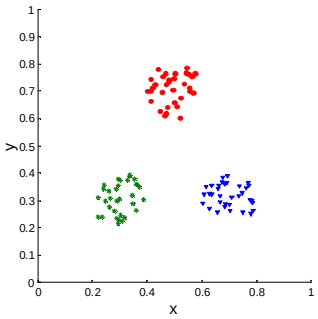
Corr = -0.9235

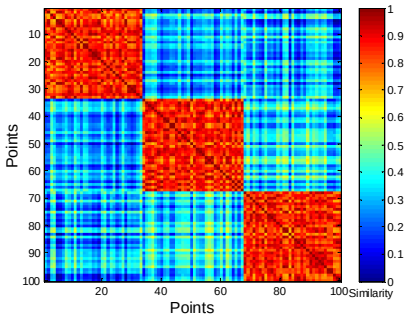


Corr = -0.5810

Misurare la validità per mezzo della matrice di similarità

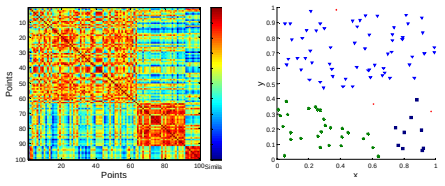
- La visualizzazione si ottiene ordinando la matrice di similarità in base ai raggruppamenti dettati dai cluster.



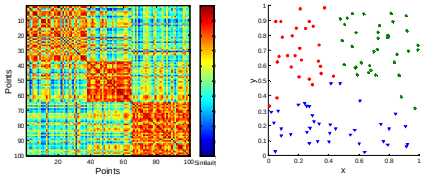


Misurare la validità per mezzo della matrice di similarità

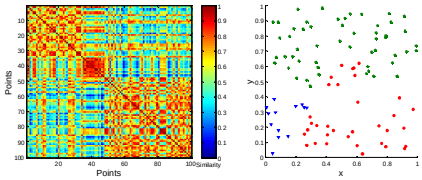
- Se i dati sono distribuiti uniformemente la matrice è più “sfumata”



DBSCAN



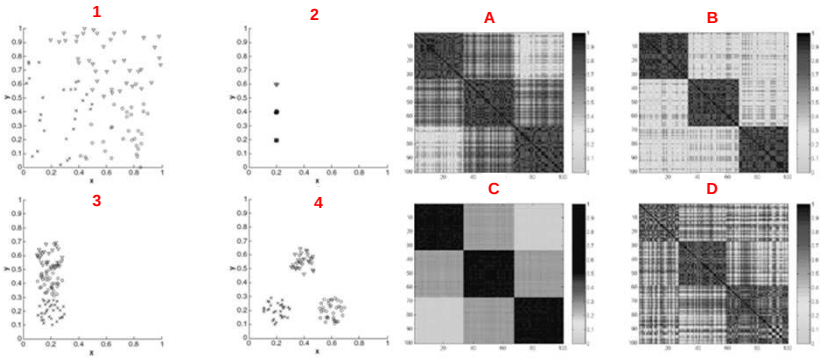
K-means



Complete link

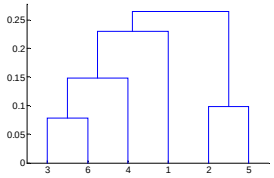
Esercizo

- Associa le matrici di similarità ai data set



Cophenetic distance

- Le misure precedenti rappresentano degli indici di validità per algoritmi di clustering partizionanti.
- Nel caso di clustering gerarchico una misura spesso utilizzata è la cophenetic distance ossia, dati due elementi, è la prossimità a cui sono posti assieme da un algoritmo di clustering agglomerativo
  - ✓ Nel dendrogramma sottostante, le coppie di punti (3,4), (6,4) hanno distanza 0.15 poichè i cluster a cui appartengono sono fusi in corrispondenza di quel valore della proximity matrix



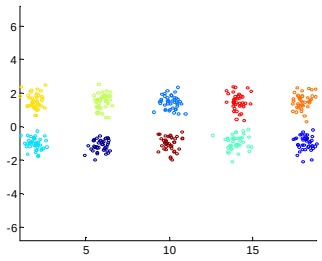
Ward's

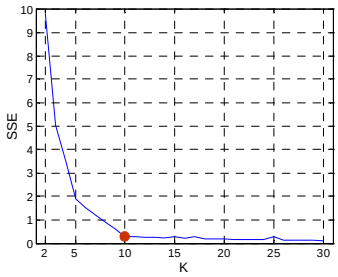
Tecnica	CPCC
Single link	0.44
Complete link	0.63
Group average	0.66
Ward's	0.64

- Calcolando la cophenetic distance per tutte le coppie di punti si ottiene una matrice che permette di calcolare il CoPhenetic Correlation Coefficient (CPCC)
  - ✓ Il CPCC è l'indice di correlazione tra la matrice delle cophenetic distance e la matrice di dissimilarità dei punti
  - ✓ Una elevata correlazione indica che l'algoritmo di clustering ha rispettato la dissimilarità tra gli elementi

Stima del numero di cluster

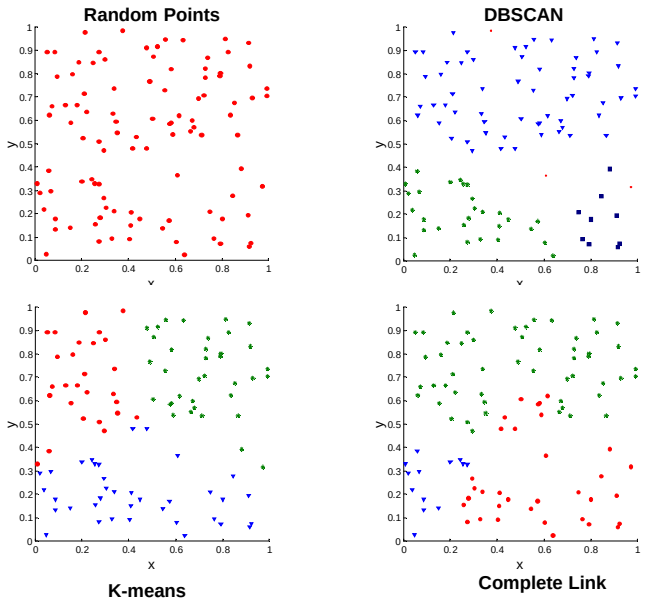
- Una stima del numero naturale di cluster può essere ottenuta tramite le misure di validazione
  - ✓ Il cambio di pendenza indica che il numero di cluster è sufficiente a rappresentare efficacemente i dati





- L'informazione rappresenta puramente una indicazione del numero corretto
  - ✓ Potrebbero esistere dei cluster non catturati dall'algoritmo utilizzato
  - ✓ I cluster naturali potrebbero essere nested

Cluster trovati in dati random



Propensione al clustering

- Un modo ovvio per verificare se un dataset presenta dei cluster è quello di clusterizzarlo
  - ✓ Come già visto gli algoritmi di clustering troveranno comunque cluster nei dati
  - ✓ Potrebbero esistere tipi di cluster non identificati dall'algoritmo applicato
    - Si potrebbero utilizzare algoritmi diversi
- In alternativa è possibile utilizzare degli indici statistici che stimano quanto i dati sono distribuiti in modo uniforme
  - ✓ Applicabili prevalentemente con una ridotta dimensionalità e in spazi euclidei
  - ✓ Uno di questi è la **statistica di Hopkins**

Statistica di Hopkins

- Dato un dataset D di *n* punti reali, si genera un dataset P di *m* punti (*m*<<*n*) distribuendoli in modo random nello spazio dei dati.
- Per ogni punto $\mathbf{p_i} \in$ P si calcola la distanza u_i dal punto al suo nearest-neighbor nel dataset reale D.
- Si campionano altri *m* punti dal data set reale $\mathbf{q_i} \in$ D e per ognuno di essi si calcola la distanza w_i dal punto al suo nearest-neighbor nel dataset reale D.
- La statistica di Hopkins è definita come:

$$H = \frac{\sum_{i=1}^{m} w_i}{\sum_{i=1}^{m} u_i + \sum_{i=1}^{m} w_i}$$

  - ✓ Se la distanza dal nearest neighbor è più o meno la stessa sia per i punti generati sia per quelli campionati allora H≈0.5. Quindi il data set ha una distribuzione random
  - ✓ Se H ≈ 1 (valori di $\mathbf{u_i}$ molto piccoli) i dati sono ben clusterizzati
  - ✓ Se H ≈ 0 (valori di $\mathbf{w_i}$ molto piccoli) i dati sono distribuiti in modo regolare (equi spaziati).

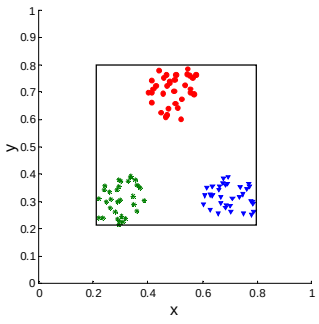
Verificare empiricamente che posizionando regolarmente i punti si minimizza la media delle distanze

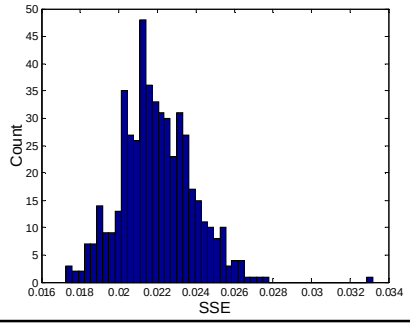
Framework statistico per la validazione del clustering

- Le tecniche precedenti restituiscono spesso delle misure che devono essere interpretate
  - ✓ Un valore pari a 10 è buono o cattivo?
  - ✓ L'analisi del valore rispetto ai valori minimi e massimi ottenibili danno qualche indicazione ma non risolvono il problema
- La statistica può fornire una metodologia adeguata per valutare la bontà di una misura
  - ✓ Siamo alla ricerca di pattern non casuali, quindi più "atipico" sarà il risultato che otterremo, più probabile sarà che esso rappresenti una struttura non casuale nei dati
  - ✓ L'idea è quindi quella di comparare il valore della misura con quello ottenuto a partire da dati random.
    - Se il valore della misura ottenuto sui dati è improbabile sui dati random allora il clustering è valido
- La problematica dell'interpretazione del valore della misura è meno pressante quando si compara il risultato di due clustering
  - ✓ In ogni caso è importante capire quanto la differenza dei valori implichi una sostanziale differenza tra i clustering

Framework statistico: un esempio

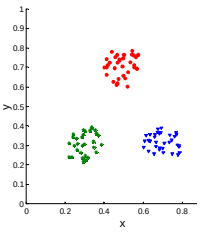
- Il valore per SSE dei tre cluster sotto riportati è 0.005
- L'istogramma riporta la distribuzione di SSE per dati random
  - ✓ 500 punti random con valori di x e y tra 0.2 e 0.8
  - ✓ I dati sono clusterizzati in 3 cluster con K-means

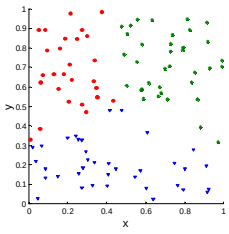


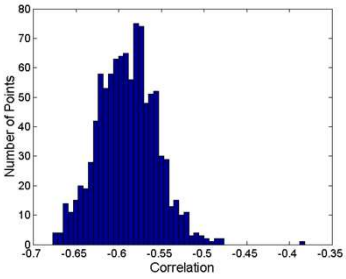


Framework statistico: un esempio

- Il valore di correlazione dei tre cluster sotto riportati è -0.9235
- L'istogramma riporta la distribuzione dell'indice di correlazione tra la matrice di incidenza e la matrice di prossimità per dati random
  - ✓ 500 punti random con valori di x e y tra 0.2 – 0.8
  - ✓ I dati sono clusterizzati in 3 cluster con K-means







Corr = -0.9235 **Corr = -0.5810**

Misure esterne per la validazione del clustering: Entropia e Purezza

- Le informazioni esterne sono solitamente le etichette di classe degli oggetti su cui si esegue il clustering
  - ✓ Permettono di misurare la corrispondenza tra l'etichetta calcolata del cluster e l'etichetta della classe
- Se si hanno a disposizione le etichette di classe perchè eseguire il clustering?
  - ✓ Comparare il risultato di tecniche di clustering diverse
  - ✓ Valutare la possibilità di ottenere automaticamente una classificazione altrimenti manuale
- Due approcci sono possibili
  - ✓ Orientati alla classificazione: valutano in che misura i cluster contengono oggetti appartenenti alla stessa classe
    - Entropia, purezza, F-measure
  - ✓ Orientati alla similarità: misurano con che frequenza due oggetti che appartengono allo stesso cluster appartengono alla stessa classe
    - Utilizzano misure di similarità per dati binari: Jaccard

Misure esterne per la validazione del clustering: Entropia e Purezza

- Entropia: per ogni cluster *i* sia $p_{ij} = m_{ij} / m_i$ la probabilità che un membro del cluster *i* appartenga alla classe *j*
  - ✓ m_i = num di oggetti nel cluster *i*
  - ✓ m_{ij} =num oggetti della classe *j* nel cluster *i*
- L'entropia per il cluster *i* e per l'intero clustering

$$e_i = -\sum_{j=1}^{L} p_{ij} \log_2 p_{ij} \qquad e = -\sum_{i=1}^{K} \frac{m_i}{m} e_i$$

  - ✓ *L*=num classi *K*=num cluster

- Purezza è calcolata come:

$$p_i = \max_j p_{ij} \qquad purity = \sum_{i=1}^{K} \frac{m_i}{m} p_i$$

  - ✓ p_i misura quanto è "forte" la relazione tra il cluster e una classe.
    - E' minima quando i punti appartenti al cluster sono equamente distribuiti tra le classi

Misure esterne per la validazione del clustering: Entropia e Purezza

K-means Clustering Results for LA Document Data Set

Cluster	Entertainment	Financial	Foreign	Metro	National	Sports	Entropy	Purity
1	3	5	40	506	96	27	1.2270	0.7474
2	4	7	280	29	39	2	1.1472	0.7756
3	1	1	1	7	4	671	0.1813	0.9796
4	10	162	3	119	73	2	1.7487	0.4390
5	331	22	5	70	13	23	1.3976	0.7134
6	5	358	12	212	48	13	1.5523	0.5525
Total	354	555	341	943	273	738	1.1450	0.7203

Impostare il calcolo dell'entropia per il cluster 1

Commento finale sull'analisi della validità dei cluster

"The validation of clustering structures is the most difficult and frustrating part of cluster analysis.

Without a strong effort in this direction, cluster analysis will remain a black art accessible only to those true believers who have experience and great courage."

Algorithms for Clustering Data, Jain and Dubes