Computer Science Department

University of Verona

A.A. 2016-17

Pattern Recognition

Unsupervised classification

Tassonomia

Nota preliminare

- Esistono moltissimi algoritmi di clustering
- Non esiste un'unica tassonomia
  - esistono diverse suddivisioni

- In questa lezione si adotta il punto di vista di Jain
  - Jain, Dubes, Alghoritms for clustering data, 1988
  - Jain et al., Data Clustering: a review, ACM Computing Surveys, 1999

Classi di approcci

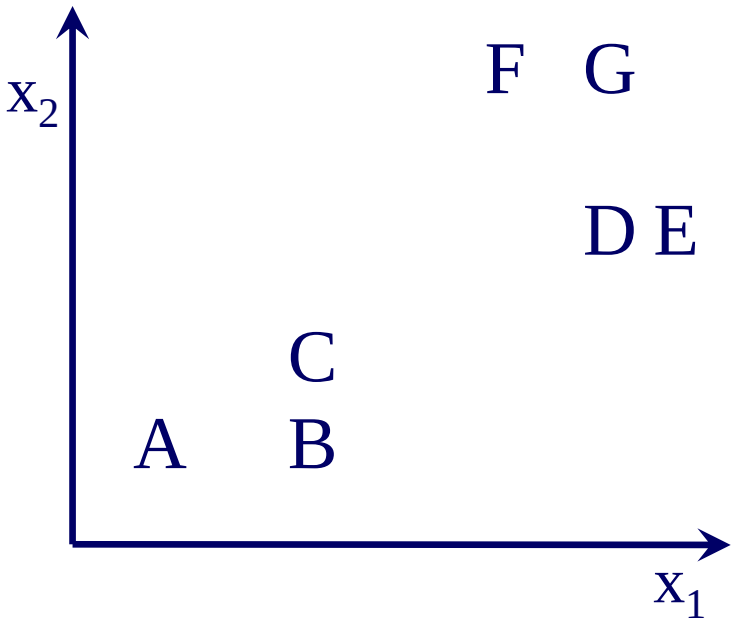
A seconda del punto di vista possiamo avere differenti classi:

- Gerarchico vs partizionale
- Hard clustering vs soft clustering
- Agglomerativo vs divisivo
- Seriale (sequenziale) vs simultaneo
- Monothetic vs polythetic
- Graph Theory vs matrix algebra
- Incrementale vs non incrementale
- Deterministico vs stocastico

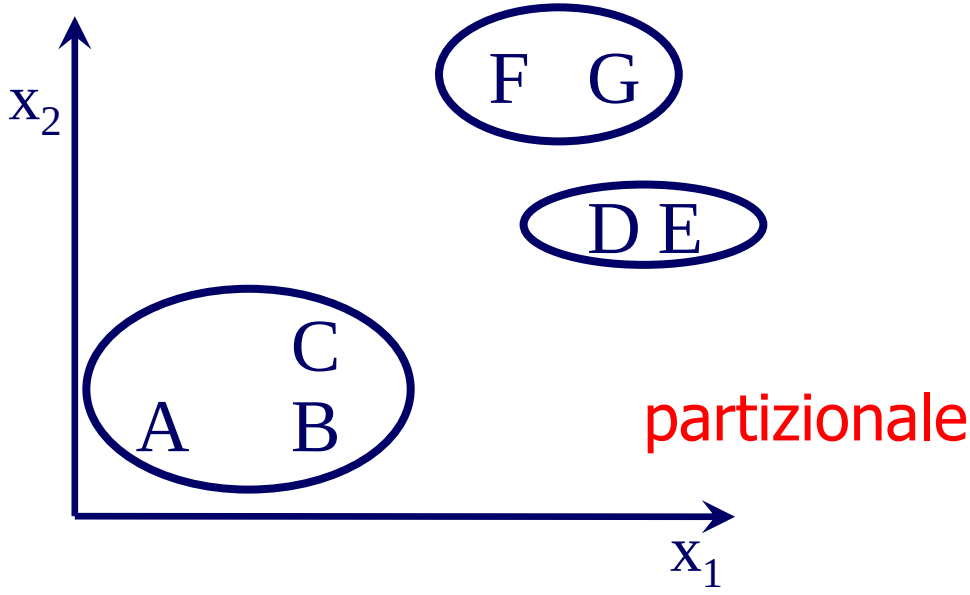
Gerarchico vs partizionale

PUNTO DI VISTA: il tipo di risultato dell'operazione di clustering

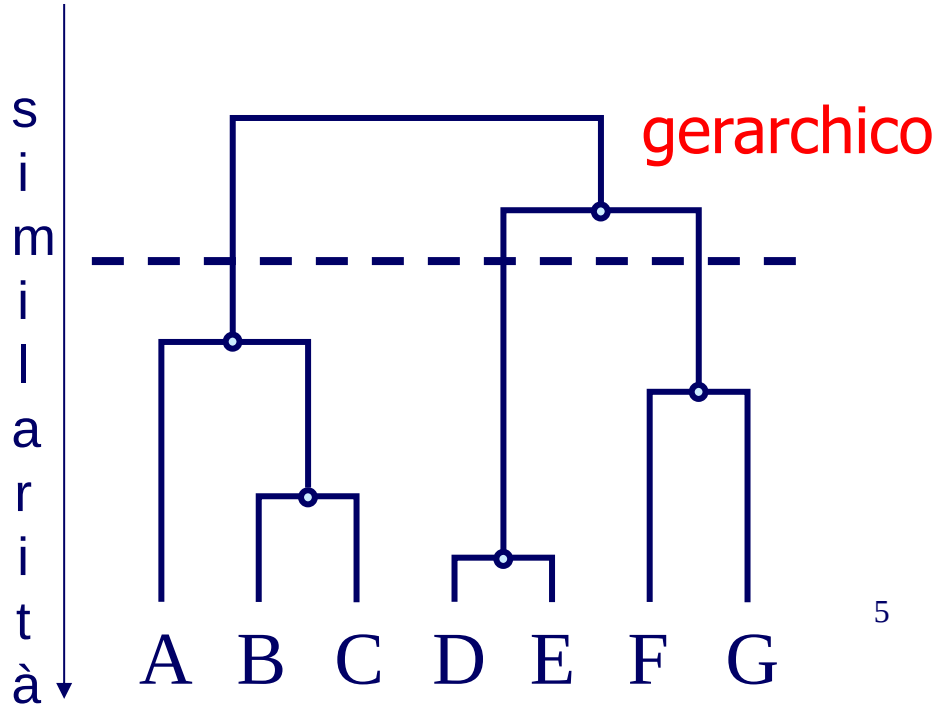
- Clustering Partizionale: il risultato è una singola partizione dei dati (tipicamente il numero di cluster deve essere dato a priori)
  - mira ad identificare i gruppi naturali presenti nel dataset
  - tipicamente richiede che i dati siano rappresentati in forma vettoriale
  - genera una partizione (insieme di cluster disgiunti la cui unione ritorna il data set originale)
- Clustering Gerarchico: il risultato è una serie di partizioni innestate (un "dendrogramma")
  - mira ad evidenziare le relazioni tra i vari pattern del dataset
  - tipicamente richiede una matrice di prossimità



problema originale



partizionale



gerarchico

Gerarchico vs partizionale

Ulteriori dettagli

- Partizionale:
  - ottimo per dataset grandi
  - scegliere il numero di cluster è un problema (esistono metodi per determinare in modo automatico il numero di cluster)
  - tipicamente il clustering è il risultato di un procedimento di ottimizzazione, definito sia localmente (in un sottoinsieme dei pattern) che globalmente (in tutti i pattern)
  - Esempi: K-means (e sue varianti), Mean Shift,...
- Gerarchico
  - non è necessario settare a priori il numero di cluster… non li da'!
  - più informativo del partizionale, è improponibile per dataset grandi
  - Esempi: Complete Link, Single Link, Ward Link,...

Hard clustering vs soft clustering

PUNTO DI VISTA: la natura dei cluster risultanti

- Hard clustering:
  - un pattern viene assegnato ad un unico cluster
    - sia durante l'esecuzione dell'algoritmo che nel risultato
  - detto anche clustering "esclusivo"
- Soft clustering:
  - un pattern può essere assegnato a diversi clusters
  - detto anche "fuzzy clustering" o "clustering non esclusivo"
  - ci può essere una funzione di "membership"
  - può essere trasformato in hard guardando la massima membership
- ESEMPIO:
  - raggruppare persone per età è esclusivo
  - raggrupparle per malattia è non esclusivo

Agglomerativo vs divisivo

PUNTO DI VISTA: come vengono formati i cluster

- Agglomerativo:
  - costruisce i cluster effettuando operazioni di “merge”
  - inizia con un cluster per ogni pattern, e successivamente fonde cluster assieme fino al raggiungimento di una determinata condizione
- Divisivo:
  - costruisce i cluster effettuando operazioni di “split”
  - inizia con un unico cluster contenente tutti i dati, e successivamente divide i cluster fino al raggiungimento di una determinata condizione
- Commenti:
  - tipo di procedura piuttosto che tipo di clustering
  - è naturalmente applicabile al clustering gerarchico, ma funziona anche per il clustering partizionale

Sequenziale vs simultaneo

PUNTO DI VISTA: in che modo vengono processati i pattern

- Sequenziale: i pattern vengono processati uno alla volta
- Simultaneo: i pattern vengono processati tutti assieme
- ESEMPIO sequenziale: prende un pattern alla volta e lo assegna ad un cluster

Monothetic vs polythetic

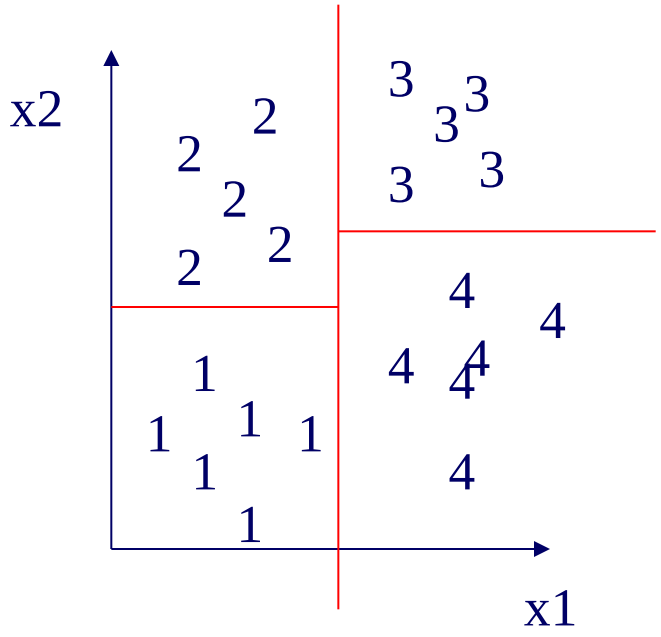
PUNTO DI VISTA: come vengono utilizzate le features

- Monothetic: viene utilizzata una feature alla volta per fare clustering
- Polythetic: vengono utilizzate tutte le features simultaneamente per fare clustering

- La maggior parte delle tecniche di clustering sono polythetic.

Monothetic vs polythetic

- Esempio di algoritmo monothetic [Anderberg 73]
  - si divide il data set in due clusters utilizzando una sola feature
  - ognuno di questi data sets viene poi diviso in due utilizzando la seconda feature
  - si procede così fino alla fine
- Svantaggi:
  - d features 2^d clusters
  - d grande, troppo frammentato



Graph Theory vs matrix algebra

PUNTO DI VISTA: come viene formulato matematicamente l'algoritmo

- graph theory: gli algoritmi sono formulati in termini di teoria dei grafi, con l'utilizzo di definizione e proprietà dei grafi (ad esempio connettività)

- matrix algebra: gli algoritmi sono espressi in termini di formule algebriche (ad esempio l'errore quadratico medio)

Incrementale vs non incrementale

PUNTO DI VISTA: cosa succede se arrivano nuovi dati

- Incrementale: il clustering può essere "aggiornato" (è costruito in modo incrementale)
- Non incrementale: all'arrivo di nuovi dati occorre riesaminare l'intero data set

- Caratteristica cruciale in questi anni: database sempre più grossi e sempre in espansione!

Deterministico vs stocastico

PUNTO DI VISTA: come viene ottimizzata la funzione di errore

- deterministico: ottimizzazione classica (discesa lungo il gradiente)
- stocastico: ricerca stocastica nello spazio degli stati della soluzione (Monte Carlo)

- Tipico problema nel clustering partizionale che deve ottimizzare una funzione di errore (come lo scarto quadratico medio)

Clustering partizionale

- Classi di approcci:
  - clustering sequenziale:
    - approccio di clustering molto semplice e intuitivo
    - tipicamente i pattern vengono processati poche volte
    - in generale, il risultato finale dipende dall'ordine con cui vengono presentati i pattern
    - funzionano bene per cluster convessi
  - center-based clustering:
    - ogni cluster è rappresentato da un centro
    - metodi efficienti per clusterizzare database grandi
    - l'obiettivo è minimizzare una funzione di costo
    - funzionano bene per cluster convessi

Clustering partizionale

- search based clustering
  - l'idea è quella di minimizzare la funzione di costo in modo "globale"

- model based clustering
  - l'idea è quella di creare dei modelli per i dati (tipicamente probabilistici)
  - tipicamente si assume che i dati siano generati da una mistura di distribuzioni di probabilità in cui ogni componente identifica un cluster

Clustering sequenziale

BSAS: Basic Sequential Algorithmic Scheme

- algoritmo di clustering sequenziale facile e intuitivo
- esempio classico: region growing per segmentazione

Assunzioni/Idee

- i pattern vengono processati una volta sola, in ordine
- ogni pattern processato viene assegnato ad un cluster esistente oppure va a creare un nuovo cluster
- il numero di cluster non è conosciuto a priori ma viene stimato durante il processo

BSAS: algoritmo

Notazione/parametri:

- $\mathbf{x}_i$: vettore di punti, $\{\mathbf{x}_1, \ldots \mathbf{x}_N\}$ dataset da clusterizzare
- C_j: j-esimo cluster
- $d(\mathbf{x}, C)$: distanza tra un punto e un insieme (un cluster) (simile alla distanza tra insiemi)
  - Max: distanza massima
  - Min: distanza minima
  - Average: distanza media
  - center-based: distanza dal "rappresentante"
- Θ: soglia di dissimilarità
- m: numero di cluster trovati ad un determinato istante

BSAS: algoritmo

Algoritmo:

m=1
C_m = $\mathbf{x}_1$
for i = 2 to N
    trova C_k tale che $d(\mathbf{x}_i, C_k) = min_{1 \leq j \leq m} d(\mathbf{x}_i, C_j)$
    if $d(\mathbf{x}_i, C_k)$ > θ
        m = m+1
        C_m = $\{\mathbf{x}_i\}$
    else
        $C_k = C_k \cup \{\mathbf{x}_i\}$
        (se necessario aggiornare i rappresentanti)
    end if
end for

BSAS: algoritmo

- Se la distanza $d(\mathbf{x}, C) = d(\mathbf{x}, \mathbf{m}_C)$ (distanza dalla media del cluster), allora l'aggiornamento dei rappresentanti può essere fatto on-line
- Notazioni
  - m_{Ck} è la media del cluster k
  - x è il punto aggiunto al cluster C_k
  - n_{Ck} è la cardinalità del cluster C_k

$$m_{C_k}^{new} = \frac{(n_{C_k}^{new} - 1)m_{C_k}^{old} + x}{n_{C_k}^{new}}$$

Clustering sequenziale

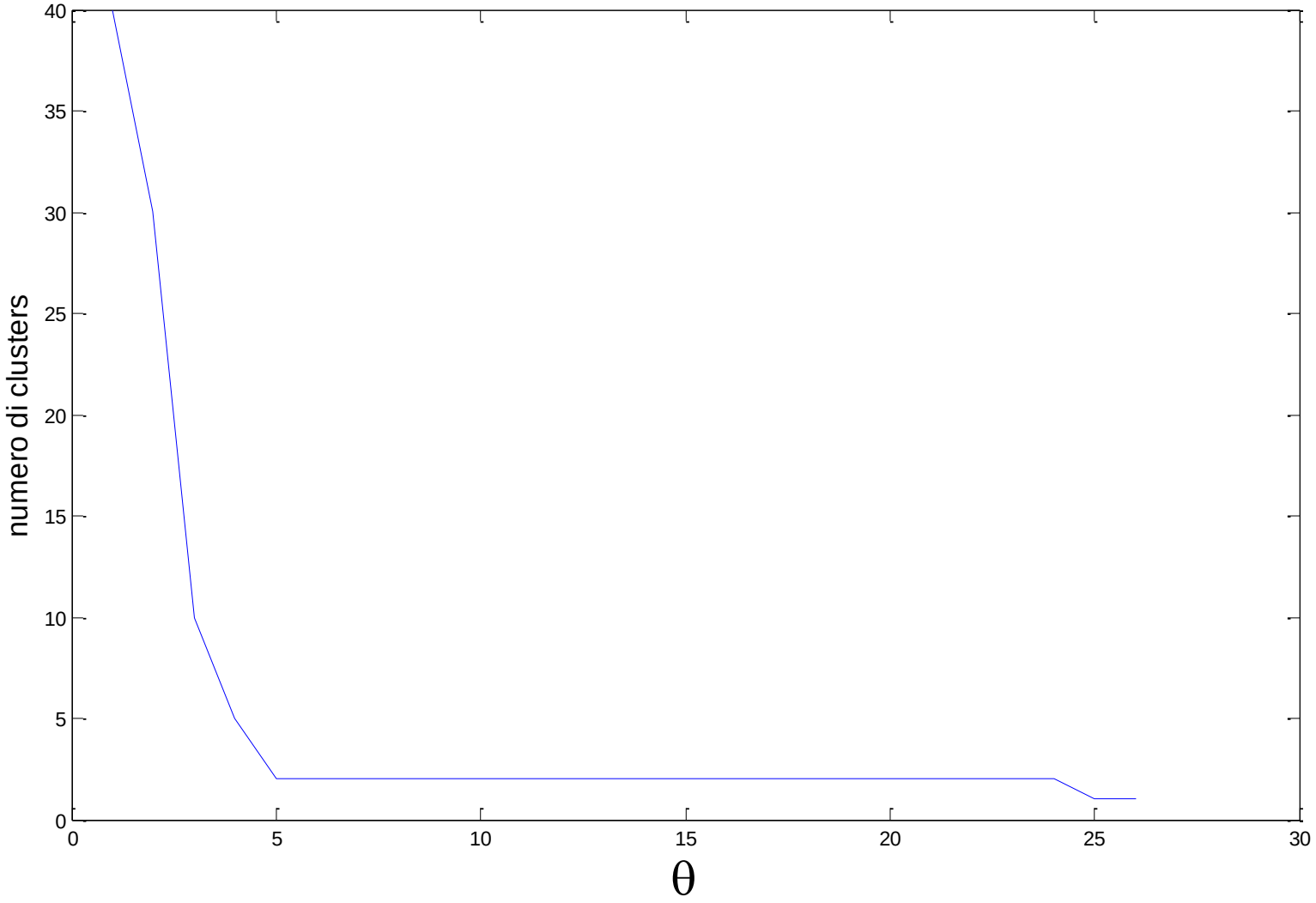
Commenti su BSAS:

- si può osservare che l'ordine con cui vengono processati i pattern è cruciale
  - ordini diversi possono produrre risultati diversi
- la scelta della soglia θ è cruciale
  - θ troppo piccola, vengono determinati troppi cluster
  - θ troppo grande, troppo pochi cluster
- si può scambiare la dissimilarità con la similarità (cambiando min con max e > con <)
- con i rappresentanti (con le medie) i cluster che escono sono compatti

Clustering sequenziale

- Metodo per calcolare il numero ottimale di clusters:
  - for θ = *a* to *b* step *c*
    - Eseguire *s* volte l'algoritmo BSAS, ogni volta processando i pattern con un ordine differente
    - stimare m_θ come il numero più frequente di cluster
  - end for
  - visualizzare il numero di cluster m_θ vs il parametro θ
  - il numero di cluster ottimale è quello della regione "piatta" più lunga
- dettagli
  - *a* è la distanza minima tra i punti, *b* la distanza massima
  - assumiamo che "esista" un clustering

Clustering sequenziale



Center-based clustering

K-means

- Algoritmo più famoso di clustering partizionale
- IDEE:
  - minimizza una funzione di errore
  - ogni cluster è rappresentato dalla sua media
  - si parte da una clusterizzazione iniziale, ed ad ogni iterazione si assegna ogni pattern alla media più vicina
  - si riaggiornano le medie
  - si continua fino a convergenza

Center-based clustering

- Commenti
  - il numero di cluster deve essere fissato a priori
  - l'ottimizzazione spesso porta ad un ottimo "locale"
    - l'inizializzazione è cruciale: una cattiva inizializzazione porta ad un clustering pessimo
  - è molto efficiente nel clusterizzare dataset grandi, perché la sua complessità computazionale è linearmente dipendente dalla dimensione del data set
  - i cluster ottenuti hanno una forma convessa
  - lavora solo su dati vettoriali numerici (deve calcolare la media)
  - tipicamente viene utilizzata la distanza euclidea

Center based clustering

Varianti del K-means

- cercare di migliorare l'inizializzazione ([Anderberg 1973])
- ISODATA (*Iterative Self-Organizing Data Analysis Techniques*)
  - permettere lo splitting e il merging dei cluster risultanti
  - Ad ogni iterazione effettua dei controlli sui cluster risultanti:
    - un cluster viene diviso se la sua varianza è sopra una soglia prefissata, oppure se ha troppi punti
    - due cluster vengono uniti se la distanza tra i due relativi centroidi è minore di un'altra soglia prefissata, oppure se hanno troppo pochi punti
  - la scelta delle soglie è cruciale, ma fornisce anche una soluzione alla scelta automatica del numero di cluster

Center based clustering

Varianti del K-means

- utilizzo della distanza di Mahalanobis come distanza per i punti ([Mao Jain 1996])
  - vantaggio: posso anche trovare cluster ellissoidali
  - svantaggio: devo calcolare ogni volta la matrice di covarianza
- PAM (Partitioning around the medoids)
  - l'idea è quella di utilizzare come "centri" del K-means i medoidi (o i punti più centrali) invece che le medie
    - non introduco nuovi elementi nel dataset
    - più robusto agli outliers
    - posso lavorare anche con dati non vettoriali (data una funzione di distanza tra questi dati)

Model-based clustering

- IDEE:
  - utilizzare un insieme di modelli per i cluster
  - l'obiettivo diventa quello di massimizzare il fit tra i modelli e i dati
  - si assume che i dati siano generati da una mistura di funzioni di probabilità differenti, ognuna delle quali rappresenta un cluster
  - ovviamente il metodo di clustering funziona bene se i dati sono conformi al modello
- Due approcci al model based clustering
  - classification likelihood approach
  - mixture likelihood approach

Model-based clustering

- Commenti:
  - mixture likelihood approach: l'algoritmo più utilizzato è l'EM
  - esistono molti risultati sulla determinazione del numero ottimale di clusters
  - esistono alcuni approcci anche per inizializzare l'EM

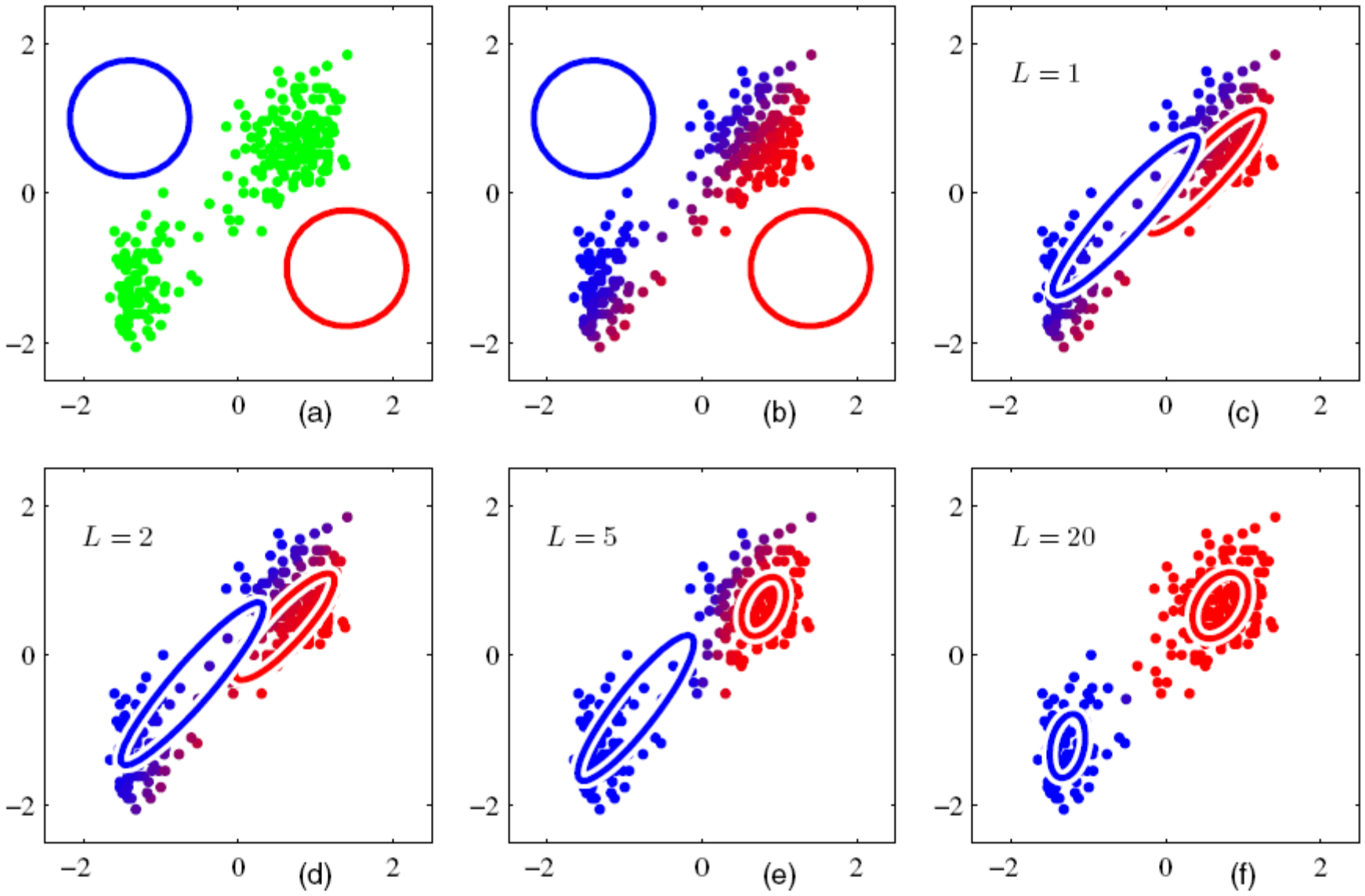
  - classification likelihood approach: è equivalente al K-means assunto che:
    - matrice di covarianza uguale per tutti i cluster
    - matrice di covarianza proporzionale all'identità

Model based clustering

Gaussian Mixture Models (GMM) Clustering

- tecnica di soft clustering molto utilizzata (mixture likelihood approach)
- la mistura è composta da Gaussiane
- il modello è stimato utilizzando Expectation-Maximization (EM)

Esempio



Model based clustering

Commenti:

- molto utilizzato in svariati contesti
- l'inizializzazione è un problema
- Numero di cluster: il problema può essere visto come un problema di model selection:
  - qual'è la miglior dimensione del modello dati i dati?

Clustering gerarchico

- Algoritmi di clustering che generano una serie di partizioni innestate
- Rappresentazione di un clustering gerarchico: il dendrogramma

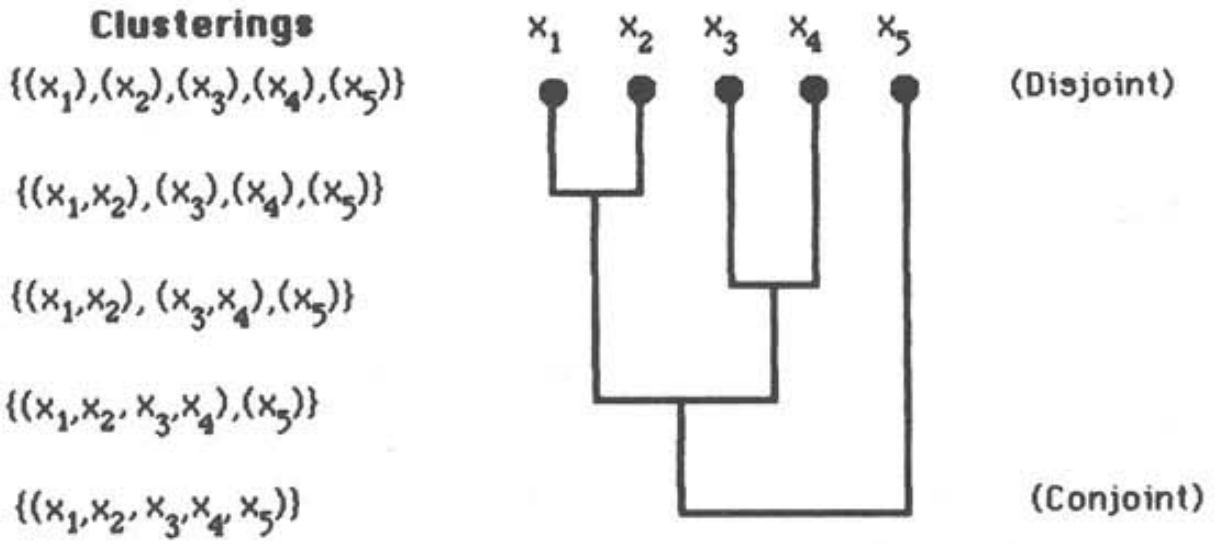


Figure 3.2 Example of dendrogram.

Clustering gerarchico

- Clustering gerarchico agglomerativo:
  - si parte da una partizione in cui ogni cluster contiene un solo elemento
  - si continua a fondere i cluster più "simili" fino ad avere un solo cluster
  - definizioni diverse del concetto di "cluster più simili" genera algoritmi diversi
- Approcci più utilizzati:
  - single link
  - complete link
  - formulazione con le matrici

Clustering gerarchico

- Clustering gerarchico agglomerativo:
  - si parte da una partizione in cui ogni cluster contiene un solo elemento
  - si continua a fondere i cluster più "simili" fino ad avere un solo cluster
  - definizioni diverse del concetto di "cluster più simili" genera algoritmi diversi
- Approcci più utilizzati:
  - single link
  - complete link
  - formulazione con le matrici

Clustering gerarchico

SL $d[(k), (r, s)] = \min \{d[(k), (r)], d[(k), (s)]\}$

CL $d[(k), (r, s)] = \max \{d[(k), (r)], d[(k), (s)]\}$

	1	2	3	4	5
1	0	2.3	3.4	1.2	3.7
2		0	2.6	1.8	4.6
3			0	4.2	(0.7)
4				0	4.4
5					0

	1	2	3,5	4
1	0	2.3	3.4	(1.2)
2		0	2.6	1.8
3,5			0	4.2
4				0

	1,4	2	3,5
1,4	0	(1.8)	3.4
2		0	2.6
3,5			0

	1,2,4	3,5
1,2,4	0	(2.6)
3,5		0

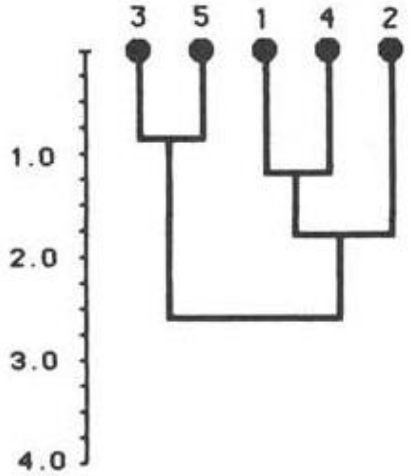
single link

	1	2	3,5	4
1	0	2.3	3.7	(1.2)
2		0	4.6	1.8
3,5			0	4.4
4				0

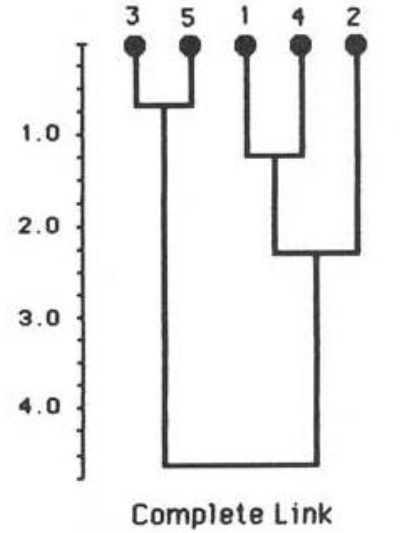
	1,4	2	3,5
1,4	0	(2.3)	4.4
2		0	4.6
3,5			0

	1,2,4	3,5
1,2,4	0	(4.6)
3,5		0

complete link



Single Link



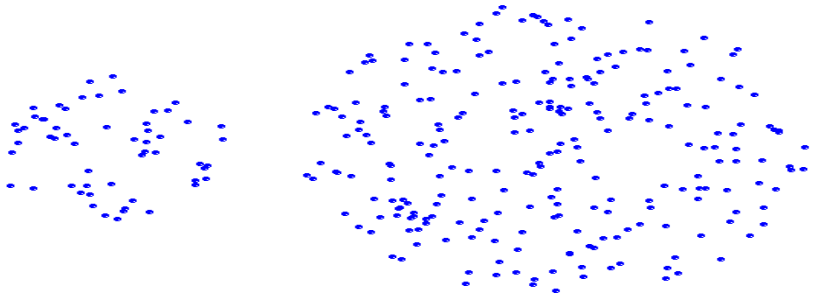
Complete Link

Clustering gerarchico

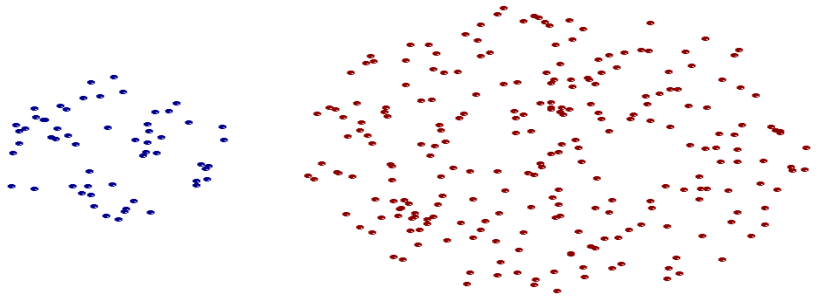
Commenti:

- Single link unisce due cluster se esiste un solo edge
  - tende a formare cluster allungati
- Complete link unisce due cluster se tutti gli elementi sono connessi
  - più conservativo, tende a formare cluster convessi
- In generale è stato dimostrato che Complete Link funziona meglio

Strengths of single-link clustering



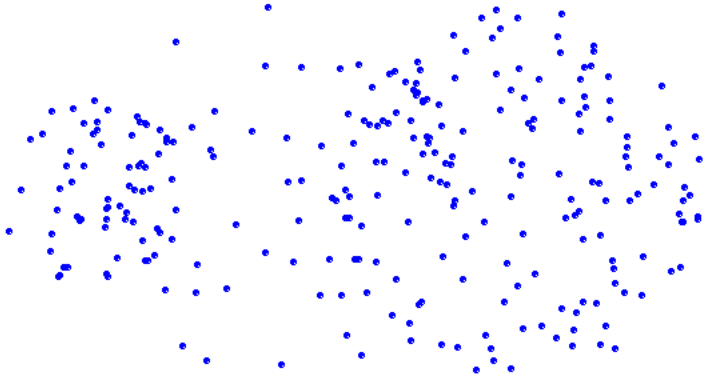
Original Points



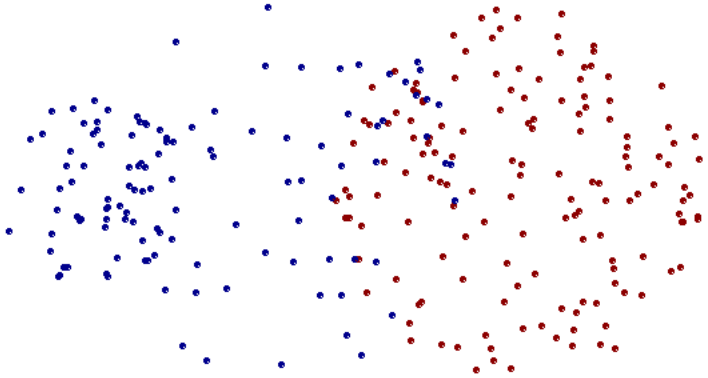
Two Clusters

- **Can handle non-elliptical shapes**

Limitations of single-link clustering



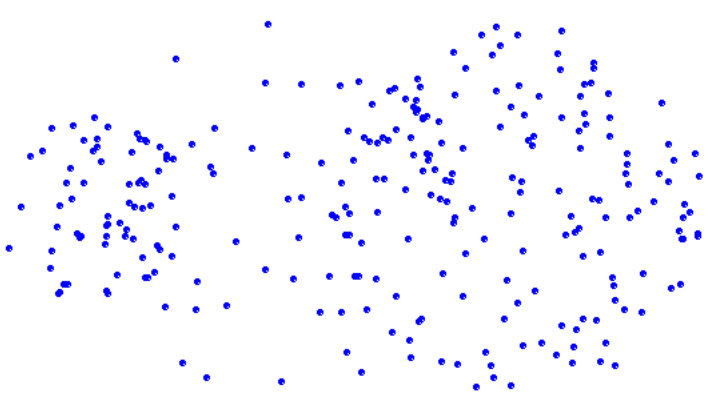
Original Points



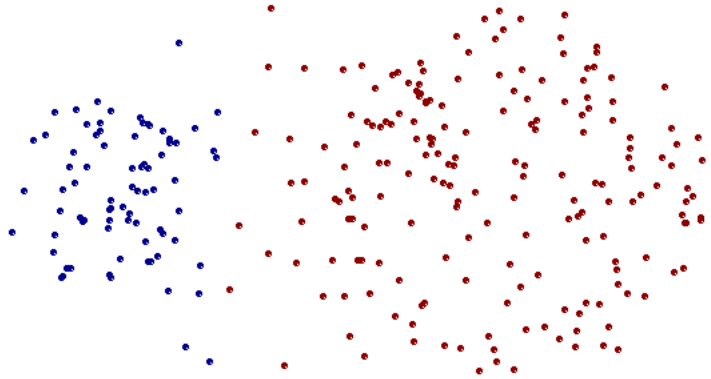
Two Clusters

- **Sensitive to noise and outliers**
- **It produces long, elongated clusters**

Strengths of complete-link clustering



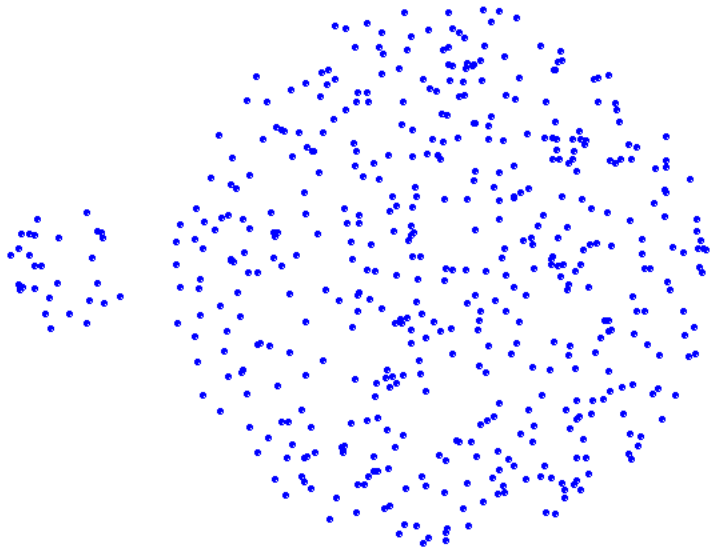
Original Points



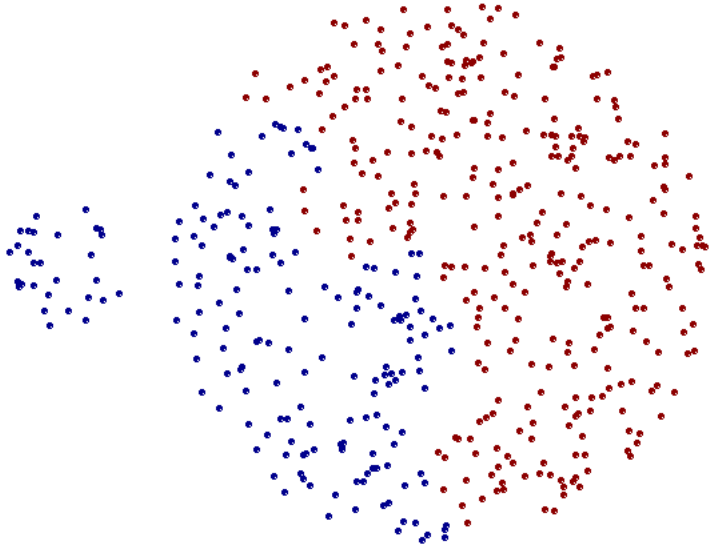
Two Clusters

- **More balanced clusters (with equal diameter)**
- **Less susceptible to noise**

Limitations of complete-link clustering



Original Points



Two Clusters

- **Tends to break large clusters**
- **All clusters tend to have the same diameter – small clusters are merged with larger ones**