

## Data Mining

**Studiare e preparare i dati**

# 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

**Oggetti**

- Un **attributo** è una proprietà o una caratteristica di un oggetto

- ✓ Sinonimi di attributo sono: variabile, campo, caratteristica

**Attributi**

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

- E' necessario conoscere le caratteristiche degli attributi per effettuare analisi sensate
- 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  $=, \neq$
  - ✓ Ordinamento  $<, \leq, >, \geq$
  - ✓ Additività  $+, -$
  - ✓ Moltiplicabilità  $*, /$
- 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

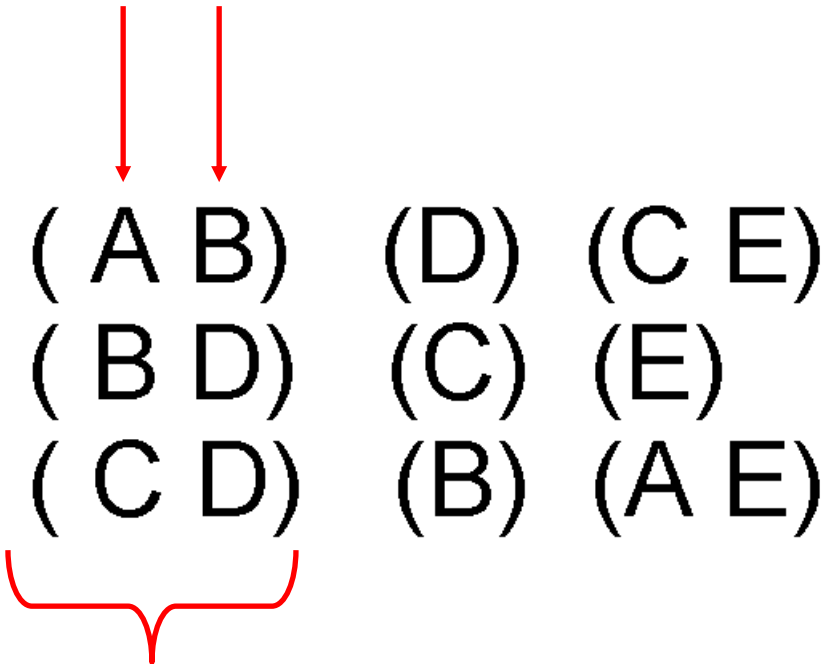
- Un tipo speciale di record in cui
  - ✓ Ogni record (transazione) coinvolge più item
  - ✓ Per esempio in un supermercato l'insieme dei prodotti comprati da un cliente durante una visita al negozio costituisce una transazione, mentre i singoli prodotti acquistati sono gli item.
  - ✓ Il numero degli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

<i>TID</i>	<i>Items</i>
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

Item/Eventi



( A B )	( D )	( C E )
( B D )	( C )	( E )
( C D )	( B )	( A E )

Un elemento di  
una sequenza





# Dati ordinati

- Sequenze di dati genomici

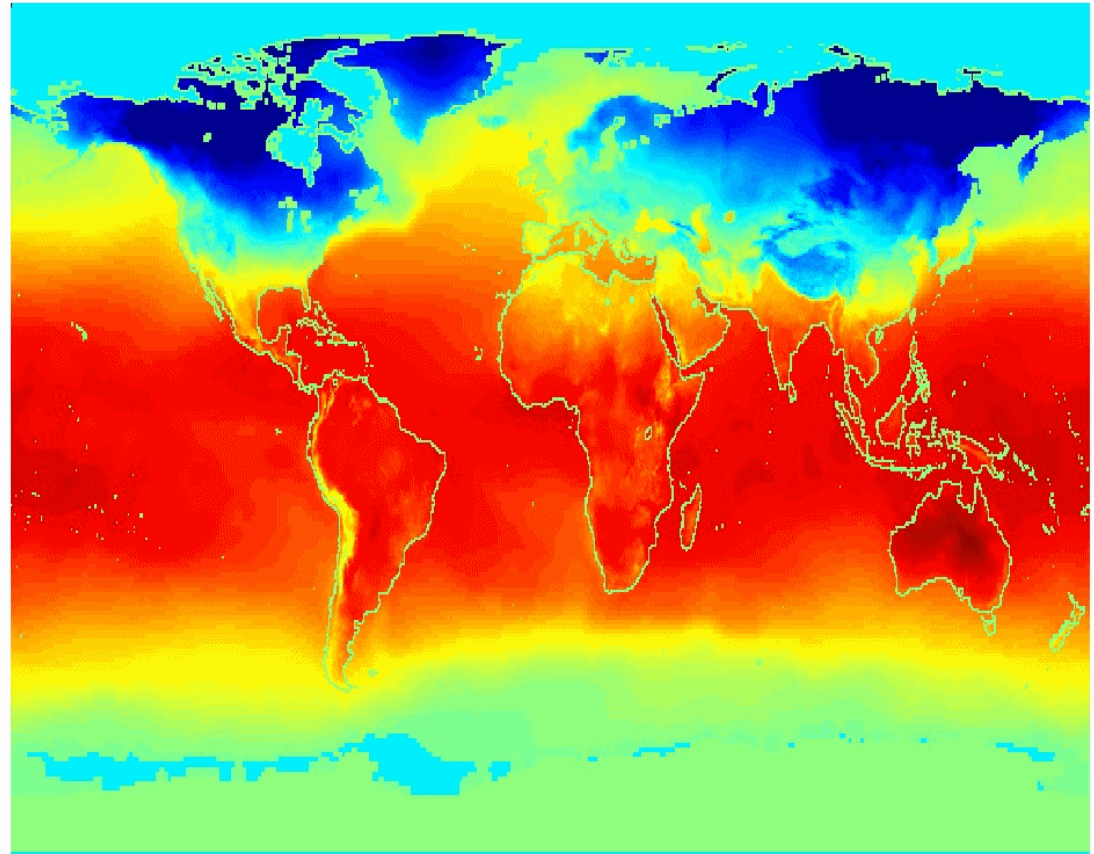
```
GGTTCCGCCTTCAGCCCCGCGCC  
CGCAGGGCCCGCCCCGCGCCGTC  
GAGAAGGGCCCGCCTGGCGGGCG  
GGGGGAGGCGGGGCCGCCCGAGC  
CCAACCGAGTCCGACCAGGTGCC  
CCCTCTGCTCGGCCTAGACCTGA  
GCTCATTAGGCGGCAGCGGACAG  
GCCAAGTAGAACACGCGAAGCGC  
TGGGCTGCCTGCTGCGACCAGGG
```

# Dati ordinati

## ■ Dati Spazio-Temporali

Jan

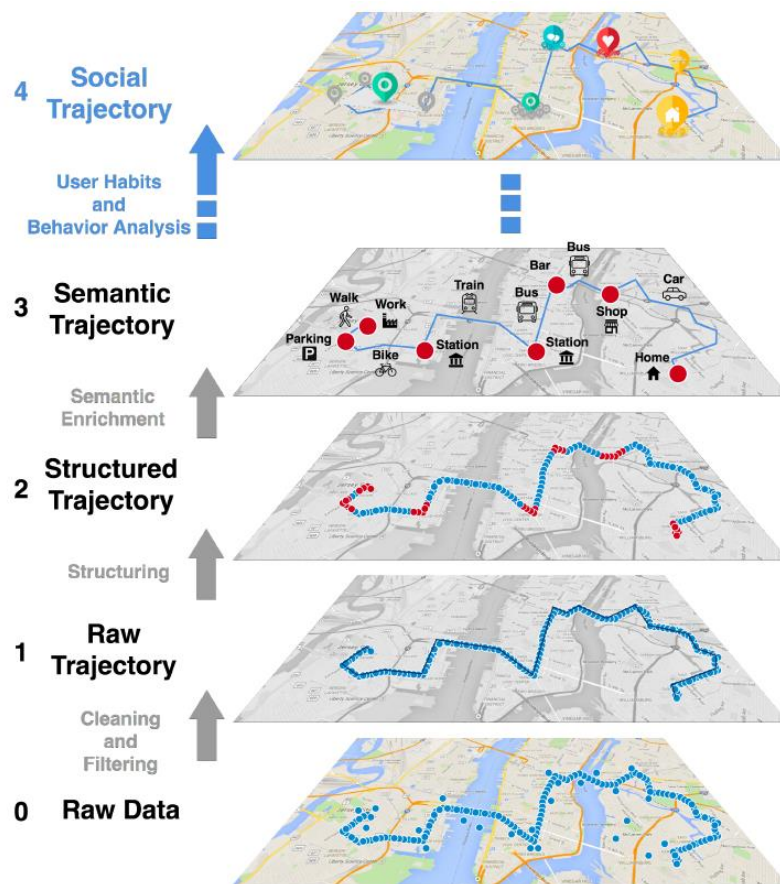
**Temperatura  
media mensile di  
terre e oceani**



# Dati di traiettoria

Semantic Trajectory  
sequenza di punti  
(*Time, Place*)

Raw data  
sequenza di punti  
(*Time, Lat, Lon*)





# Esplorazione dei dati

- Un'analisi preliminare dei dati finalizzata a individuarne le principali caratteristiche
  - ✓ Aiuta a scegliere il tool migliore per il preprocessing e l'analisi
  - ✓ Permette di utilizzare le capacità umane per individuare pattern
    - Un analista umano può individuare velocemente pattern non individuabili dai tool di analisi
- 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 ( $\sim 3.7 \times 10^5 / 6 \times 10^7$ ).
- 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 ( $\sim 2.7 \times 10^6 / 6 \times 10^7$ ).
- Le nozioni di frequenza e moda sono normalmente utilizzate per attributi categorici

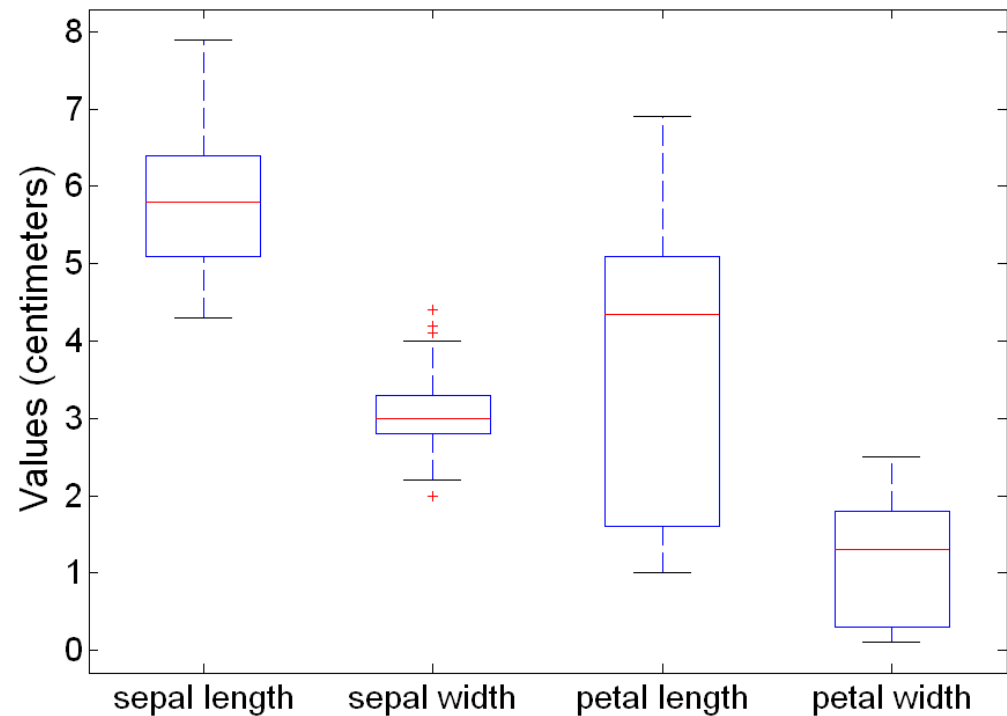
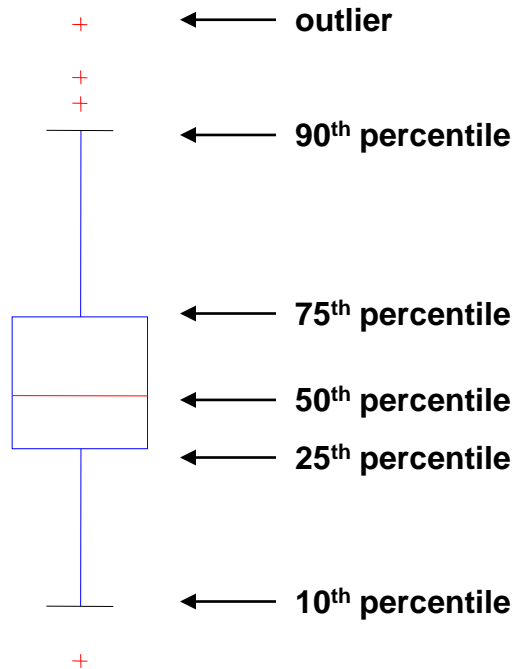


# Percentili

- 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”.

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

- ✓ 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.

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

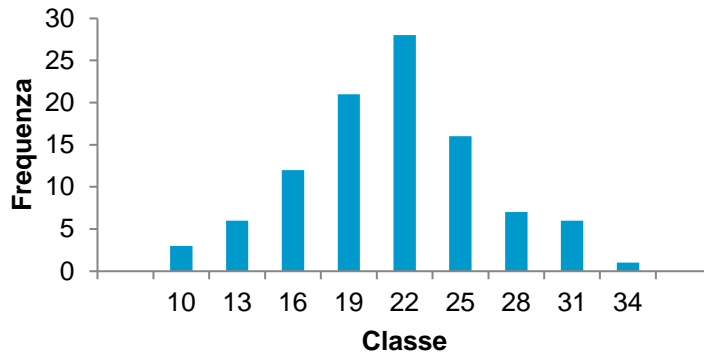
- Varianza e scarto quadratico medio sono sensibili agli outlier poichè sono legati quadraticamente al concetto di media
- Altre misure meno sensibili a questo problema sono:

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

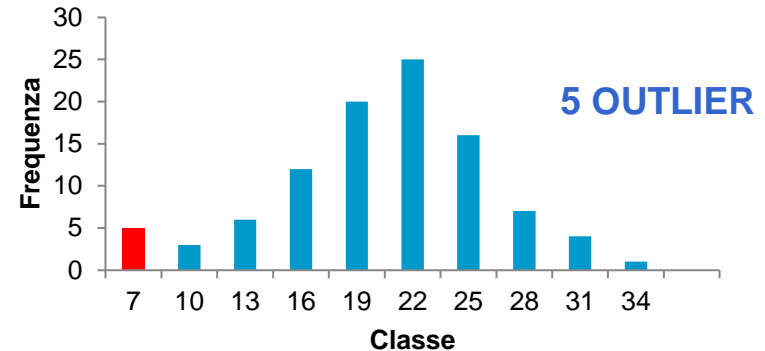
$$\text{MedianAbsoluteDeviation} \quad MAD(\mathbf{x}) = \text{mediana}(\{|x_1 - \bar{x}|, \dots, |x_n - \bar{x}|\})$$

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

# 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 statistici  
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

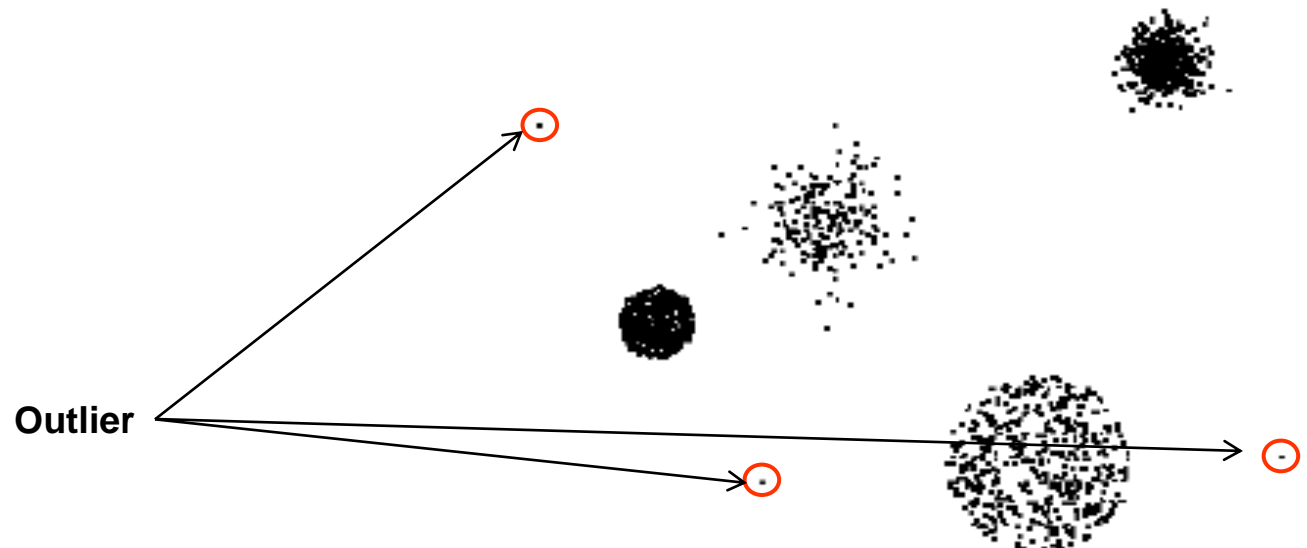


# Rumore

- 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



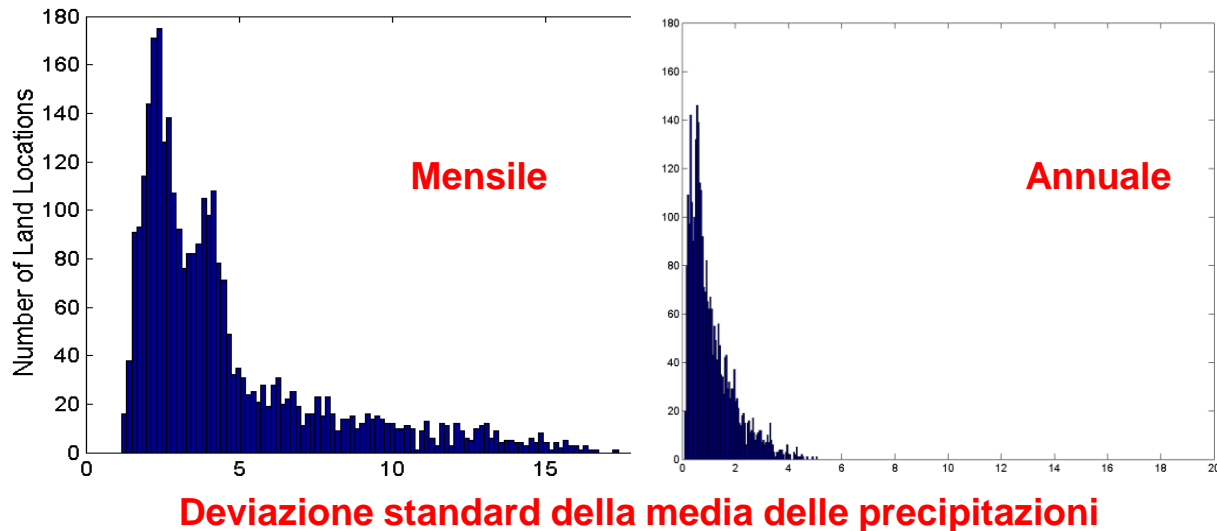
# Preprocessing del data set

- 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
  - ✓ 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



# 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

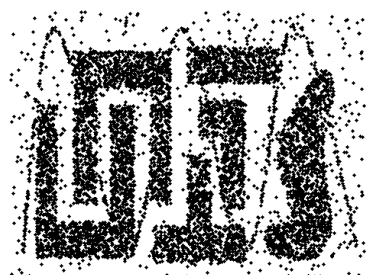
- ✓ C'è la stessa probabilità di selezionare ogni elemento
- ✓ Campionamento senza reimbussolamento
  - Gli elementi selezionati sono rimossi dalla popolazione
- ✓ Campionamento con reimbussolamento
  - Gli elementi selezionati non sono rimossi dalla popolazione
  - In questo caso un elemento può essere selezionato più volte.
  - Dà risultati simili al precedente se la cardinalità del campione è  $\ll$  di quella della popolazione
  - E' più semplice da esaminare poiché la probabilità di scegliere un elemento non cambia durante il processo

## ■ Campionamento stratificato:

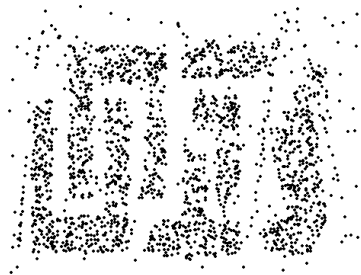
- ✓ Si suddividono i dati in più partizioni, quindi si usa un campionamento casuale semplice su ogni partizione.
- ✓ Utile nel caso in cui la popolazione sia costituita da tipi diversi di oggetti con cardinalità differenti. Un campionamento casuale può non riuscire a fornire un'adeguata rappresentazione dei gruppi meno frequenti

# La dimensione del campione

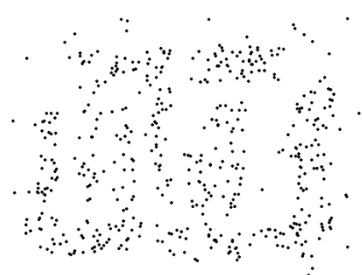
- Scelta la modalità di campionamento è necessario fissare la dimensione del campione al fine di limitare la perdita di informazione



8000 punti

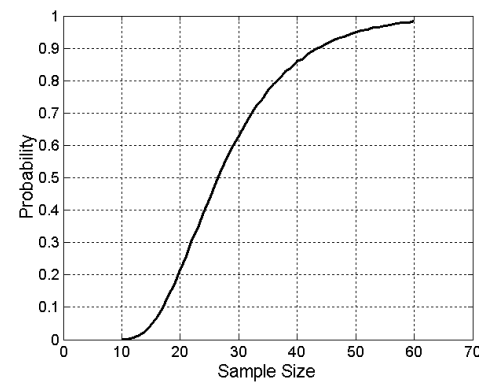


2000 punti



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:

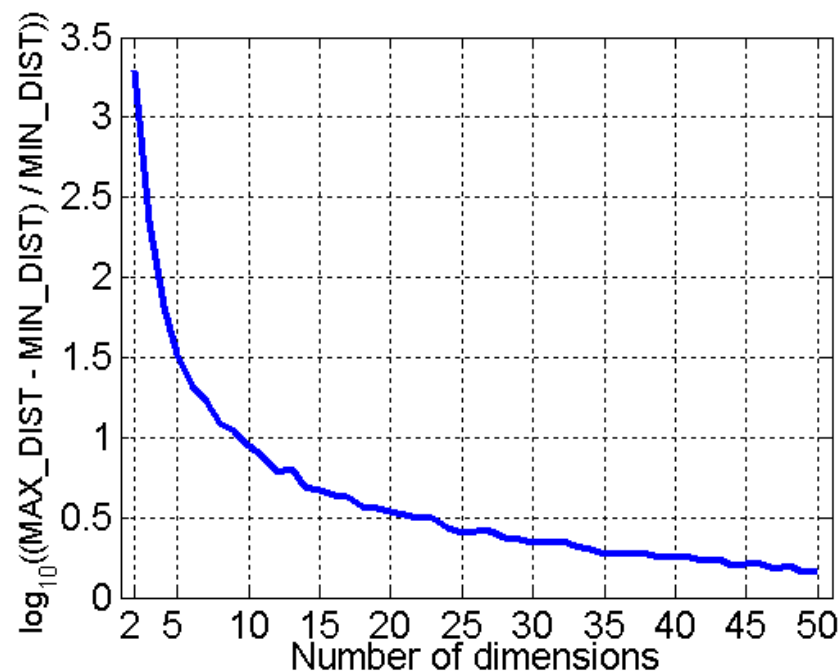
- ✓ Evitare la “*curse of dimensionality*”: la maledizione della dimensionalità
- ✓ Ridurre la quantità di tempo e di memoria utilizzata dagli algoritmi di data mining (riduzione dello spazio di ricerca)
- ✓ Semplificare la visualizzazione dei dati
- ✓ Eliminare attributi non rilevanti ed eliminare il 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 outlier 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

# 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:

- ✓ **Approcci 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

# 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

- One-hot encoding**

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



# 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  $[0,\infty]$

- La similarità/dissimilarità tra due oggetti con più attributi è tipicamente definita combinando opportunamente le similarità/dissimilarità tra le coppie di attributi corrispondenti

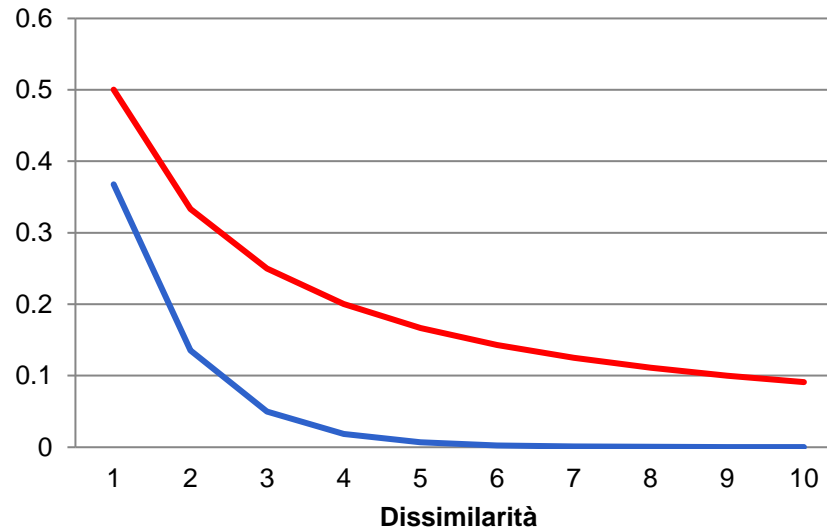
# Similarità e dissimilarità

- Il significato cambia in base al tipo di attributo considerato

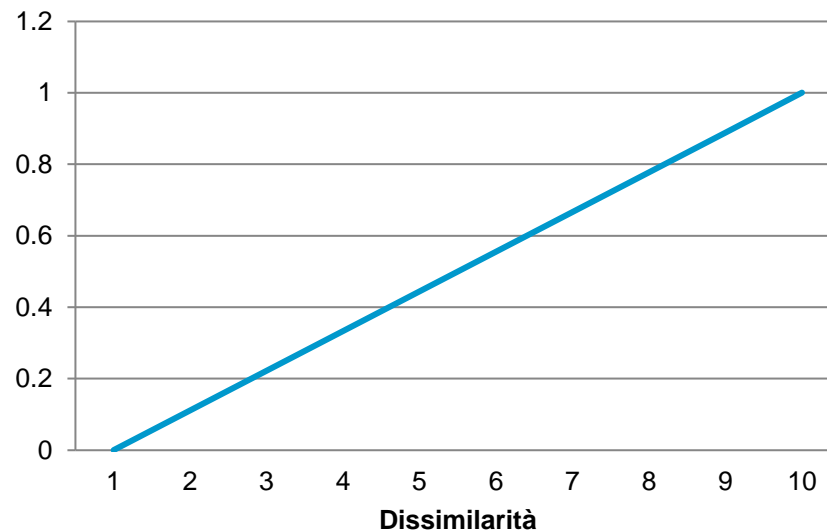
Tipo		Dissimilarità	Similarità
Categorici (qualitativi)	Nominale	$d = \begin{cases} 0 & \text{se } x = y \\ 1 & \text{se } x \neq y \end{cases}$	$s = \begin{cases} 1 & \text{se } x = y \\ 0 & \text{se } x \neq y \end{cases}$
	Ordinale (con valori mappati in $[0, n-1]$ )	$d = \frac{ x - y }{n - 1}$	$s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

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

# Similarità e dissimilarità



$$\text{Red line: } s = \frac{1}{1+d}$$
$$\text{Blue line: } s = e^{-d}$$



$$\text{Cyan line: } s = \frac{d - \text{MinD}}{\text{MaxD} - \text{MinD}}$$

# 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 Minkowski

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

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

# 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

- ✓  $SMC = \text{numero di match} / \text{numero di attributi}$   
 $= (M_{11} + M_{00}) / (M_{01} + M_{10} + M_{11} + M_{00})$
- ✓ Utile per misurare quali studenti hanno risposto in modo simile alle domande di un test VERO/FALSO
- ✓ Non utilizzabile in presenza di attributi **asimmetrici**

- Coefficiente di Jaccard

- ✓  $J = \text{\#corrispondenze 11} / \text{\#attributi con valori diversi da 00}$   
 $= (M_{11}) / (M_{01} + M_{10} + M_{11})$
- ✓ 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

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

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

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

$$\begin{aligned} \text{SMC} &= (M_{11} + M_{00}) / (M_{01} + M_{10} + M_{11} + M_{00}) \\ &= (0 + 7) / (2 + 1 + 0 + 7) = 0.7 \end{aligned}$$

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

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

# Similarità Coseno

- 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  $d$ .
$$\|d\| = \sqrt{d \bullet d} = \sqrt{\sum_{k=1}^n d_k^2}$$
  - ✓ La similarità coseno è effettivamente una misura dell'angolo tra i due vettori ed è quindi 0 se l'angolo è  $90^\circ$ , ossia se non condividono alcun elemento comune

# Similarità Coseno: un esempio

$$d_1 = 3 \ 2 \ 0 \ 5 \ 0 \ 0 \ 0 \ 2 \ 0 \ 0$$

$$d_2 = 1 \ 0 \ 0 \ 0 \ 0 \ 0 \ 0 \ 1 \ 0 \ 2$$

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

$$\begin{aligned} \|d_1\| &= (3*3 + 2*2 + 0*0 + 5*5 + 0*0 + 0*0 + 0*0 + 2*2 + 0*0 + 0*0)^{0.5} = (42)^{0.5} \\ &= 6.481 \end{aligned}$$

$$\begin{aligned} \|d_2\| &= (1*1 + 0*0 + 0*0 + 0*0 + 0*0 + 0*0 + 0*0 + 1*1 + 0*0 + 2*2)^{0.5} = (6)^{0.5} \\ &= 2.245 \end{aligned}$$

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

*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

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

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

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

- 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  $x_k = ay_k + b$

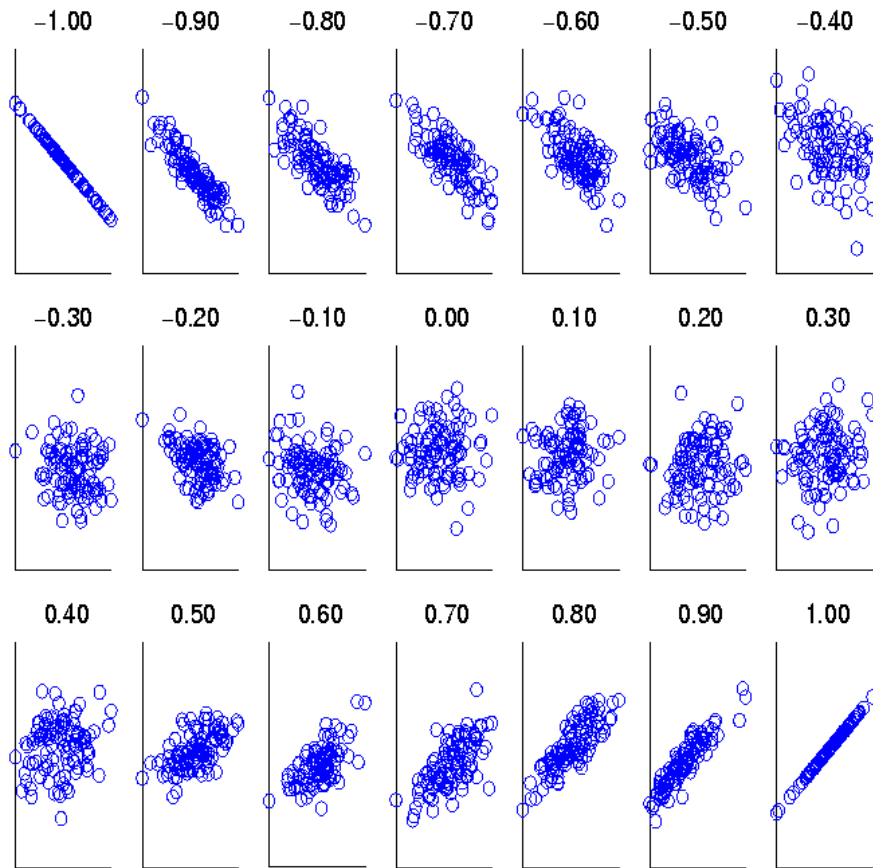
# Correlazione

$$\mathbf{x}=(-3, 6, 0, 3, -6) \quad \mathbf{y}=(1,-2, 0, -1, 2) \quad \text{Corr}(\mathbf{x},\mathbf{y})=-1$$

$$\mathbf{x}=(3, 6, 0, 3, 6) \quad \mathbf{y}=(1,2, 0, 1, 2) \quad \text{Corr}(\mathbf{x},\mathbf{y})=1$$

- 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  $\text{Corr}(\mathbf{x},\mathbf{y})=0$ 
$$\mathbf{x}=(-3, -2, -1, 0, 1, 2, 3) \quad \mathbf{y}=(9, 4, 1, 0, 1, 4, 9)$$
- 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

