Riconoscimento e recupero dell'informazione per bioinformatica

Rappresentazione dati e visualizzazione

Manuele Bicego

Corso di Laurea in Bioinformatica
Dipartimento di Informatica - Università di Verona

Sommario

- Rappresentazione dei dati:
  - campionamento
  - estrazione delle features: tipo di dato e tipo di pattern
  - rappresentazione preprocessing (scaling, riduzione del rumore, riduzione della dimensionalità)
  - (visualizzazione)

2

La rappresentazione dei dati

- Rappresentazione dei dati: il problema di come rappresentare gli oggetti del problema in esame
  - Rappresentazione quantitativa/qualitativa
  - Rappresentazione utilizzabile per calcolare la similarità/distanza, per costruire il modello, per fare il testing
  - scelta cruciale

3

La rappresentazione dei dati

Fasi:

1. campionamento (acquisizione dati)
2. estrazione delle features
3. rappresentazione
4. preprocessing:
   - scaling
   - riduzione del rumore
   - riduzione della dimensionalità
5. (visualizzazione)

4

Fase 1: campionamento

⇨ Esempio 1: modellare pesci

campionamento

immagine (dato grezzo)

⇨ L'immagine può essere vista come un punto in uno spazio MxN dimensionale (immagine di dimensione MxN)

5

Fase 1: campionamento

⇨ Problemi da tenere in considerazione:

- ⇨ tipo di sensore: caratteristiche tecniche
- ⇨ frequenza di campionamento: capacità di modellare evoluzione temporale
- ⇨ risoluzione (dipende dal sensore): quanti dettagli si riesce a recuperare
- ⇨ capacità di gestire cambiamenti di condizioni al contorno (eg. cambi di illuminazione)

6

Fase 1: campionamento

- ⇨ Scelta del sensore è la prima scelta cruciale
  - ⇨ Esempio: filogenesi di microrganismi
    - ⇨ l'obiettivo è sequenziare il gene scelto per fare filogenesi nei microrganismi in esame
    - ⇨ far crescere i batteri (scelta del terreno di crescita,...)
    - ⇨ estrarre il DNA e amplificare quello dei geni prescelti per la filogenesi (scelta dei primer, scelta dei parametri della PCR)
    - ⇨ sequenziare
  - ⇨ Esempio: riconoscimento di facce

telecamera a infrarosso:

- funziona anche al buio
- risoluzione bassa

telecamera tradizionale:

-sensibile ai cambi di illuminazione

-alta risoluzione

7

Fase 2: estrazione delle features

- ⇨ Trovare una rappresentazione "quantitativa" più compatta e più descrittiva dell'oggetto
- ⇨ Definizione di feature:
  - ⇨ caratteristica del problema in esame
    - ⇨ rilevante
    - ⇨ discriminante
    - ⇨ quantificabile a partire dai dati grezzi
    - ⇨ (interpretabile)
- ⇨ La scelta della rappresentazione è chiaramente cruciale

8

Fase 2: estrazione delle features

⇨ Opzione 1:

Altezza e larghezza

Tipo di dato risultante:

vettore di due dimensioni [a,l]

⇨ Opzione 2:

Contorno dell'oggetto, a partire dalla coda

Tipo di dato risultante:

sequenza di vettori di due dimensioni (le coordinate di ogni punto del contorno)

x1,y1
x2,y2
x3,y3
...
xn,yn

9

⇨ Opzione 1: vantaggi:
- ⇨ rappresentazione compatta
- ⇨ ogni oggetto è un punto in uno spazio vettoriale bidimensionale
- ⇨ non è difficile da calcolare

⇨ Opzione 1: svantaggi
- ⇨ troppo semplificata: non riesce a modellare la forma del pesce, il colore

⇨ Opzione 2: vantaggi
- ⇨ rappresentazione più ricca, modella la forma del pesce

⇨ Opzione 2: svantaggi
- ⇨ più complicata da calcolare
- ⇨ il pattern risultante è una sequenza (che può essere di dimensione diversa a seconda del pesce). Non siamo quindi più in uno spazio vettoriale

10

Fase 3: Rappresentazione

Due concetti:

⇨ Tipo di dato:

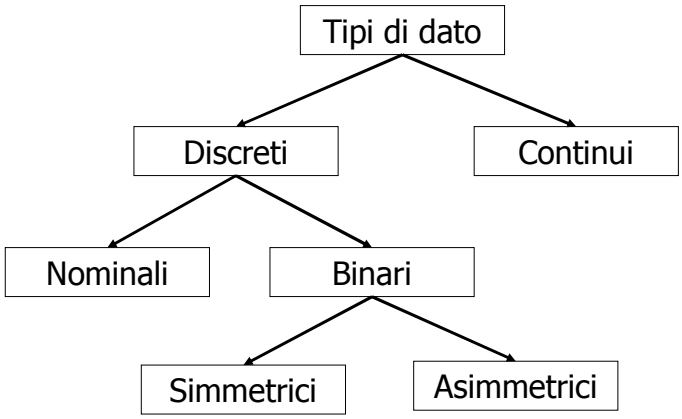
- ⇨ come viene codificata ogni singola feature
- ⇨ Esempio: numero intero, un valore di verità

⇨ Tipo di pattern:

- ⇨ come sono messe assieme le varie feature per un singolo oggetto
- ⇨ Esempio: un vettore di numeri interi (altezza-larghezza)

11

Tipi di dato



12

Tipi di dato

- Continui: il valore della feature può assumere un numero infinito di valori (e.g. numeri reali)
  - attenzione: campionamento!
- Discreti: il valore della feature può essere solo uno di un insieme finito di valori possibili
- Nominali: il valore della feature può essere uno di un insieme di nomi (o di simboli) – tipo di dato discreto
- EXE: sequenza di DNA:
  - T Timina
  - A Adenina
  - G Guanina
  - C Citosina

13

Tipi di dato

- Dati binari: dati che possono assumere solo due valori:
  - 0/1, vero/falso
- Dati binari simmetrici: i due valori sono ugualmente importanti
  - esempio: maschio/femmina
- Dati binari asimmetrici: uno dei due valori porta più informazione dell'altro
  - esempio:
    - sì: presenza di un attributo
    - no: assenza

14

Tipi di pattern: i vettori

Dati vettoriali (formato più diffuso)

⇨ per ogni oggetto viene estratto un insieme prefissato di features

⇨ Esempio:

Altezza e larghezza

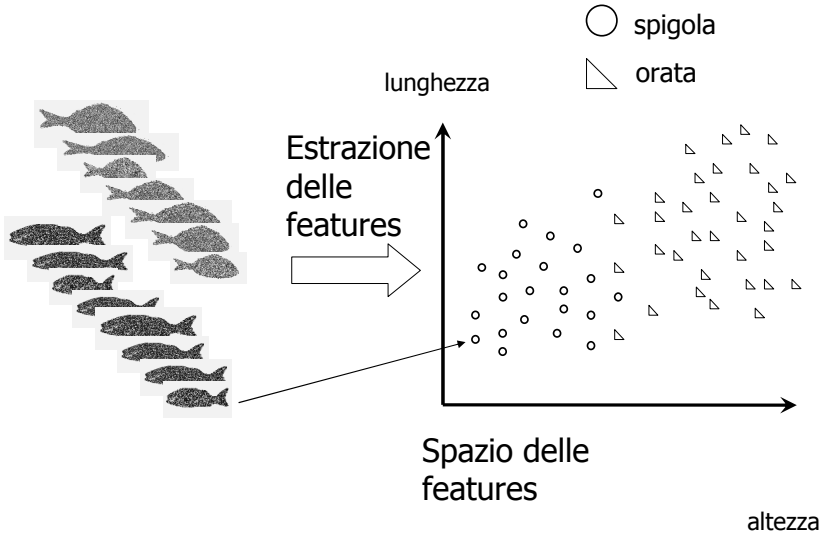
⇨ il vettore è ordinato

⇨ l'oggetto viene proiettato in un punto in uno spazio d-dimensionale, detto "spazio delle features"

⇨ "d" è il numero delle features

15

Tipi di pattern: i vettori



16

Commenti su spazi vettoriali

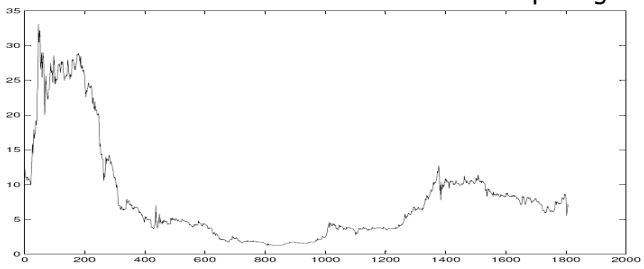
- La scelta delle features è cruciale
- Utilizzando molte features gli spazi diventano estremamente grandi
  - problema della curse of dimensionality
  - problema di visualizzare i dati
- Esistono metodi che riducono la dimensionalità di questi spazi (vedremo in seguito)
- Il problema della "scalatura" dello spazio

17

Tipi di pattern: le sequenze

- Sequenze: dati che si presentano in forma ordinata e sequenziale (uno dopo l'altro): è importante l'ordine

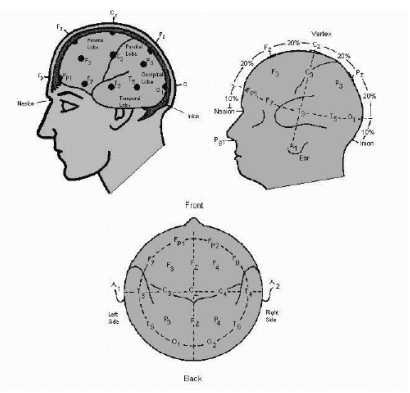
  $x_1, x_2, x_3, x_4 \ldots x_T$
- Sequenze temporali:
  - sequenza di features (di diverso tipo) che rappresentano la misurazione di un fenomeno a intervalli di tempo regolari

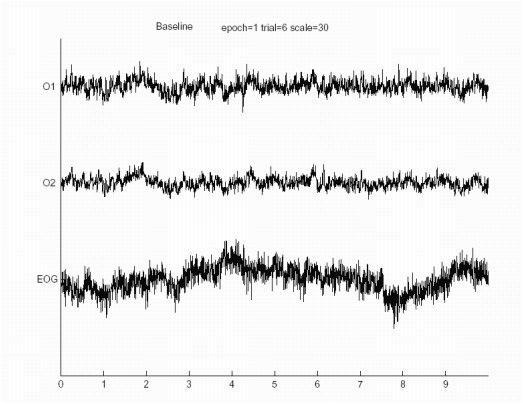


Esempio 1: indici di mercato (DowJones)

18

Tipi di pattern: le sequenze





⇨ Esempio 2: segnali elettro encefalografici (EEG)

19

Tipi di pattern: le sequenze

⇨ Esempio 3: sequenze di immagini (video)

⇨ Goal: capire cosa succede nella scena, ad esempio i diversi comportamenti delle persone: clustering di traiettorie

20

Tipi di pattern: le sequenze

⇨ Sequenze non temporali:

⇨ sequenze dove l'ordine non è dato dal tempo

⇨ Esempio 1:

⇨ sequenze nulceotidiche

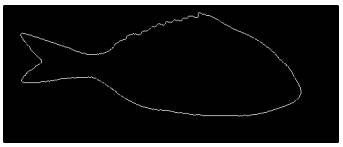
```
atgcgatcgatcgatcgatcaggcgcgctacgagcggcgagga
cctcatcatcgatcag
```

⇨ sequenze aminoacidiche

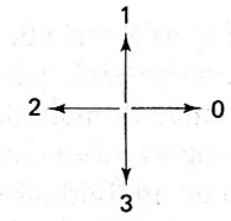
```
MRPQAPGSLVDPNEDELRMAPWYWGRISREEAKSILHGKPDGSFLVRDALSMKGEYTLTLMK
DGCEKLIKICHMDRKYGFIETDLFNSVVEMINYYKENSLSMYNKTLDITLSNPIVRAREDEE
SQPHGDLCLLSNEFIRTCQLLQNLEQNLENKRNSFNAIREELQEKKLHQSVFGNTEKIFRNQ
IKLNESFMKAPADA……
```


Tipi di pattern: le sequenze

⇨ Esempio 2: codifica di un contorno di una forma 2D



contorno della forma

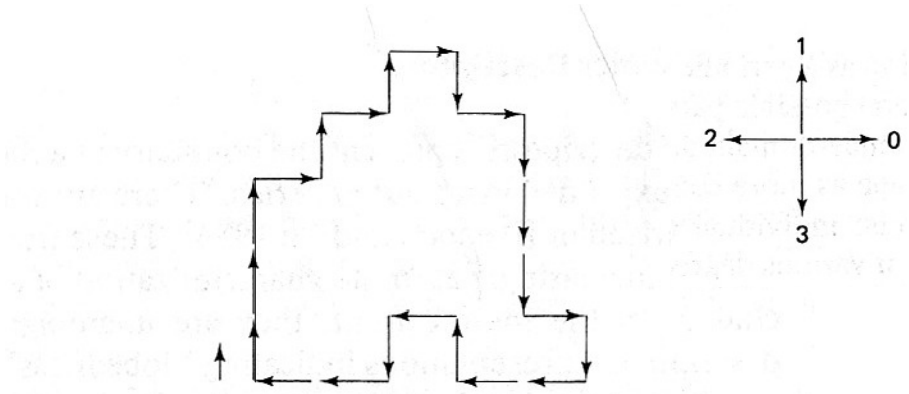


Chain code: specifica la direzione del contorno ad ogni punto di edge

le direzioni sono quantizzate in 4/8 valori

Rappresentazione: coordinate del punto iniziale e una sequenza di chain code che segue il contorno.

Tipi di pattern: le sequenze



(b) Chain code: 1 1 1 0 1 0 1 0 3 0 3 3 3 0 3 2 2 1 2 3 2 2
(c) Derivative: 1 0 0 3 1 3 1 3 3 1 3 0 0 1 3 3 0 3 1 1 3 0

23

Tipi di pattern: le sequenze

Riassumendo le caratteristiche delle sequenze:

⇨ esiste un concetto di "sequenzialità" dei dati

⇨ l'ordine è importante

- ⇨ xy non è la stessa cosa di yx

⇨ la lunghezza di due sequenze diverse potrebbe essere diversa

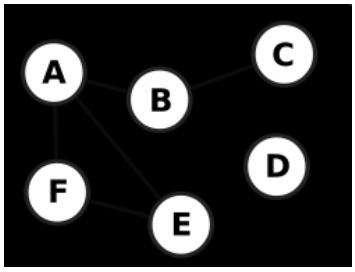
- ⇨ e.g. lunghezza contorno di un oggetto
- ⇨ e.g. lunghezza gene

⇨ Sequenze di lunghezza diversa non possono essere rappresentate in uno spazio vettoriale (gli spazi necessari sarebbero di dimensione diversa)

24

Tipi di pattern: i grafi

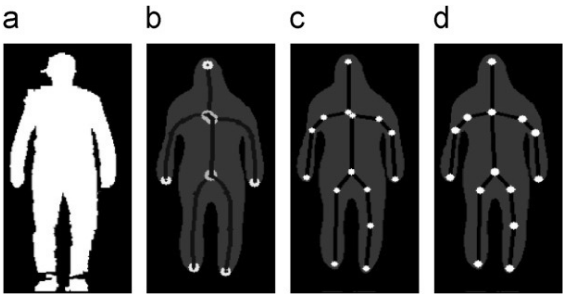
- ⇨ I grafi (e gli alberi) rappresentano un insieme di nodi collegati da archi (vedi il corso di Algoritmi)
- ⇨ Codificano la relazione tra parti
  - ⇨ esempio: vicinanza, connettività, etc

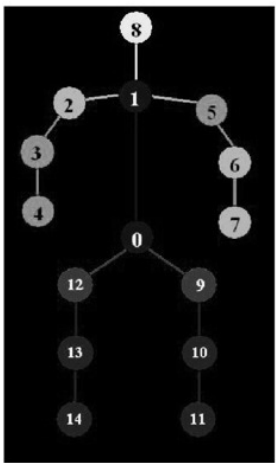


25

Tipi di pattern: i grafi

- ⇨ Esempio 1: modellare le diverse parti di un corpo umano

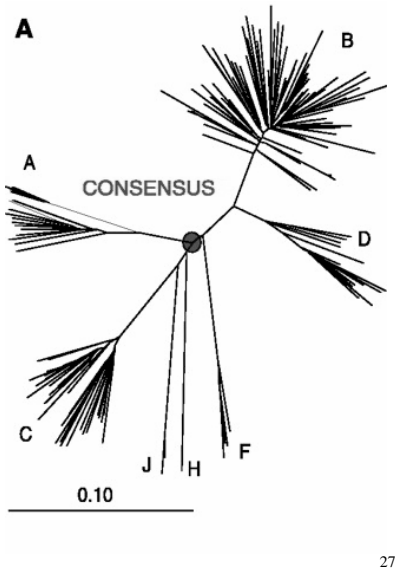




26

Tipi di pattern: i grafi

- ⇨ ESEMPIO 2: alberi filogenetici (sono un tipo particolare di grafo):
  - ⇨ si può cercare di trovare l'albero "consenso" tra un insieme elevato di alberi
  - ⇨ ognuno è un pattern si cerca il modello migliore



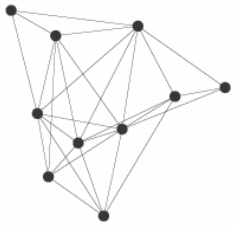
27

Tipi di pattern: i grafi

- ⇨ Esempio 3: Protein-Protein Interaction Networks
- ⇨ Grafi dove vengono visualizzate le interazioni tra le proteine
  - ⇨ fondamentali in biologia

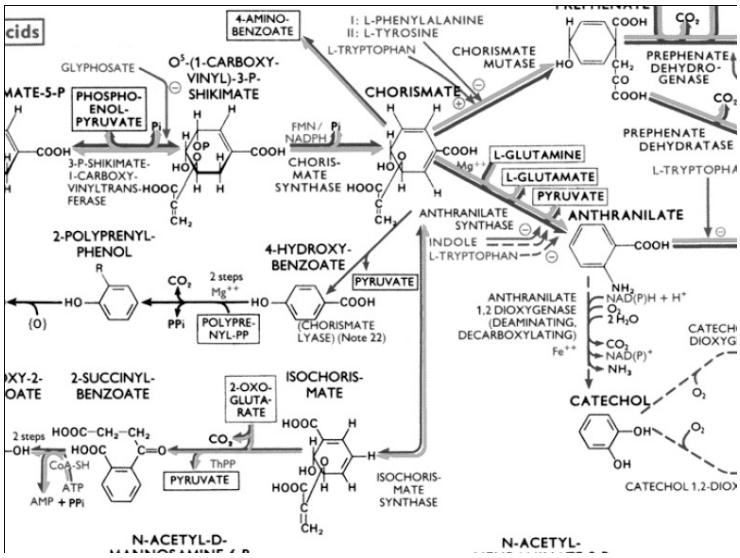
- list of N **proteins** (nodes)
- list of **protein pairs** (edges)

This is an **undirected, unweighted graph**



28

⇨ Esempio 4: percorsi metabolici: insieme di reazioni di composti chimici per svolgere un processo cellulare



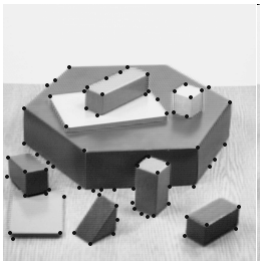
29

Tipi di pattern: gli insiemi

⇨ insiemi: collezione non ordinata di dati a cardinalità variabile

⇨ insieme di descrittori tutti relativi alla stessa entità, ma non ordinati

⇨ Esempio 1: angoli in un'immagine



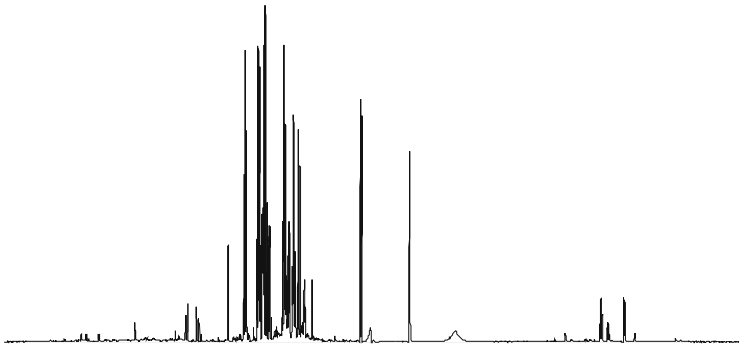
non c'è un ordine

in ogni immagine ce ne può essere un numero variabile

30

Tipi di pattern: gli insiemi

- Esempio 2: picchi in uno spettro NMR:
  - numero di picchi può essere diverso in due diversi spettri
  - i picchi non sono ordinati (si possono forse ordinare per ampiezza, o per ppm, ma in generale questo non è possibile)



31

Altri tipi di pattern

Transactions

- Dato un insieme di oggetti, una transaction è un sottoinsieme di questi oggetti
- Esempio 1: market basket analysis
  - analisi dei prodotti acquistati da un consumatore
  - dati N prodotti disponibili
  - solo M (M<N) vengono acquistati
- Esempio 2: dato l'insieme di geni, caratterizziamo quelli espressi ("attivati") in una determinata condizione (attenzione, livello di espressione)

- tipicamente rappresentati da un vettore binario lungo N
  - 0 indica l'assenza dell'oggetto
  - 1 indica la presenza di un oggetto

32

Problema

- ⇨ Molte delle tecniche di Pattern Recognition Statistica funzionano nel caso di spazi vettoriali
- ⇨ In caso di dati non vettoriali la maggior parte delle assunzioni non vale
- ⇨ Soluzioni:
  - ⇨ creazione di metodi di classificazione/clustering che lavorano con questi tipi di dato
    - ⇨ definizione di misure di similarità che riescano a tenere in considerazione la struttura del pattern
  - ⇨ incapsulamento (embedding) in uno spazio vettoriale

33

Dettagli

- ⇨ Metodi di classificazione/clustering capaci di lavorare con questi tipi di dato
  - ⇨ metodi che non assumono uno spazio vettoriale
  - ⇨ definizione di misure di similarità che riescano a tenere in considerazione la struttura non vettoriale del pattern
    - ⇨ Esempio: distanza dynamic time warping per riconoscimento del parlato
- ⇨ incapsulamento (embedding) in uno spazio vettoriale:
  - ⇨ L'idea è quella di riportare il problema in uno spazio vettoriale

34

Dettagli

Primo metodo: estrazione di un insieme predeterminato di features dall'oggetto non vettoriale

ESEMPIO

⇨ Sequenza nucleotidica

⇨ Metodo di estrazione di features: si conta la frequenza di A, T, C, G

```
atgcgatcgatcgatcgatcaggcgcgctacg
agcggcgaggacctcatcatcgatcag
```

[14,10,17,18]

Adesso ogni sequenza è un punto in uno spazio 4-dimensionale

Vantaggi: spazio vettoriale, facile

Svantaggi: che feature occorre estrarre? Quanta informazione si perde?

35

Dettagli

Secondo metodo: incapsulamento in uno spazio dove vengono preservate le caratteristiche geometriche degli oggetti

ESEMPIO: Multidimensional Scaling

⇨ si calcolano le distanze tra gli oggetti non vettoriali

⇨ si "creano" dei punti in uno spazio vettoriale in modo che sia preservata la distanza

⇨ (si vedrà meglio in seguito)

36

Dettagli

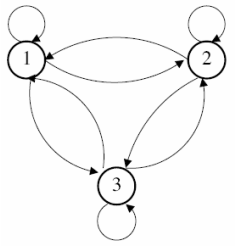
Terzo metodo: embedding generativo

⇨ Si assume un modello statistico per il dato non vettoriale

⇨ Si calcolano i parametri del modello

⇨ Si estraggono delle features legate al modello statistico

ESEMPIO: Hidden Markov Models



⇨ features = Probabilità media di essere nello stato Si osservando la sequenza

37

Fase 4: preprocessing

Preprocessing

⇨ Concetto di scala

⇨ Data standardization

⇨ Data transformation

⇨ Riduzione della dimensionalità

⇨ Riduzione del rumore

(ci concentriamo sul caso generico di spazi vettoriali)

38

Scala

⇨ Definizione di scala: significatività relativa dei numeri

ESEMPIO: si considerino due numeri 10, 12

⇨ sono molto simili in una scala [0-100]

⇨ molto diversi in una scala [10-13]

⇨ Riconoscere la scala di un problema è fondamentale:

- ⇨ nel calcolare la relazione tra due pattern
- ⇨ nell'interpretare i risultati (in particolare del clustering)

39

Scala

Tipi di scala:

⇨ qualitativa: le misure non hanno significato

- ⇨ nominal scale
- ⇨ ordinal scale

⇨ quantitativa: le misure hanno significato

- ⇨ interval scale
- ⇨ ratio scale

40

Scala

- ⇨ Nominal scale:
  - ⇨ non è una scala vera e propria, perché i numeri sono utilizzati come nomi
  - ⇨ Esempio: (sì/no) può essere codificata come (1,0), (0,1) o (50,100).
  - ⇨ I numeri da soli non hanno nessun significato quantitativo
- ⇨ Ordinal scale:
  - ⇨ scala numerica più povera
  - ⇨ i numeri hanno un senso solo in relazione agli altri
  - ⇨ Esempio: (1,2,3), (10,20,30), (1,20,300) sono equivalenti da un punto di vista ordinale
  - ⇨ Esempio: lista di accesso all'università (non è importante il valore assoluto della prova d'ingresso ma il valore relativo – essere nei primi 100)

41

Scala

- ⇨ Interval scale:
  - ⇨ la separazione tra i numeri ha un senso.
  - ⇨ Esiste un'unità di misura e l'interpretazione dei numeri dipende da questa unità di misura
  - ⇨ Esempio: voto di un compito
    - ⇨ 9 è un buon voto se la scala è [0-10]
    - ⇨ è un pessimo voto se la scala è [0-30]
  - ⇨ Esempio 2:
    - ⇨ temperatura a 90° è diversa a seconda che la scala sia Celsius o Fahrenheit

42

Scala

- ⇨ Ratio scale: i numeri hanno un valore assoluto.
  - ⇨ Esistono uno zero assoluto e un'unità di misura, ma il rapporto tra due numeri ha sempre lo stesso significato.
- ⇨ ESEMPIO: la distanza tra due città
  - ⇨ può essere misurata in metri, centimetri o kilometri
  - ⇨ se raddoppio la distanza raddoppio anche il tempo che ci metto ad andare da una città all'altra (indipendentemente dall'unità di misura)

43

Problema

- ⇨ Problema: a volte le variabili che descrivono un oggetto non sono nella stessa scala
- ⇨ Ci sono metodi che soffrono se le features sono a scala diversa
- ⇨ Alla lavagna: esempio: calcolo della distanza euclidea tra due punti
- ⇨ Soluzioni:
  - ⇨ data standardization
  - ⇨ data transformation

44

Data standardization

- La standardizzazione dei dati produce dati "senza dimensionalità"
  - tutta la conoscenza su scala e locazione dei dati viene persa dopo la standardizzazione
  - creazione di nuovi dati "in formato standard" – confrontabili
- E' necessario standardizzare i dati!
  - esempio distanza euclidea visto in precedenza
- Approcci di standardizzazione:
  - approcci globali standardizzano l'intero data set
  - approcci intra-gruppo standardizzano ogni gruppo
    - problema nel caso del clustering: i cluster non sono noti
    - possibile soluzione: standardizzazione iterativa (prima si calcolano i cluster, poi si standardizza, poi si ricalcolano i cluster e così via)

45

Approcci di data standardization

- NOTA: la scelta dell'approccio da utilizzare dipende dal data set e dal campo di applicazione

- Alla lavagna:
  - notazione
  - formulazione generale
  - esempi di standardizzazione
  - intuizioni sugli effetti

46

Data transformation

⇨ In qualche modo legato alla standardizzazione dei dati
  ⇨ serve per migliorare la rappresentazione

⇨ Data standardization: le operazioni sono implementate dimensione per dimensione

⇨ Data transformation: le operazioni agiscono su tutte le dimensioni contemporaneamente (tipicamente in modo lineare)

47

Data transformation

In genere effettuata con i seguenti obiettivi:

1. ridurre la dimensionalità dello spazio delle features
   ⇨ per visualizzare il dataset
   ⇨ per ridurre il carico computazionale delle tecniche applicate
   ⇨ per alleviare il problema della "curse of dimensionality"
   ⇨ per eliminare la ridondanza di alcune direzioni del dataset
2. mettere in evidenza particolari strutture o migliorare le capacità discriminative dello spazio

48

Data transformation

⇨ Approccio classico: trasformazione lineare dello spazio delle features

$$Y = A'X$$

⇨ SCOPO: ridurre la dimensionalità dello spazio mantenendo la maggior quantità di informazione possibile

⇨ "Informazione": concetto che assume significati diversi a seconda della tecnologia utilizzata

49

Diversi approcci

⇨ Approcci non supervisionati:
- ⇨ si utilizzano solo i dati
- ⇨ Esempio: Principal Component Analysis

⇨ Approcci supervisionati:
- ⇨ si utilizzano altre informazioni (ad esempio le etichette)
- ⇨ Esempio: trasformata di Fisher

⇨ La seconda classe di approcci si può utilizzare solo in caso di problema supervisionato (classificazione)
- ⇨ L'idea è che lo spazio viene "ridotto" tenendo conto del task finale di classificazione

50

Principal Component Analysis

- ⇨ Approccio lineare non supervisionato alla riduzione della dimensionalità
- ⇨ Obiettivo: mantenere la maggior aderenza ai dati originali
  - ⇨ Minimizza lo scarto quadratico medio tra i dati originali e quelli ricostruiti
  - ⇨ estrae le direzioni di massima varianza dei dati

- ⇨ IDEA:
  - ⇨ trasformazione lineare delle variabili principali che proietta i dati in uno spazio così costruito:
    - ⇨ la prima direzione è quella di massima varianza
    - ⇨ la seconda direzione viene scelta tra le rimanenti, in modo che sia ortogonale alla prima, e di massima varianza
    - ⇨ si continua così fino alla fine

51

Principal Component Analysis

- ⇨ come si realizza (alla lavagna)

52

Principal Component Analysis

- Vantaggi:
  - migliore tecnica lineare di compressione dei dati
    - migliore in senso di "errore quadratico medio"
  - i parametri del modello possono essere ricavati direttamente dai dati
  - compressione e decompressione sono operazioni molto veloci (moltiplicazione di matrici)
- Svantaggi:
  - alto costo computazionale per il calcolo dei parametri del modello (soprattutto in caso di dimensionalità elevata)
  - non è chiaro come questa tecnica possa gestire il caso di dati incompleti
  - PCA non tiene conto della densità di probabilità dello spazio considerato (viene considerata solo la vicinanza del vettore trasformato al vettore originale)
  - non è detto in tutti i casi che le direzioni a varianza maggiore siano le direzioni ottimali

53

Principal Component Analysis

- Problema: come calcolare il numero di componenti principali ottimali?
- Soluzione 1. scegliere il numero di componenti che arrivano a coprire una certa percentuale di varianza (esempio 95%)
  - Come si misura? concetto di "importanza" di una componente principale

$$\text{imp(i)} = \frac{\lambda_i}{\sum_{i=1}^{d} \lambda_i} \qquad \text{varianza(1,..,L)} = \sum_{i=1}^{L} \text{imp(i)} = \frac{\sum_{i=1}^{L} \lambda_i}{\sum_{i=1}^{d} \lambda_i}$$

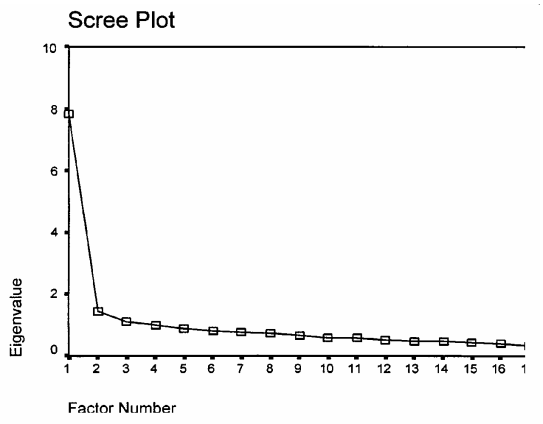
54

Principal Component Analysis

⇨ Soluzione 2: Cattell Scree test

⇨ plottare gli autovalori

⇨ trovare dove la decrescita degli autovalori non è più così marcata



55

Altri metodi

⇨ Independent Component Analysis (ICA):

⇨ Un metodo computazionale per distinguere le diverse componenti additive di un segnale

⇨ si assume l'indipendenza statistica tra le varie componenti

$$x_i = a_{i,1}s_1 + a_{i,2}s_2 + \ldots + a_{i,m}s_m \qquad x = As$$

⇨ ESEMPIO: Source separation

⇨ Problema in signal processing: vengono mescolati assieme diversi segnali, l'obiettivo è quello di identificare i segnali originali

⇨ Esempio classico: "**cocktail party problem**": un numero di persone stanno parlando simultaneamente in una stanza e una persona cerca di seguire una sola delle conversazioni

56

Altri metodi

- Il problema viene risolto cercando di massimizzare l'indipendenza statistica tra le componenti stimate.
- Dipendentemente dalla scelta del concetto di "indipendenza statistica" abbiamo versioni differenti
  - Esempio: mutua informazione
- Modelli di ICA
  - ICA lineare senza rumore $x = As$
  - ICA lineare con rumore $x = As + \eta$
  - ICA non lineare $x = f(s \mid \theta) + \eta$

57

Metodi supervisionati

- IDEA: utilizzo l'informazione del task finale
  - non sempre l'informazione "classicamente" estratta corrisponde alla migliore informazione per risolvere il problema
  - ESEMPIO: PCA vs Fisher
- Linear Discriminant Analysis (Fisher)
  - tecnica classica di riduzione della dimensionalità che mira a massimizzare la separabilità tra le classi nello spazio risultante
  - spesso utilizzata anche come classificatore lineare (dettagli in seguito)
  - anche chiamata:
    - Discriminant Analysis
    - Fisher Linear Analysis

58

Linear Discriminant Analysis

⇨ come si realizza (alla lavagna)

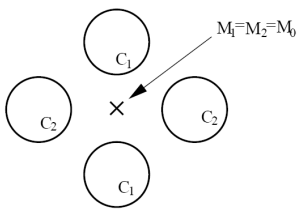
59

Linear Discriminant Analysis

⇨ Vantaggi: massimizza la separabilità tra le classi

⇨ Svantaggi:

- ⇨ Principale: dato un problema a C classi, la massima dimensionalità dello spazio risultante è c-1
  - ⇨ se abbiamo un problema binario allora possiamo proiettare i dati in uno spazio monodimensionale (molto restrittivo!)
- ⇨ Secondario: il criterio di Fisher non funziona se le classi sono multimodali e condividono la stessa media



60

Feature Selection

- Approccio alternativo alla riduzione della dimensionalità
  - rid. dimensionalità: considera tutte le feature e le trasforma
  - feature selection: sceglie solo alcune feature (in base ad un criterio di ottimalità)
  - Vantaggio: spesso alcune features sono irrilevanti / dannose
  - Svantaggio: computazionalmente oneroso
- Problemi da risolvere
  - scegliere il criterio di ottimalità: informazione (come si misura?), varianza, capacità di classificazione (solo per problemi supervisionati)
  - come trovare il sottoinsieme ottimale senza provarli tutti (troppo oneroso computazionalmente)
- Esempio: Sequential Forward Feature Selection

61

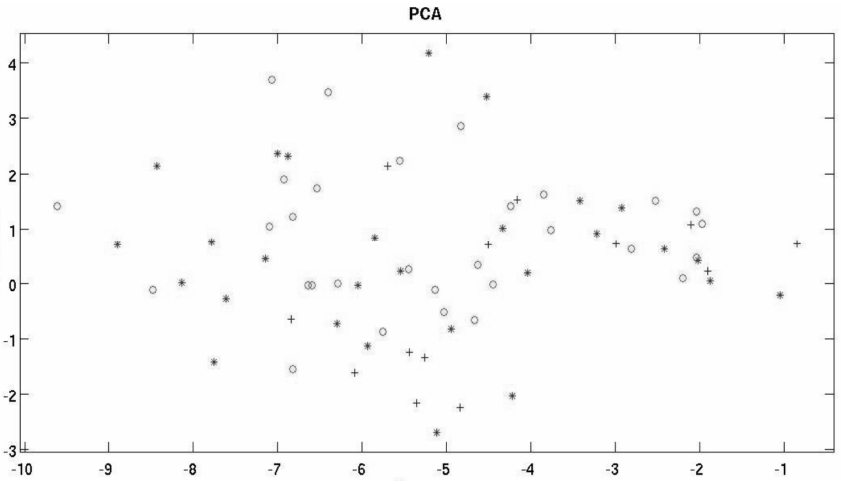
Feature Selection

- Forward Sequential Feature Selection
  - Algoritmo greedy per trovare l'insieme ottimale di features
- Schema:
  - si valuta il criterio per tutte le features singolarmente
  - si sceglie la feature che massimizza il criterio (chiamata f1)
  - si valuta il criterio per tutte le coppie (f1,fx) – cioè tenendo fissata f1
  - si sceglie la coppia che massimizza il criterio (la coppia (f1,f2))
  - si valuta il criterio per tutte le terne (f1,f2,fx) – cioè tenendo fissate f1 e f2
  - ...
- Algoritmo greedy – ad ogni istante la scelta migliore
  - computazionalmente efficiente
  - sub ottimale

62

Esempio: neuropatologie

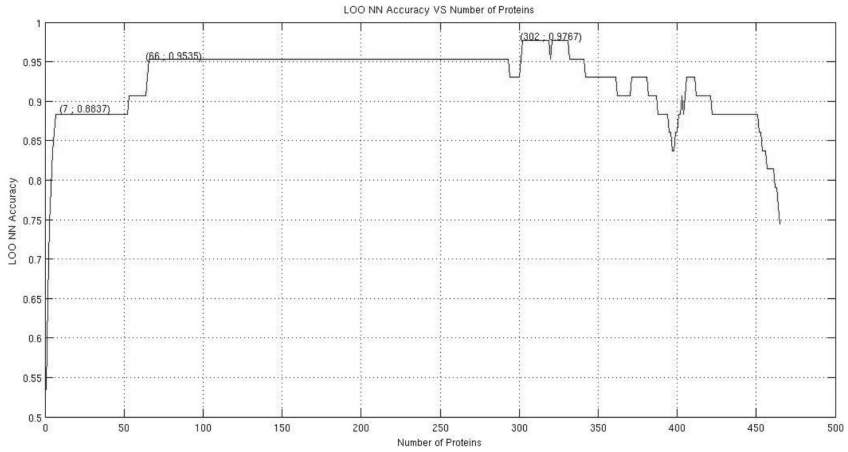
⇨ Usando tutte le features e applicando la PCA



63

Esempio: neuropatologie

⇨ feature selection



64

Visualizzazione dei dati

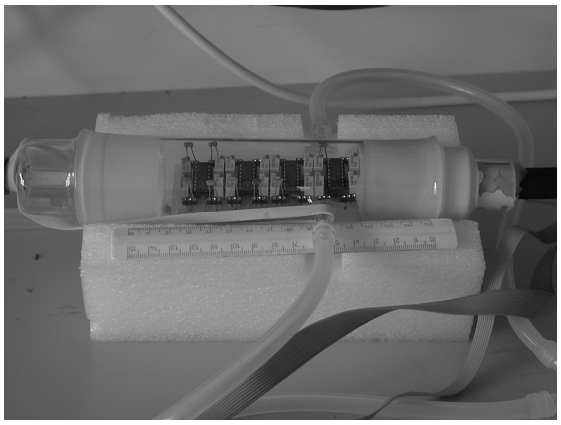
- ⇨ Studio di metodologie per visualizzare dati/informazioni in modo grafico
- ⇨ Fondamentale per:
  - ⇨ validazione/interpretazione dei risultati
- ⇨ Il nostro punto di vista:
  - ⇨ rappresentare dati in uno spazio vettoriale a dimensione 2/3 per poter vedere la relazione tra i dati
- ⇨ Approccio classico: ridurre la dimensionalità con le tecniche viste precedentemente, tipo PCA

65

Esempio

Riconoscimento di odori

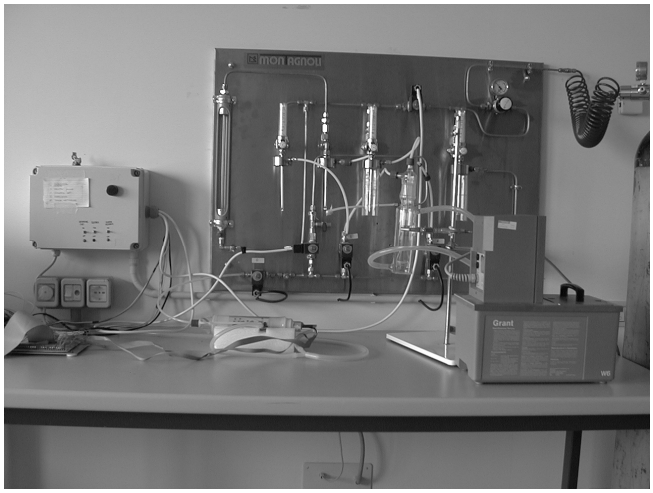
- ⇨ Array di sensori chimici, ognuno sensibile a composti diversi



66

Esempio

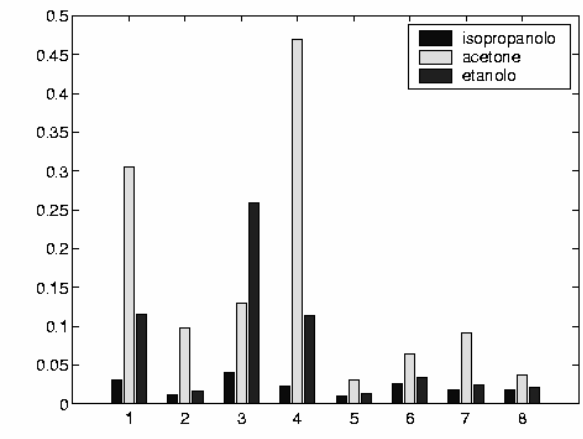
⇨ Dato un composto da analizzare, viene fatto passare sopra il "naso elettronico"



67

Esempio

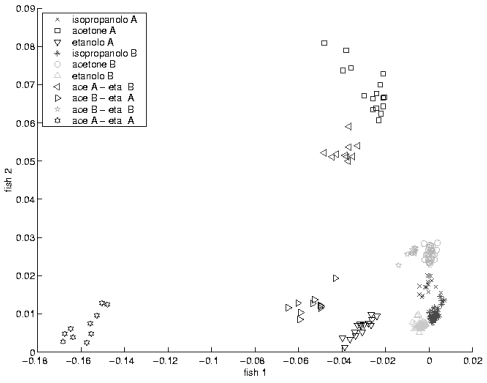
⇨ Il risultato è un punto in uno spazio 8-dimensionale



68

Esempio

⇨ Per visualizzare tutti gli esperimenti: PCA + plot



69

Multi dimensional Scaling (MDS)

⇨ insieme di tecniche statistiche utilizzate per esplorare similarità e dissimilarità nei dati

⇨ Il punto di partenza è una matrice di prossimità

⇨ Rappresentazione proximity-based vs. rappresentazione feature-based

⇨ Idea: proiettare questi dati in uno spazio vettoriale dove le distanze vengono preservate

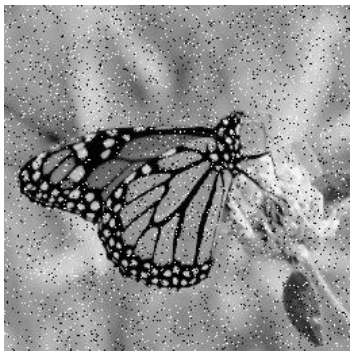
⇨ ESEMPIO: Sammon's mapping (alla lavagna)

70

Riduzione del rumore

⇨ Rumore: informazione irrilevante (o dannosa) nei dati

⇨ Per rimuovere il rumore vengono utilizzate tecniche di filtraggio (molto utilizzate in signal/image processing)



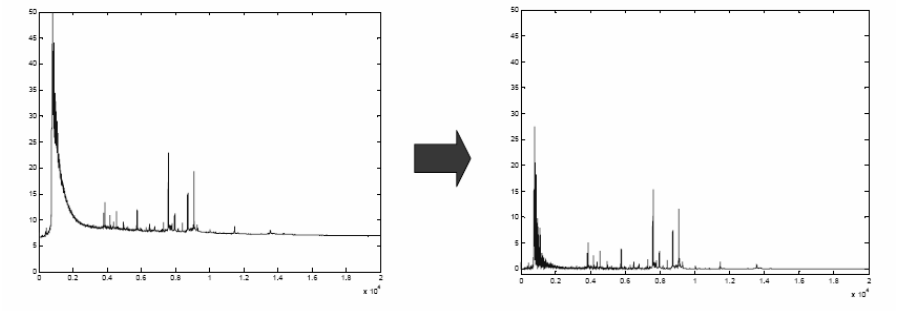
rumore sale e pepe

immagine "ripulita"

71

Riduzione del rumore

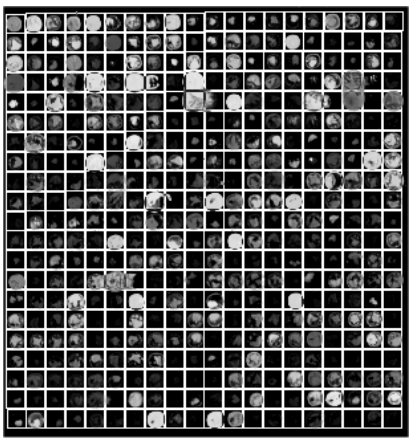
⇨ Mass spectrometry: esiste un bias di intensità sistematico per il quale il profilo osservato differisce da zero

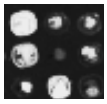


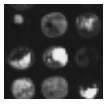
72

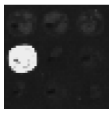
Riduzione del rumore

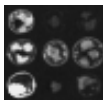
⇨ Microarray: errori nelle immagini degli spot




dimensione


rotondità


intensità


distribuzione dei pixel

73

Riduzione del rumore

⇨ NOTA: E' necessario eliminare il rumore preservando l'informazione

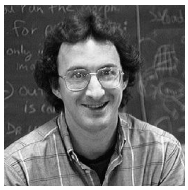
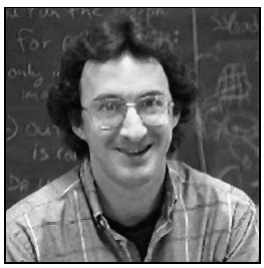
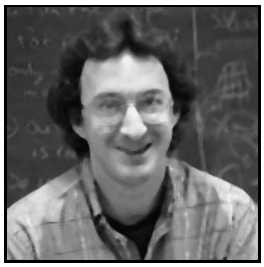


immagine originale


filtro mediano 3x3


filtro mediano 5x5


filtro mediano 7x7

74