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
  - (dettagli alla lavagna)

29

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

30

Model based clustering

Gaussian Mixture Models (GMM) Clustering

⇨ tecnica di soft clustering molto utilizzata (mixture likelihood approach)

⇨ la mistura è composta da Gaussiane

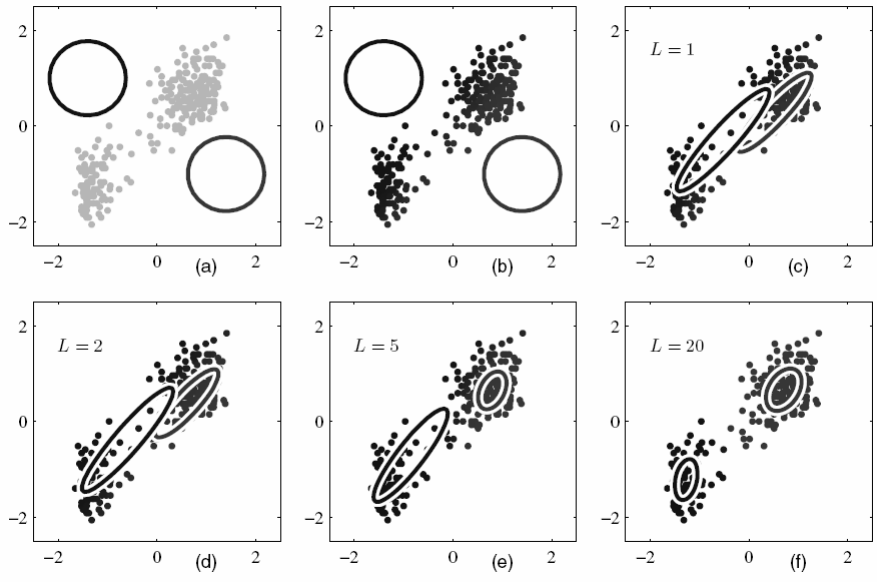
⇨ il modello è stimato utilizzando Expectation-Maximization (EM)

⇨ (alla lavagna)

- ⇨ definizione di GMM
- ⇨ GMM come modello a variabili nascoste
- ⇨ ottimizzazione della likelihood (algoritmo EM)

31

Esempio



EM in generale

⇨ Obiettivo EM: trovare una soluzione maximum likelihood per modelli con variabili latenti (nascoste)

⇨ Abbiamo un insieme di dati osservati X, un insieme di variabili latenti Z, un insieme di parametri θ

$$\ln p(\mathbf{X}|\boldsymbol{\theta}) = \ln\left\{\sum_{\mathbf{Z}} p(\mathbf{X},\mathbf{Z}|\boldsymbol{\theta})\right\}$$

33

⇨ data set "completo" {**X**,**Z**}

⇨ data set "incompleto" **X**

⇨ Noi abbiamo solo il data set incompleto **X**, la nostra conoscenza sui valori delle variabili nascoste **Z** è data solo dalla prob. a posteriori p(**Z**|**X**,**θ**)

⇨ Non possiamo utilizzare la log likelihood dei dati completi

⇨ Ma possiamo calcolare il suo valore atteso sotto la posterior (E step), massimizzando poi questo valore atteso (M step)

34

- ⇨ Nell' E step, utilizziamo i parametri correnti $\boldsymbol{\theta}^{\text{old}}$ per calcolare la posterior delle variabili nascoste $p(\mathbf{Z}|\mathbf{X},\boldsymbol{\theta}^{\text{old}})$.
- ⇨ Poi utilizziamo questa posterior per trovare il valore atteso della log likelihood dei dati completi (funzione che chiamiamo Q)

$$\mathcal{Q}(\boldsymbol{\theta}, \boldsymbol{\theta}^{\text{old}}) = \sum_{\mathbf{Z}} p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta}^{\text{old}}) \ln p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta})$$

- ⇨ Nell'M step, calcoliamo la nuova stima dei parametri $\boldsymbol{\theta}^{\text{new}}$ massimizzando

$$\boldsymbol{\theta}^{\text{new}} = \arg\max_{\boldsymbol{\theta}} \mathcal{Q}(\boldsymbol{\theta}, \boldsymbol{\theta}^{\text{old}})$$

- ⇨ Ogni ciclo dell'EM aumenta la log likelihood dei dati incompleti

35

The General EM Algorithm

Given a joint distribution $p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta})$ over observed variables $\mathbf{X}$ and latent variables $\mathbf{Z}$, governed by parameters $\boldsymbol{\theta}$, the goal is to maximize the likelihood function $p(\mathbf{X}|\boldsymbol{\theta})$ with respect to $\boldsymbol{\theta}$.

1. Choose an initial setting for the parameters $\boldsymbol{\theta}^{\text{old}}$.
2. **E step** Evaluate $p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta}^{\text{old}})$.
3. **M step** Evaluate $\boldsymbol{\theta}^{\text{new}}$ given by

$$\boldsymbol{\theta}^{\text{new}} = \arg\max_{\boldsymbol{\theta}} \mathcal{Q}(\boldsymbol{\theta}, \boldsymbol{\theta}^{\text{old}}) \tag{9.32}$$

where

$$\mathcal{Q}(\boldsymbol{\theta}, \boldsymbol{\theta}^{\text{old}}) = \sum_{\mathbf{Z}} p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta}^{\text{old}}) \ln p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta}). \tag{9.33}$$

4. Check for convergence of either the log likelihood or the parameter values. If the convergence criterion is not satisfied, then let

$$\boldsymbol{\theta}^{\text{old}} \leftarrow \boldsymbol{\theta}^{\text{new}} \tag{9.34}$$

and return to step 2.

36

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?
- ⇨ vedi slides su HMM (equivalente alla determinazione del miglior numero di stati di un HMM)

37

Clustering gerarchico

⇨ Algoritmi di clustering che generano una serie di partizioni innestate

⇨ Rappresentazione di un clustering gerarchico: il dendrogramma

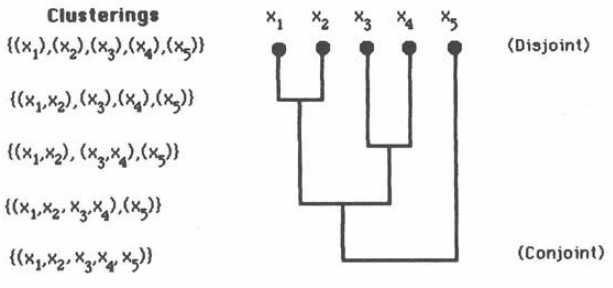


Figure 3.2 Example of dendrogram.

38

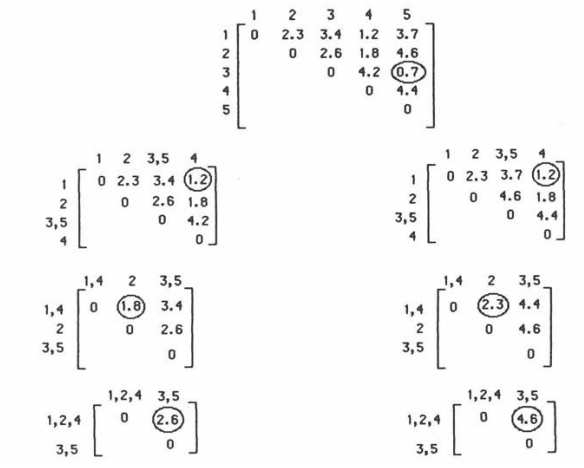
Clustering gerarchico

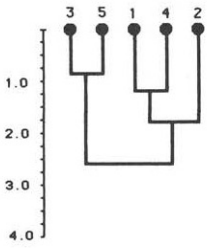
- ⇨ Clustering gerarchico agglomerativo:
  - ⇨ si parte da una partizione in cui ogni cluster contiene un solo elemento
  - ⇨ si continua a fondere i cluster più "simili" fino ad avere un solo cluster
  - ⇨ definizioni diverse del concetto di "cluster più simili" genera algoritmi diversi
- ⇨ Approcci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

Clustering gerarchico

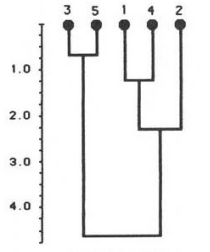
SL $d[(k), (r, s)] = \min \{d[(k), (r)], d[(k), (s)]\}$

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





Single Link



Complete Link

single link

complete link

Clustering gerarchico

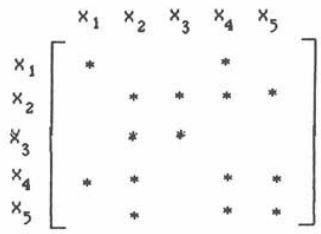
Single link / Complete link:
formulazione con i grafi (cenni)

⇨ Definizione "threshold graph" G(v)

- ⇨ grafo in cui i nodi sono tutti gli elementi
- ⇨ esiste un arco tra *i* e *j* se *d(i,j) <= v*

ESEMPIO: G(5)

$$\mathcal{D}_1 = \begin{array}{c} \\ x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \end{array} \begin{array}{c} \begin{array}{ccccc} x_1 & x_2 & x_3 & x_4 & x_5 \end{array} \\ \begin{bmatrix} 0 & 6 & 8 & 2 & 7 \\ 6 & 0 & 1 & 5 & 3 \\ 8 & 1 & 0 & 10 & 9 \\ 2 & 5 & 10 & 0 & 4 \\ 7 & 3 & 9 & 4 & 0 \end{bmatrix} \end{array}$$



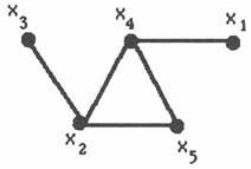


Figure 3.3 Binary relation and threshold graph for threshold 5.

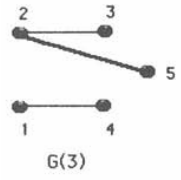
Clustering gerarchico

⇨ Si calcolano tutti i possibili grafi G(v)

- ⇨ Assunzioni: no ties
- ⇨ v varia tra gli interi (scala ordinale)

⇨ Single Link:

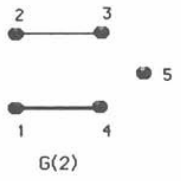
- ⇨ i cluster sono le componenti connesse (sottografi massimamente connessi)



G(3)

⇨ Complete Link

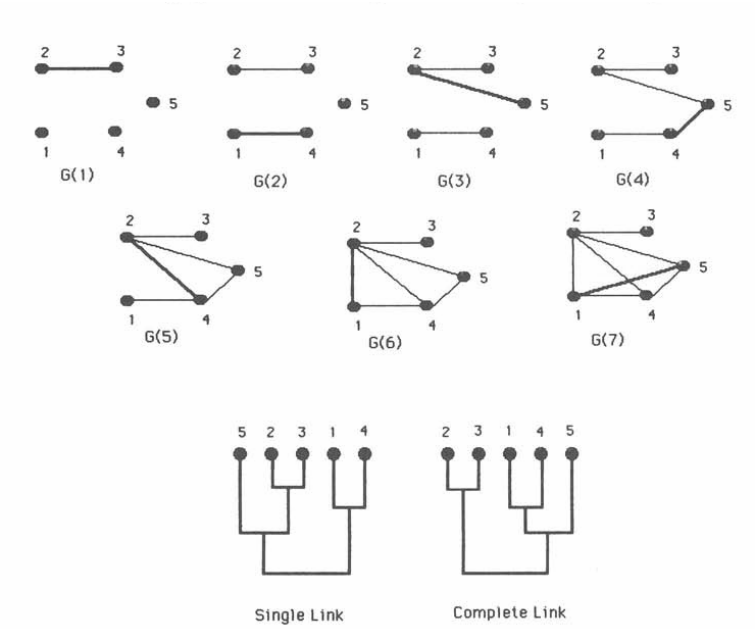
- ⇨ i cluster sono clique tra cluster precedentemente determinati
  - ⇨ clique: sottografo massimamente completo – tutti con tutti



G(2)

42

Clustering gerarchico



43

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

44

Clustering gerarchico

Altri criteri di unione dei cluster

⇨ UPGMA (Unweighted pair group method using arithmetic averages)

- ⇨ la distanza tra cluster è definita come la media delle distanze di tutte le possibili coppie formate da un punto del primo e un punto del secondo
- ⇨ utilizzato nel periodo iniziale della filogenesi

⇨ Metodo di Ward

- ⇨ fonde assieme i cluster che portano alla minima perdita di informazione
- ⇨ informazione intesa in termini di varianza

45

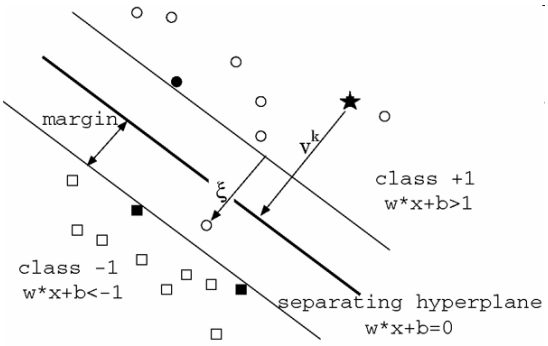
Altre metodologie

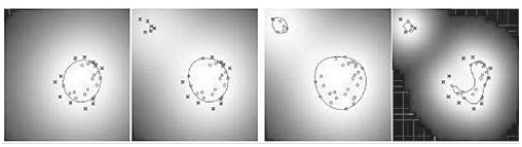
Clustering con Kernel Machines

⇨ Idea, utilizzo delle One Class Support Vector Machines

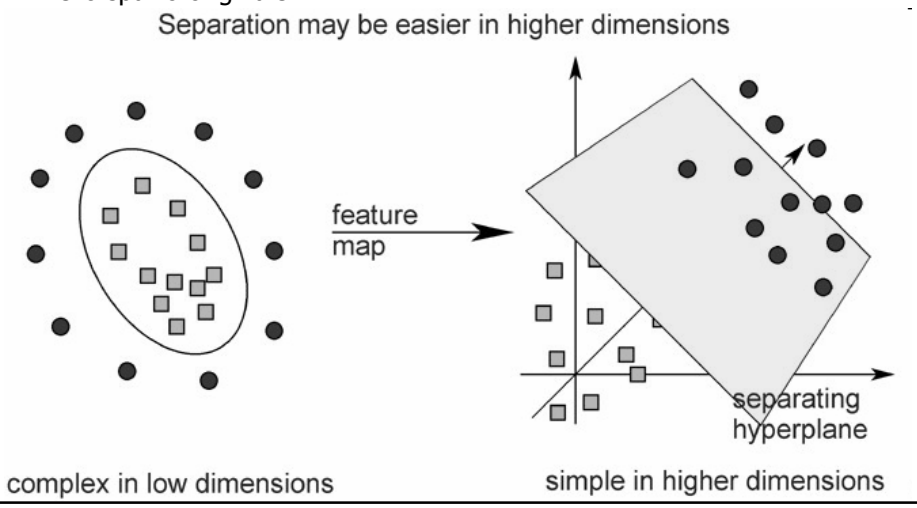
⇨ One Class Support Vector Machines

- ⇨ SVM che si addestrano con una sola classe
- ⇨ SVM trovano il miglior iperpiano, OCSVM trova la miglior ipersfera che racchiude i dati





- possibilità di gestire bene gli outliers (la funzione da minimizzare tiene conto anche di eventuali errori)
- con il trucco dei kernel riesco ad ottenere un'ipersfera nello spazio altamente dimensionale che corrisponde ad una superficie complessa nello spazio originale



Altre metodologie

- Idea del clustering:
  - rappresentare ogni cluster con una OC-SVM
  - utilizzare un algoritmo iterativo come K-means
  - distanza = distanza dal centro della sfera (si può calcolcare)
  - ricalcolo dei rappresentanti (addestramento di ogni OC-SVM con i punti assegnati al cluster corrispondente)
  - si ottengono cluster di qualsiasi forma
  - Problemi: determinare i parametri ottimali della OC-SVM
  - Altre versioni: versione soft clustering

48

Altre metodologie

- ⇨ Fuzzy clustering:
  - ⇨ Tecniche di clustering che si basano sulla teoria dei Fuzzy Sets (Zadeh 1965)
  - ⇨ Nella classica teoria degli insiemi, l'appartenza di un punto ad un insieme è una variabile binaria (o appartiene o non appartiene)
  - ⇨ La teoria degli insiemi fuzzy permette che un elemento appartenga a più di un insieme contemporaneamente
  - ⇨ questo è descritto tramite una funzione di membership
    - ⇨ a valori nell'intervallo [0, 1].
- ⇨ Logica Fuzzy:
  - ⇨ logica dove le variabili non hanno un valore binario ma un valore nell'intervallo [0,1]
    - ⇨ 0.8 -> quasi vero

49

Altre metodologie

Differenza tra probabilità e funzione di membership

- ⇨ Sottile e controversa
- ⇨ Più accettata:
  - ⇨ probabilità: approccio frequentista
    - ⇨ un oggetto appartiene ad una sola classe (e.g. testa o croce)
    - ⇨ la probabilità misura quanto spesso l'oggetto appartiene ad una classe
    - ⇨ approccio basato sulla misura (ripetizioni)
  - ⇨ funzione di membership
    - ⇨ un oggetto può appartenere a più classi
    - ⇨ più appropriato per concetti sfumati e soggettivi (esempio il concetto di caldo, freddo, alto, basso)

50

Altre metodologie

- Clustering con logica fuzzy
  - re-implementazione di diverse tecniche di clustering con la logica fuzzy
  - Esempio: fuzzy K-means
  - Minimizza la funzione

$$J_q(U,V) = \sum_{i=1}^{n}\sum_{j=1}^{k} u_{ij}^{q} d^2(\mathbf{x}_i, \mathbf{V}_j)$$

  - u funzione di membership
  - q controlla la "fuzziness" del clustering risultante

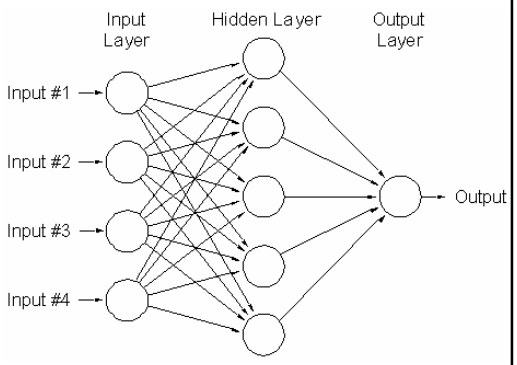
51

Altre metodologie

Clustering con le reti neurali

Reti neurali: modello di calcolo che replica il meccanismo del cervello umano

- Tante piccole unità di calcolo elementari (i neuroni)
- I neuroni sono collegati assieme a formare una rete (livelli nascosti)
- L'uscita di un neurone dipende dalla somma pesata dei suoi ingressi (pesi = sinapsi)
- Esiste una funzione di attivazione che determina l'eventuale "accendersi" del neurone



52

Altre metodologie

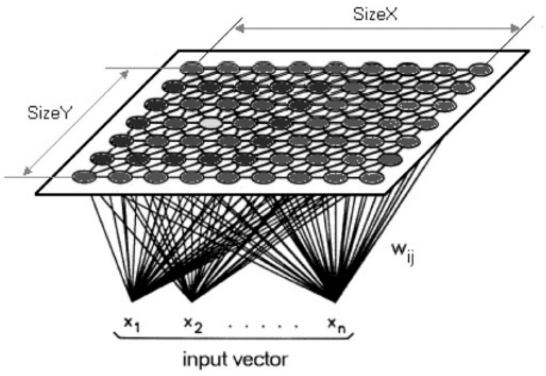
- ⇨ Commenti:
  - ⇨ parametri da settare: numero di neuroni per livello, numero di strati nascosti, funzione di attivazione
    - ⇨ esistono alcuni teoremi ma tipicamente la scelta è euristica
  - ⇨ Addestramento del modello:
    - ⇨ calcolare i pesi della rete dato un training set (ingressi-uscite)
    - ⇨ funzione difficile da minimizzare
    - ⇨ tipicamente molti parametri da settare
  - ⇨ Problema: scatola nera

53

Altri metodi

- ⇨ Clustering con le reti neurali: le SOM (Self-Organizing Maps)
- ⇨ SOM:

- ⇨ reti neurali feed forward ad un singolo strato, dove l'uscita è tipicamente bi-dimensionale
- ⇨ ogni ingresso è connesso a tutti i vettori in uscita
- ⇨ ogni neurone ha un vettore di pesi della stessa dimensionalità dell'ingresso



54

Altri metodi

Funzionamento:

Partendo da una configurazione iniziale, ripetere fino a convergenza

⇨ Processo competitivo: dato un vettore in ingresso, viene calcolato il neurone più "simile" (neurone vincente)

- ⇨ il neurone che si "attiva" di più
- ⇨ il neurone il cui vettore di pesi è il più simile al vettore in ingresso

⇨ Processo adattivo

- ⇨ viene modificato il vettore dei pesi del neurone vincente e di tutto il suo vicinato (tipicamente definito con una gaussiana)
- ⇨ viene modificato in modo da assomigliare al vettore di ingresso

55

Altri metodi

Interpretazione:

⇨ nella fase di addestramento i pesi di tutto il vicinato sono spostati nella stessa direzione

- ⇨ elementi simili tendono ad eccitare neuroni adiacenti.

⇨ le SOM formano una mappa semantica dove campioni simili vengono mappati vicini e i dissimili distanti.

⇨ Rappresentano un ottimo modo di visualizzare dati altamente dimensionali

56