Comprendere e preparare i dati

Prof. Matteo Golfarelli
Alma Mater Studiorum - Università di Bologna

Cosa sono i dati?

- Nelle applicazioni di data mining i dati sono composti da collezioni di ***oggetti*** descritti da un insieme di attributi
  - ✓ Sinonimi di oggetto sono record, punto, caso, esempio, entità, istanza, elemento

- Un ***attributo*** è una proprietà o una caratteristica di un oggetto
  - ✓ Sinonimi di attributo sono: variabile, campo, caratteristica

Attributi

Oggetti

Tid	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

Tipi di attributi

- **It is imperative to know the attribute properties to carry out meaningful operations and research with them**
- Un impiegato è descritto da un ID e dall'età, ma non ha senso calcolare l'ID medio degli impiegati!
- Il tipo dell'attributo ci dice quali proprietà dell'attributo sono riflesse nel valore che usiamo come misura
- Un modo semplice per caratterizzare i vari tipi di attributi si basa sul *tipo di operatore* che ha senso applicare ai valori che esso assume:
  - ✓ Diversità =, ≠
  - ✓ Ordinamento <, ≤, >, ≥
  - ✓ Additività +, -
  - ✓ Moltiplicatività *, /
- Si determinano così 4 tipi di dati: : nominali, ordinali, di intervallo, e di rapporto

Tipi di attributi

Tipo		Descrizione	Esempio	Operatori statistici
Categorici (qualitativi)	Nominale	Nomi diversi dei valori. Possiamo solo distinguerli	Sesso, colore degli occhi, codici postali, ID	Moda, correlazione
	Ordinale	I valori ci consentono di ordinare gli oggetti in base al valore dell'attributo	Voto, Durezza di un minerale	Mediana, percentile
Numerici (quantitativi)	Di Intervallo	La differenza tra i valori ha un significato, ossia esiste una unità di misura	Date, temperatura in Celsius e Fahrenheit	Media, varianza
	Di Rapporto	Il rapporto tra i valori ha un significato	Età, massa, lunghezza, quantità di denaro, temperatura espressa in Kelvin	Media geometrica, media armonica

Tipi di attributi: altre classificazioni

- Binari, discreti e continui
  - ✓ Un attributo discreto ha un numero finito o un insieme infinito numerabile di valori normalmente rappresentati mediante interi o etichette
  - ✓ Un attributo continuo assume valori reali
  - ✓ Gli attributi nominali e ordinali sono tipicamente discreti o binari, mentre quelli di intervallo e di rapporto sono continui
- Attributi asimmetrici: hanno rilevanza solo le istanze che assumono valori diversi da zero:
  - ✓ Es. Consideriamo i record relativi agli studenti: in cui ogni attributo rappresenta un corso dell'Ateneo che può essere seguito (1) o meno (0) dallo studente. Visto che gli studenti seguono una frazione molto ridotta dei corsi dell'Ateneo se si comparassero le scelte degli studenti sulla base di tutti i valori degli attributi il loro comportamento apparirebbe molto simile.

Documenti

- I documenti sono gli oggetti dell'analisi, sono descritti da un vettore di termini
  - ✓ Ogni termine è un attributo del documento
  - ✓ Il valore degli attributi indica il numero di volte in cui il corrispondente termine compare nel documento.

	team	coach	play	ball	score	game	win	lost	timeout	season
Document 1	3	0	5	0	2	6	0	2	0	2
Document 2	0	7	0	2	1	0	0	3	0	0
Document 3	0	1	0	0	1	2	2	0	3	0

Che tipo di dato è?

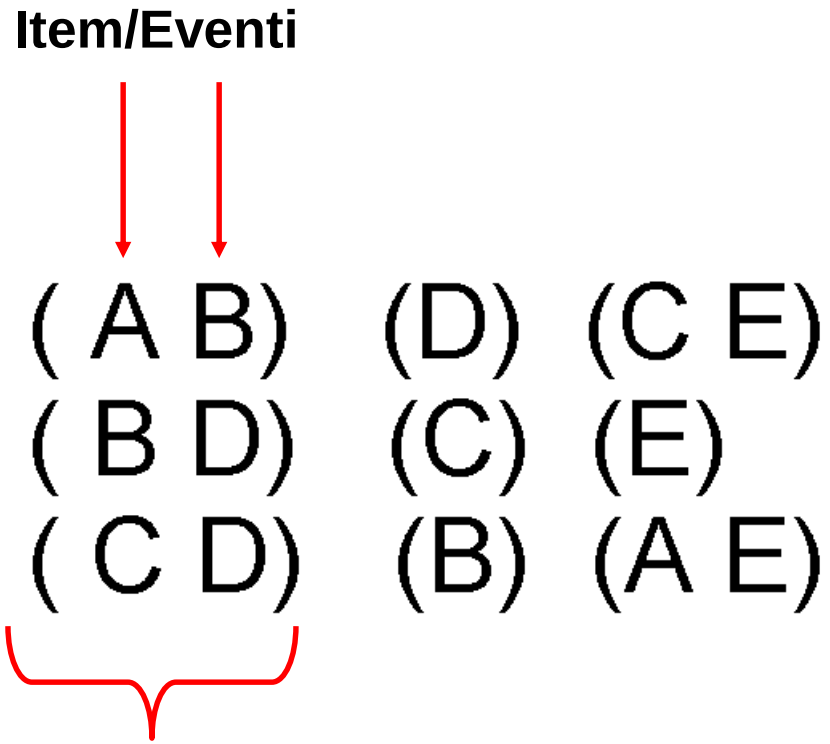
Transazioni

- Un tipo speciale di record in cui
  - ✓ Ogni record (transazione) coinvolge più item
  - ✓ Per esempio in un supermercato l'insieme dei prodotti comprati da un cliente durante una visita al negozio costituisce una transazione, mentre i singoli prodotti acquistati sono gli item.
  - ✓ Il numero degli item può variare da transazione a transazione

TID	*Items*
1	Bread, Coke, Milk
2	Beer, Bread
3	Beer, Coke, Diaper, Milk
4	Beer, Bread, Diaper, Milk
5	Coke, Diaper, Milk

Dati ordinati

- Sequenze di transazioni



Dati ordinati

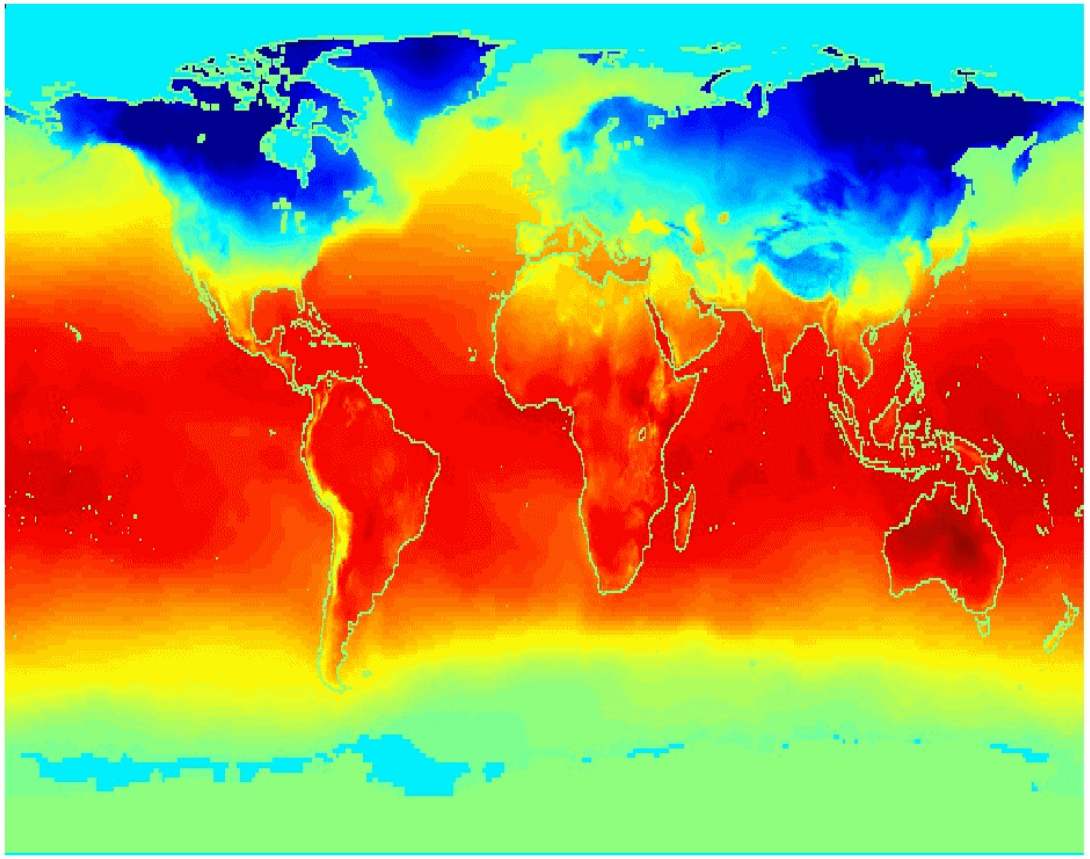
- Sequenze di dati genomici

```
GGTTCCGCCTTCAGCCCCGCGCC
CGCAGGGCCCGCCCCGCGCCGTC
GAGAAGGGCCCGCCTGGCGGGCG
GGGGGAGGCGGGGCCGCCCGAGC
CCAACCGAGTCCGACCAGGTGCC
CCCTCTGCTCGGCCTAGACCTGA
GCTCATTAGGCGGCAGCGGACAG
GCCAAGTAGAACACGCGAAGCGC
TGGGCTGCCTGCTGCGACCAGGG
```

Dati ordinati

- Dati Spazio-Temporali

Temperatura media mensile di terre e oceani



Esplorazione dei dati

- A preliminary data analysis aimed at identifying the main features
  - ✓ It helps you choose the best tool for preprocessing and analysis
    Allows you to use human skills to locate patterns
  - ✓ A human domain expert can quickly locate unidentifiable patterns from analysis tools
- L'esplorazione dei dati sfrutta
  - ✓ Visualizzazione
  - ✓ Indici statistici
  - ✓ OLAP e Data Warehousing

Moda e Frequenza

- La frequenza del valore di un attributo è la percentuale di volte in cui quel valore compare nel data set
  - ✓ Dato L'attributo 'Comune di residenza' per il data set dei cittadini italiani, il valore 'Bologna' compare circa nello 0.6% dei casi (~3.7×10^5 / 6×10^7).
- La moda di un attributo è il valore che compare più frequentemente nel data set
  - ✓ La moda per l'attributo 'Comune di residenza' per il data set dei cittadini è 'Roma' che compare circa nel 4.5% dei casi (~2.7×10^6 / 6×10^7).
- Le nozioni di frequenza e moda sono normalmente utilizzate per attributi categorici

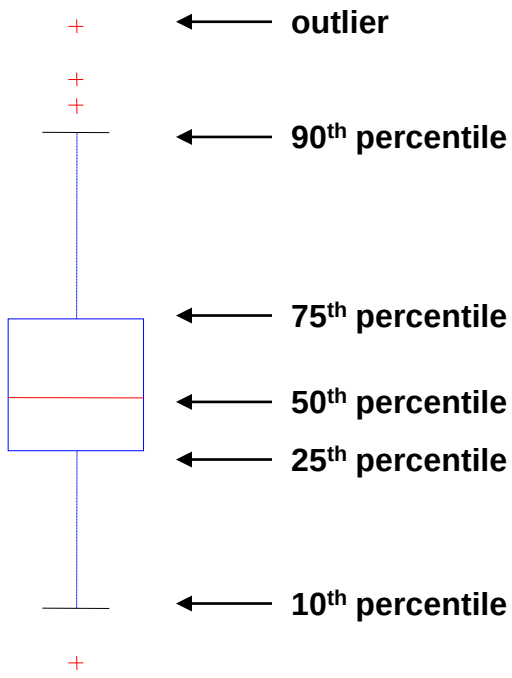
Percentili

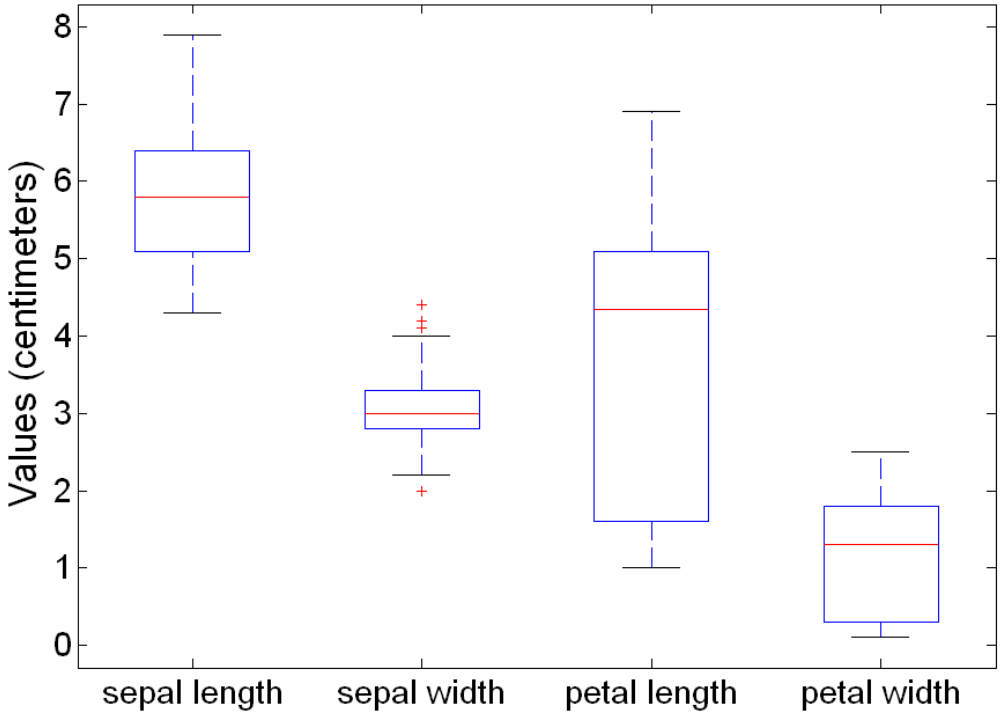
- Dato un attributo ordinale o continuo x e un numero *p* compreso tra 0 e 100, il *p-esimo* percentile è il valore di x_p di x tale che *p*% dei valori osservati per x sono inferiori x_p.
  - ✓ Per l'attributo "altezza in centimetri" per la popolazione dei neonati italiani femmine a un anno di vita è:
    - 50-esimo percentile= 78 cm -> la metà delle bambine è più alta di 78 cm
    - 97-esimo percentile= 81 cm -> solo il 3% delle bambine è più alta di 81 cm

- Le informazioni sui percentili sono spesso rappresentate mediante box plot

Tecniche di visualizzazione: Box Plot

- Permettono di rappresentare una distribuzione di dati
- Possono essere utilizzati per comparare più distribuzioni quando queste hanno grandezze omogenee





Misure di posizione: media e mediana

- La media è la più comune misura che permette di localizzare un insieme di punti

$$mean(\mathbf{x}) = \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

- Purtroppo la media è molto sensibile agli outlier
- In molti casi si preferisce utilizzare la mediana o una media “controllata”.

$$mediana(\mathbf{x}) = \begin{cases} x_{m+1} & se\ n \text{ è dispari } n = 2m+1 \\ (x_m + x_{m+1})/2 & se\ n \text{ è pari } n = 2m \end{cases}$$

✓ In un insieme *n* di dati disposti in ordine crescente la mediana è il termine che occupa il posto centrale, se i termini sono dispari, se i termini sono pari la mediana è la media aritmetica dei 2 termini centrali.

Misure di dispersione: Range e Varianza

- Il range è la differenza tra i valori minimi e massimi assunti dall'attributo
- Varianza e deviazione standard (o scarto quadratico medio) sono le più comuni misure di dispersione di un data set.

$$Varianza(\mathbf{x}) = s_{\mathbf{x}}^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 \qquad DevStandard(\mathbf{x}) = s_{\mathbf{x}} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

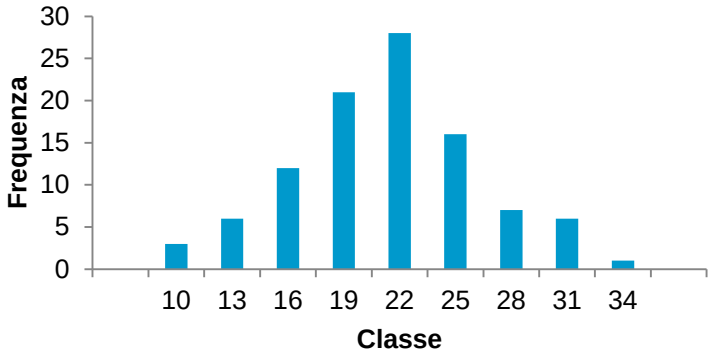
- Varianza e scarto quadratico medio sono sensibili agli outlier poichè sono legati quadraticamente al concetto di media
- Altre misure meno sensibili a questo problema sono:

$$\text{AbsoluteAverageDeviation} \quad AAD(\mathbf{x}) = \frac{1}{n}\sum_{i=1}^{n}|x_1 - \bar{x}|$$

$$\text{MedianAbsoluteDeviation} \quad MAD(\mathbf{x}) = mediana\left(\left\{|x_1 - \bar{x}|,\ldots,|x_n - \bar{x}|\right\}\right)$$

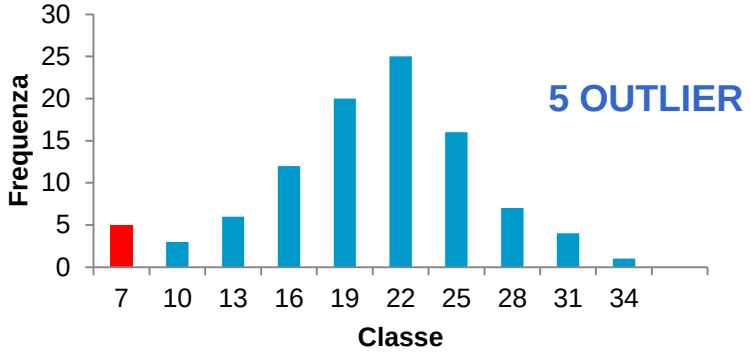
$$\text{InterquartileRange} \quad RI(\mathbf{x}) = x_{75\%} - x_{25\%}$$

Misure di dispersione: Range e Varianza



- Media 19,82617
- Mediana 19,65625
- 25% quartile 16,79252
- 75% quartile 22,75032

- Varianza 25,31324
- DevStandard 5,031227
- RI 5,957806
- AAD 3,857429
- MAD 2,979841



- Media 18,67617
- Mediana 19,27243
- 25% quartile 15,25606
- 75% quartile 22,55218

- **Varianza 37,58087**
- **DevStandard 6,130324**
- RI 7,29612
- AAD 4,579804
- MAD 3,095489

Calcolare i precedenti indici statisitici per X={5, 7, 2, 9, 8, 7, 5, 1, 1, 5}

Qualità dei dati

- La qualità dei dataset utilizzati incide profondamente sulle possibilità di trovare pattern significativi.

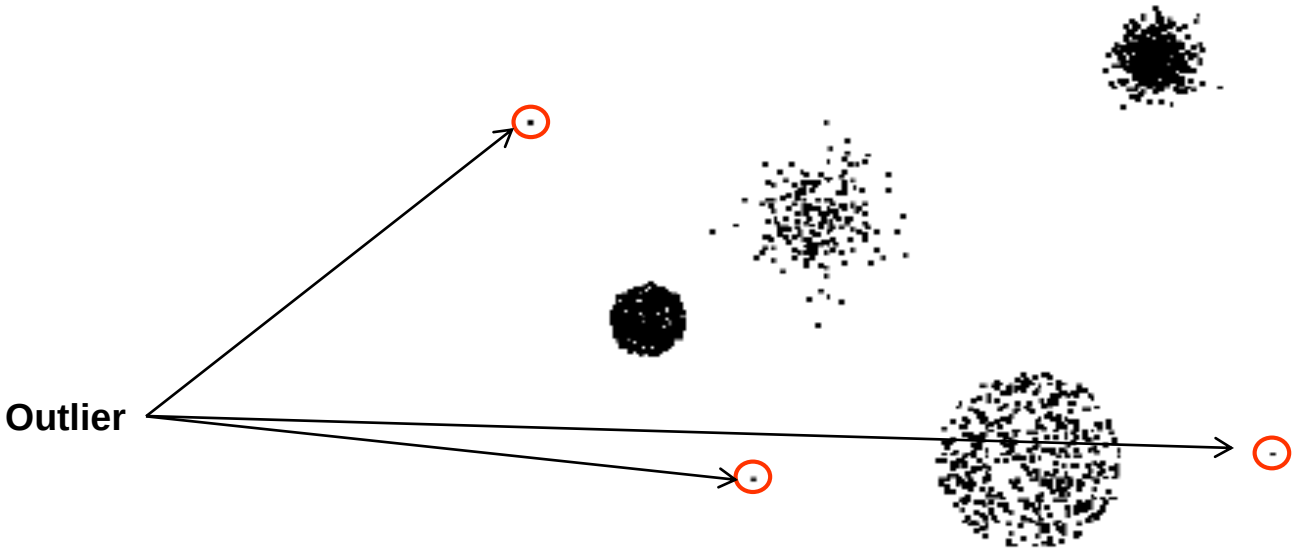
- I problemi più frequenti che deteriorano la qualità dei dati sono
  - ✓ Rumore e outlier
  - ✓ Valori mancanti
  - ✓ Valori duplicati

Rumore

- Indica il rilevamento di valori diversi da quelli originali
  - ✓ Distorsione della voce di una persona quando registrata attraverso un microfono di scarsa qualità
  - ✓ Registrazione approssimata di valori degli attributi
  - ✓ Registrazione errata di valori degli attributi

Outlier

- Outlier sono oggetti con caratteristiche molto diverse da tutti gli altri oggetti nel data set che complicano la determinazione delle sue caratteristiche essenziali
  - ✓ Sono normalmente rari
  - ✓ Potrebbero essere l'oggetto della ricerca



Valori mancanti

- Motivazioni per la mancata registrazione
  - ✓ L'informazione non è stata raccolta (es. l'intervistato non indica la propria età e peso)
  - ✓ L'attributo non è applicabile a tutti gli oggetti (es. il reddito annuo non ha senso per i bambini)

- Come gestire i dati mancanti?
  - ✓ Eliminare gli oggetti che li contengono (se il dataset è sufficientemente numeroso)
  - ✓ Ignorare i valori mancanti durante l'analisi
  - ✓ Compilare manualmente i valori mancanti
    - In generale è noioso, e potrebbe essere non fattibile
  - ✓ Compilare automaticamente i valori mancanti

Valori mancanti

- Come gestire i dati mancanti?
  - ✓ Stimare i valori mancanti
    - **usare la media** dell'attributo al posto dei valori mancanti
    - per problemi di classificazione, usare la media dell'attributo per tutti i campioni della stessa classe
    - **predire** il valore dell'attributo mancante sulla base degli altri attributi noti. Si usano algoritmi di data mining per preparare i dati in input ad altri algoritmi di data mining.
  - ✓ Usare un valore costante come “Unknown” oppure 0 (a seconda del tipo di dati).
    - potrebbe alterare il funzionamento dell'algoritmo di analisi, meglio allora ricorrere ad algoritmi che gestiscono la possibilità di dati mancanti
    - È utile se la mancanza di dati ha un significato particolare di cui tener conto

Dati duplicati

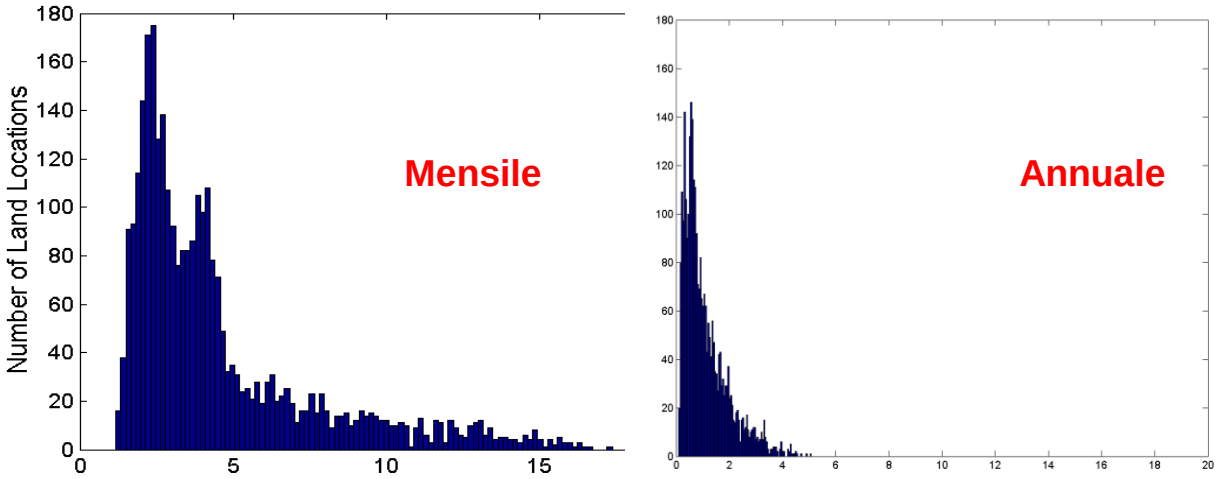
- Il data set potrebbe includere oggetti duplicati
  - ✓ Problema primario quando il data set è il risultato della fusione di più sorgenti dati
  - ✓ Esempi: stessa persona con più indirizzi e-mail; stesso cliente registrato due volte
- Può essere necessario introdurre una fase di data cleaning al fine di individuare ed eliminare la ridondanza

Preprocessing del data set

- Raramente il dataset presenta le caratteristiche ottimali per essere trattato al meglio dagli algoritmi di data mining. E' quindi necessario mettere in atto una serie di azioni volte a consentire il funzionamento degli algoritmi di interesse
  - ✓ Aggregazione
  - ✓ Campionamento
  - ✓ Riduzione della dimensionalità
  - ✓ Selezione degli attributi
  - ✓ Creazione degli attributi (*feature engineering*)
  - ✓ Discretizzazione e binarizzazione
  - ✓ Trasformazione degli attributi

Aggregazione

- Combina due o più attributi (oggetti) in un solo attributo (oggetto) al fine di:
  - ✓ Ridurre la cardinalità del data set
  - ✓ Effettuare un cambiamento di scala
    - Le città possono essere raggruppate in regioni e nazioni
  - ✓ Stabilizzare i dati
    - I dati aggregati hanno spesso una minore variabilità



Deviazione standard della media delle precipitazioni

Campionamento

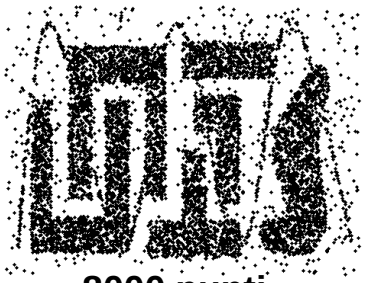
- E’ la tecnica principale utilizzata per selezionare i dati
  - ✓ E’ spesso utilizzata sia nella fase preliminare sia nell’analisi finale dei risultati.
- Gli statistici campionano poiché ottenere l’intero insieme di dati di interesse è spesso troppo costoso o richiede troppo tempo.
- Il campionamento è utilizzato nel data mining perché processare l’intero dataset è spesso troppo costoso o richiede troppo tempo.
- Il principio del campionamento è il seguente:
  - ✓ Se il campione è rappresentativo il risultato sarà equivalente a quello che si otterrebbe utilizzando l’intero dataset
  - ✓ Un campione è rappresentativo se ha approssimativamente le stesse proprietà (di interesse) del dataset originale

Tipi di campionamento

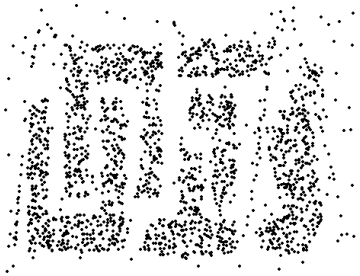
- Campionamento casuale semplice
  - ✓ C’è la stessa probabilità di selezionare ogni elemento
  - ✓ Campionamento senza reimbussolamento
    - Gli elementi selezionati sono rimossi dalla popolazione
  - ✓ Campionamento con reimbussolamento
    - Gli elementi selezionati non sono rimossi dalla popolazione
    - In questo caso un elemento può essere selezionato più volte.
    - Dà risultati simili al precedente se la cardinalità del campione è << di quella della popolazione
    - E’ più semplice da esaminare poiché la probabilità di scegliere un elemento non cambia durante il processo
- Campionamento stratificato:
  - ✓ Si suddividono i dati in più partizioni, quindi si usa un campionamento casuale semplice su ogni partizione.
  - ✓ Utile nel caso in cui la popolazione sia costituita da tipi diversi di oggetti con cardinalità differenti. Un campionamento casuale può non riuscire a fornire un’adeguata rappresentazione dei gruppi meno frequenti

La dimensione del campione

- Scelta la modalità di campionamento è necessario fissare la dimensione del campione al fine di limitare la perdita di informazione



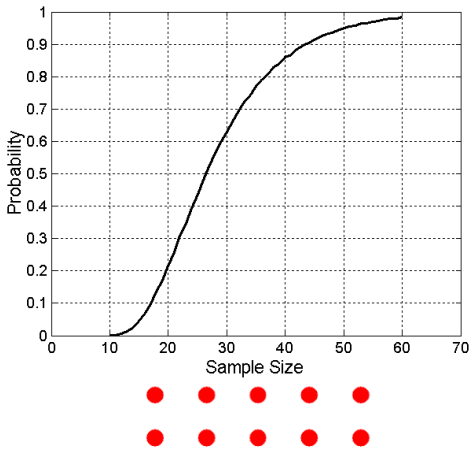
8000 punti



2000 punti

500 punti

- La probabilità di avere rappresentanti di tutta la popolazione aumenta in modo non lineare rispetto alla dimensione del campione
  - ✓ Nell'esempio si vuole ottenere un campione per ognuno dei 10 gruppi

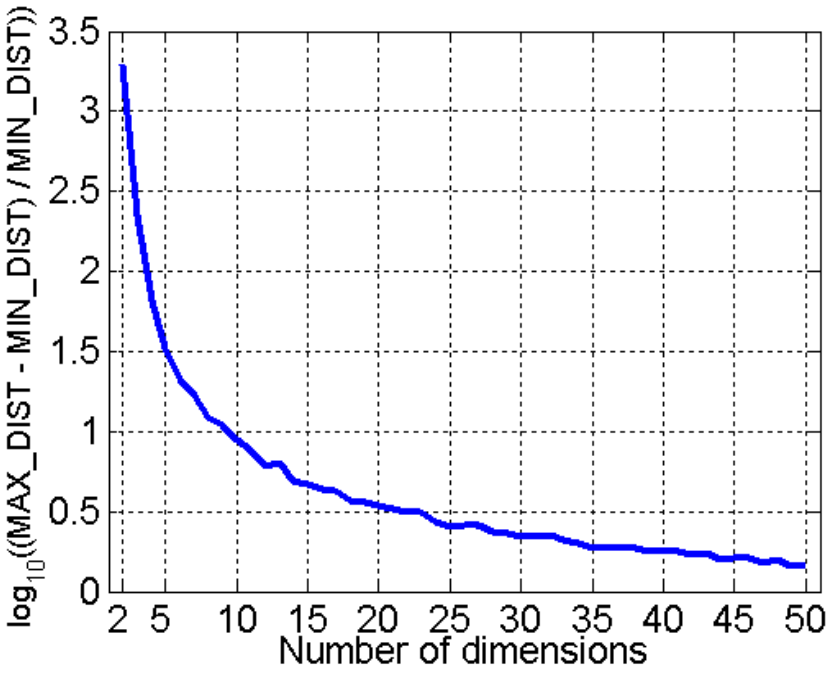


Riduzione della dimensionalità

- Obiettivi:
  - ✓ Evitare la "*curse of dimensionality*": la maledizione della dimensionalità
  - ✓ Ridurre la quantità di tempo e di memoria utilizzata dagli algoritmi di data mining (riduzione dello spazio di ricerca)
  - ✓ Semplificare la visualizzazione dei dati
  - ✓ Eliminare attributi non rilevanti ed eliminare il rumore sui dati
- Tecniche
  - ✓ Principle Component Analysis
  - ✓ Singular Value Decomposition
  - ✓ Selezione degli attributi con tecniche supervisionate

Curse of Dimensionality

- Al crescere della dimensionalità i dati diventano progressivamente più sparsi
- Molti algoritmi di clustering e di classificazione trattano con difficoltà dataset a elevata dimensionalità
- Le definizioni di densità e di distanza tra i punti che sono essenziali per esempio per il clustering e per l'individuazione degli outilier diventano meno significativi



- **500 punti generati in modo casuale**
- **Il grafico mostra una misura della differenza tra la distanza minima e la distanza massima di ogni coppia di punti**

ML

Principal Component Analysis

- E’ un metodo di proiezione che trasforma gli oggetti appatenenti a uno spazion p-dimensionale in uno spazio k-dimensionale (con k < p) in modo da conservare il massimo dell’informazione (l’informazione è misurata come totale varianza del dataset) nelle dimensioni iniziali.
- Svariati gli utilizzi nell’ambito del data mining:
  - ✓ Studio e visualizzazione della correlazione tra le variabili al fine di limitare il numero di variabili da considerare
  - ✓ Ottenere dimensioni non correlate che siano combinazioni lineari delle variabili iniziali in modo da utilizzare queste dimensioni al posto delle originali.
  - ✓ Visualizzare osservazioni in uno spazio bi/tri dimensionale al fine di identificare gruppi di istanze omogenee

Selezione degli attributi

- E' una modalità per ridurre la dimensionalità dei dati. La selezione mira solitamente a eliminare:
  - ✓ Attributi ridondanti
    - Duplicano in gran parte le informazioni contenute in altri attributi a causa di una forte correlazione tra le informazioni
    - Esempio: l'importo dell'acquisto e l'importo dell'IVA
  - ✓ Caratteristiche irrilevanti
    - Alcune caratteristiche dell'oggetto possono essere completamente irrilevanti ai fini del mining
    - Esempio: la matricola di uno studente è spesso irrilevante per predire la sua media

Per quale tipo di pattern può essere utile la matricola assumendo che questa sia un numero positivo che non è azzerato negli anni?

Modalità di selezione degli attributi

- Approccio esaustivo:
  - ✓ Prova tutti i possibili sottoinsiemi di attributi e scegli quello che fornisce i risultati migliori sul test set utilizzando l'algoritmo di mining come funzione di bontà black box
  - ✓ Dati n attributi il numero di possibili sottoinsiemi è 2^n-1
- Approcci non esaustivi:
  - ✓ Aprrocci embedded
    - La selezione degli attributi è parte integrante dell'algoritmo di data mining. L'algoritmo stesso decide quali attributi utilizzare (es. alberi di decisione)
  - ✓ Approcci di filtro:
    - La fase di selezione avviene prima del mining e con criteri indipendenti dall'algoritmo usato (es. si scelgono insiemi di attributi le cui coppie di elementi presentano il più basso livello di correlazione)
  - ✓ Approcci euristici:
    - Approssimano l'approccio esaustivo utilizzando tecniche di ricerca euristiche.

Creazione di attributi

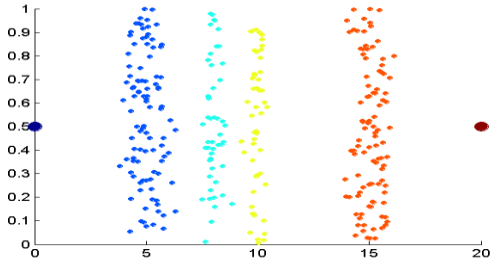
- Può essere utile creare nuovi attributi che meglio catturino le informazioni rilevanti in modo più efficace rispetto agli attributi originali
  - ✓ Estrazione di caratteristiche
    - Utilizzano normalmente tecniche diverse da dominio a dominio
    - Impronte digitali → minuzie
  - ✓ Mapping dei dati su nuovi spazi
    - Trasformata di Fourier
    - PCA
  - ✓ Combinazione di attributi

Discretizzazione

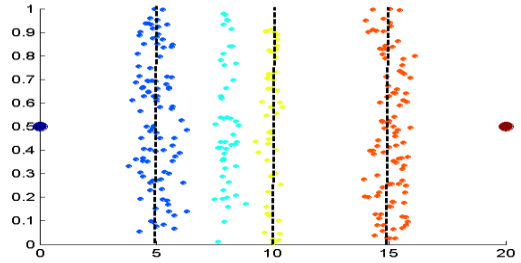
- Trasformazione di attributi a valori continui in attributi a valori discreti
  - ✓ Indispensabile per utilizzare alcune tecniche di mining (es. regole associative)
- Può essere utilizzata anche per ridurre il numero di classi di un attributo discreto
- La discretizzazione richiede di:
  - ✓ Individuare il numero più idoneo di intervalli
  - ✓ Definire come scegliere gli *split point*
- Le tecniche di discretizzazione sono:
  - ✓ Non supervisionate: non sfruttano la conoscenza sulla classe di appartenenza degli elementi
  - ✓ Supervisionate: sfruttano la conoscenza sulla classe di appartenenza degli elementi

Discretizzazione non supervisionata

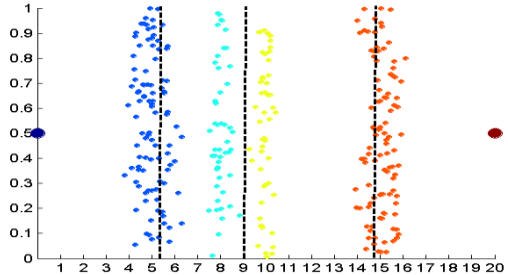
- Equi-larghezza: il range è suddiviso in intervalli di uguale lunghezza
- Equi-frequenza: il range è suddiviso in intervalli con un similare numero di elementi
- K-mediani: sono individuati k raggruppamenti in modo da minimizzare la distanza tra i punti appartenenti allo stesso raggruppamento



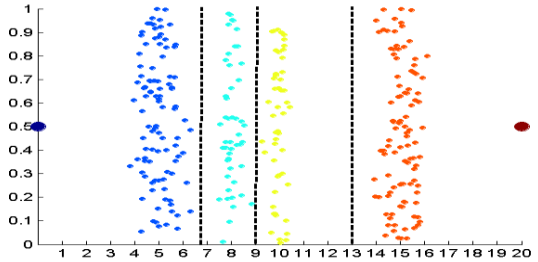
Dati



Equi-width



Equi-frequency



K-mediani

Discretizzazione supervisionata

- Gli intervalli di discretizzazione sono posizionati in modo da massimizzare la "purezza" degli intervalli.
- Si ricade in un problema di classificazione in cui a partire da classi (intervalli) composte da (contenenti) un solo elemento si fondono ricorsivamente classi attigue.
  - ✓ Una misura statistica della purezza è l'entropia degli intervalli
- Ogni valore v di un attributo A è una possibile frontiera per la divisione negli intervalli $A \leq v$ e $A > v$.
- Scelgo il valore che fornisce il maggiore **guadagno di informazione**, ossia la maggior riduzione di entropia
  - ✓ Il processo si applica ricorsivamente ai sotto-intervalli così ottenuti, fino a che non si raggiunge una condizione di arresto ad esempio, fino a che il guadagno di informazione che si ottiene diventa inferiore a una certa soglia d

Entropia e guadagno informativo

- Interpretazione “fisica”: misura del disordine
- Teoria dell’Informazione: è la misura dell’*incertezza* sul risultato di un esperimento modellabile mediante una variabile aleatoria X
  - ✓ X variabile aleatoria p(X) è la distribuzione di probabilità di X
  - ✓ Se X assume valori discreti x_i $1 \le i \le k$

$$H(X) = -\sum_{i=1}^{k} p(X = x_i) \log_2 p(X = x_i) = -\sum_{i=1}^{k} p_i \log_2 p_i$$

  - ✓ L’entropia di un evento certo è zero!!
  - ✓ L’entropia di *k* eventi **equiprobabili** è $\log_2 k$ (massimo dell’incertezza)
  - ✓ ATTENZIONE $0 \cdot \log_2 0 = 0$
- Entropia di una classificazione, sia:
  - ✓ D l’insieme di eventi (A,C) da suddividere in *n* intervalli
  - ✓ |D| = *m* il numero degli eventi da suddividere in intervalli in base al valore dell’attributo A
  - ✓ |C| = *k* l’insieme delle possibili etichette delle classi

Entropia e guadagno informativo

- L'entropia di una suddivisione in *n* intervalli è definita da:

$$e = \sum_{i=1}^{n} w_i e_i = -\sum_{i=1}^{n} \frac{m_i}{m} \sum_{j=1}^{k} p_{ij} \log_2 p_{ij} = -\sum_{i=1}^{n} \frac{m_i}{m} \sum_{j=1}^{k} \frac{m_{ij}}{m_i} \log_2 \frac{m_{ij}}{m_i}$$

- ✓ *wi* = peso dell'intervallo o classe (dipende dal numero di elementi che contiene)
- ✓ *ei* = entropia dell'intervallo o classe (dipende dalla confusione in esso presente ossia da quanti elementi di classi diverse contiene)
- ✓ *mi* = numero di eventi nell'intervallo i-esimo
- ✓ mij = numero di eventi di classe j-esima nell'intervallo i-esimo
- ✓ *mi/m* = peso dell'i-esimo intervallo
- ✓ *mij/mi* = probabilità(frazione) di eventi della classe j-esima nell'intervallo i-esimo

Entropia e guadagno informativo

- Supponiamo di avere i seguenti 6 eventi di tipo (A,C):

  (0, s), (2,n), (30,n), (31,n), (32,s), (40,s)

  - ✓ L'entropia dell'insieme non discretizzato è data da:

$$e = -\frac{3}{6}\log_2\frac{3}{6} - \frac{3}{6}\log_2\frac{3}{6} = 1$$

  - ✓ Il guadagno informativo che si ottiene fissando il limite (≤) dell'intervallo sui diversi valori di A è:

$$GI(A,0) = 1 - \left(-\frac{1}{6}\left(\frac{1}{1}\log_2\frac{1}{1} + \frac{0}{1}\log_2\frac{0}{1}\right) - \frac{5}{6}\left(\frac{3}{5}\log_2\frac{3}{5} + \frac{2}{5}\log_2\frac{2}{5}\right)\right) = 1 - \frac{5}{6}(0.44 + 0.53) = 0.19$$

GI(A,2)=0 GI(A,30)=0.08 GI(A,31)=0.46
GI(A,32)=0.19 GI(A,40)=0

- Il guadagno di informazione maggiore si ha generando i sottointervalli A≤31 e A>31

Calcolare GI(A,31)

Binarizzazione

- La rappresentazione di un attributo discreto mediante un insieme di attributi binari è invece detta binarizzazione

Categoria	Valore intero	X1	X2	X3
Gravemente insuff.	4	0	0	0
Insuff.	5	0	0	1
Suff.	6	0	1	0
Discreto	7	0	1	1
Buono	8	1	0	0

- Questa soluzione può portare la tecnica di data mining a inferire una relazione tra “Suff” e “Discreto” poiché entrambi hanno il bit X2=1

- Questa soluzione utilizza attributi asimmetrici binari

Categoria	Valore intero	X1	X2	X3	X4	X5
Gravemente insuff.	4	1	0	0	0	0
Insuff.	5	0	1	0	0	0
Suff.	6	0	0	1	0	0
Discreto	7	0	0	0	1	0
Buono	8	0	0	0	0	1

Trasformazione di attributi

- Una funzione che mappa l'intero insieme di valori di un attributo in un nuovo insieme in modo tale che a ogni valore nell'insieme di partenza corrisponda un unico valore in quello di arrivo
  - ✓ Funzioni semplici: x^k, log(x), e^x, |x|
  - ✓ Standardizzazione e normalizzazione

Trasformazione di attributi

- Funzioni semplici: sono utilizzate per
  - ✓ Enfatizzare alcune proprietà dei dati
    - Particolari distribuzioni dei dati
  - ✓ Ridurre range di variabilità troppo elevate
    - La quantità di byte trasferiti in una sessione varia da 1 a un bilione! Utilizzando una trasformazione logaritmica in base 10 si riducono le differenze tra file di dimensione 10^8 e 10^9 per enfatizzare che entrambi riguardano trasferimenti di file di grandi dimensione. Tale differenza sarà maggiore a quella tra 10 (10^1) e 1000 (10^3) che potrebbero modellare due tipi di operazioni differenti in rete.
- Attenzioni alle proprietà della trasformazione
  - ✓ 1/X riduce i valori maggiori di 1 ma incrementa quelli minori di 1 quindi inverte l'ordinamento di un insieme di eventi

Trasformazione di attributi

- Normalizzazione: permette all'intero set di valori di rispettare una certa proprietà
  - ✓ Necessaria per poter combinare variabili con differenti intervalli di variazione
    - Si pensi per esempio a dover combinare l'età di una persona con il suo reddito
- Max-Min normalization: si riscala l'attributo A in modo che i nuovi valori cadano tra $NewMin_A$ e $NewMax_A$.

$$x' = \frac{(x - Min_A)}{(Max_A - Min_A)}(NewMaxA - NewMin_A) + NewMin_A$$

  - Molto sensibile agli outlier
  - Richiede di conoscere minimo e massimo
  - ✓ Z-score normalization: fa si che una distribuzione statistica abbia media 0 e deviazione standard 1

$$x' = (x - \bar{x}) / s_x$$

    - Meno sensibile agli outlier
    - I valori riscalati non rientrano in un intervallo predefinito

Similarità e dissimilarità

- Similarità
  - ✓ Una misura numerica che esprime il grado di somiglianza tra due oggetti
  - ✓ E' tanto maggiore quanto più gli oggetti si assomigliano
  - ✓ Normalmente assume valori nell'intervallo [0,1]
- Dissimilarità o distanza
  - ✓ Una misura numerica che esprime il grado di differenza tra due oggetti
  - ✓ E' tanto minore quanto più gli oggetti si assomigliano
  - ✓ Il range di variazione non è fisso, normalmente assume valori nell'intervallo [0,1] oppure [0,∞]
- La similarità/dissimilarità tra due oggetti con più attributi è tipicamente definita combinando opportunamente le similarità/dissimilarità tra le coppie di attributi corrispondenti

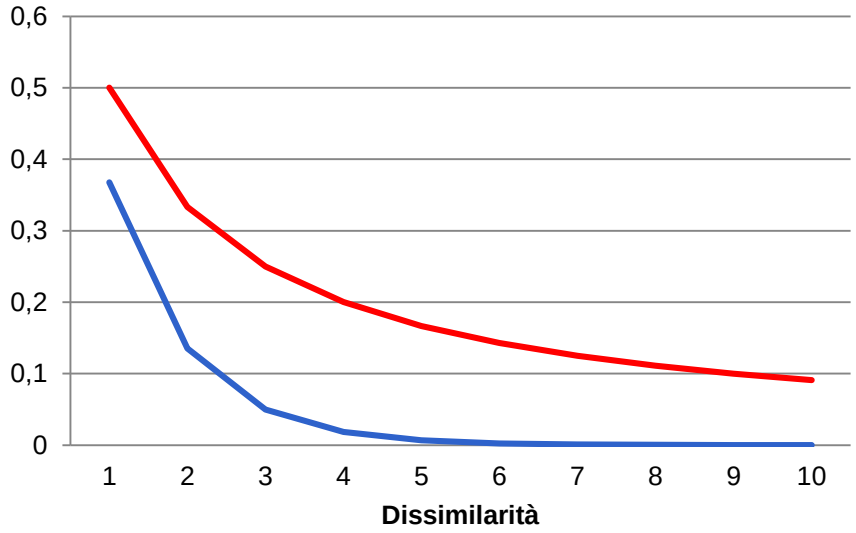
Similarità e dissimilarità

- Il significato cambia in base al tipo di attributo considerato

Tipo		Dissimilarità	Similarità
Categorici (qualitativi)	Nominale	$d = \begin{cases} 0 & se\ x = y \\ 1 & se\ x \neq y \end{cases}$	$s = \begin{cases} 1 & se\ x = y \\ 0 & se\ x \neq y \end{cases}$
	Ordinale (con valori mappati in [0,n-1])	$d = \frac{\lvert x - y \rvert}{n-1}$	$s = 1\text{-}d$
Numerici (quantitativi)	Di Intervallo o Di Rapporto	$d = \lvert x - y \rvert$	$s = -d$ $\quad s = \frac{1}{1+d}$ $\quad s = e^{-d}$ $s = 1 - \frac{d - MinD}{MaxD - MinD}$

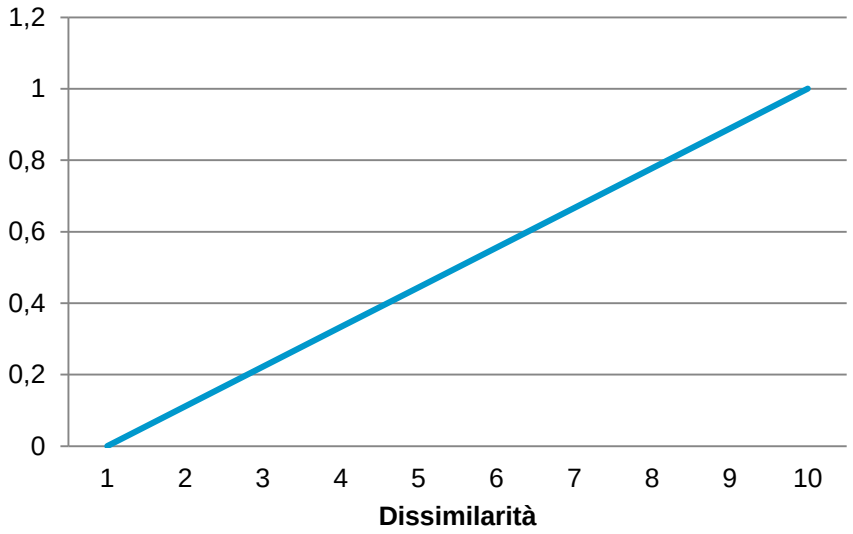
- La similarità in giallo non è vincolata al range [0,..,1] e quindi si preferiscono usare i rapporti anche se forniscono misure non lineari

Similarità e dissimilarità



$$s = \frac{1}{1+d}$$

$$s = e^{-d}$$



$$s = \frac{d - MinD}{MaxD - MinD}$$

Distanze

- Sono dissimilarità con particolari proprietà
- Distanza euclidea

$$d(\mathbf{x},\mathbf{y}) = \sqrt{\sum_{k=1}^{n}(x_k - y_k)^2}$$

  - ✓ *n* è il numero degli attributi (dimensioni) coinvolte

- Distanza di Minkowsi

$$d(\mathbf{x},\mathbf{y}) = \left(\sum_{k=1}^{n}|\, x_k - y_k\,|^r\right)^{1/r}$$

  - ✓ *r* =1 City block
  - ✓ *r* =2 Distanza euclidea
  - ✓ *r* =∞ Lmax ossia la massima differenza tra tutte le coppie di attributi corrispondenti

Distanze

- Sono dissimilarità con particolari proprietà
- Distanza euclidea

$$d(\mathbf{x},\mathbf{y}) = \sqrt{\sum_{k=1}^{n}(x_k - y_k)^2}$$

  - ✓ *n* è il numero degli attributi (dimensioni) coinvolte

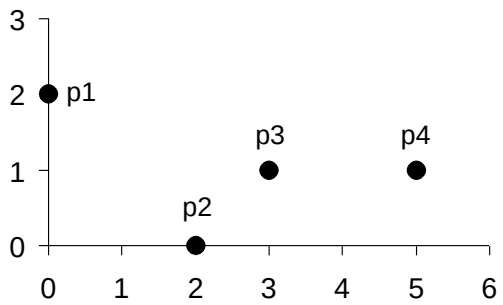
- Distanza di Minkowsi

$$d(\mathbf{x},\mathbf{y}) = \left(\sum_{k=1}^{n}|\, x_k - y_k\,|^r\right)^{1/r}$$

  - ✓ *r* =1 City block
  - ✓ *r* =2 Distanza euclidea
  - ✓ *r* =∞ Lmax ossia la massima differenza tra tutte le coppie di attributi corrispondenti

Distanza di Minkowski

Punti	x	y
p1	0	2
p2	2	0
p3	3	1
p4	5	1



L1	p1	p2	p3	p4
p1	0	4	4	6
p2	4	0	2	4
p3	4	2	0	2
p4	6	4	2	0

L2	p1	p2	p3	p4
p1	0	2.828	3.162	5.099
p2	2.828	0	1.414	3.162
p3	3.162	1.414	0	2
p4	5.099	3.162	2	0

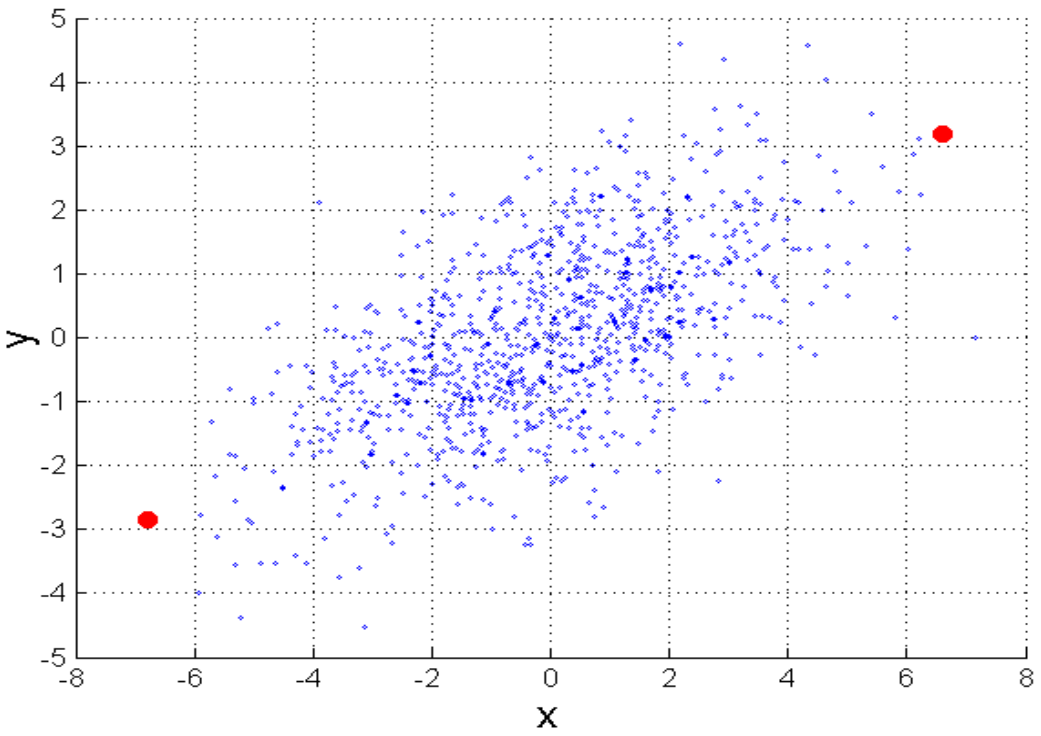
L_∞	p1	p2	p3	p4
p1	0	2	3	5
p2	2	0	1	3
p3	3	1	0	2
p4	5	3	2	0

Matrice delle distanze

ML

Distanza di Mahalanobis

$$mahalanobis(p,q) = (p-q)\Sigma^{-1}(p-q)^T$$

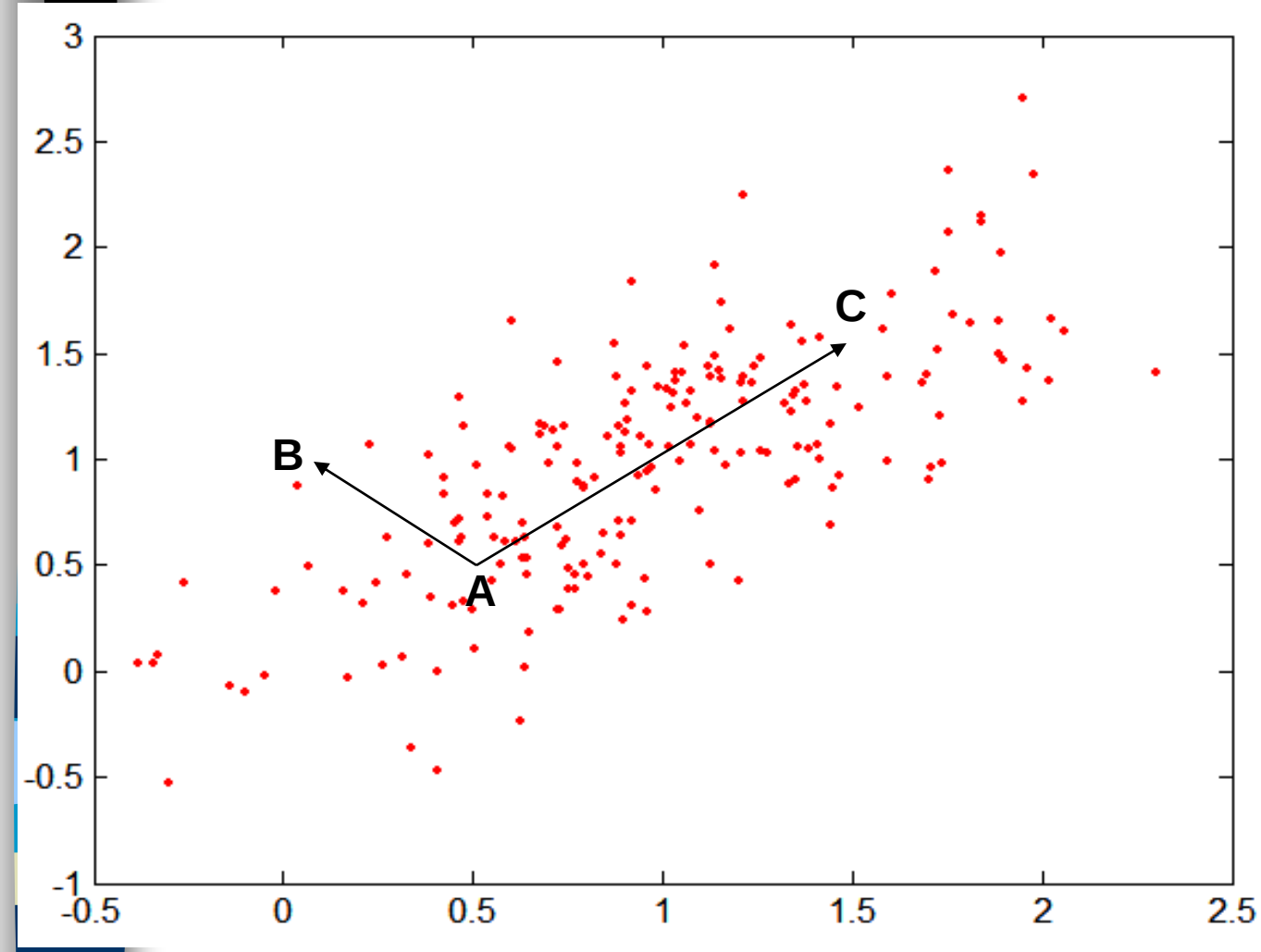


Σ è la matrice di covarianza dei punti

$$\Sigma_{j,k} = \frac{1}{n-1}\sum_{i=1}^{n}(X_{ij} - \overline{X}_j)(X_{ik} - \overline{X}_k)$$

Per i punti in rosso, la distanza euclidea è 14.7, quella di Mahalanobis è 6.

Mahalanobis Distance



Covariance Matrix:

$$\Sigma = \begin{bmatrix} 0.3 & 0.2 \\ 0.2 & 0.3 \end{bmatrix}$$

A: (0.5, 0.5)

B: (0, 1)

C: (1.5, 1.5)

Mahal(A,B) = 5

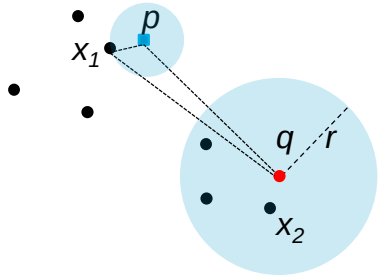
Mahal(A,C) = 4

Proprietà delle distanze

- Dati due oggetti *p* e *q* e una misura di dissimilarità *d*()
  1. $d(p, q) \geq 0$ (Positività I)
  2. $d(p, q) = 0$ solo se $p = q$ (Positività II)
  3. $d(p, q) = d(q, p)$ (Simmetria)
  4. $d(p, r) \leq d(p, q) + d(q, r)$ (Disuguaglianza triangolare)

- Una distanza che soddisfi tutte le proprietà è detta **metrica**
- Le proprietà delle distanze rendono più agevole (o possibile) l'utilizzo di alcuni algoritmi (es. clustering).

Query di range e disuguaglianza triangolare



Sia dato un insieme di punti $P=\{x_1, x_2,\ldots, x_n\}$ una range query con raggio r da un punto q. Siano note inoltre le distanze $d(x_i,p)$ con $x_i \in P$

Sfruttando la disuguaglianza triangolare è possibile limitare il numero di distanze $d(x_i,q)$ da calcolare per rispondere alla query

- $d(p,q)<=d(p, x_i)+d(q, x_i)$ → $d(q, x_i)>= d(p, q)-d(p, x_i)$ → $d(q, x_i)>=$LB tutti i punti x_i per cui $d(p,q)-d(p, x_i) > r$ devono essere scartati senza valutarli
- $d(p, x_i)<=d(p, q)+d(q, x_i)$ → $d(q, x_i) <= d(p, x_i)-d(p, q)$ → $d(q, x_i)>=$UB tutti i punti x_i per cui $d(p, x_i)-d(p,q) < r$ devono essere accettati senza valutarli

Dissimilarità non metriche

- Set difference
  - ✓ La differenza tra due insiemi A e B non gode della proprietà di simmetricità
  - ✓ *A* = {1,2,3,4} *B* = {2,3,4} A-B = {1} B-*A*=∅

- Tempo

$$d(t_1,t_2)=\begin{cases} t_2-t_1 & se\ \ t_1<t_2 \\ 24+t_2-t_1 & se\ \ t_1\geq t_2 \end{cases}$$

  - ✓ Non rispetta la simmetricità
    - La distanza d(1pm,2pm)=1
    - La distanza d(2pm,1pm)=23

Esiste una misura di similarità tra insiemi che sia una metrica?

Proprietà delle similarità

- Anche le misure di similarità hanno delle proprietà comuni
- Dati due oggetti *p* e *q* e una misura di similarità *s*( )
  1. *s(p, q) = 1* solo se *p = q*.
  2. *s(p, q) = s(q, p)* (Simmetria)

- Non esiste per le misure di similarità un concetto equivalente alla disuguaglianza triangolare
- Talvolta le misure di similarità possono essere convertite in metriche (es. similarità Coseno e Jaccard)

Similarità tra vettori binari

- E' frequente che gli attributi che descrivono un oggetto contengano solo valori binari. Dati quindi i due vettori p e q, si definiscono le seguenti grandezze
  - ✓ M_{01} = Il numero di attributi in cui p = 0 e q = 1
  - ✓ M_{10} = Il numero di attributi in cui p = 1 e q = 0
  - ✓ M_{00} = Il numero di attributi in cui p = 0 e q = 0
  - ✓ M_{11} = Il numero di attributi in cui p = 1 e q = 1
- Simple Matching coefficient
  - ✓ SMC = numero di match / numero di attributi
    = (M11 + M00) / (M01 + M10 + M11 + M00)
  - ✓ Utile per misurare quali studenti hanno risposto in modo similare alle domande di un test VERO/FALSO
  - ✓ Non utilizzabile in presenza di attributi asimmetrici
- Coefficiente di Jaccard
  - ✓ J = num. di corrispondenze 11/num. attributi con valori diversi da 00
    = (M11) / (M01 + M10 + M11)
  - ✓ Non considera i casi le corrispondenze 00

SMC versus Jaccard: un esempio

- Siano p e q i vettori che descrivono le transazioni di acquisto di due clienti. Ogni attributo corrisponde a uno dei prodotti in vendita

$p = 1\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0$

$q = 0\ 0\ 0\ 0\ 0\ 0\ 1\ 0\ 0\ 1$

$M_{01} = 2 \quad M_{10} = 1 \quad M_{00} = 7 \quad M_{11} = 0$

$$SMC = (M_{11} + M_{00})/(M_{01} + M_{10} + M_{11} + M_{00})$$
$$= (0+7) / (2+1+0+7) = 0.7$$

$$J = (M_{11}) / (M_{01} + M_{10} + M_{11}) = 0 / (2 + 1 + 0) = 0$$

- Con SMC gli attributi a 0 dominano l'informazione derivante dagli attributi a 1

Similarità Coseno

ML

- Come l'indice di Jaccard non considera le corrispondenze 00, ma permette inoltre di operare con vettori non binari
  - ✓ Codifica di documenti in cui ogni attributo del vettore codifica il numero di volte in cui la parola corrispondente compare nel testo

- Siano d_1 e d_2 sono due vettori non binari

  $$\cos(d_1, d_2) = (d_1 \bullet d_2) / \|d_1\| \|d_2\| ,$$

  dove $\bullet$ indica il prodotto scalare dei vettori e $\| d \|$ è la lunghezza del vettore d.

  $$\|\mathbf{d}\| = \sqrt{\mathbf{d} \cdot \mathbf{d}} = \sqrt{\sum_{k=1}^{n} d_k^2}$$

  - ✓ La similarità coseno è effettivamente una misura dell'angolo tra i due vettori ed è quindi 0 se l'angolo è 90°, ossia se non condividono alcun elemento comune

Similarità Coseno: un esempio

ML

d_1 = **3 2 0 5 0 0 0 2 0 0**

d_2 = **1 0 0 0 0 0 0 1 0 2**

$d_1 \bullet d_2$= 3*1 + 2*0 + 0*0 + 5*0 + 0*0 + 0*0 + 0*0 + 2*1 + 0*0 + 0*2 = 5

$\|d_1\|$ = (3*3+2*2+0*0+5*5+0*0+0*0+0*0+2*2+0*0+0*0)$^{0.5}$ = (42) $^{0.5}$

= 6.481

$\|d_2\|$ = (1*1+0*0+0*0+0*0+0*0+0*0+0*0+1*1+0*0+2*2) $^{0.5}$ = (6) $^{0.5}$

= 2.245

$$\cos(d_1, d_2) = 0.343$$

La similarità coseno è spesso utilizzata per calcolare la similarità tra i documenti: a ogni elemento del vettore corrisponde un termine. Documenti con lunghezze diverse avranno vettori con lunghezze diverse.
Che tipo di normalizzazione può essere necessaria per confrontare documenti di lunghezza diversa?

Correlazione

- La correlazione tra coppie di oggetti descritti da attributi (binari o continui) è una misura dell'esistenza di una relazione lineare tra i suoi attributi

$$Corr(\mathbf{x},\mathbf{y}) = \frac{Cov(\mathbf{x},\mathbf{y})}{StDev(\mathbf{x}) \cdot StDev(\mathbf{y})}$$

$$Cov(\mathbf{x},\mathbf{y}) = \frac{1}{n}\sum_{k=1}^{n}(x_k - \bar{x})(y_k - \bar{y}) \qquad StDev(\mathbf{x}) = \sqrt{\frac{1}{n}\sum_{k=1}^{n}(x_k - \bar{x})^2}$$

- La correlazione varia tra [-1,1].
  - ✓ Una correlazione di 1(-1) significa che gli attributi possono essere vicendevolmente espressi da una relazione lineare del tipo $x_k = ay_k + b$

Correlazione

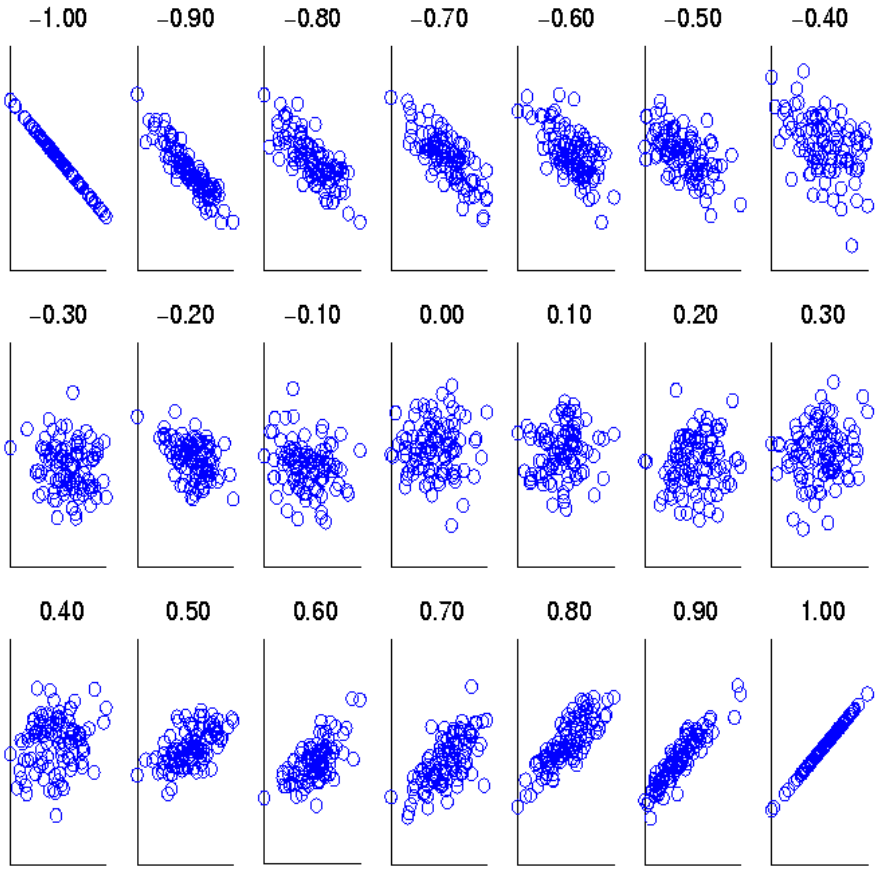
x=(-3, 6, 0, 3, -6) **y=(1,-2, 0, -1, 2)** ***Corr*(x,y)=-1**

x=(3, 6, 0, 3, 6) **y=(1,2, 0, 1, 2)** ***Corr*(x,y)=1**

- Potrebbero comunque esistere tra i dati relazioni non lineari che non sarebbero quindi non catturate!
  - ✓ Tra i seguenti oggetti esiste una correlazione del tipo $x_k=y_k^2$ ma *Corr*(**x**,**y**)=0

    x=(-3, -2, -1, 0, 1, 2, 3) **y**=(9, 4, 1, 0, 1, 4, 9)
- La correlazione può essere utile anche per scartare attributi che non portano informazioni aggiuntive
  - ✓ In questo caso x e y rappresentano due attributi distinti e i loro elementi le istanze dei due attributi nei diversi oggetti del data set

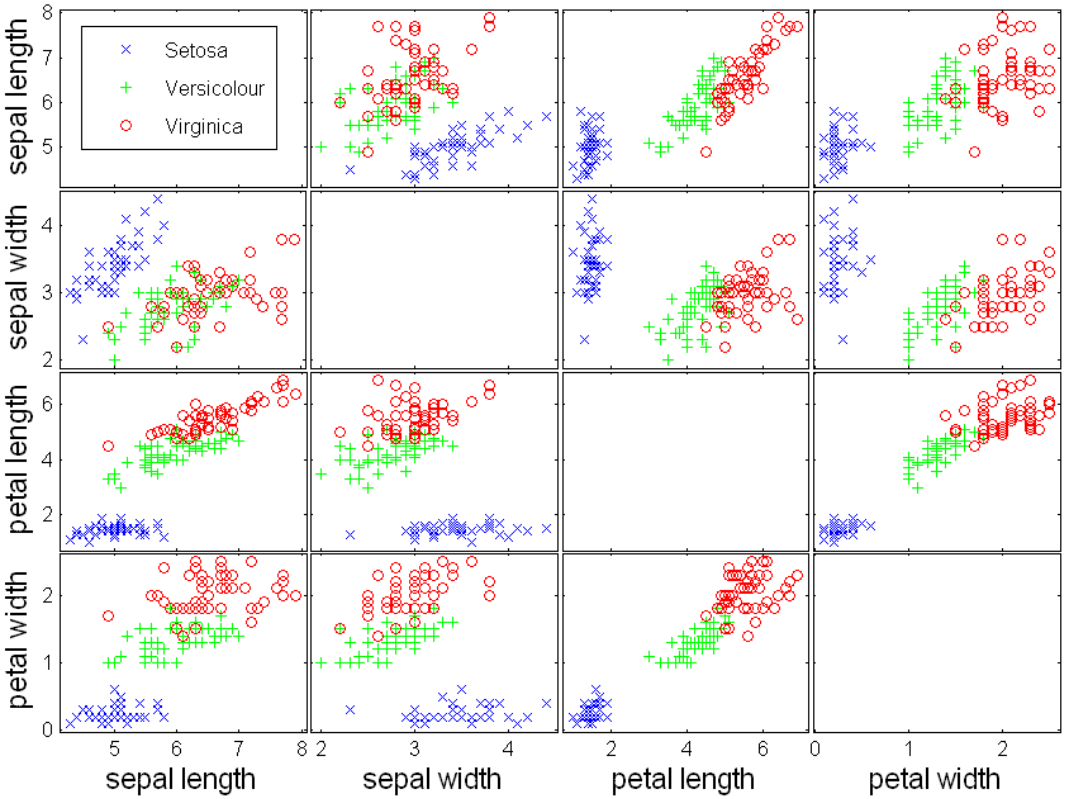
Visualizzazione della correlazione



- ✓ **x** e **y** sono due oggetti descritti da 30 attributi continui.
- ✓ In ogni grafico i valori degli attributi sono stati generati con livelli diversi di correlazione
- ✓ Ogni cerchio rappresenta uno dei trenta attributi di **x** e **y**. La sua ascissa corrisponde a x_k mentre l'ordinata a y_k

Visualizzazione della correlazione: grafici a dispersione

- ✓ Permette di determinare se alcuni degli attributi sono correlati
  - ✓ Utile per ridurre il numero di attributi considerati
- ✓ Quando le etichette sono disponibili, permette di determinare se è possibile classificare gli oggetti in base ai valori di due attributi
- ✓ Un grafico per ogni coppia di attributi utilizzati per descrivere i fiori



Similarità in presenza di attributi eterogenei

- I precedenti approcci considerano oggetti descritti da attributi dello stesso tipo
- In presenza di attributi eterogenei è necessario calcolare separatamente le similarità e quindi combinarle in modo che il loro risultato appartenga al range [0;1]
  - ✓ Se uno o più degli attributi è asimmetrico è necessario escluderli dal computo qualora il loro match sia di tipo 00
  - ✓ Se gli attributi hanno una rilevanza diversa è possibile aggiungere un peso w_k nel calcolo della similarità complessiva. E' consigliabile che la somma dei pesi sia 1

$$Sim(\mathbf{x},\mathbf{y}) = \frac{\sum_{k=1}^{n} w_k \delta_k s_k(\mathbf{x},\mathbf{y})}{\sum_{k=1}^{n} \delta_k}$$

$$\delta_k \begin{cases} 0 \text{ se l'attributo k è asimmetrico e il match è 00} \\ 1 \text{ altrimenti} \end{cases}$$