Modello generatore dei dati

- Dato un insieme di1-D puntiX = {x_1, …, x_n} per la clustering analysis ed assumendo che sono generati da una distribuzione Gaussiana:

$$\mathcal{N}(\mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

- La probabilità che una unità $x_i \in X$ sia generata dal modello:

$$P(x_i|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}}$$

- La verosimiglianza che *X* sia generata dal modello:

$$L(\mathcal{N}(\mu, \sigma^2) : X) = P(X|\mu, \sigma^2) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}}$$

- L'attività di apprendere dal modello generatore dei dati: trovare I parametri μ e σ^2 tali che

the maximum likelihood

$$\mathcal{N}(\mu_0, \sigma_0^2) = \arg\max\{L(\mathcal{N}(\mu, \sigma^2) : X)\}$$

Clustering probabilistico gerarchico

- Per un insieme di unità ripartito in *m* gruppi $C_1, \ldots, C_m$, la qualità può essere misurata da:

$$Q(\{C_1, \ldots, C_m\}) = \prod_{i=1}^{m} P(C_i)$$

  Dove *P*() è la massima verosimiglianza

- Distanza tra I gruppi C_1 and C_2: $dist(C_i, C_j) = -\log \frac{P(C_1 \cup C_2)}{P(C_1)P(C_2)}$
- Algoritmo: Prograssivamente unisci punti e gruppi

  Input: $D = \{o_1, \ldots, o_n\}$: un data set contenente n unità

  Output: Una gerarchia di gruppi

  Metodo

  Creare un gruppo per ogni unità $C_i = \{o_i\}$, $1 \le i \le n$;

  For i = 1 to n {

  Trova coppie di gruppi C_i and C_j tali che

  $C_i, C_j = \text{argmax}_{i \neq j} \{\log (P(C_i \cup C_j)/(P(C_i)P(C_j))\}$;

  If $\log (P(C_i \cup C_j)/(P(C_i)P(C_j)) > 0$ then merge C_i and C_j }

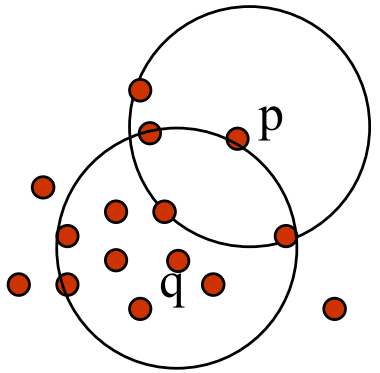
Metodi di Clustering basati sulle distribuzioni dei dati osservati

- Raggruppamento basato sulla densità (criterio cluster locale)
- Principali caratteristiche:
  - Individuare i gruppi di forma arbitraria
  - Tenere conto del'errore di misura
  - Una sola scansione
  - Necessita di parametri di densità come condizione di stop
- Diversi lavori interessanti:
  - DBSCAN: Ester, et al. (KDD'96)
  - OPTICS: Ankerst, et al (SIGMOD'99).
  - DENCLUE: Hinneburg e D. Keim (KDD'98)
  - CLIQUE: Agrawal, et al. (SIGMOD'98) (più basato sulla griglia)

Concetti di base dei metodi basati sulle distribuzioni dei dati osservati

- 2 parametri*:*
  - *Eps*: raggio Massimo del vicinato
  - *MinPts*: Numero minimo di unità nello Eps-neighbourhood di quella unità
- $N_{Eps}(p)$: {q appartiene a D | dist(p,q) ≤ Eps}
- Direttamente raggiungibile per densità: Una unità *p* è direttamente raggiungibile per densitàda una unità *q* w.r.t. *Eps*, *MinPts* se
  - *p* appartine a $N_{Eps}(q)$
  - Condizione del punto centrale:

$$|N_{Eps}(q)| \geq MinPts$$

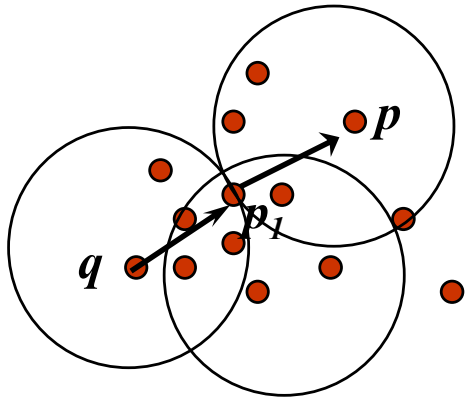


MinPts = 5

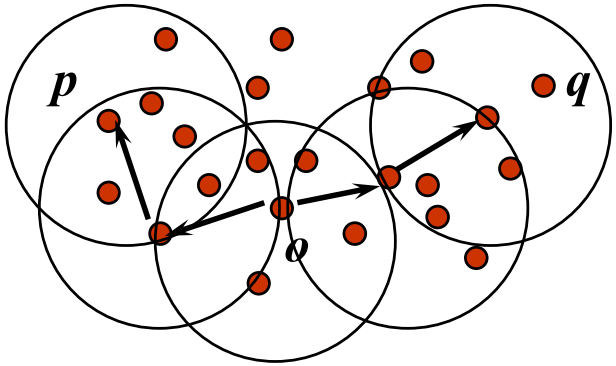
Eps = 1 cm

Concetti di base dei metodi basati sulle distribuzioni dei dati osservati

- Raggiungibile per densità:
  - Una unità *p* è raggiungibile per densità da una unità *q* w.r.t. *Eps*, *MinPts* se esiste una sequenza di unità $p_1, \ldots, p_n$, $p_1 = q$, $p_n = p$ tali che p_{i+1} è direttamente raggiungibile per densità da p_i

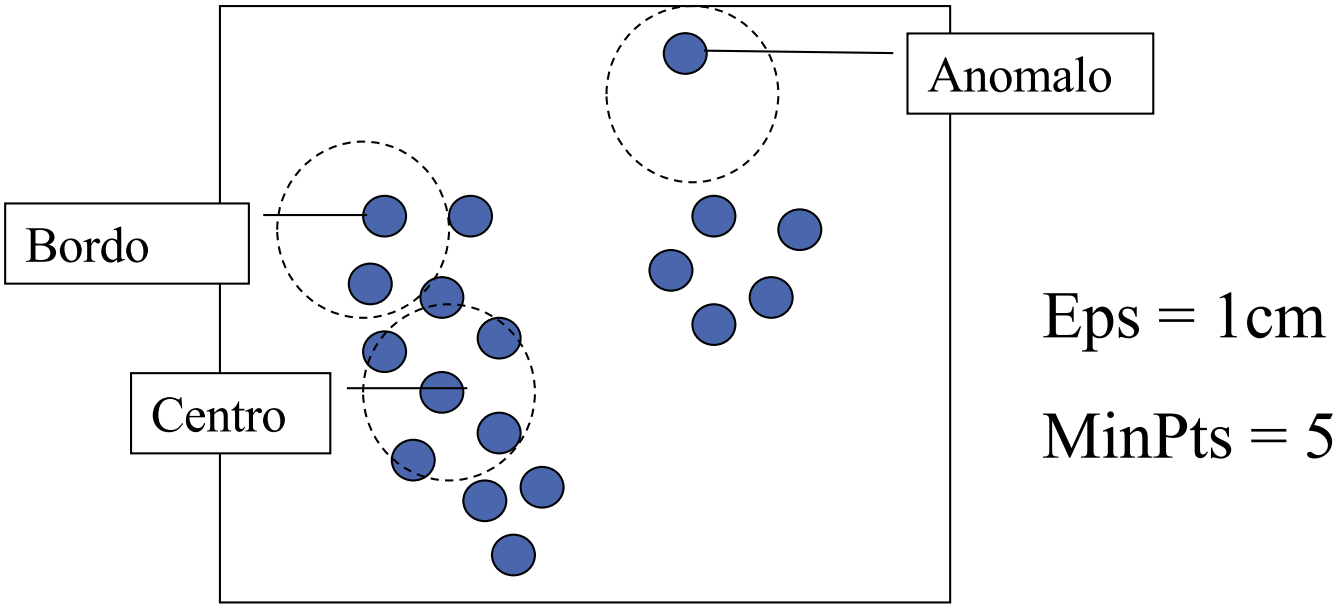


- Connesso per densità
  - Una unità *p* è connessa per densità ad una unità *q* w.r.t. *Eps*, *MinPts* se esiste una unità *o* tale che entrambe, *p* e *q* siano direttamente raggiungibili per densità da *o* w.r.t. *Eps* e *MinPts*



DBSCAN: Raggruppamento spaziale basato su densità per applicazioni con dati affetti da errore di misura

- Si basa su una nozione di gruppo basata sulla densità: un gruppo è definito come un insieme massimo di punti connessi alla densità
- Individua i gruppi di forma arbitraria nei database spaziali con il rumore



DBSCAN: l'algoritmo

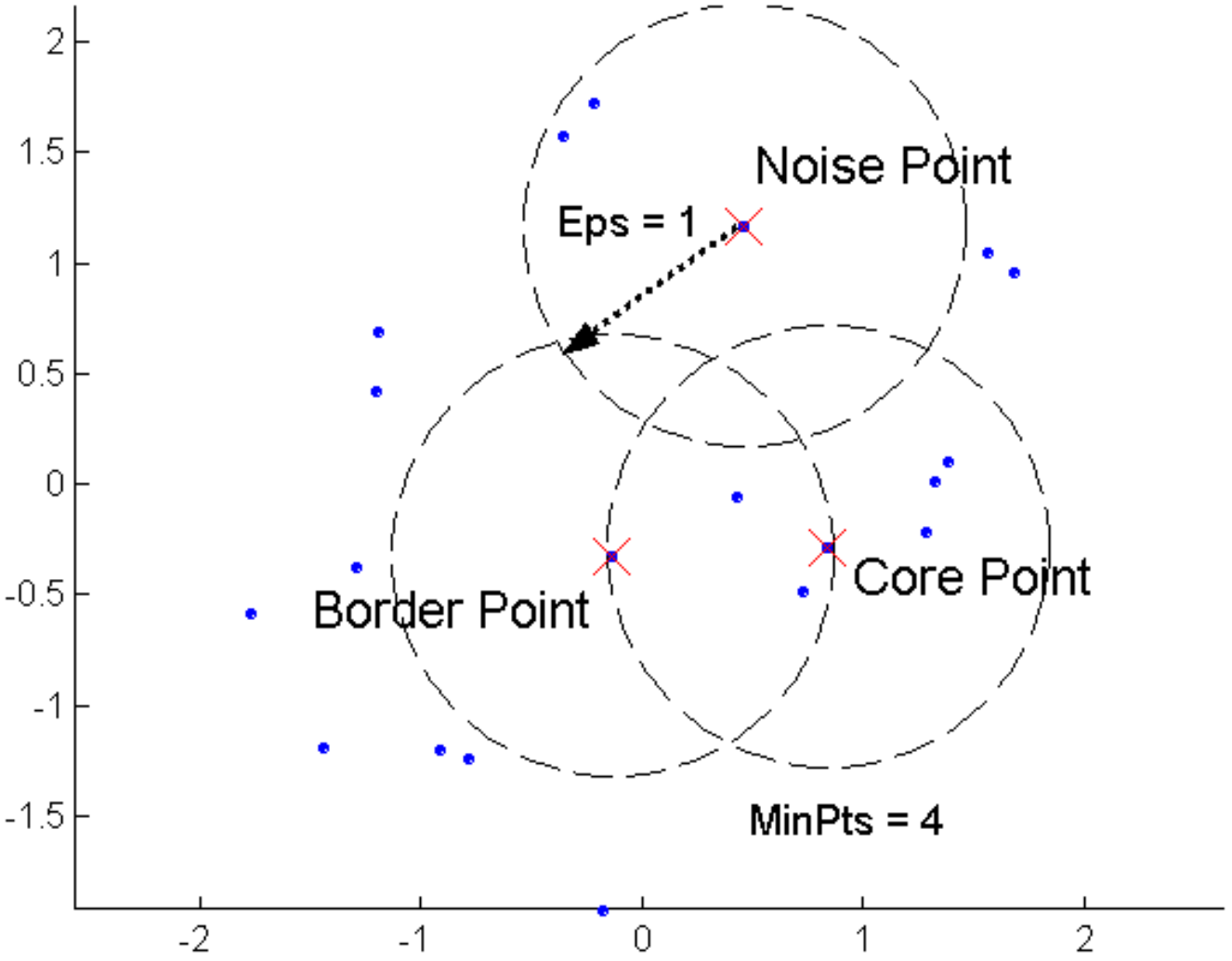
- Arbitrario seleziona una unità *p*
- Recupera tutti i punti con densità raggiungibile da *p* w.r.t. *Eps* e *MinPts*
- Se *p* è una unità centrale, viene formato un gruppo
- Se *p* è una unità di confine, nessuna unità è raggiungibile in densità da *p* e DBSCAN visita l'unità successiva del database
- Continua il processo fino a quando tutte le unità sono state elaborate

DBSCAN: l'algoritmo

- DBSCAN : density-based algorithm.
  - Densità = numero di punti entro un raggio specificato (Eps)
  - Un punto è definito core point se ha più di un numero specificato di punti (MinPts) entro Eps
    - Punti all'interno del cluster
  - Un border point ha meno punti di MinPts entro Eps, ma è nel vicinato di un core point
  - Un noise point è ogni altro punto che non è né core point o border point.

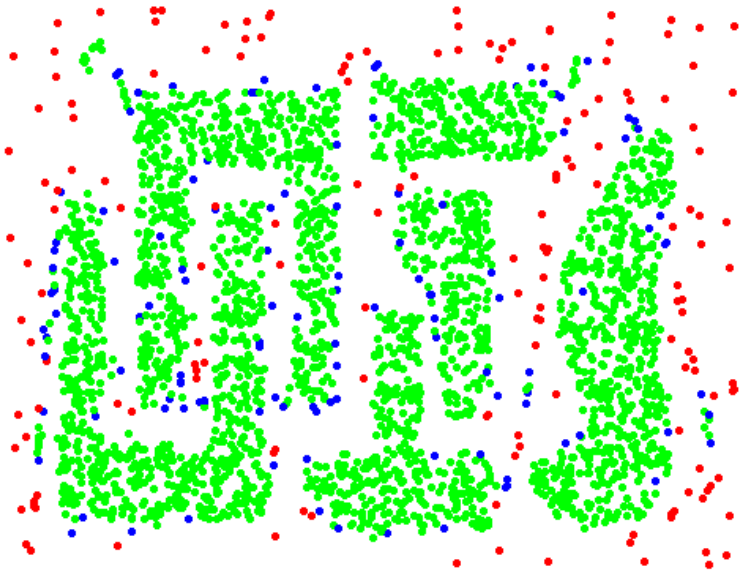
- Elimina i noise points
- Aggrega in cluster i punti rimanenti
  - Gruppi formati da core point e border point collegati tra loro

DBSCAN: Core, Border, e Noise Points



DBSCAN: Core, Border, e Noise Points

Punti originari

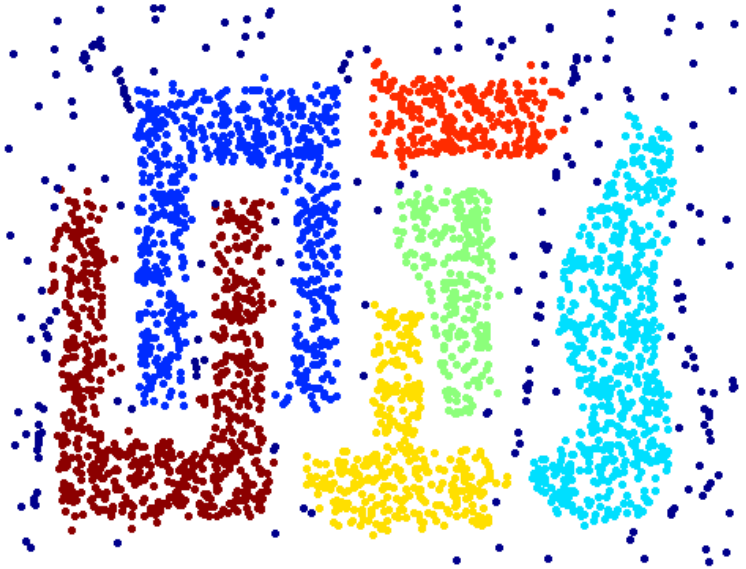


Tipi: core, border e noise

Eps = 10, MinPts = 4

DBSCAN: quando funziona

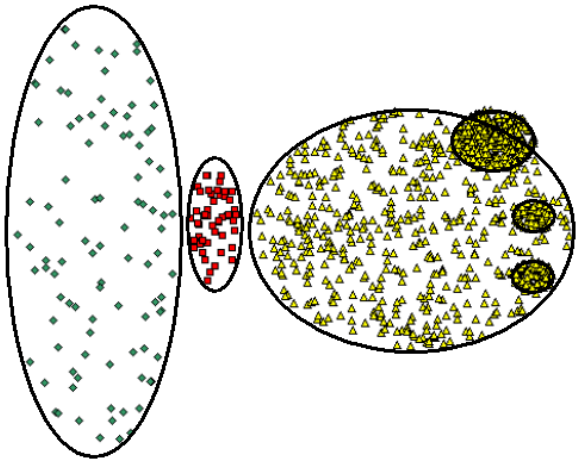
Punti originari



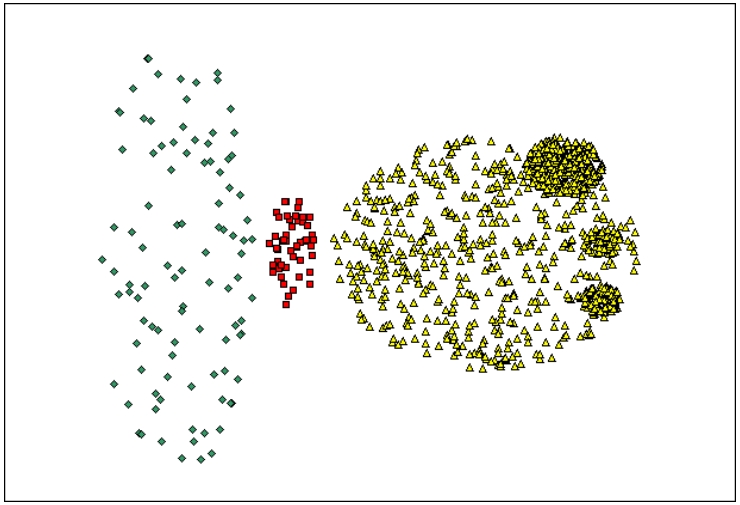
Cluster

- Resistente al rumore statistico
- Può gestire gruppi di forma e dimensione differenti

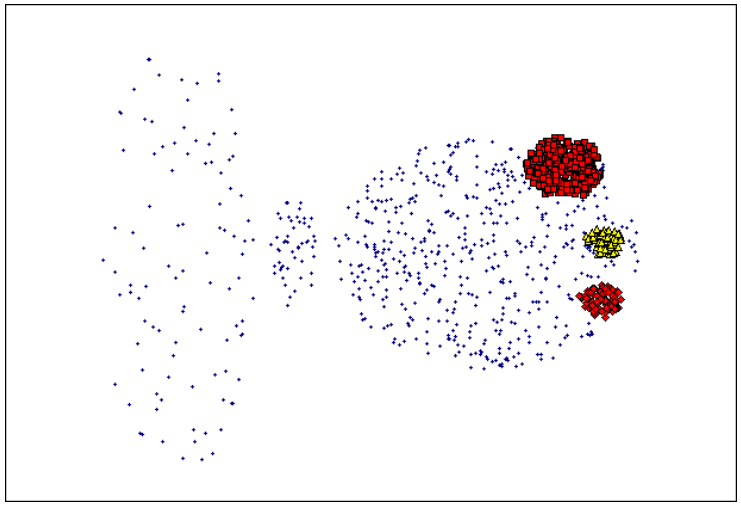
DBSCAN: quando NON funziona



Punti originari



(MinPts=4, Eps=9.75).

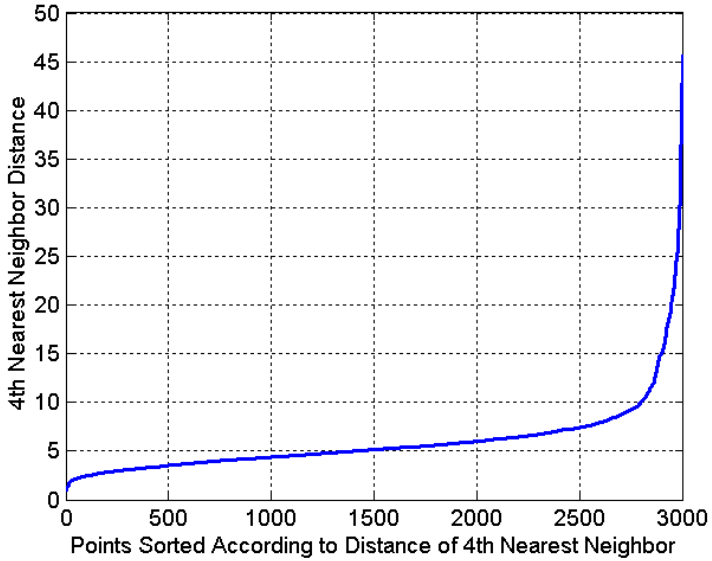


(MinPts=4, Eps=9.92)

- Densità variabili
- Dati ultra-dimensionali

DBSCAN: Determinare EPS e MinPts

- Idea: per i punti in un cluster, il loro k^{th} nearest neighbor è approssimativamente alla stessa distanza
- I Noise points hanno il k^{th} nearest neighbor a distanza maggiore
- Disegna su un grafico la distanza di ogni punto dal suo k^{th} nearest neighbor

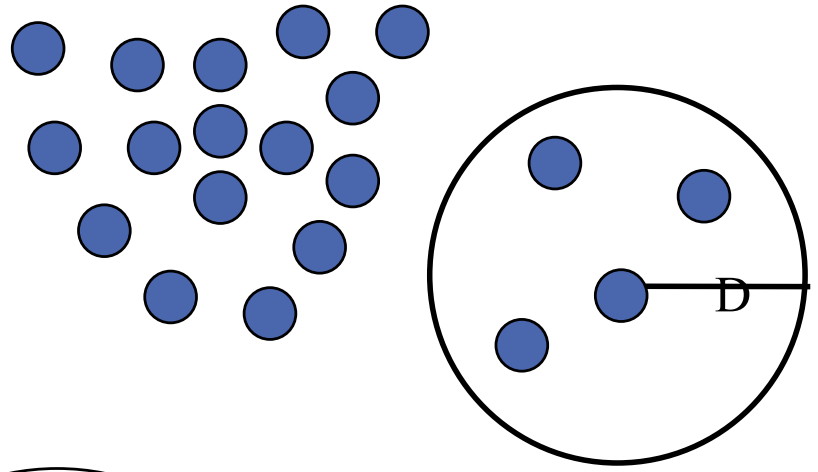


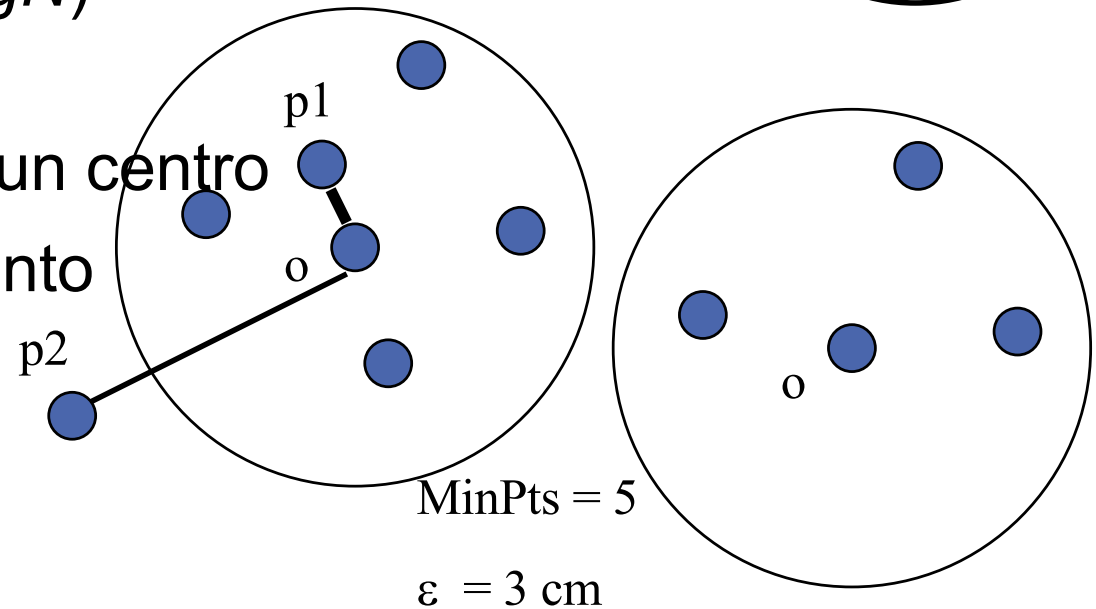
OPTICS: Un Metodo per ordinamento-raggruppamento

- OPTICS: punti di ordinamento per identificare la struttura dei gruppi
- Ankerst, Breunig, Kriegel e Sander (SIGMOD'99)
- Produce un ordine speciale del database rispetto alla sua struttura in gruppi basata sulla densità
- Questo ordinamento-gruppo contiene informazioni equivalenti ai raggruppamenti basati sulla densità corrispondenti ad un ampio intervallo di impostazioni dei parametri
- Ottimo per l'analisi automatica e interattiva dei gruppi, inclusa la ricerca di una struttura di gruppi intrinseca
- Può essere rappresentato graficamente o utilizzando tecniche di visualizzazione

OPTICS: Un Metodo per ordinamento-raggruppamento

- Index-based:
  - k = numero di dimensioni
  - N = 20
  - p = 75%
  - M = N(1-p) = 5
  - Complessità: O(*NlogN*)
- Distanza centrale:
  - min eps s.t. l'unità è un centro
- Distanza di raggiungimento

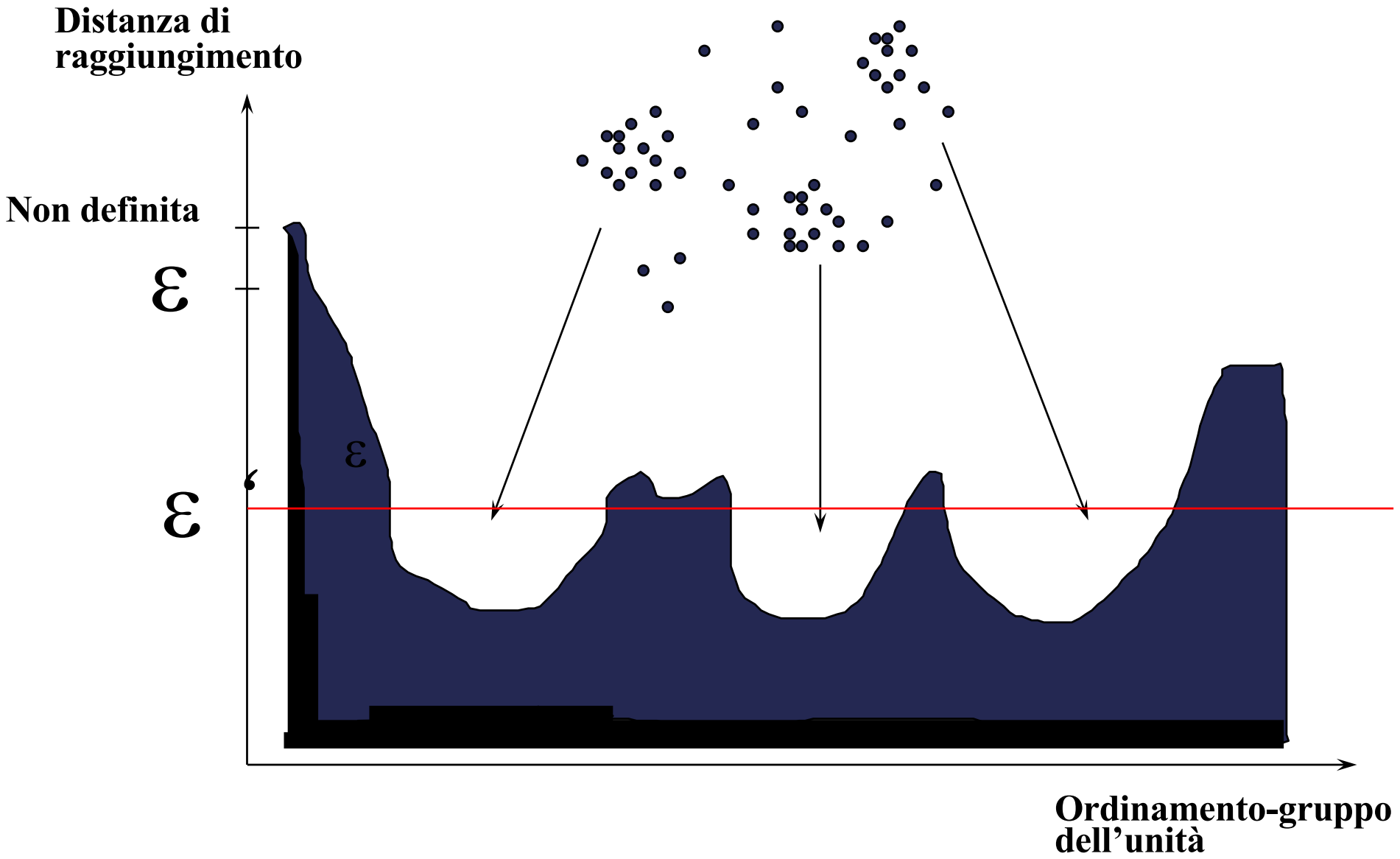




Max (core-distance (o), d (o, p))

r(p1, o) = 2.8cm. r(p2,o) = 4cm

OPTICS: Un Metodo per ordinamento-raggruppamento



DENCLUE

- DENsity-based CLUstEring by Hinneburg & Keim (KDD'98)
- Usa la funzione di densità:

$$f_{Gaussian}(x,y) = e^{-\frac{d(x,y)^2}{2\sigma^2}}$$

Influenza di y su x

$$f^D_{Gaussian}(x) = \sum_{i=1}^{N} e^{-\frac{d(x,x_i)^2}{2\sigma^2}}$$

Influenza totali su x

$$\nabla f^D_{Gaussian}(x,x_i) = \sum_{i=1}^{N} (x_i - x) \cdot e^{-\frac{d(x,x_i)^2}{2\sigma^2}}$$

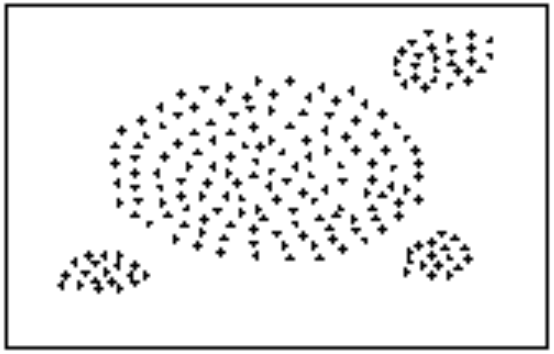
Gradiente di x nella direzione di x_i

- Principali caratteristiche
  - Solide basi matematiche
  - Buono per i dataset con grandi errori di misura
  - Consente una descrizione matematica compatta di gruppi di forma arbitraria in serie di dati di alta dimensione
  - Significativamente più veloce altri algoritmi esistenti (ad es. DBSCAN)
  - Ma ha bisogno di un gran numero di parametri

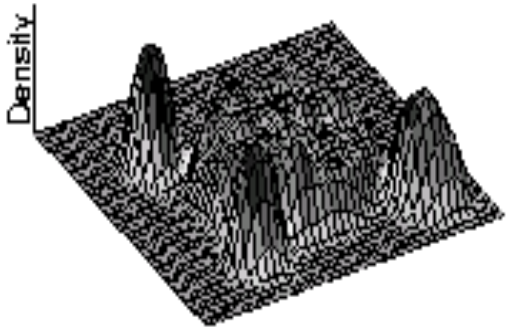
DENCLUE: Technical Essence

- Utilizza una griglia ma conserva solo informazioni sulle celle della griglia che contengono effettivamente unità statistiche e gestiscono queste celle in una struttura di accesso basata su alberi
- Funzione Influenza: descrive l'impatto di una unità all'interno del suo vicinato
- La densità complessiva dello spazio dei dati può essere calcolata come somma della funzione di influenza di tutte le unità
- I gruppi possono essere determinati matematicamente identificando gli attrattori di densità
- Gli attrattori di densità sono il massimo locale della funzione di densità complessiva
- Gruppi definiti dal centro: assegnare ad ogni attrattore di densità la densità dei unità da esso attratte
- Gruppi di forma arbitraria: unisce gli attrattori a densità che sono collegati attraverso percorsi ad alta densità (> threshold)

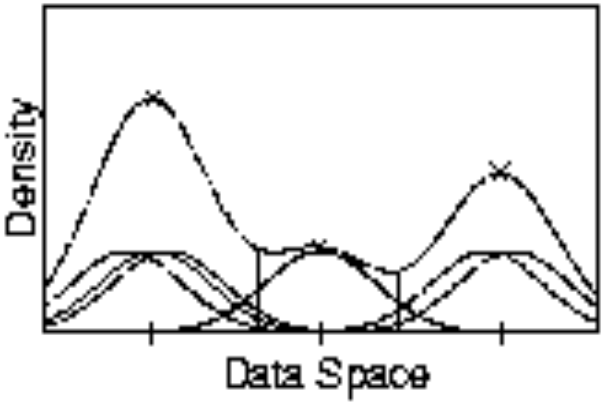
DENCLUE



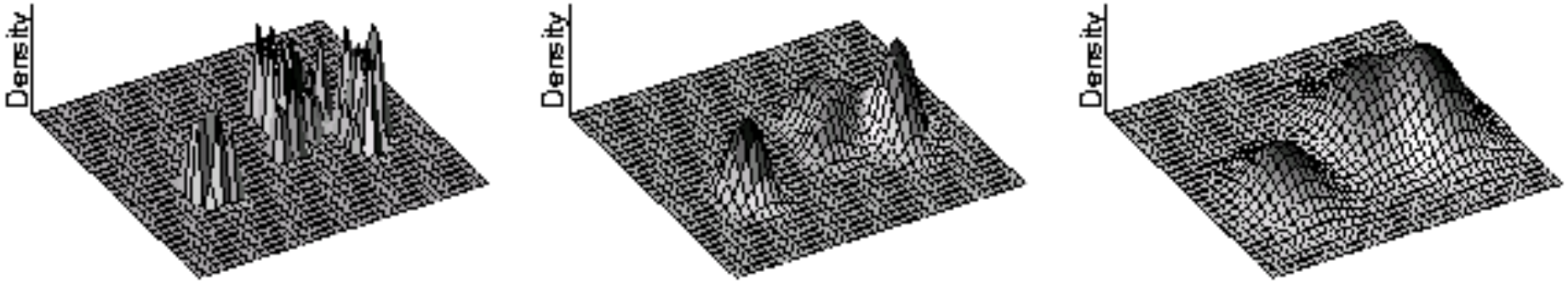
(a) Data Set



(c) Gaussian

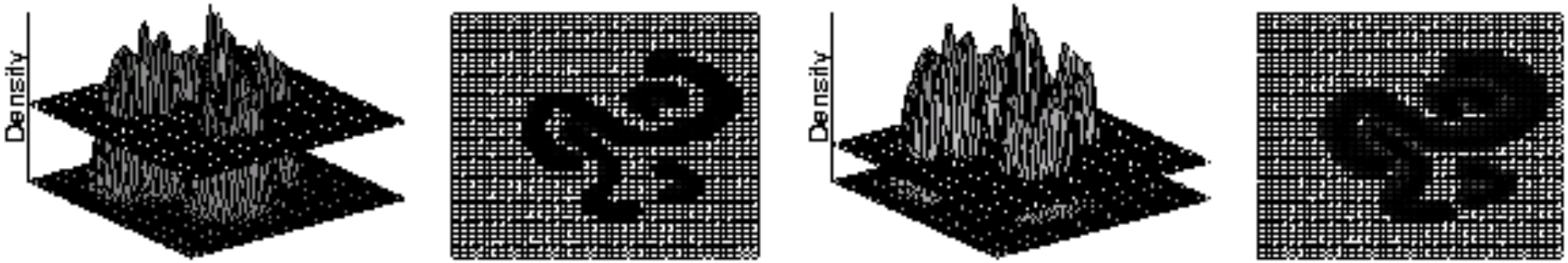


DENCLUE



(a) $\sigma = 0.2$ (b) $\sigma = 0.6$ (d) $\sigma = 1.5$

Figure 3: Example of Center-Defined Clusters for different σ



(a) $\xi = 2$ (b) $\xi = 2$ (c) $\xi = 1$ (d) $\xi = 1$

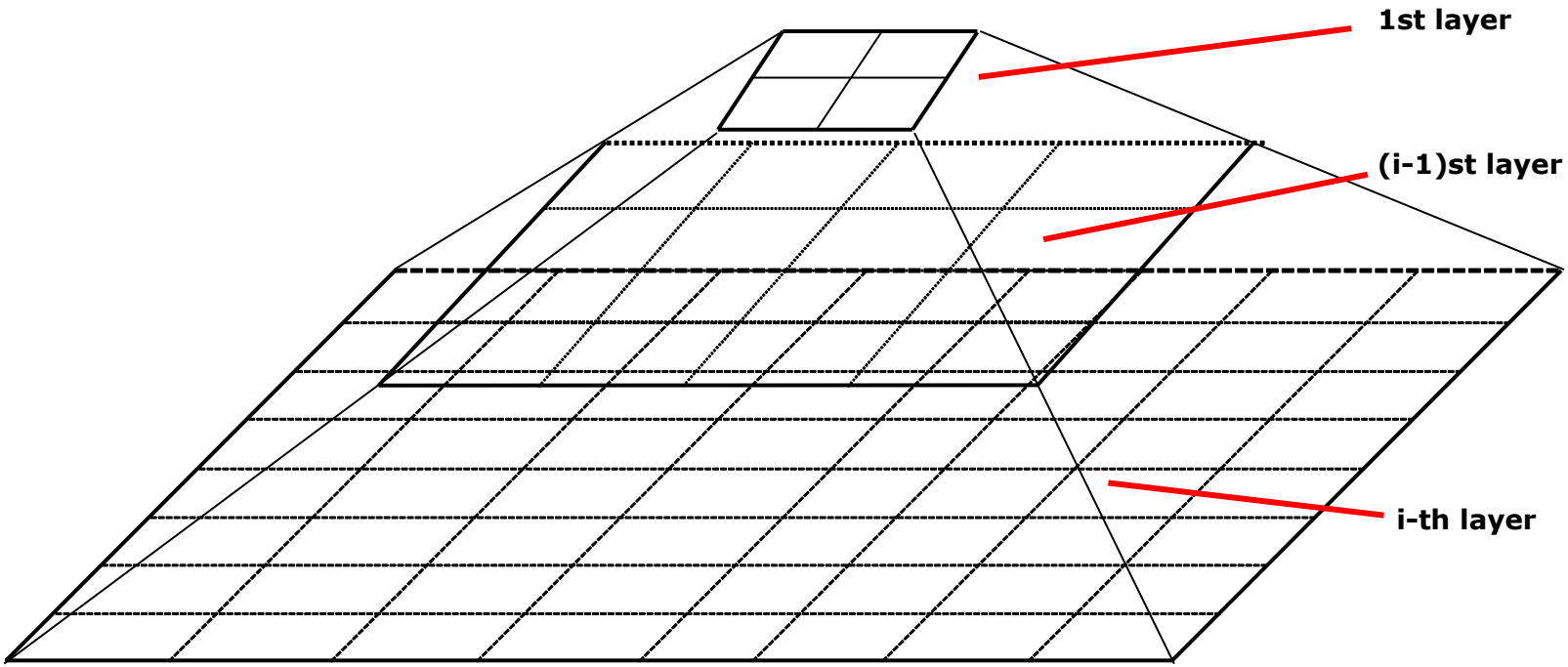
Figure 4: Example of Arbitray-Shape Clusters for different ξ

Metodi basati su Griglie (Grid)

- Utilizzo della struttura dati della griglia multi-risoluzione
- Diversi metodi interessanti
  - STING (approccio alla griglia informaticamente avanzato) di Wang, Yang e Muntz (1997)
  - WaveCluster di Sheikholeslami, Chatterjee e Zhang (VLDB'98)
    - Un approccio di clustering multi-risoluzione che utilizza il metodo wavelet
  - CLIQUE: Agrawal, et al. (SIGMOD'98)
    - Raggruppamento di griglia e sottosistema

Metodi basati su Griglie (Grid)

- Wang, Yang and Muntz (VLDB’97)
- L'area spaziale è divisa in celle rettangolari
- Esistono diversi livelli di celle corrispondenti a diversi livelli di risoluzione



Metodi basati su Griglie (Grid)

- Ogni cella ad un livello più alto viene suddivisa in un numero di celle più piccole nel livello inferiore successivo
- Le informazioni statistiche di ogni cella vengono calcolate e archiviate in precedenza e vengono utilizzate per rispondere alle queries
- I parametri delle celle di livello superiore possono essere facilmente calcolati dai parametri della cella di livello inferiore
  - *conteggi, medie, s, min, max*
  - tipo di distribuzione: *normale, uniforme*, ecc.
- Utilizzare un approccio top-down per rispondere a query di dati spaziali
- Inizia da un livello preselezionato, in genere con un numero ridotto di celle
- Per ogni cella del livello corrente calcola l'intervallo di confidenza

Metodi basati su Griglie (Grid)

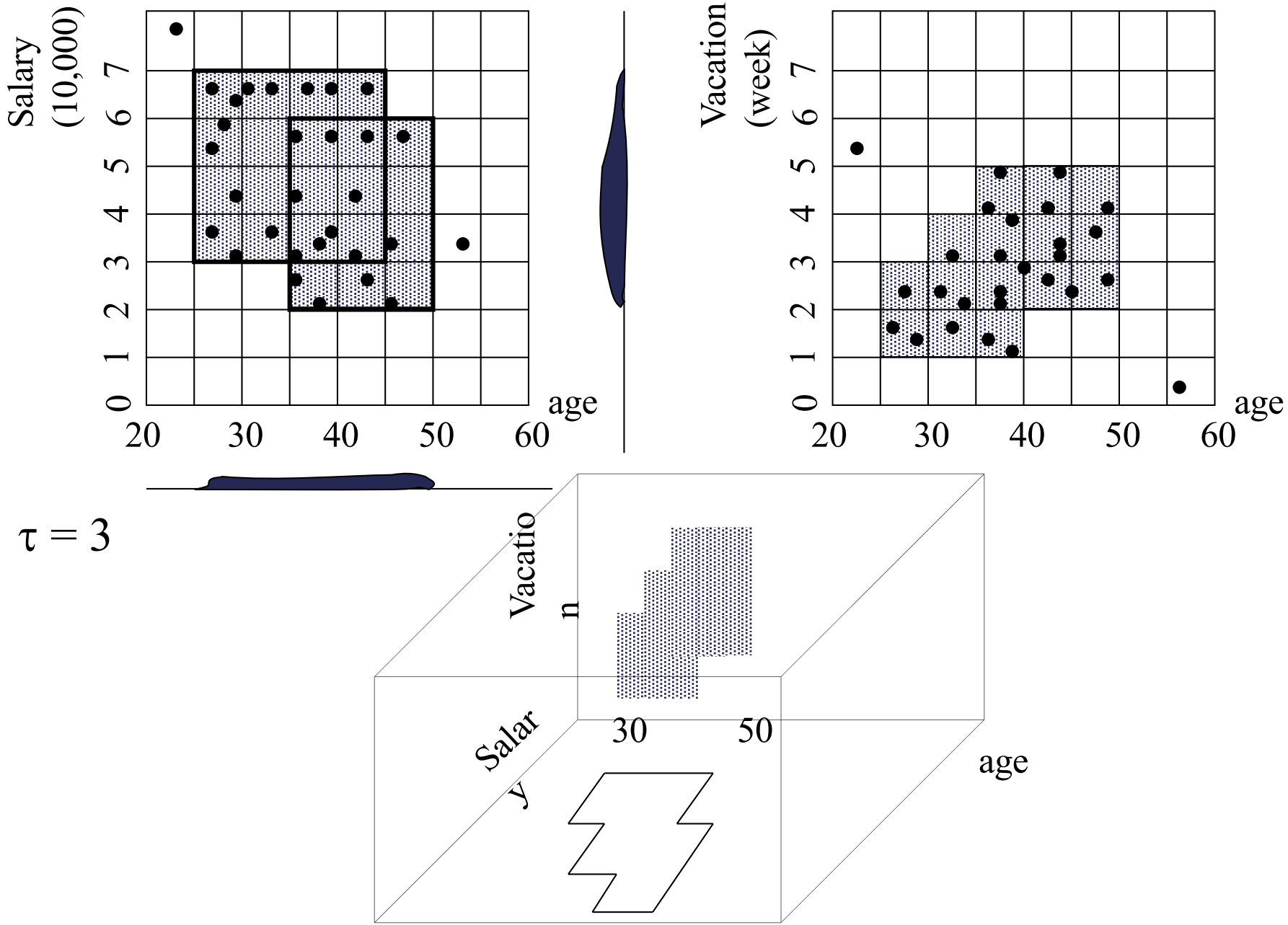
- Rimuovere le celle irrilevanti da ulteriori analisi
- Al termine dell'esame del livello corrente, passa al livello inferiore successivo
- Ripetere questo processo fino a quando non viene raggiunto lo strato inferiore
- vantaggi:
  - Aggiornamento incrementale indipendente da query, facile da parallelizzare
  - O (K), dove K è il numero di celle della griglia al livello più basso
- svantaggi:
  - Tutti i limiti dei gruppi sono orizzontali o verticali e non viene rilevato alcun limite diagonale

CLIQUE (Clustering In QUEst)

- Agrawal, Gehrke, Gunopulos, Raghavan (SIGMOD'98)
- Identificazione automatica di sottospazi di uno spazio dati ad alta dimensione che consente un clustering migliore rispetto allo spazio originale
- CLIQUE può essere considerato sia basato sulla densità sia basato sulla griglia
  - Separa ogni dimensione nello stesso numero di intervalli di lunghezza uguale
  - Divide uno spazio dati m-dimensionale in unità rettangolari non sovrapposte
  - Un'unità è densa se la frazione dei punti di dati totali contenuti nell'unità supera il parametro del modello di input
  - Un gruppo è un insieme massimo di unità densamente collegate all'interno di un sottospazio

CLIQUE: Gli Step principali

- Partizionare lo spazio dati e trovare il numero di punti che si trovano all'interno di ciascuna cella della partizione.
- Identificare i sottospazi che contengono cluster usando il principio *a-priori*
- Identificare i gruppi
  - Determina unità dense in tutti i sottospazi di interesse
  - Determina le unità dense collegate in tutti i sottospazi di interesse.
- Genera una descrizione minima per i gruppi
  - Determinare le regioni massime che coprono un gruppo di unità densamente collegate per ciascun cluster
  - Determinazione della copertura minima per ciascun gruppo


Salary
(10,000)
0 1 2 3 4 5 6 7
20 30 40 50 60
age
Vacation
(week)
0 1 2 3 4 5 6 7
20 30 40 50 60
age
τ = 3
Vacatio
n
Salar
y
30
50
age

Punti di forza e debolezza di *CLIQUE*

- Forza
  - trova automaticamente sottospazi della massima dimensionalità in modo tale che i gruppi ad alta densità esistano in tali sottospazi
  - insensibile all'ordine delle unità in input e non assume la distribuzione di dati
  - scala *linearmente* con la dimensione dell'input e ha una buona scalabilità al crescere del numero di dimensioni nei dati
- Debolezza
  - La precisione del risultato del clustering può essere ridotta a scapito della semplicità del metodo

Metodi più sofisticati di clustering

- Clustering basato sulle densità di probabilità
  - Ogni oggetto può avere una probabilità di appartenere ad un gruppo
- Clustering di grafi e dati di rete
  - Misurazioni di similarità e metodi di clustering per grafi e reti

Fuzzy Set e Fuzzy Clustering

- Metodi di raggruppamento discussi finora
  - Ogni unità è assegnata esattamente ad un gruppo
- Alcune applicazioni potrebbero necessitare dell'assegnazione di gruppi fuzzy o soft
  - Ad esempio un videogame potrebbe appartenere sia all'intrattenimento che al software
- Metodi: cluster fuzzy e cluster basati su modelli probabilistici
- Cluster fuzzy: set fuzzy S: F_S: X → [0, 1] (valore compreso tra 0 e 1)
- Esempio: la popolarità delle telecamere è definita come una mappatura fuzzy

Telecamera	Vendite	F(Telecamera)
A	50	0,05
B	1320	1,00
C	860	0,86
D	270	0,27

- Dove F_S=min(1,Vendite/1000)

Fuzzy Clustering

- Esempio: supponiamo che le caratteristiche del problema siano
  - C_1: "fotocamera digitale" e "obiettivo"
  - C_2: "computer"

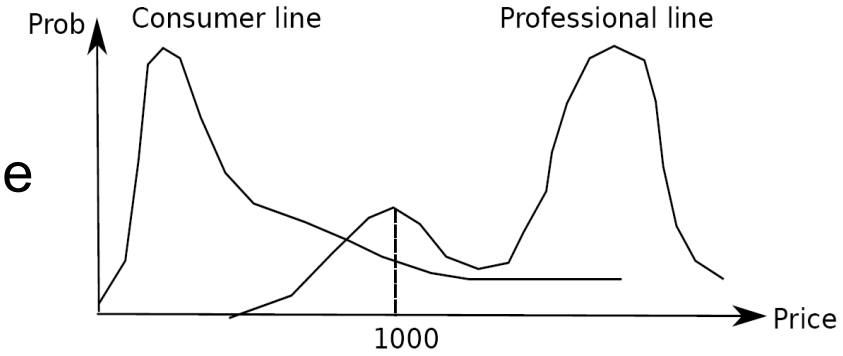
Review-id	Keywords
R_1	digital camera, lens
R_2	digital camera
R_3	lens
R_4	digital camera, lens, computer
R_5	computer, CPU
R_6	computer, computer game

$$= \begin{bmatrix} 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ \frac{2}{3} & \frac{1}{3} \\ 0 & 1 \\ 0 & 1 \end{bmatrix}$$

- Raggruppamento Fuzzy (sfocato)
- k cluster fuzzy C_1, ..., C_k, rappresentato come matrice di partizione M = $[w_{ij}]$
- P1: per ogni unità o e gruppo C_j, $0 \leq w_{ij} \leq 1$ (set fuzzy) $\sum_{i=1}^{k} w_{ij} = 1$
- P2: per ciascun unità o_i, uguale partecipazione al raggruppamento $0 < \sum_{i=1}^{n} w_{ij} < n$
- P3: per ogni cluster Cj, assicura che non ci sia un gruppo vuoto
- Supponiamo siano c_1, ..., c_k i centri dei k gruppi
- Per una unità o_i, la somma dell'errore quadratico (SSE), p è un parametro:
- Per un gruppo C_i, il SSE: $\mathrm{SSE}(C_j) = \sum_{i=1}^{n} w_{ij}^p dist(o_i, c_j)^2$ $\mathrm{SSE}(o_i) = \sum_{j=1}^{k} w_{ij}^p dist(o_i, c_j)^2$
- Misura quanto bene un clustering si adatta ai dati: $\mathrm{SSE}(\mathcal{C}) = \sum_{i=1}^{n} \sum_{j=1}^{k} w_{ij}^p dist(o_i, c_j)^2$

Clustering Probabilistico Model-Based

- L'analisi del cluster consiste nel trovare dei parametri nascosti, le categorie indici del gruppo di appartenenza.
- Una categoria nascosta (cioè un cluster probabilistico) è una distribuzione sullo spazio dei dati, che può essere rappresentata matematicamente utilizzando una funzione di densità di probabilità (o funzione di distribuzione).
  - Esempio: vengono vendute 2 tipi di fotocamere digitali
  - linea di consumo contro linea professionale
  - funzioni di densità f_1, f_2 per C_1, C_2
  - ottenuto da clustering probabilistico



  - Un modello di MISTURE presuppone che un insieme di dati osservati sia una miscela di istanze provenienti da più gruppi probabilistici e concettualmente ogni unità osservata viene generato indipendentemente
  - Il compito: deduce un insieme di k cluster probabilistici che è più probabile che generino D usando il processo di generazione dei dati sopra descritto

Model-Based Clustering

- Un insieme C di *k* gruppi C_1, ..., C_k con funziona di densità di probabilità f_1, ..., f_k, rispettivamente, e le loro probabilità ω_1, ..., ω_k.
- La probabilità di una unità *i* di essere generata dal cluster *Cj* è: $P(i|C_j)=w_j f_j(i)$
- La probabilità di *i* di essere generata dall'insieme dei gruppi **C** è:

  $P(i|\boldsymbol{C})=\Sigma_{j=1,\ldots,k} w_j f_j(i)$

  - Poiché si presume che le unità siano generate in modo indipendente, per un set di dati D = {o_1, …, o_n}, abbiamo,

$$P(D|\boldsymbol{C}) = \prod_{i=1}^{n} P(o_i|\boldsymbol{C}) = \prod_{i=1}^{n} \sum_{j=1}^{k} \omega_j f_j(o_i)$$

  - Compito: trova k gruppi di ***C*** con I quali si massimizza P (D | ***C***)
  - Massimizzare P (D | ***C***) è spesso impossibile (o difficile) poiché la funzione di densità di probabilità di un gruppo può assumere una forma arbitrariamente complessa
  - Per renderlo computazionalmente fattibile, supponiamo che le funzioni di densità di probabilità siano alcune distribuzioni parametriche

Mistura finita di Gaussiane

- Siano O = {o_1, ..., o_n} (n unità osservate), Θ = {θ_1, ..., θ_k} (parametri delle distribuzioni k), e P_j (o_i | θ_j) è la probabilità che o_i sia generato dal *j*-esima la distribuzione usando il parametro θ_j, abbiamo

$$P(o_i|\mathbf{\Theta}) = \sum_{j=1}^{k} \omega_j P_j(o_i|\Theta_j) \qquad P(\mathbf{O}|\mathbf{\Theta}) = \prod_{i=1}^{n} \sum_{j=1}^{k} \omega_j P_j(o_i|\Theta_j)$$

- Mistura di gaussiane univariata
- Supponiamo che la funzione di densità di probabilità di ciascun gruppo segua una distribuzione gaussiana (μ,σ). Supponiamo che ci siano k gruppi.
- La funzione di densità di probabilità di ciascun cluster è

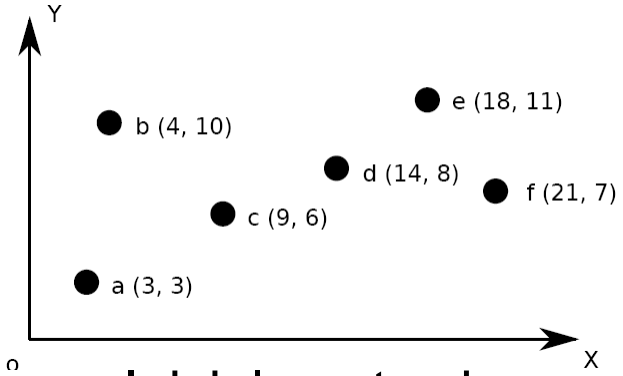
$$P(o_i|\Theta_j) = \frac{1}{\sqrt{2\pi}\sigma_j} e^{-\frac{(o_i-\mu_j)^2}{2\sigma^2}} \qquad P(o_i|\mathbf{\Theta}) = \sum_{j=1}^{k} \frac{1}{\sqrt{2\pi}\sigma_j} e^{-\frac{(o_i-\mu_j)^2}{2\sigma^2}}$$

$$P(\mathbf{O}|\mathbf{\Theta}) = \prod_{i=1}^{n} \sum_{j=1}^{\kappa} \frac{1}{\sqrt{2\pi}\sigma_j} e^{-\frac{(o_i-\mu_j)^2}{2\sigma^2}}$$

Algoritmo EM

- L'algoritmo **kmeans** ha due passaggi a ogni iterazione:
  - **Expectation Step (E-step)**: dati gli attuali centri del gruppi, ogni unità è assegnata al gruppo il cui centro è più vicino all'unità: una unità dovrebbe appartenere al gruppo più vicino (questa è la nostra aspettativa)
  - **Step di ottimizzazione (passo M)**: data l'assegnazione del gruppo, per ciascun gruppo, l'algoritmo modifica il centro in modo tale che la somma della distanza dalle unità assegnate a questo gruppoe il nuovo centro sia minima
- L'algoritmo **EM**: una regola per trovare la massima verosimiglianza o le stime a posteriori dei parametri nei modelli statistici.
  - **E-step** assegna le unità ai gruppi in base all'attuale cluster fuzzy o ai parametri dei cluster probabilistici
  - **M-step** trova il nuovo clustering o i parametri che massimizzano la somma dell'errore quadratico (SSE) o la verosimiglianza prevista

Cal



Iteration	E-step	M-step
1	$M^T = \begin{bmatrix} 1 & 0 & 0.48 & 0.42 & 0.41 & 0.47 \\ 0 & 1 & 0.52 & 0.58 & 0.59 & 0.53 \end{bmatrix}$	$c_1 = (8.47, 5.12)$, $c_2 = (10.42, 8.99)$
2	$M^T = \begin{bmatrix} 0.73 & 0.49 & 0.91 & 0.26 & 0.33 & 0.42 \\ 0.27 & 0.51 & 0.09 & 0.74 & 0.67 & 0.58 \end{bmatrix}$	$c_1 = (8.51, 6.11)$, $c_2 = (14.42, 8.69)$
3	$M^T = \begin{bmatrix} 0.80 & 0.76 & 0.99 & 0.02 & 0.14 & 0.23 \\ 0.20 & 0.24 & 0.01 & 0.98 & 0.86 & 0.77 \end{bmatrix}$	$c_1 = (6.40, 6.24)$, $c_2 = (16.55, 8.64)$

- Inizialmente si pone c_1 = a and c_2 = h
- 1st E-step: assegnare o a c_1

$$\frac{\frac{1}{dist(o,c_1)^2}}{\frac{1}{dist(o,c_1)^2}+\frac{1}{dist(o,c_2)^2}} = \frac{dist(o,c_2)^2}{dist(o,c_1)^2+dist(o,c_2)^2}$$

$$w_{c,c_1} = \frac{41}{45+41} = 0.48$$

- 1st M-step: ricalcolare I centroidi secondo la matrice di partizione, minimizzando la sommadei quadrati degli errori (SSE)

$$c_j = \frac{\sum_{\text{each point } o} w_{o,c_j}^2 o}{\sum_{\text{each point } o} w_{o,c_j}^2} \quad c_1 = \left(\frac{1^2\times3+0^2\times4+0.48^2\times9+0.42^2\times14+0.41^2\times18+0.47^2\times21}{1^2+0^2+0.48^2+0.42^2+0.41^2+0.47^2}, \frac{1^2\times3+0^2\times10+0.48^2\times6+0.42^2\times8+0.41^2\times11+0.47^2\times7}{1^2+0^2+0.48^2+0.42^2+0.41^2+0.47^2}\right) = (8.47, 5.12)$$

- Ripetere iterativamente finché I centri dei gruppi non convergono o il cambiamento è ritenuto abbastanza piccolo

Misture finite con l'EM

- Date le n unità O = {o_1, …, o_n}, vogliamo stimare I parametri Θ = {θ_1, …, θ_k} in modo che P(**O**|**Θ**) sia massimizzata, dove $\theta_j = (\mu_j, \sigma_j)$ sono medie e varianze delle distribuzioni Gaussiane univariate.
- Inizialmente assegnamo dei valori random ai parametri θ_j, quindi iterativamente attraverso gli step E- and M- fino a convergenza
- Nello E-step, per ogni unità o_i, calcolare la probabilità che o_i appartenga a ciascuna distribuzione:

$$P(\Theta_j|o_i, \mathbf{\Theta}) = \frac{P(o_i|\Theta_j)}{\sum_{l=1}^{k} P(o_i|\Theta_l)}$$

- Nell'M-step, modificare il parametro $\theta_j = (\mu_j, \sigma_j)$ in modo da massimizzare la verosimiglianza P(**O**|**Θ**)

$$\mu_j = \sum_{i=1}^{n} o_i \frac{P(\Theta_j|o_i, \mathbf{\Theta})}{\sum_{l=1}^{n} P(\Theta_j|o_l, \mathbf{\Theta})} = \frac{\sum_{i=1}^{n} o_i P(\Theta_j|o_i, \mathbf{\Theta})}{\sum_{i=1}^{n} P(\Theta_j|o_i, \mathbf{\Theta})} \quad \sigma_j = \sqrt{\frac{\sum_{i=1}^{n} P(\Theta_j|o_i, \mathbf{\Theta})(o_i - u_j)^2}{\sum_{i=1}^{n} P(\Theta_j|o_i, \mathbf{\Theta})}}$$

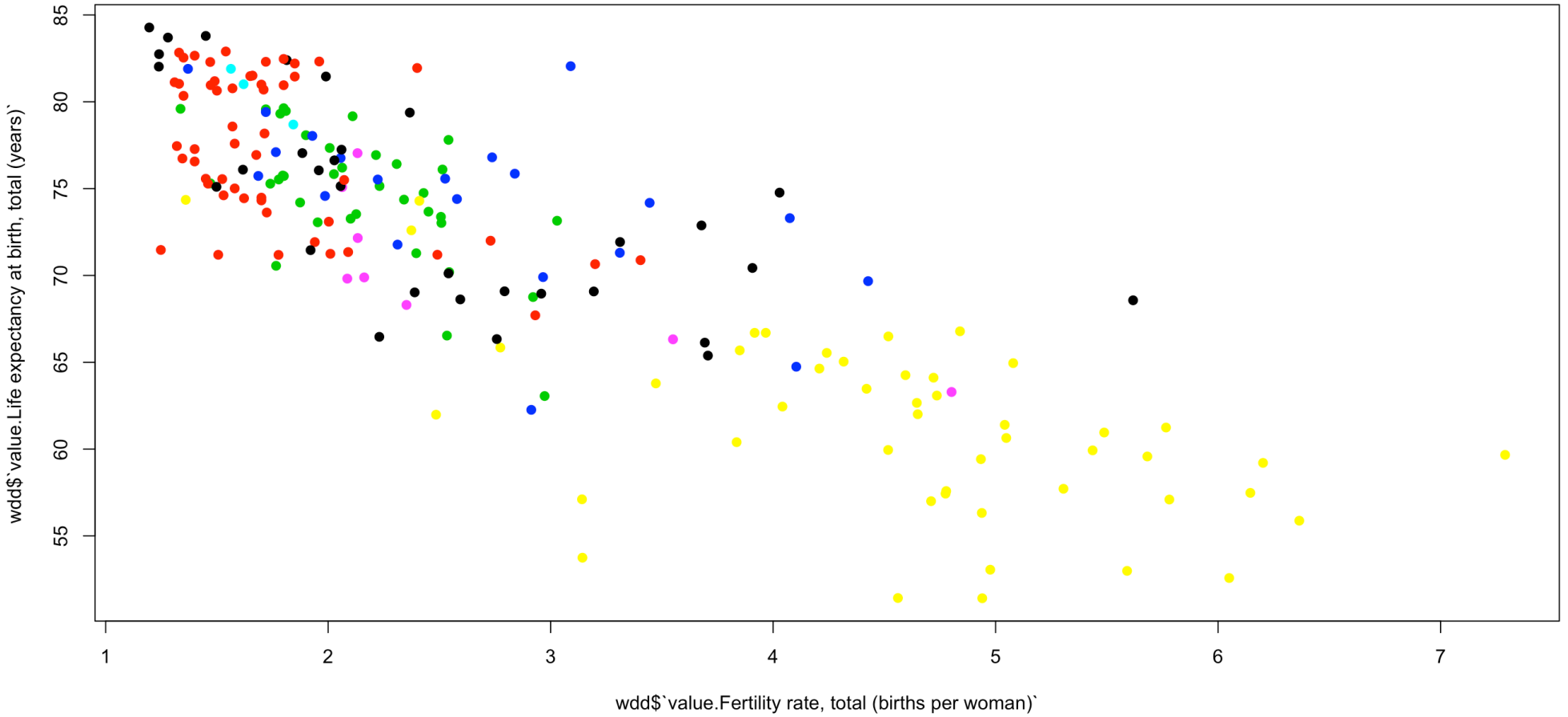
Pregi e Difetti delle Misture finite

- Forza
  - I modelli di mistura sono più generali del partizionamento e del clustering fuzzy
  - I gruppi possono essere caratterizzati da un numero limitato di parametri
  - I risultati soddisfano le ipotesi dei processi che li hanno generati
- Debolezza
  - Converge in ottimi locali (possibile soluzione: eseguire l'inizializzazione casuale più volte)
  - Computazionalmente costoso se il numero di distribuzioni (gruppi) è elevato o se dataset contiene pochi dati
  - Hai bisogno di grandi set di dati
  - Difficile stimare il numero di gruppi

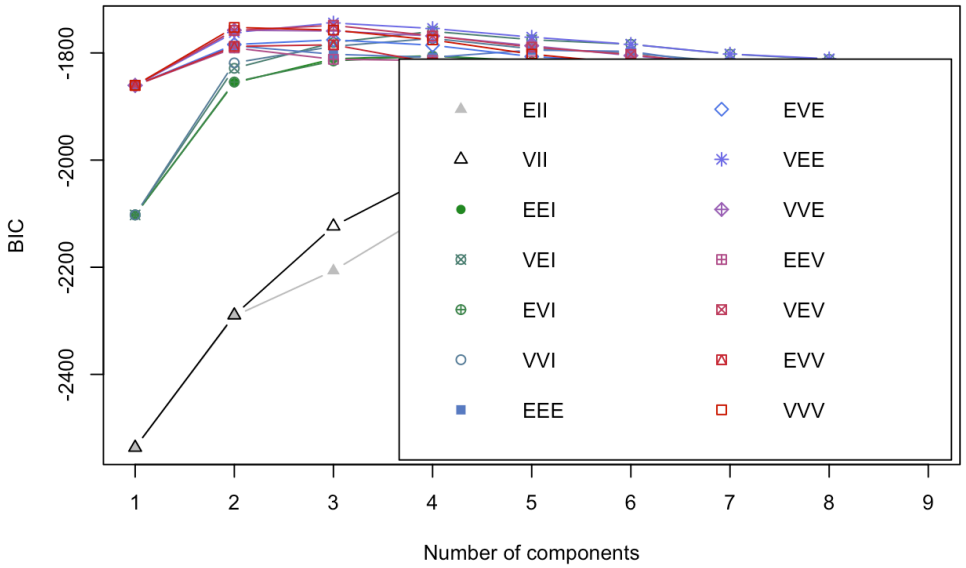
Fertilità ed aspettativa di vita 2015

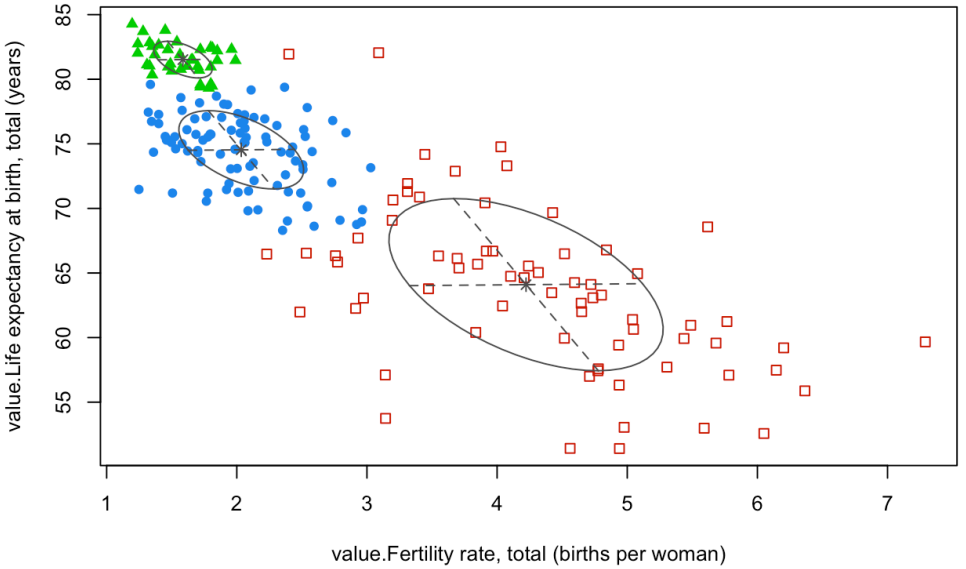
https://www.gapminder.org/data/

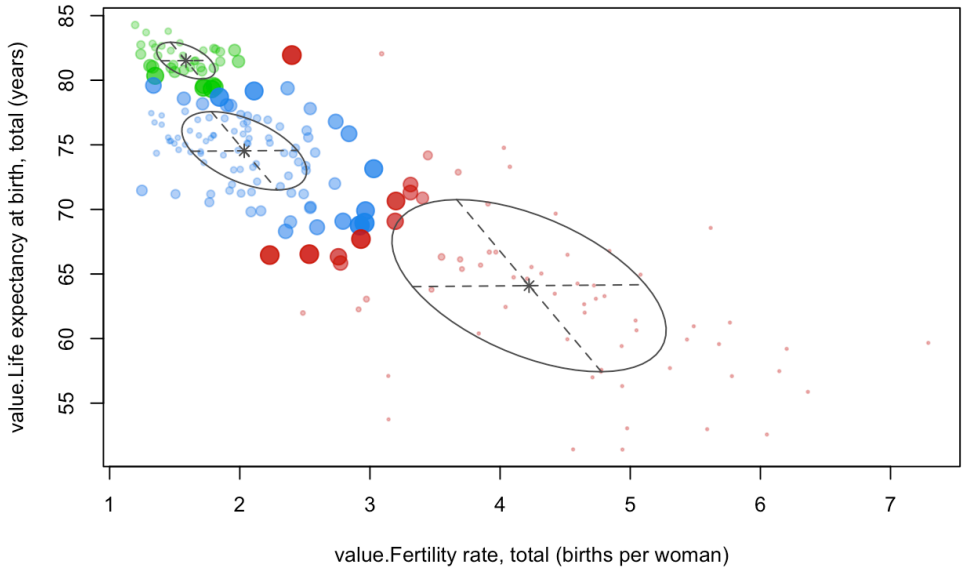
```
library(mclust)
risgap <- Mclust(na.omit(wdd[,3:4]),G=1:9)
plot(risgap)
```

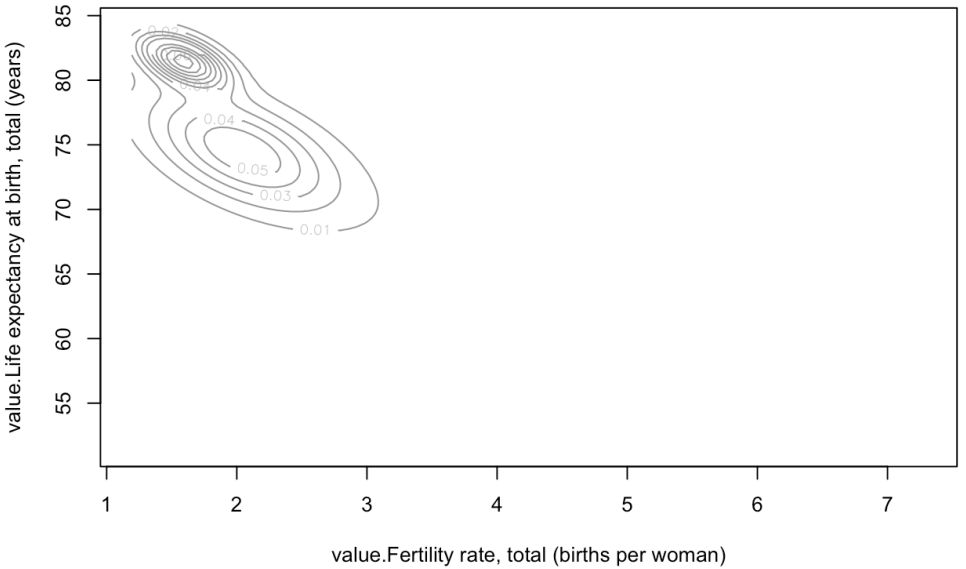


Fertilità ed aspettativa di vita 2015







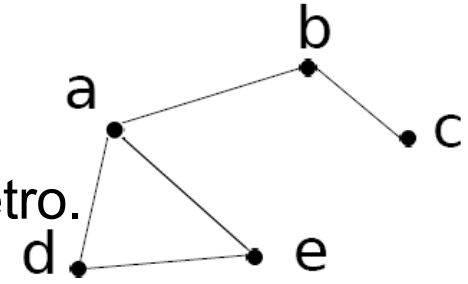


Clustering Graphs and Network Data

- Applicazioni
  - Grafici bipartiti, ad es. Clienti e prodotti, autori e conferenze
  - Motori di ricerca Web, ad esempio, fare clic su grafici e grafici Web
  - Reti sociali, grafici di amicizia / coautori
- Misure di similarità
  - Distanze geodetiche
  - Distanza basata su Random Walk (SimRank)
- Metodi di clustering su grafi
  - Tagli minimi: FastModularity (Clauset, Newman & Moore, 2004)
  - Cluster basato sulla densità: SCAN (Xu et al., KDD'2007)

Distanze su Grafo

- Distanza geodetica (A, B): lunghezza (cioè, # di archi) del percorso più breve tra A e B (se non connesso, definito come infinito)
- Eccentricità di v, eccen (v): la più grande distanza geodetica tra v e qualsiasi altro vertice u ∈ V - {v}.
  - Ad esempio, eccen (a) = eccen (b) = 2; eccen (c) = eccen (d) = eccen (e) = 3
- Raggio del grafico G: L'eccentricità minima di tutti i vertici, cioè la distanza tra "il punto più centrale" e il "bordo più lontano"
  - $r = \min_{v \in V}$ eccen (v)
  - Ad esempio, raggio (g) = 2
- Diametro del grafico G: la massima eccentricità di tutti i vertici, cioè la massima distanza tra qualsiasi coppia di vertici in G
  - $d = \max_{v \in V}$ eccen (v)
  - Ad esempio, diametro (g) = 3
- Un vertice periferico è un vertice che raggiunge il diametro.
  - Ad esempio, i vertici c, d ed e sono vertici periferici



Distanze su Grafo

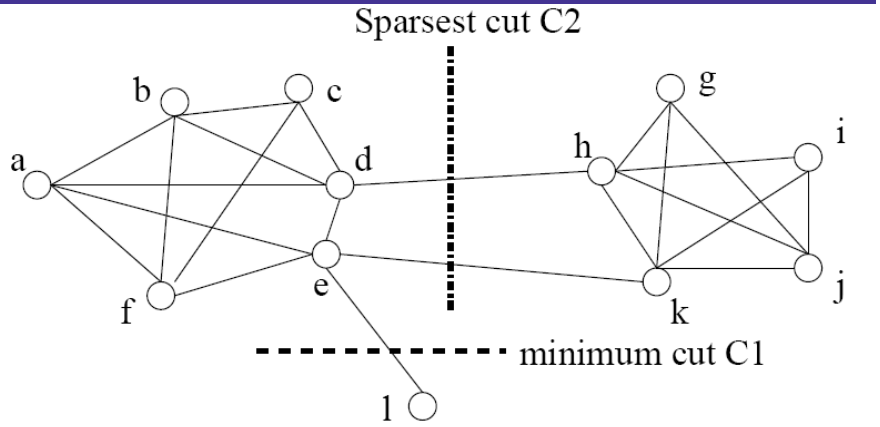
- SimRank: similarità su contesto strutturale, basata sulla similarità dei suoi vicini
- In un grafo diretto G = (V,E),
  - *Individui in-neighborhood* di v: I(v) = {u | (u, v) ∈ E}
  - *Individui out-neighborhood* di v: O(v) = {w | (v, w) ∈ E}
- Similarità in SimRank: $$s(u,v) = \frac{C}{|I(u)||I(v)|} \sum_{x \in I(u)} \sum_{y \in I(v)} s(x,y)$$
  - Inizializza: $$s_0(u,v) = \begin{cases} 0 & \text{if } u \neq v \\ 1 & \text{if } u = v \end{cases}$$
  - Quindi si calcola s_{i+1} da s_i basata sulla definizione
- Similarità basata su r$$d(u,v) = \sum_{t:u \rightsquigarrow v} P[t] l(t)$$onenti che seguono l'entropia
  - Distanza attesa: $$m(u,v) = \sum P[t] l(t)$$
  - Distanza attesa di incontro: $$p(u,v) = \sum_{t:(u,v) \rightsquigarrow (x,x)} P[t] C^{l(t)}$$
  - Probabilità attesa di incontro:

$$P[t] = \begin{cases} \prod_{i=1}^{k-1} \frac{1}{|O(w_i)|} & \text{if } l(t) > 0 \\ 0 & \text{if } l(t) = 0. \end{cases}$$

P[t] è la probabilità del tour

Graph Clustering: Sparsest Cut

- G = (V, E). L'insieme di tagli di un taglio è l'insieme di spigoli {(u, v) ∈ E | u ∈ S, v ∈ T} e S e T sono in due partizioni
- Dimensione del taglio: # di bordi nel set di taglio
- Min-cut (ad es., C_1) non è una buona partizione
- Una misura migliore: **Sparsity**: $\Phi = \frac{\text{the size of the cut}}{\min\{|S|, |T|\}}$



  - Un taglio è più **Sparsest** se la sua scarsità non è maggiore di quella di qualsiasi altro taglio
    - Esempio Taglia C_2 = ({a, b, c, d, e, f, l}, {g, h, i, j, k}) è il taglio più corto
  - Per i k gruppi, la **modularità** di un clustering valuta la qualità del clustering: $Q = \sum_{i=1}^{k} (\frac{l_i}{|E|} - (\frac{d_i}{2|E|})^2)$

    l_i: # edges between vertices in the i-th cluster
    d_i: the sum of the degrees of the vertices in the i-th cluster
  - La **modularità** di un clustering di un grafo è la differenza tra la frazione di tutti i bordi che cadono nei singoli gruppi e la frazione che lo farebbe se i vertici del grafico fossero collegati casualmente

Clustering su grafo: la sfida dei buoni tagli

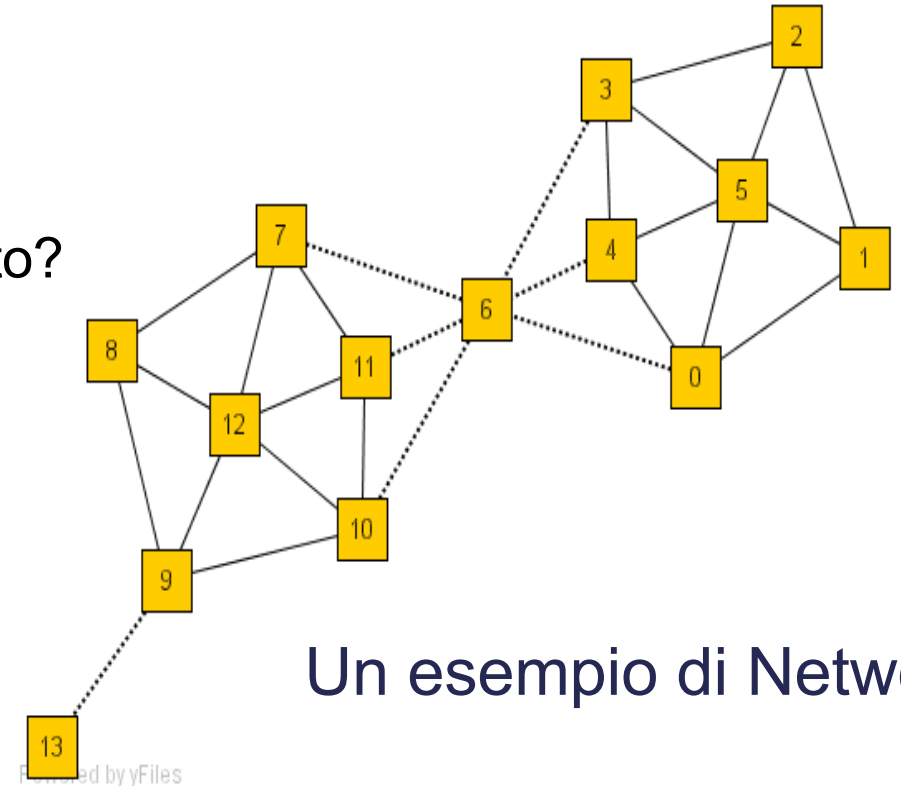
- Elevato costo computazionale
  - Molti problemi di taglio del grafo sono computazionalmente costosi
  - Il problema più corto è NP-hard
  - Necessità di un compromesso tra efficienza / scalabilità e qualità
- Grafi sofisticati
  - Può comportare pesi e / o cicli.
- Alta dimensionalità
  - Un grafo può avere molti vertici. In una matrice di similarità, un vertice è rappresentato come un vettore (una riga nella matrice) la cui dimensionalità è il numero di vertici nel grafo
- Sparsità
  - Un grande grafo è spesso sparso, il che significa che in media ogni vertice si connette solo a un piccolo numero di altri vertici
  - Una matrice di similarità di un grafo sparso può anche essere sparsa

2 approcci al Clustering su grafo

- Due approcci per il clustering dei dati su grafo
  - Utilizzare metodi di clustering generici per dati ad alta dimensione
  - Progettato specificamente per il clustering su grafo
- Utilizzo di metodi di clustering per dati ad alta dimensionalità
  - Estrarre una matrice di similarità da un grafo utilizzando una misura di somiglianza
  - Un metodo di clustering generico può quindi essere applicato alla matrice di similarità per trovarei gruppi
- Metodi specifici per i grafici
- Cerca nel grafo per trovare componenti ben collegati come cluster
  - Ex. SCAN (Structural Clustering Algorithm for Networks)
    - X. Xu, N. Yuruk, Z. Feng, and T. A. J. Schweiger, “SCAN: A Structural Clustering Algorithm for Networks”, KDD'07

SCAN: Density-Based Clustering of Networks

- Quanti cluster?
- Che taglia dovrebbero essere?
- Qual è il miglior partizionamento?
- Dovrebbero essere segregati alcuni punti?

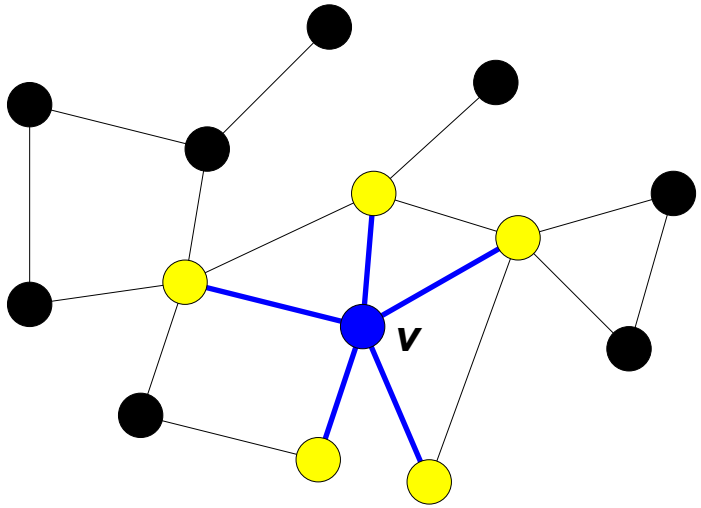


Un esempio di Network

- Applicazione: data la semplice informazione di chi si associa a chi, si potrebbero identificare gruppi di individui con interessi comuni o relazioni speciali (famiglie, clique, cellule terroristiche)?

Un modello di Social Network

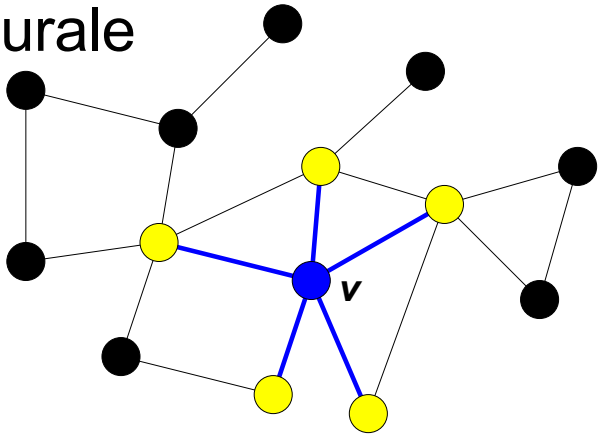
- Clique, hubs e valori anomali
- Gli individui in un ristretto gruppo sociale, o clique, conoscono molte delle stesse persone, indipendentemente dalle dimensioni del gruppo
- Gli individui che sono hub conoscono molte persone in gruppi diversi ma non appartengono a nessun singolo gruppo. I politici, ad esempio, collegano più gruppi
- Gli individui che sono outlier risiedono ai margini della società. Gli eremiti, ad esempio, conoscono poche persone e non appartengono a nessun gruppo
- Il vicinato di un vertice
  - Definisci C (v) come l'immediato vicinato di un vertice (cioè l'insieme di persone che un individuo conosce)



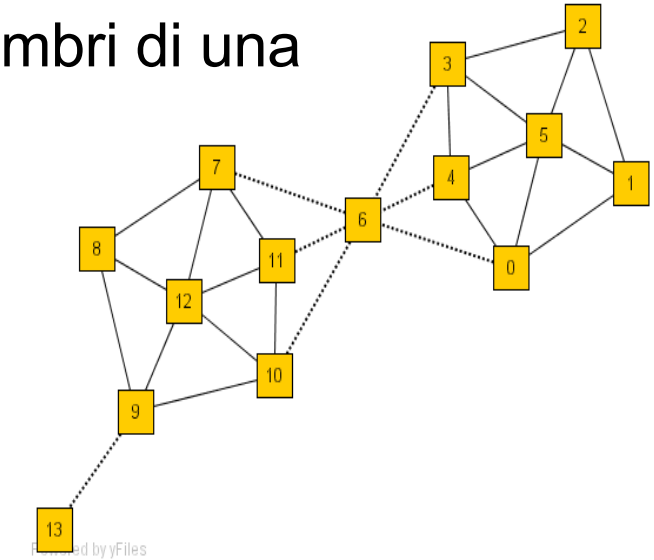
Similarità Strutturale

- Le caratteristiche desiderate tendono ad essere catturate da una misura che chiamiamo similarità strutturale

$$\sigma(v,w) = \frac{|\Gamma(v) \bigcap \Gamma(w)|}{\sqrt{|\Gamma(v)||\Gamma(w)|}}$$



- La somiglianza strutturale è grande per i membri di una clique e piccola per gli hub e i valori anomali



Structural Connectivity

- *ε-Neighbour:* $N_\varepsilon(v) = \{w \in \Gamma(v) \mid \sigma(v,w) \geq \varepsilon\}$
- *Core:* $CORE_{\varepsilon,\mu}(v) \Leftrightarrow |N_\varepsilon(v)| \geq \mu$
- *Struttura diretta raggiungibile:*

$$DirRECH_{\varepsilon,\mu}(v,w) \Leftrightarrow CORE_{\varepsilon,\mu}(v) \wedge w \in N_\varepsilon(v)$$

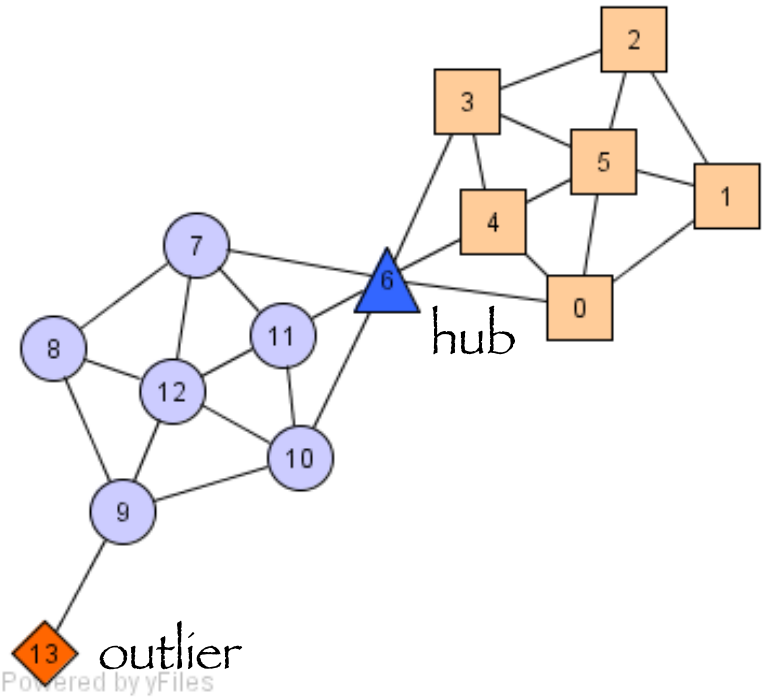
- *Struttura raggiungibile: chiusura transitiva della raggiungibilità della struttura diretta*
- *Struttura connessa:*

$$CONNECT_{\varepsilon,\mu}(v,w) \Leftrightarrow \exists u \in V : RECH_{\varepsilon,\mu}(u,v) \wedge RECH_{\varepsilon,\mu}(u,w)$$

M. Ester, H. P. Kriegel, J. Sander, & X. Xu (KDD'96) "A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases"

Gruppi Structure-Connected

- Cluster stucture-connected a C
  - Connettività: $\forall v, w \in C : CONNECT_{\varepsilon,\mu}(v, w)$
  - massimalità: $\forall v, w \in V : v \in C \wedge REACH_{\varepsilon,\mu}(v, w) \Rightarrow w \in C$
- Hubs:
  - Non appartiene a nessun cluster
  - Ponte a molti gruppi
- Valori anomali:
  - Non appartiene a nessun cluster
  - Connetti a meno cluster



Valutare il clustering ottenuto

- Le misure numeriche usate per valutare la bontà di un clustering, sono classificate in tre tipi:
  - Indici esterni: Usato per misurare fino a che punto le etichette del cluster corrispondono a quelle fornite da una fonte esterna.
    - Entropia
  - Indici interni: Usati per misurare la bontà di un clustering senza riferimento a informazione esterna.
    - Sum of Squared Error (SSE)
  - Indici relativi: Usati per comparare due clustering or cluster.
    - Spesso un indice interno o esterno è usato in questi casi: SSE o entropia
- A volte si parla di criteri piuttosto che di indici

Misurare la validità con la correlazione

- Due matrici
  - Matrice di prossimità
  - Matrice di incidenza
    - Un riga e una colonna per ogni punto (n x n)
    - La cella della matrice è 1 se I punti associati sono nello stesso cluster. E' nulla altrimenti
- Calcola la correlazione tra le celle corrispondenti delle due matrici
  - Poichè le matrici sono simmetriche, solo n(n-1) / 2 celle sono considerate.
- Alta correlazione indica che i punti che appartengono allo stesso cluster sono vicini tra loro.
- Non è una misura che funziona bene nel caso di cluster a densità differente o definito dalla contiguità.

Misure interne: Coesione e Separazione

- Coesione dei Cluster: misura quanto sono coese le unità all'interno di un cluster
- Separazione dei cluster: indice di separazione tra cluster
- Esempio: Errore quadratico (Squared Error)
  - La coesione è misurata dalla somma dei quadrati NEI cluster (SSE)

$$WSS = \sum_i \sum_{x \in C_i} (x - m_i)^2$$

  - La separazione è misurata dalla somma dei quadrati FRA cluster

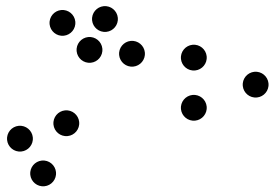
$$BSS = \sum_i |C_i|(m - m_i)^2$$

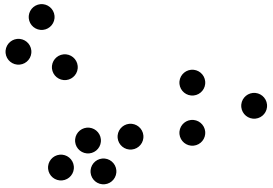
Valutare il clustering ottenuto

- Valutare se la struttura non casuale esiste nei dati misurando la probabilità che i dati siano generati da una distribuzione uniforme dei dati

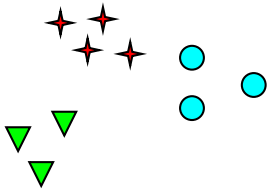
- Test di casualità spaziale con test statistico: statistica test di Hopkins
  - Dato un set di dati D considerato come un campione di una variabile aleatoria, determinare quanto sia lontano da essere uniformemente distribuito nello spazio dei dati
  - Esempio n punti, p_1, ..., p_n, uniformemente distribuiti su D. Per ogni p_i, trova il suo vicino più prossimo in D: x_i = min {dist (p_i, v)} dove v in D
  - Esempio di *n* punti, q_1, ..., q_n, uniformemente da D. Per ogni q_i, trovare il suo vicino più prossimo in D - {q_i}: y_i = min {dist (q_i, v)} dove v in D e v ≠ q_i
  - Calcola la statistica di Hopkins: $$H = \frac{\sum_{i=1}^{n} y_i}{\sum_{i=1}^{n} x_i + \sum_{i=1}^{n} y_i}$$
  - Se D è distribuito uniformemente, Σ x_i e Σ y_i saranno vicini l'uno all'altro e H è vicino a 0.5. Se D è molto concentrato, H è vicino a 0

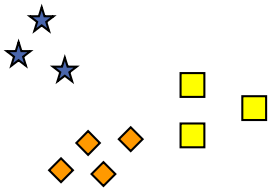
Definire il numero di gruppi



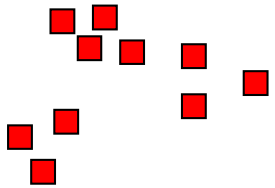


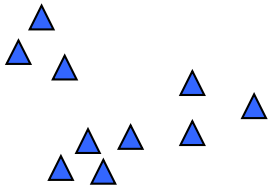
Quanti cluster?



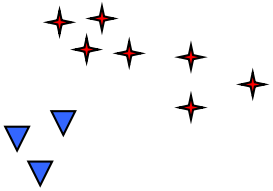


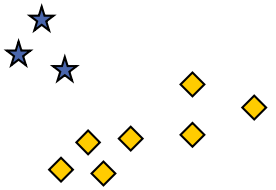
Sei Cluster





Due Cluster





Quattro Cluster

Definire il numero di gruppi

- Metodo empirico
  - # di cluster ≈√n / 2 per un set di dati di n punti
- Metodo del “gomito” (Elbow)
  - Utilizzare il punto di svolta nella curva della somma della varianza within del gruppo w.r.t il numero di cluster
- Metodo di “cross validation“
  - Dividere un dataset in *m* parti
  - Usare *m* - 1 parti per ottenere un modello di clustering
  - Utilizzare la parte rimanente per verificare la qualità del clustering
    - Ad esempio, per ciascun punto nel set di test, trovare il centroide più vicino e utilizzare la somma della distanza quadratica tra tutti i punti nell’insieme di test e i centroidi più vicini per misurare quanto bene il modello si adatta all’insieme di test
  - Per ogni k> 0, ripetilo *m* volte, confronta la misura generale di qualità w.r.t. diversi k, e trova # di gruppi che si adattano al meglio ai dati

Definire il numero di gruppi: Elbow method

- Ricordiamo che l'idea alla base dei metodi di partizionamento, come il k-means, consiste nel definire gruppi tali che la variazione totale intra-cluster [o la somma totale del quadrato (WSS) all'interno di un gruppo) sia ridotta al minimo. Il WSS totale misura la compattezza del clustering e vogliamo che sia il più piccolo possibile.
- Il metodo Elbow considera il WSS totale come una funzione del numero di gruppi: uno dovrebbe scegliere un numero di gruppi in modo che l'aggiunta di un altro gruppo non migliori molto il WSS totale.
- Il numero ottimale di gruppi può essere definito come segue:
- Calcolo dell'algoritmo di clustering per diversi valori di k. Ad esempio, variando k da 1 a 10 gruppi.
- Per ogni k, calcola la somma totale entro il cluster di square (WSS).
- Traccia la curva di WSS in base al numero di cluster k.
- La posizione di una curva (ginocchio) nella trama è generalmente considerata come un indicatore del numero appropriato di gruppi.

Definire il numero di gruppi: Average silhouette method

- L'approccio dell'indice di silhouette medio misura la qualità di un clustering. Cioè, determina quanto bene ogni unità si trova all'interno del suo gruppo. Una larghezza alta dell'indice media indica un buon raggruppamento.
- Il metodo dell'indice di silhouette media calcola l'indice medio delle osservazioni per diversi valori di k. Il numero ottimale di gruppi k è quello che massimizza la silhouette media su un intervallo di valori possibili per k (Kaufman e Rousseeuw 1990).
- L'algoritmo è simile al metodo Elbow e può essere calcolato come segue:
- Calcolo dell'algoritmo di clustering per diversi valori di k. Ad esempio, variando k da 1 a 10 gruppi.
- Per ogni k, calcola la silhouette media delle osservazioni.
- Tracciare la sua curva in funzione del numero di gruppi k.
- La posizione del massimo è considerata come il numero appropriato di gruppi.

Definire il numero di gruppi: la statistica GAP

- La statistica GAP confronta il totale all'interno della variazione intra-cluster per diversi valori di k con i loro valori previsti sotto la distribuzione di riferimento casuale dei dati. La stima dei gruppi ottimali sarà un valore che massimizza la statistica del GAP (cioè, che produce la statistica GAP più grande). Ciò significa che la struttura di clustering è lontana dalla distribuzione uniforme casuale dei punti.
- L'algoritmo segue i passi:
- Raggruppa i dati osservati, variando il numero di cluster da k = 1, ..., k_{max} e calcola il totale corrispondente all'interno della variazione W_k intra-cluster.
- Generare un dataset di riferimento, diciamo B, con una distribuzione uniforme casuale. Raggruppa ognuno di questi set di dati di riferimento con un numero variabile di cluster k = 1, … , k_{max} e calcola il totale corrispondente all'interno della variazione W_{kb} intra-cluster.
- Calcola la statistica del GAP stimata come la deviazione del valore di W_k osservato dal suo valore previsto W_{kb} sotto l'ipotesi nulla: $GAP(k) = (1/B)\Sigma_{b = 1..B} \log (W^*_{kb}) - \log (W_k)$
- Calcola anche la deviazione standard delle statistiche.
- Scegli il numero di cluster come il valore più piccolo di k tale che la statistica gap sia all'interno di una deviazione standard del GAP su k + 1: $GAP(k) \geq GAP(k+1)-s_{k+1}$.