Preparazione di Dati per Data Mining

Giuseppe Manco

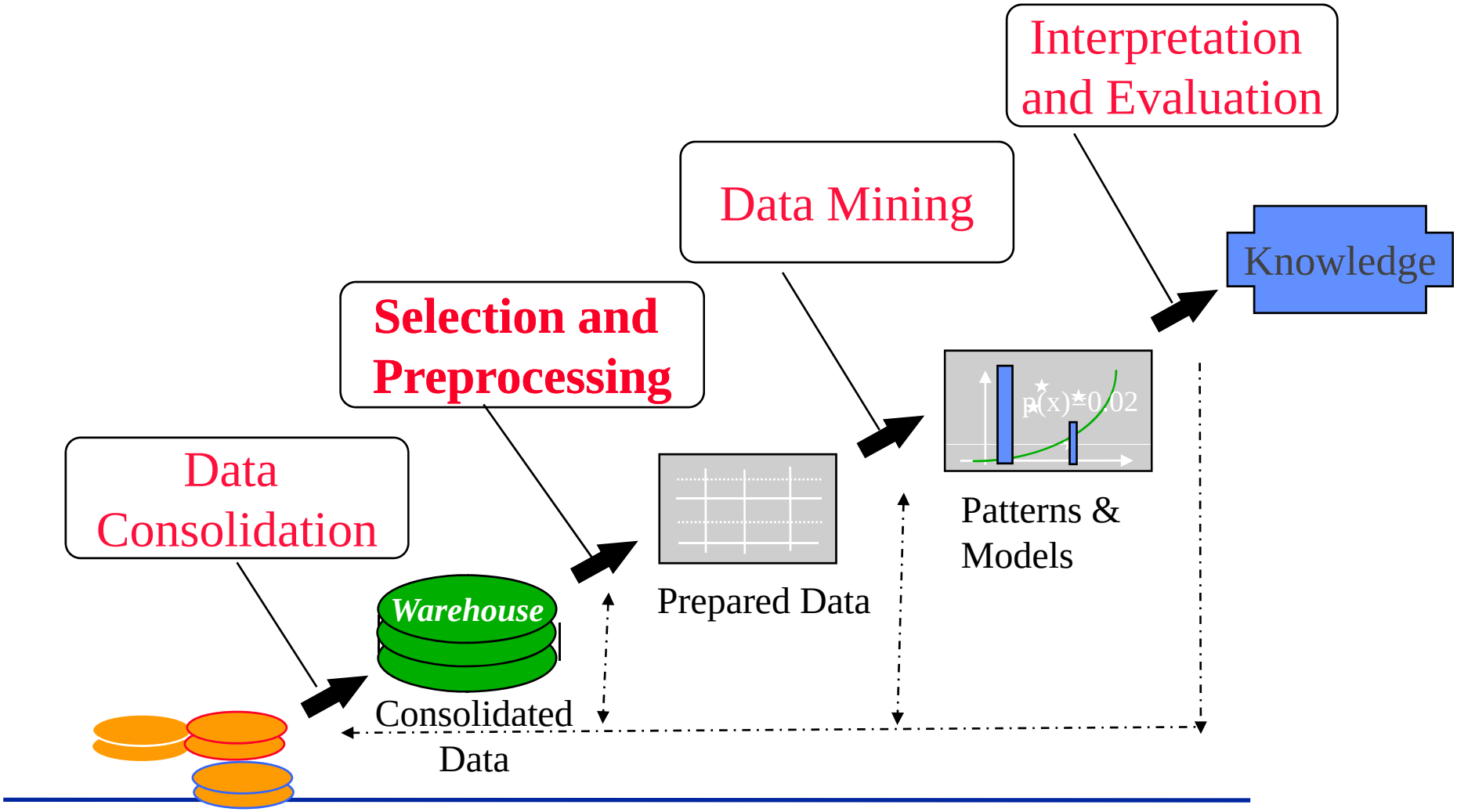
Outline del Modulo

- Introduzione e Concetti di Base
  - Motivazioni
  - Il punto di partenza: dati consolidati, Data Marts
- Information Gathering
  - Misurazioni
  - Visualizzazioni
  - Statistiche
- Data Selection
  - Manipolazione di Tabelle
- Data cleaning
  - Trattamento di valori anomali
  - Identificazione di Outliers
  - Risoluzione di inconsistenze
- Data reduction
  - Campionamento
  - Riduzione di Dimensionalità
- Data transformation
  - Normalizzazioni
  - aggregazione
  - Discretizzazione

Outline del Modulo

- Introduzione e Concetti di Base
- Data Selection
- Information Gathering
- Data cleaning
- Data reduction
- Data transformation

Il Processo di KDD



Problemi tipici

- **Troppi dati**
  - **dati sbagliati, rumorosi**
  - **dati non rilevanti**
  - **dimensione intrattabile**
  - **mix di dati numerici/simbolici**
- **Pochi dati**
  - **attributi mancanti**
  - **valori mancanti**
  - **dimensione insufficiente**

Il Data Preprocessing è un Processo

- Accesso ai Dati
- Esplorazione dei Dati
  - Sorgenti
  - Quantità
  - Qualità
- Ampliamento e arricchimento dei dati
- Applicazione di tecniche specifiche

Il Data Preprocessing dipende (ma non sempre) dall'Obiettivo

- Alcune operazioni sono necessarie
  - Studio dei dati
  - Pulizia dei dati
  - Campionamento
- Altre possono essere guidate dagli obiettivi
  - Trasformazioni
  - Selezioni

Outline del Modulo

- Introduzione e Concetti di Base
- Data Selection
- Information Gathering
- Data cleaning
- Data reduction
- Data transformation

Un tool Fondamentale: le Queries

- **Base di partenza: un datamart**
  - **Sintetizza l'obiettivo principale**
- **Dal datamart estraiamo una tabella**
  - **Contenente le informazioni che ci interessano**
- **Le informazioni (e le trasformazioni) sulla tabella permettono di effettuare data preprocessing**
  - **SELECT**
  - **UPDATE**
  - **DELETE**

SQL Queries

- **Forma principale:**

```
SELECT Attributi necessari
FROM variabili di relazioni
WHERE condizioni sulle variabili
```

- **Tabelle d’esempio:**

```
Beers(name, manf)
Bars(name, addr, license)
Drinkers(name, addr, phone)
Likes(drinker, beer)
Sells(bar, beer, price)
Frequents(drinker, bar)
```

Esempio

- **Quali sono le birre fatte da Anheuser-Busch?**
- **Tabella coinvolta:**

  Beers(name, manf)
- **Query:**

```
SELECT name
FROM Beers
WHERE manf = 'Anheuser-Busch'
```

- **Risposta:**

name
Bud
Bud Lite
Michelob

SQL per la manipolazione di Tabelle

- *** come lista di tutti gli attributi**
- **tabella coinvolta**

  Beers(name, manf)
- **Query**

```
SELECT *
FROM Beers
WHERE manf = 'Anheuser-Busch'
```

- **Risposta:**

name	manf
Bud	Anheuser-Bush
Bud Lite	Anheuser-Bush
Michelob	Anheuser-Bush

SQL per la manipolazione di tabelle

- **Rinomina delle colonne**
- **tabella coinvolta**

  Beers(name, manf)
- **Query**

```
SELECT name AS beer
FROM Beers
WHERE manf = 'Anheuser-Busch'
```

- **Risposta:**

beer
Bud
Bud Lite
Michelob

SQL per la manipolazione di tabelle

- **Espressioni come valori di colonne**
- **tabella coinvolta**

```
Sells(bar, beer, price)
```

- **Query**

```
SELECT bar, beer,
price*120 AS priceInYen
FROM Sells
```

- **Risposta**

bar	beer	priceInYen
Joe's	Bud	300
Sue's	Miller	360
…	…	…

SQL per la manipolazione di tabelle

- **Le espressioni possono anche essere costanti**
- **tabella**

  Likes(drinker, beer)
- **Query**

```
SELECT drinker,
       'likes Bud' AS whoLikesBud
FROM Likes
WHERE beer = 'Bud';
```

- **Risposta**

Drinker	whoLikesBud
Sally	Likes Bud
Fred	Likes Bud
…	…

SQL per la manipolazione di tabelle

- **Condizioni nel WHERE possono utilizzare operatori logici AND, OR, NOT**
- **Seleziona i prezzi per la birra ‘Bud’ nel bar ‘Joe’s’**

```
SELECT price
FROM Sells
WHERE bar = 'Joe''s Bar' AND
      beer = 'Bud'
```

Queries su piu' relazioni

- **Trova le birre che piacciono ai frequentatori del bar "Joe's"**
- **tabelle coinvolte**

  Likes(drinker, beer)
  Frequents(drinker, bar)
- **query**

```
SELECT beer
FROM Frequents, Likes
WHERE bar = 'Joe''s Bar' AND
      Frequents.drinker = Likes.drinker
```

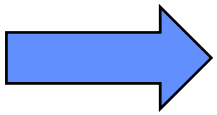
Join

- Queries che coinvolgono valori correlati in due tabelle diverse
- Inner join
  - Esempio precedente
- outer join

Risposte multiple

- **Le risposte sono bags**

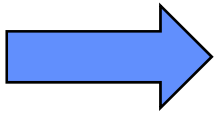
```
SELECT beer
FROM Sells
```



beer
Bud
Miller
Bud
…

- **Possiamo comunque utilizzare la parola chiave DISTINCT**

```
SELECT DISTINCT beer
FROM Sells
```



beer
Bud
Miller
…

Unioni di queries

- **Descrivi i prezzi maggiori di 100 come “alti”, tutti gli altri come “bassi”**

```
        (SELECT bar, beer,
                ‘high’ AS price
        FROM Sells
        WHERE price > 100)
UNION
        (SELECT bar, beer,
                ‘low’ AS price
        FROM Sells
        WHERE price < 100)
```

Subqueries

- **I risultati possono essere annidati**

```
SELECT *
FROM beers
WHERE price in
      (
       SELECT beer
       FROM Likes
       WHERE drinker = 'Fred'
      )
```

Aggregati

- **Trova il prezzo medio della “Bud”**

```
SELECT AVG(price)
FROM Sells
WHERE beer = 'Bud'
```

- **Contiamo ogni tupla contenente ‘Bud’ esattamente una volta**

```
SELECT COUNT(DISTINCT price)
FROM Sells
WHERE beer = 'Bud'
```

Raggruppamenti

- **Possiamo aggiungere in fondo al costrutto la parola chiave GROUP BY e una lista di attributi**
- **La relazione risultante dalle clausole FROM e WHERE é raggruppata in accordo ai valori di questi attributi**
- **Le aggregazioni vengono effettuate solo all'interno di ogni gruppo**
- **Trova il prezzo medio di ogni birra**

```
SELECT beer, AVG(price)
FROM Sells
GROUP BY beer
```

Raggruppamenti

- **Trova, per ogni bevitore, il prezzo medio della “Bud” nei bar che frequenta**

```
SELECT drinker, AVG(price)
FROM Frequents, Sells
WHERE beer = 'Bud' AND
      Frequents.bar = Sells.bar
GROUP BY drinker
```

raggruppamenti

- **La clausola HAVING permette di specificare condizioni sui gruppi generati**
- **trova il prezzo medio delle birre servite in almeno 3 bar o fabbricate da Anheuser-Busch.**

```
SELECT beer, AVG(price)
FROM Sells
GROUP BY beer
HAVING COUNT(*) >= 3 OR
      beer IN (
            SELECT name
            FROM Beers
            WHERE manf = 'Anheuser-Busch')
```

E’ sempre necessario SQL?

- I moderni tools raggruppano una serie di operazioni in maniera uniforme
- La metafora di interazione è visuale
  - Ne vedremo una in particolare
    - Weka
- SQL è più generico
  - Ma anche più difficile da usare

Outline del Modulo

- Introduzione e Concetti di Base
- Data Selection
- Information Gathering
- Data cleaning
- Data reduction
- Data transformation

Concetti, Proprietà,Misurazioni

- Il mondo reale consiste di Concetti
  - Automobili, Vigili, Norme, …
  - Nel nostro caso, ciò che deve essere appreso
- Ad ogni concetto è associabile un insieme di proprietà (features)
  - Colore, Cilindrata, Proprietario, …
- Su ogni proprietà è possibile stabilire delle misurazioni
  - Colore = rosso, Cilindrata = 50cc, Proprietario = luigi, …

La Nostra Modellazione

- La realtà di interesse è descritta da un insieme di istanze, raccolte in una tabella
- Le tuple (istanze) della tabella sono i concetti che vogliamo studiare
  - Esempi di concetti
- Le colonne (attributi) della tabella rappresentano le caratteristiche degli oggetti che vogliamo studiare
- Una variabile è un contenitore per una misurazione di una caratteristica particolare di un oggetto
  - A volte utilizzata anche per indicare un attributo

Cos'è un esempio?

- Istanza: esempio di concetto
  - La cosa da classificare, clusterizzare, associare
  - Un esempio individuale e indipendente di concetto target
  - Caratterizzato da un insieme predeterminato di attributi
- Input ad uno schema di learning: un insieme di istanze (dataset)
  - Rappresentato da una singola relazione/tabella
- Rappresentazione piuttosto povera
  - Non si possono esprimere relazioni tra oggetti
- La più tipica nel data mining

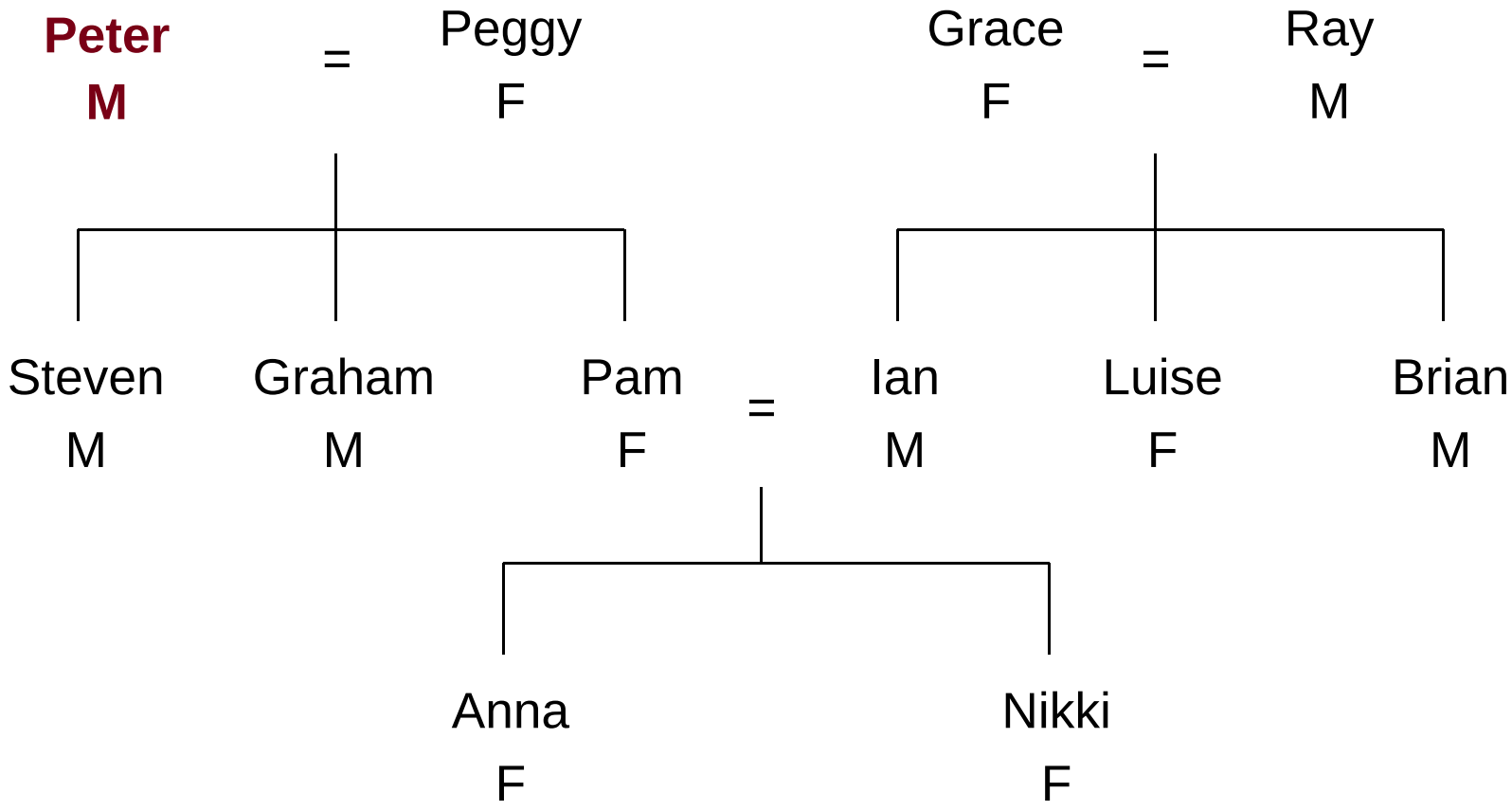
Istanze, attributi, misurazioni

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

Istanze

Un albero genealogico



L’albero genealogico rappresentato in forma tabellare

Name	Gender	Parent1	parent2
Peter	Male	?	?
Peggy	Female	?	?
Steven	Male	Peter	Peggy
Graham	Male	Peter	Peggy
Pam	Female	Peter	Peggy
Ian	Male	Grace	Ray
Luise	Female	Grace	Ray
Brian	Male	Grace	Ray
Anna	Female	Pam	Ian
Nikki	Female	Pam	Ian

La relazione “sister-of”

First person	Second person	Sister of?
Peter	Peggy	No
Peter	Steven	No
…	…	…
Steven	Peter	No
Steven	Graham	No
Steven	Pam	Yes
…	…	…
Ian	Pippa	Yes
…	…	…
Anna	Nikki	Yes
…	…	…
Nikki	Anna	yes

First person	Second person	Sister of?
Steven	Pam	Yes
Graham	Pam	Yes
Ian	Pippa	Yes
Brian	Pippa	Yes
Anna	Nikki	Yes
Nikki	Anna	Yes
All the rest		No

Assunzione di Mondo Chiuso

Una rappresentazione completa

First person				Second person				Sister of?
Name	Gender	Parent1	Parent2	Name	Gender	Parent1	Parent2	
Steven	Male	Peter	Peggy	Pam	Female	Peter	Peggy	Yes
Graham	Male	Peter	Peggy	Pam	Female	Peter	Peggy	Yes
Ian	Male	Grace	Ray	Pippa	Female	Grace	Ray	Yes
Brian	Male	Grace	Ray	Pippa	Female	Grace	Ray	Yes
Anna	Female	Pam	Ian	Nikki	Female	Pam	Ian	Yes
Nikki	Female	Pam	Ian	Anna	Female	Pam	Ian	Yes
All the rest								No

```
If second person's gender = female
   and first person's parent = second person's parent
   then sister-of = yes
```

Generazione di un file “piatto”

- Chiamato anche processo di “denormalizzazione”
  - Molte relazioni sono messe in join per costruirne un’unica
- Problematica: relazioni senza un numero di oggetti predefinito
- La denormalizzazione può produrre regolarità spurie (dipendenze funzionali)
  - Esempio: “fornitore” predice “Indirizzo fornitore”

Tipi di Misurazioni

- **Nel concreto, i valori che una variabile può assumere**
  - **La nozione di Tipo nelle basi di Dati**
- **Scala delle misurazioni**
  - **Variabili scalari**
- **Insieme di caratteristiche scalari**
  - **Più componenti**
  - **Velocità: km/h + direzione**
  - **Tempo: Ore + Minuti + Secondi**

Misure su Scale Nominali

- **Misurazioni qualitative**
- **Variabili Nominali**
  - **Un'etichetta da associare per l'identificazione**
    - **L'attributo Nome può assumere valori: giuseppe, antonio, luigi, …**
    - **Il codice fiscale di ogni persona è unico**
- **Variabili Categoriche**
  - **Per riferirsi a gruppi di cose**
  - **La misura può essere condiviso da più oggetti**
    - **Colore: rosso, giallo, blu**
  - **Anche un numero può essere una variabile categorica!**
    - **Il Codice di avviamento postale è condiviso da tutte le persone appartenenti alla stessa città**
- **Non è associata una nozione di ordine**
- **Non c'è un criterio per stabilire una differenza (quantitativa) tra gli elementi**

Misure su scale Ordinali

- Possiamo imporre un ordine sul range di valori possibili
  - Graduatoria: 1°, 2°, 3°
- Non necessariamente implicano valori numerici
  - Grado: Colonnello, Tenente, Sergente
- Non hanno associata una distanza
  - Operazioni algebriche non hanno senso

Misure su scale numeriche

- Esprimono misure quantitative
- Misurazioni Basate su intervalli
  - I valori provengono da un range continuo
  - E.g.: Temperatura,
  - Non c'è (apparente) correlazione tra i valori
- Misurazioni Ratio-Scaled
  - Le misure esprimono proprietà
    - La quantità di denaro in una macchina per il caffè è un multiplo dell'unità minima che si può inserire
    - Una misurazione fisica che dipende dal tempo t (in secondi): $e^{-\delta t}$

Variabili Binarie

- **Simmetriche (dicotomiche)**
  - **Sesso: Maschio/Femmina**
- **Asimmetriche**
  - **Responso: SI/NO, Vero/Valso, 1/0**

Riassumendo

- **Variabili Discrete (simboliche)**
  - **Solo test di uguaglianza**
  - **Nominali**
  - **Categoriche**
  - **Ordinali**
  - **Binarie**
- **Variabili Continue**
  - **Interval-Based (valori interi)**
  - **Ratio-Scaled (valori reali)**

Perché c’è bisogno di specificare i tipi?

- *D: Perché gli algoritmi di learning devono sapere il tipo degli attributi?*
- **R: per poter effettuare i confronti appropriati, e apprendere i concetti significativi**
  - Outlook > “sunny” **non ha senso, mentre**
  - Temperature > “cool” **oppure**
  - Humidity > 70 **ha senso**

Le proprietà dei dati

- **Il tipo di un attributo dipende da quali proprietà possiede:**
  - **distinguibilità: = ≠**
  - **Ordine: < >**
  - **Additività: + -**
  - **Moltiplicabilità: * /**

  - **Attributi nominali: distinguibilità**
  - **Attributi ordinali: distinguibilità, ordine**
  - **intervalli: distinguibilità, ordine, additività**
  - **Ratio: tutte le proprietà**

Utilizzare Variabili

- **Sparsità**
  - **Mancanza di valore associato ad una variabile**
    - **Un attributo è sparso se contiene molti valori nulli**
- **Monotonicità**
  - **Crescita continua dei valori di una variabile**
    - **Intervallo [-∞, ∞] (o simili)**
  - **Non ha senso considerare l'intero intervallo**
- **Outliers**
  - **Valori singoli o con frequenza estremamente bassa**
  - **Possono distorcere le informazioni sui dati**
- **Dimensionalità**
  - **Il numero di valori che una variabile può assumere può essere estremamente alto**
    - **Tipicamente riguarda valori categorici**
- **Anacronismo**
  - **Una variabile può essere contingente: abbiamo i valori in una sola porzione dei dati**

L’Influenza (bias)

- **Un fattore esterno significativo e rilevante nei dati**
  - **Comporta problemi (espliciti o impliciti) nei dati**
  - **Il valore della variabile `Velocità` in una tabella `Infrazioni` è alto**
- **Il problema è sistematico**
  - **Appare con una certa persistenza**
    - **Il misuratore della velocità è tarato male**
- **Il problema può essere trattato**
  - **Il valore è suscettibile di una distorsione, che deve essere considerata**
    - **Considera solo i valori che vanno oltre una certa tolleranza**

Comprensione dei dati
Quantità

- **Numero di istanze**
  - ***Regola empirica: 5,000 o più***
  - **Se sono di meno, i risultati sono meno affidabili**
- **Numero di attributi**
  - ***Regola pratica: per ogni campo,, 10 (o più) istanze***
  - **Se ci sono più campi, si deve utilizzare riduzione e selezione di dimensionalità**
- **Numero di esempi (nella classificazione)**
  - ***Regola pratica: >100 per ogni concetto***
  - **Se i dati sono sbilanciati, si può (deve) utilizzare il campionamento stratificato**

Esistono altri tipi di dati?

- Si
  - Testo
  - Grafi
  - Dati spazio-temporali
- In genere, molti di questi possono essere riportati nel formato descritto in precedenza
  - Non è vero (o conveniente) in generale

Document Data

- **Ogni documento diventa un vettore di termini,**
  - **Un termine è un attributo (componente) del vettore**
  - **Il valore di ogni componente è la frequenza del termine 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

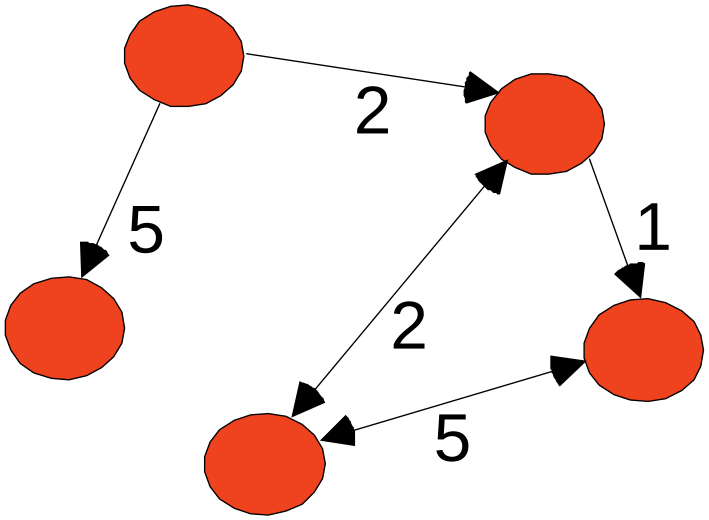
Dati transazionali

- **Coinvolge insiemi**
  - **Si può trasformare nel formato tabellare**

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

Grafi

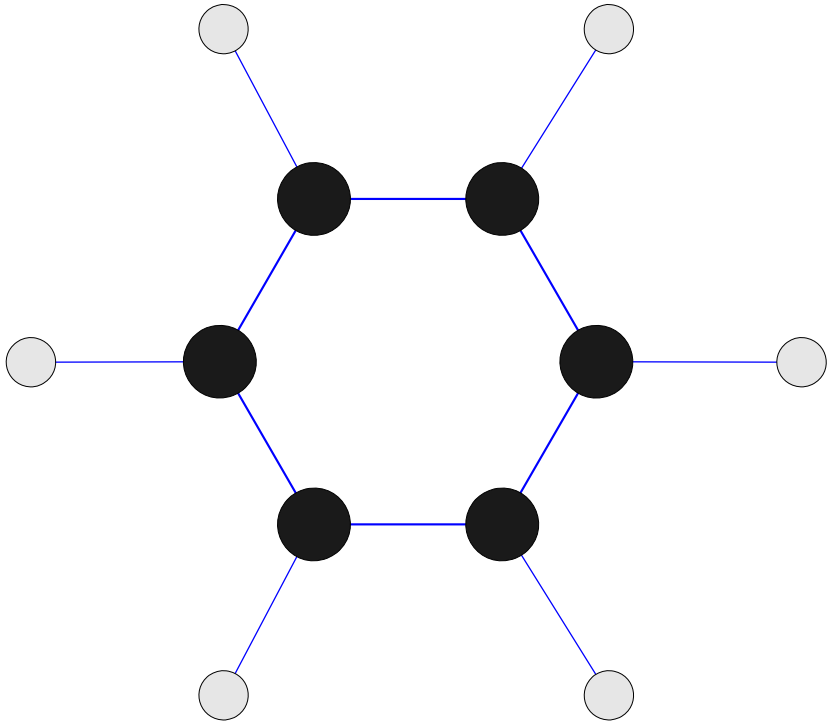
- **Grafo dei links HTML**




```
<a href="papers/papers.html#bbbb">
Data Mining </a>
<li>
<a href="papers/papers.html#aaaa">
Graph Partitioning </a>
<li>
<a href="papers/papers.html#aaaa">
Parallel Solution of Sparse Linear System of Equations </a>
<li>
<a href="papers/papers.html#ffff">
N-Body Computation and Dense Linear System Solvers
```

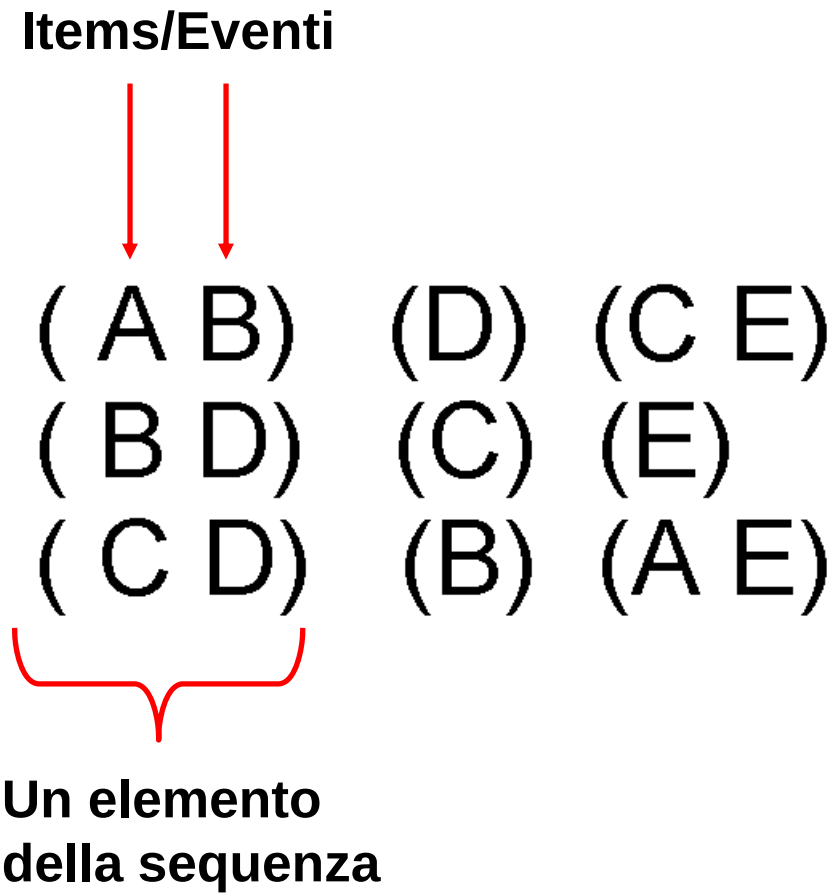
Dati chimici

- **Molecola del benzene: C_6H_6**



Dati ordinati

- **Sequenze di transazioni**



Dati ordinati

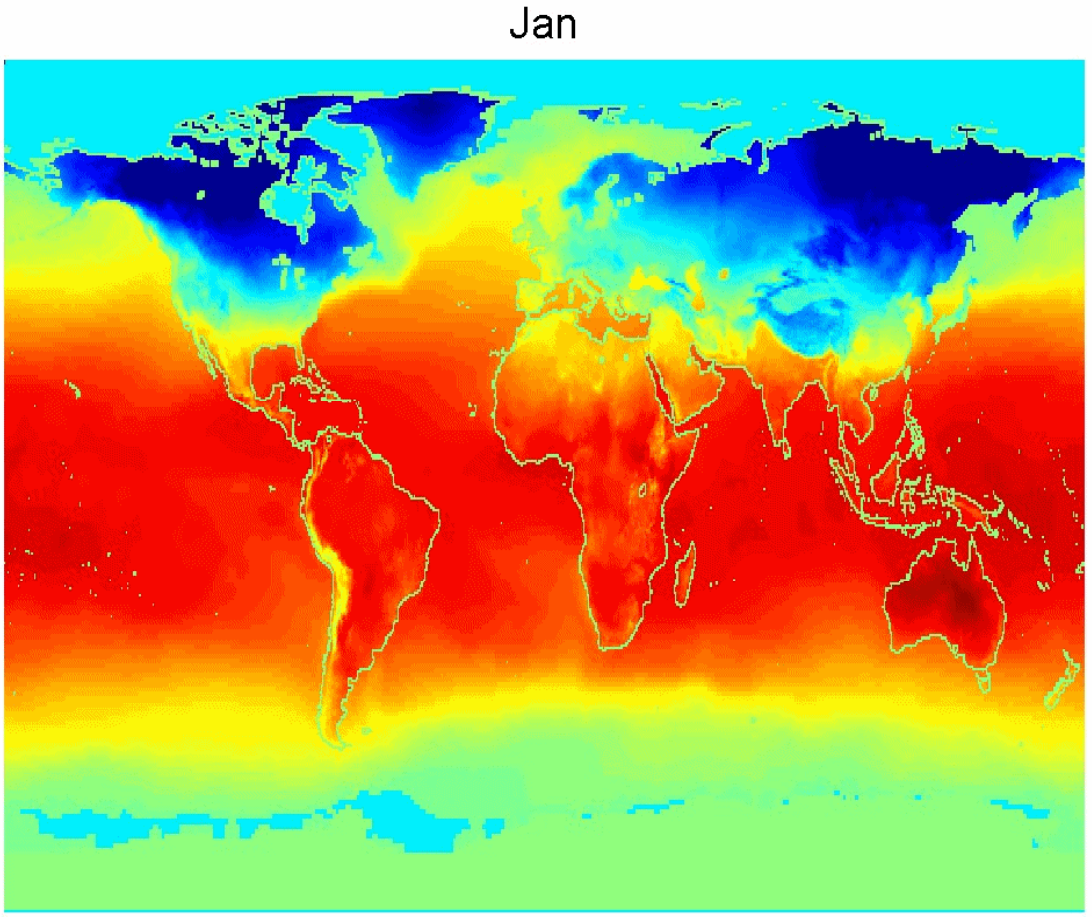
- **Sequenze genomiche**

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

Dati ordinati

- **Dati spazio-temporali**

Temperatura media mensile



Analisi esplorativa dei dati

- **Due approcci:**
  - **Parametrica**
    - **Conosco la distribuzione che regola il mio campione, ma non ne conosco i parametri**
      - Stimo i parametri
  - **Non parametrica**
    - **Non conosco la distribuzione**
      - Cerco di capire qual'è la distribuzione e quali sono i suoi parametri

Misure descrittive dei dati

- **Distribuzioni,frequenze**
  - **offre una lettura rapida delle caratteristiche più importanti dei di dati**
  - **Media, varianza, deviazione standard**
  - **Tendenze**
- **variabilità, dispersione**
  - **mediana, quartili**
  - **forma**
  - **simmetria**
  - **curtosi**

Visualizzazione

- **Conversione dei dati in formato visuale/tabellare**
  - **Utile per analizzare le caratteristiche tra i dati e le relazioni tra i valori/attributi**
- **Strumento estremamente potente**
  - **Si possono analizzare datasets di grosse dimensioni**
  - **Si possono determinare tendenze e statistiche generali**
  - **Si possono determinare outliers/patterns inusuali**

Dati qualitativi

- **Valori mutuamente esclusivi,descrizione esaustiva**
- **Distribuzione della Frequenza**

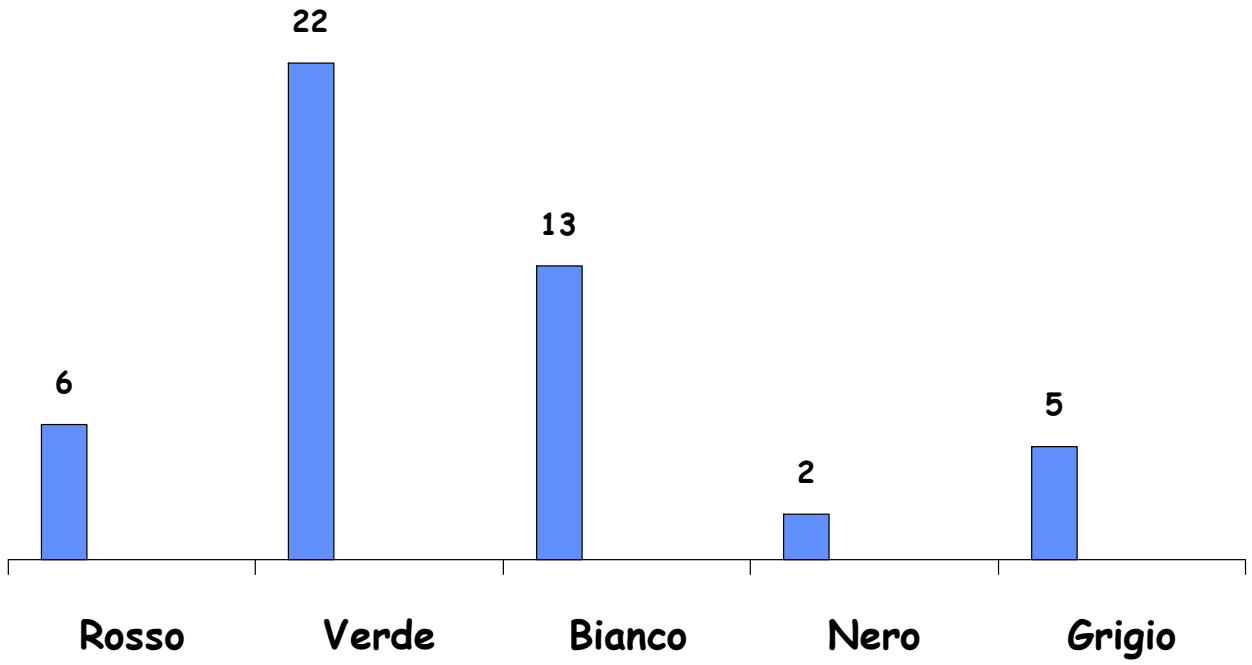
Colore	***Frequenza***		
	Assoluta	*Relativa*	*Cumulativa*
Rosso	6	0,125	0.125
Verde	22	0,458	0.583
Bianco	13	0,271	0.854
Nero	2	0,042	0.896
Grigio	5	0,104	1.000
Totale	48	1,000	1.000

Visualizzazione

- Diagrammi a barre
- Dot Diagrams
- Stem and leaf
- Box Plots
- Studi di distribuzioni

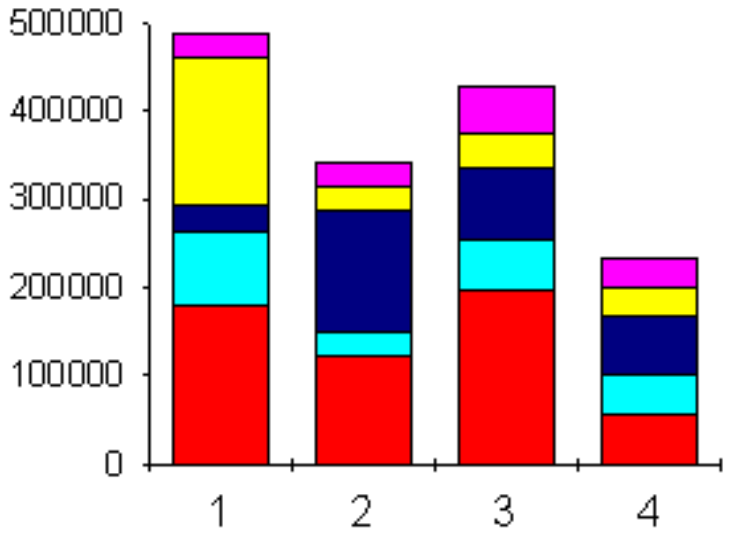
Diagrammi di Pareto

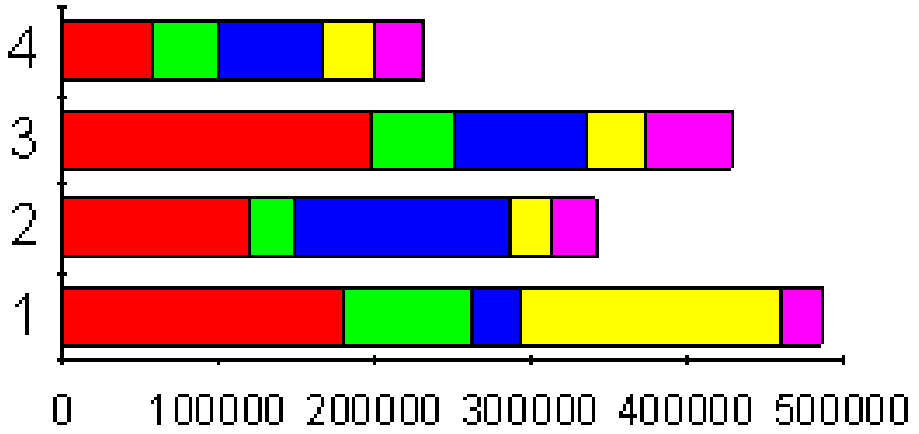
- Diagammi a barre distanziate
- Un assortimento di eventi presenta pochi picchi e molti elementi comuni



Ortogrammi

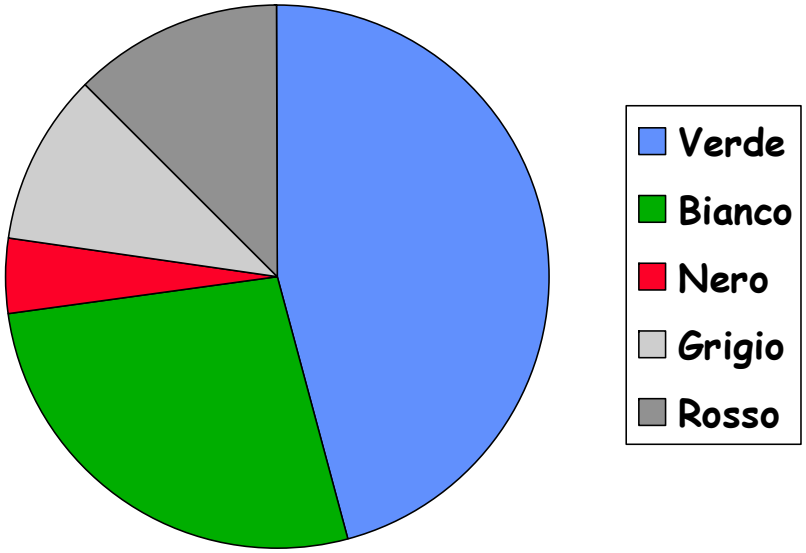
- **Ogni colonna indica la la distribuzione interna per un dato valore e la frequenza**





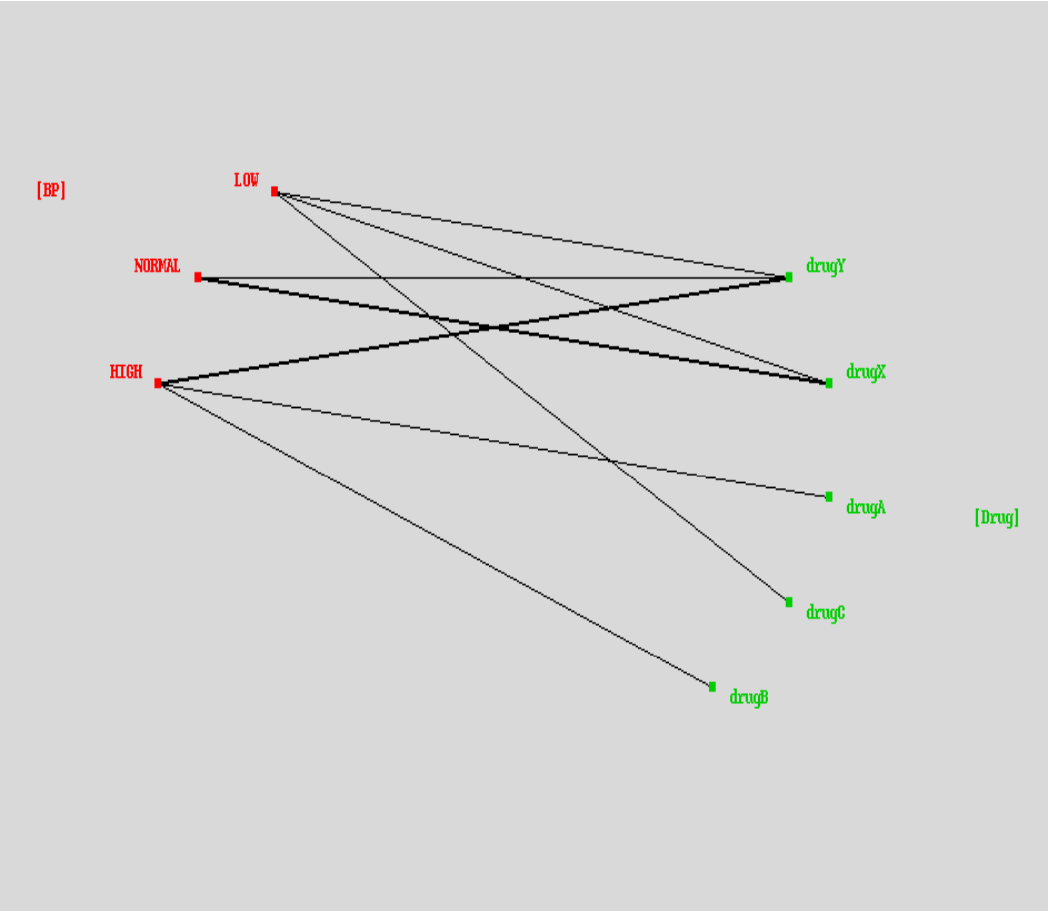
Aerogrammi

- **Rappresentazioni a torta**
- **frequenza della distribuzioni**

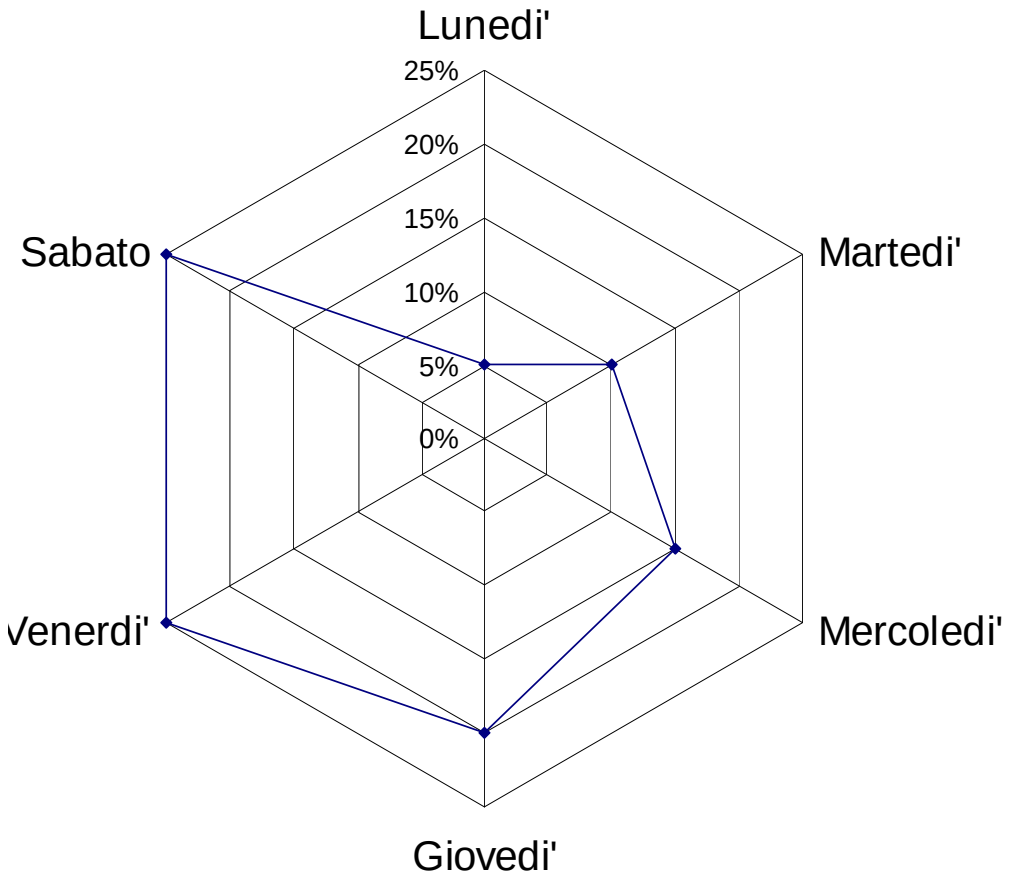


Web

- **Visualizzano correlazioni tra valori simbolici**



Diagrammi polari



- **Rappresentano fenomeni ciclici**
  - **E.g., concentrazione delle vendite nell'arco settimanale**

Dati Quantitativi

- **Istogrammi**
- **Poligoni**
- **Diagrammi cartesiani**
- **Diagrammi quantili**

Esempio: Iris Data

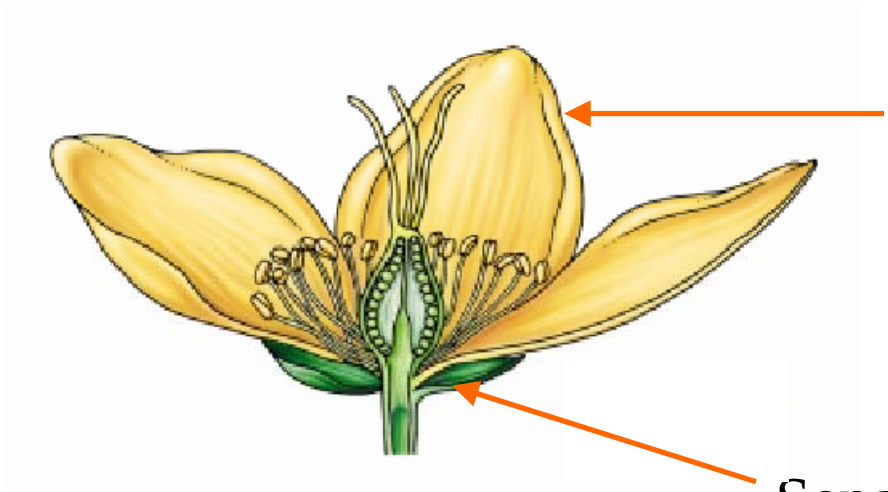
Iris setosa

sepal length	sepal width	petal length	petal width
5.1	3.5	1.4	0.2
4.9	3	1.4	0.2
...	...	...	...
5.9	3	5.1	1.8

Iris versicolor

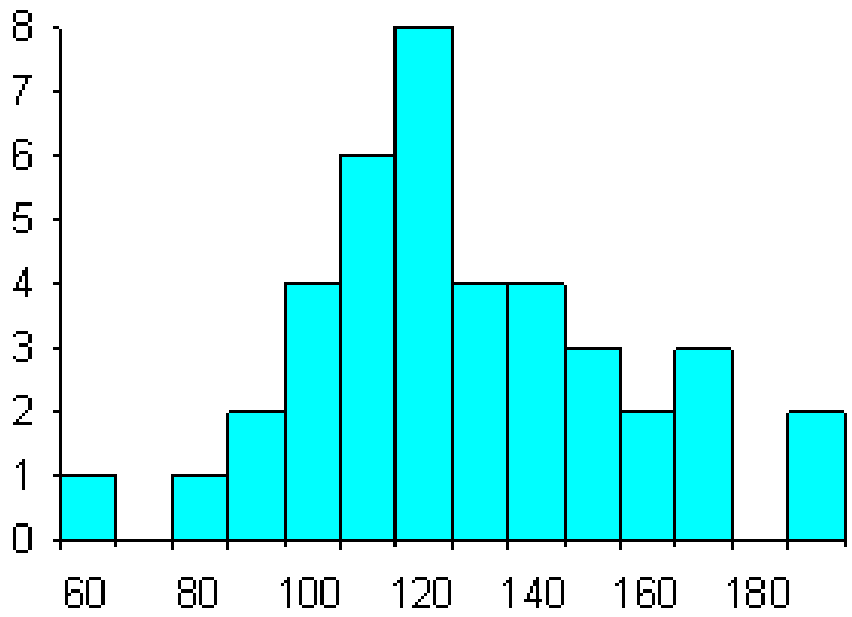
Iris virginica

Parti del fiore



- **Dataset standard**
  - **UCI Machine Learning Repository**
    http://www.ics.uci.edu/~mlearn/MLRepository.html
  - **Offerto da Douglas Fisher**
  - **Tre tipi di fiori (classi):**
    - **Setosa**
    - **Virginica**
    - **Versicolour**
  - **Quattro attributi**
    - **Sepal width/length**
    - **Petal width/length**

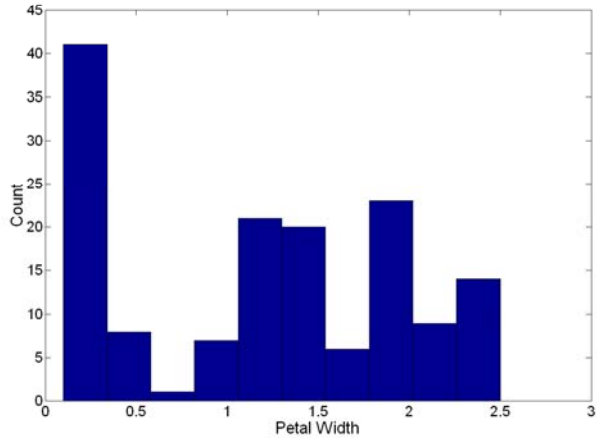
Istogrammi



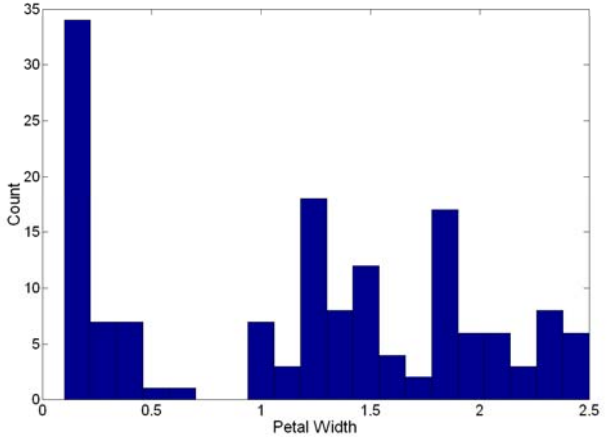
- Rappresentazioni a barre
- Evidenziano la frequenza su intervalli adiacenti
  - La larghezza di ogni rettangolo misura l'ampiezza degli intervalli
  - Quale larghezza?
- Utili per determinare
  - Il centro dei dati
  - La variabilità e la dispersione
  - I picchi
  - La presenza di outliers
  - La presenza di valori modali multipli

Istogramma di Petal Width

- **10 bins**

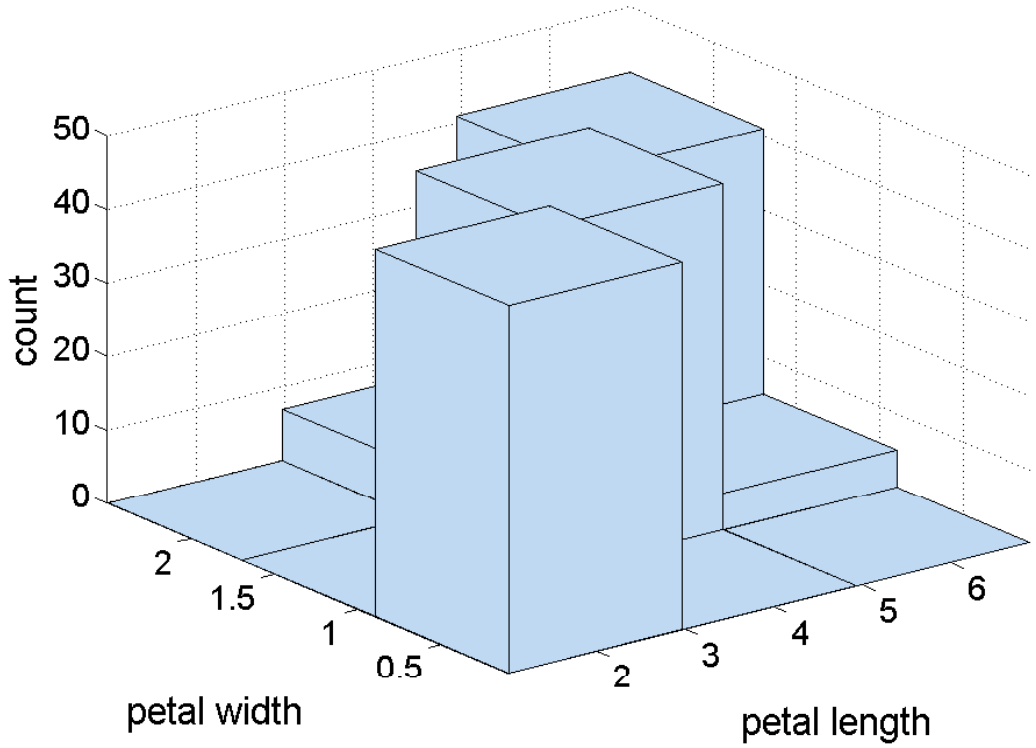


- **20 bins**



Istogrammi a due dimensioni

- **Mostrano la distribuzione congiunta di due attributi**
- **Esempio: petal width e petal length**
  - **Ci indica qualcosa?**



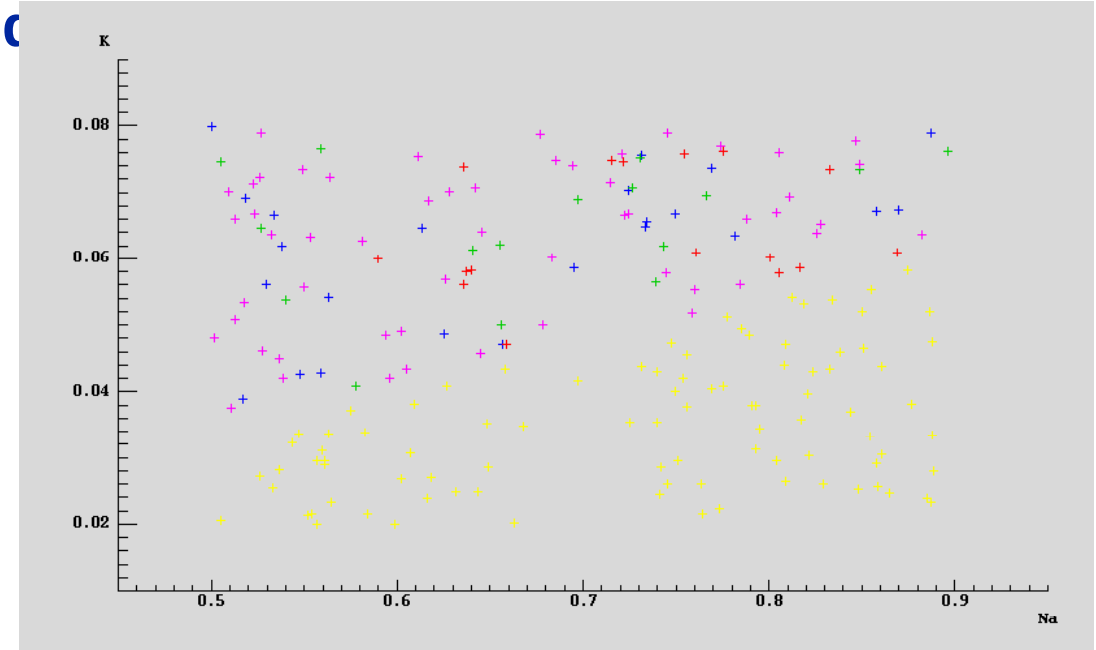
Rappresentazione “Stem & Leaf”

- **Simile a istogrammi**
- **Per evitare perdita di informazione**
- **Utile per pochi dati**

10-19	2 7 5
20-29	9 19 5 3 4 7 1 8
30-39	4 9 2 4 7
40-49	4 8 2
50-59	3

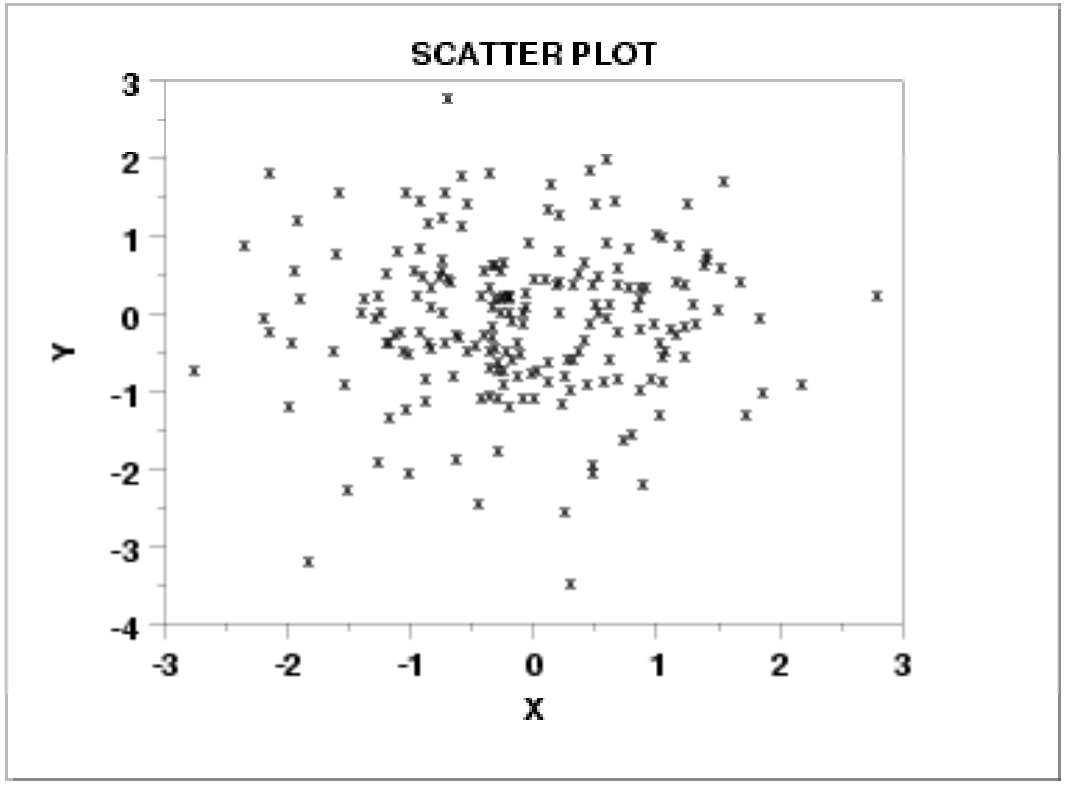
Dot Diagrams, Scatters

- **Visualizza la Dispersione**
  - **Esiste una correlazione tra X e Y?**
  - **Esiste una correlazione lineare/nonlineare?**
  - **Come varia la densità d in funzione di X?**
  - **Ci sono outliers?**



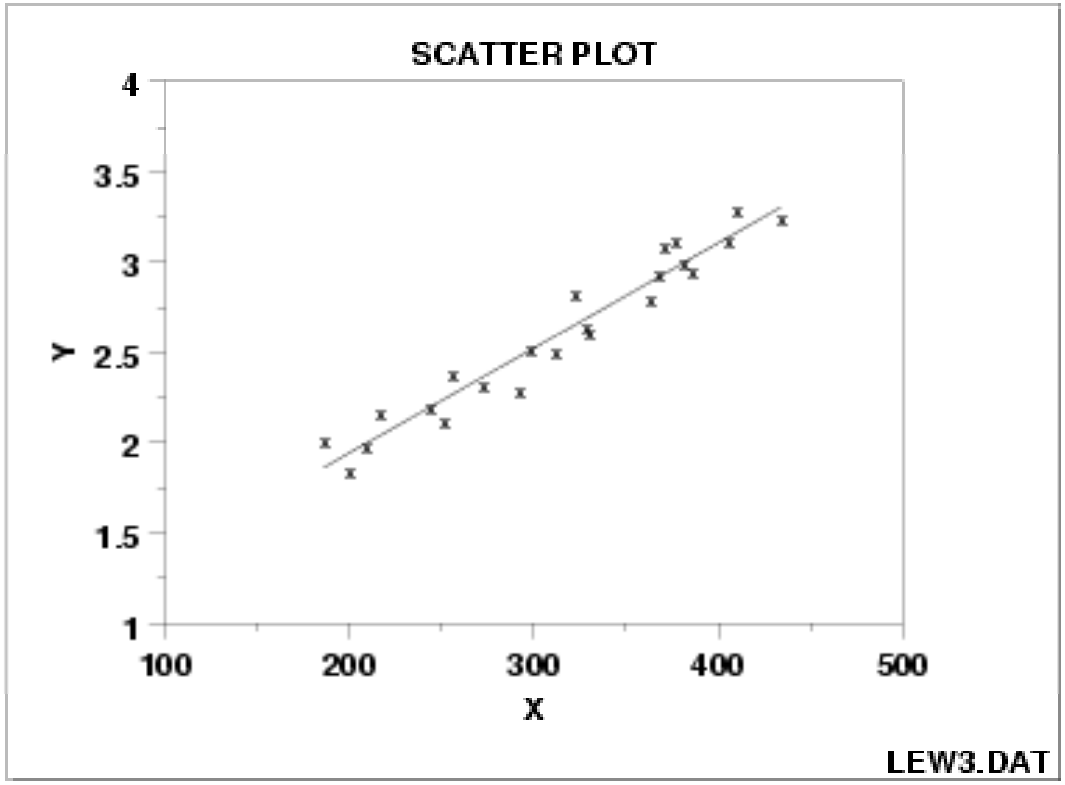
Scatter plots

- **Nessuna correlazione**



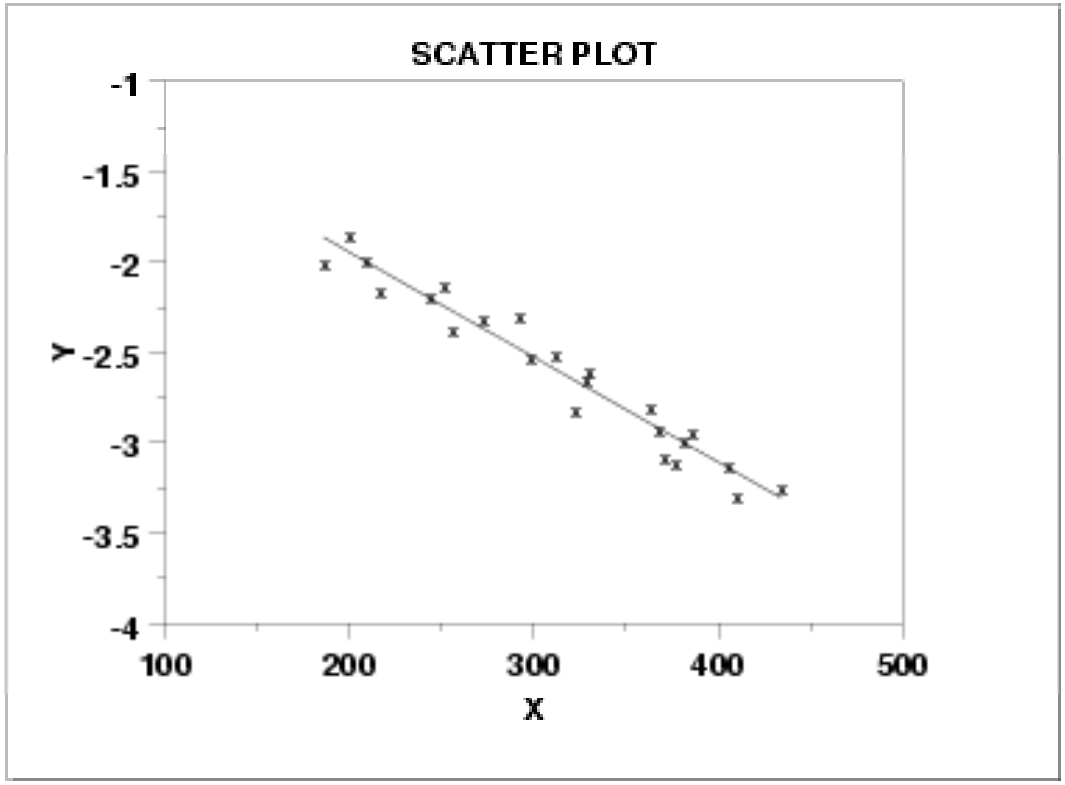
Scatter plots

- **Correlazione lineare positiva**



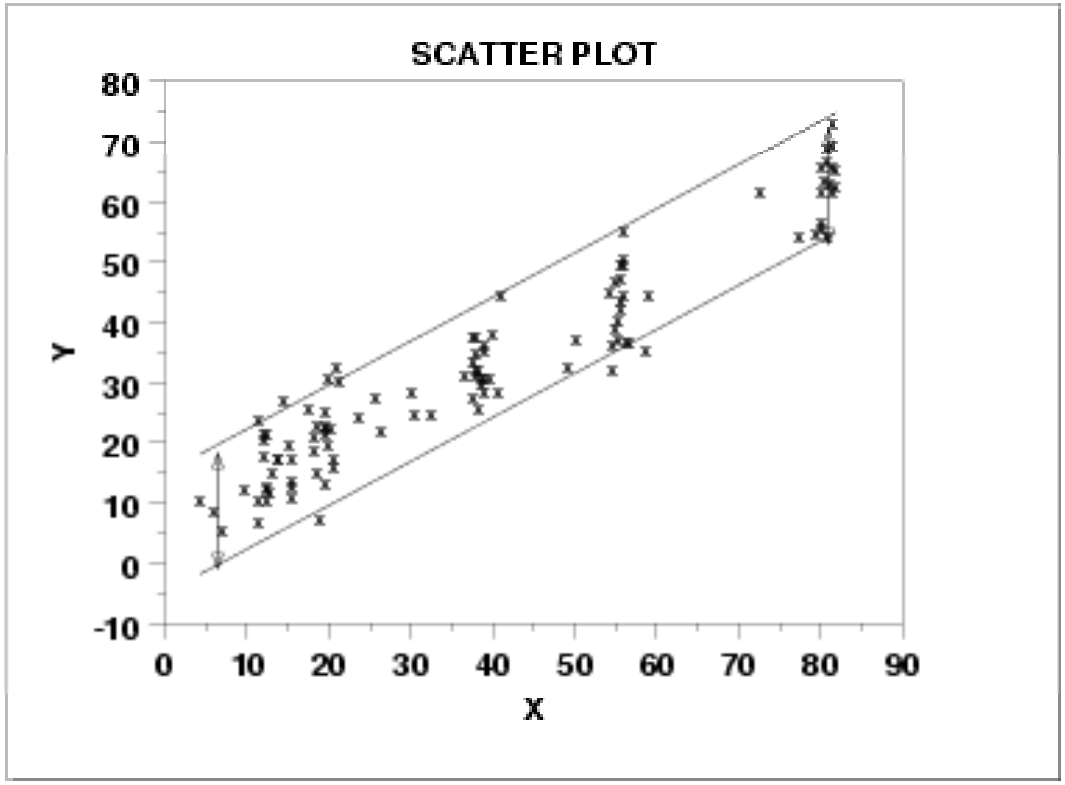
Scatter plots

- **Correlazione lineare negativa**



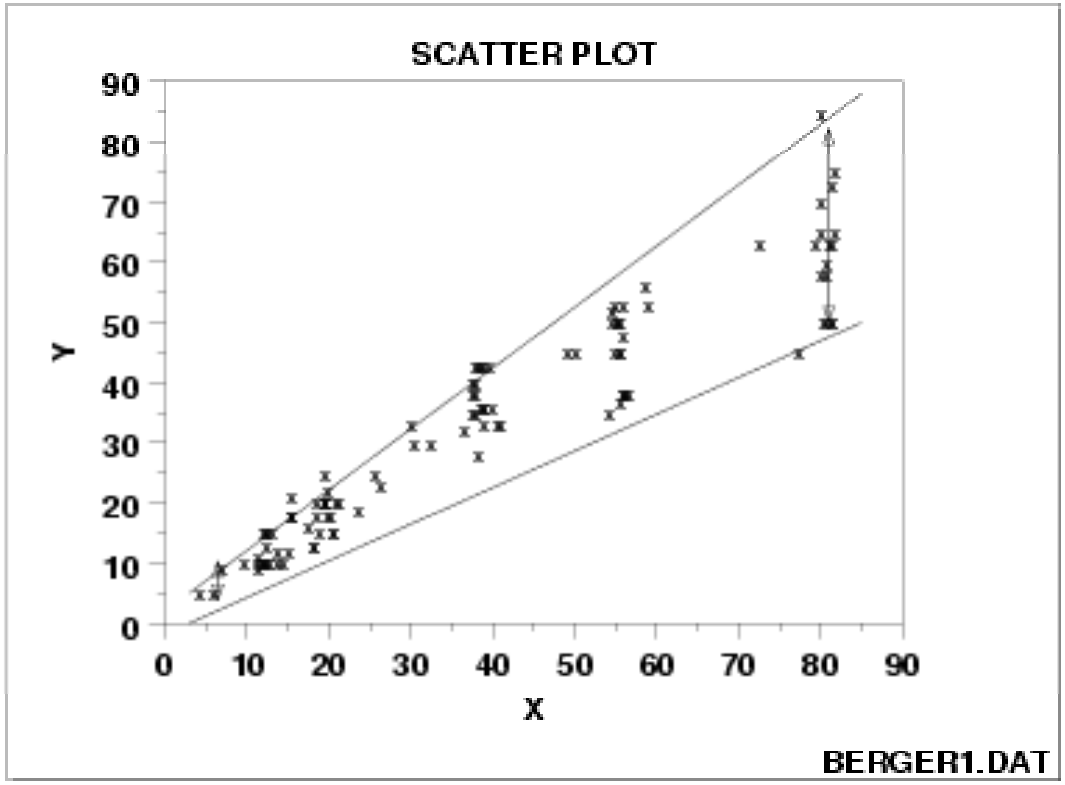
Scatter plots

- **La variabilità di Y non dipende da X**

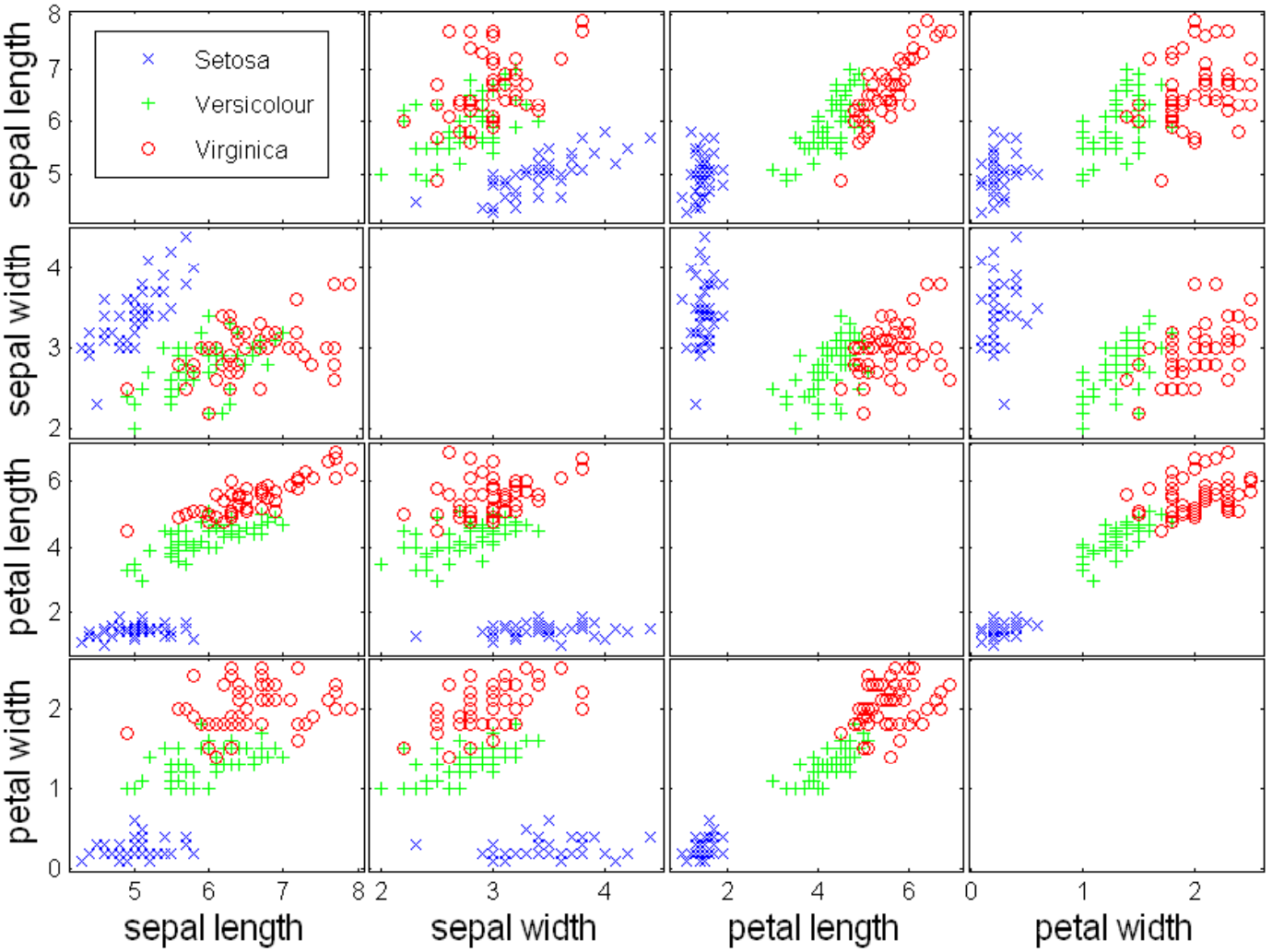


Scatter plots

- **La variabilità di Y dipende da X**



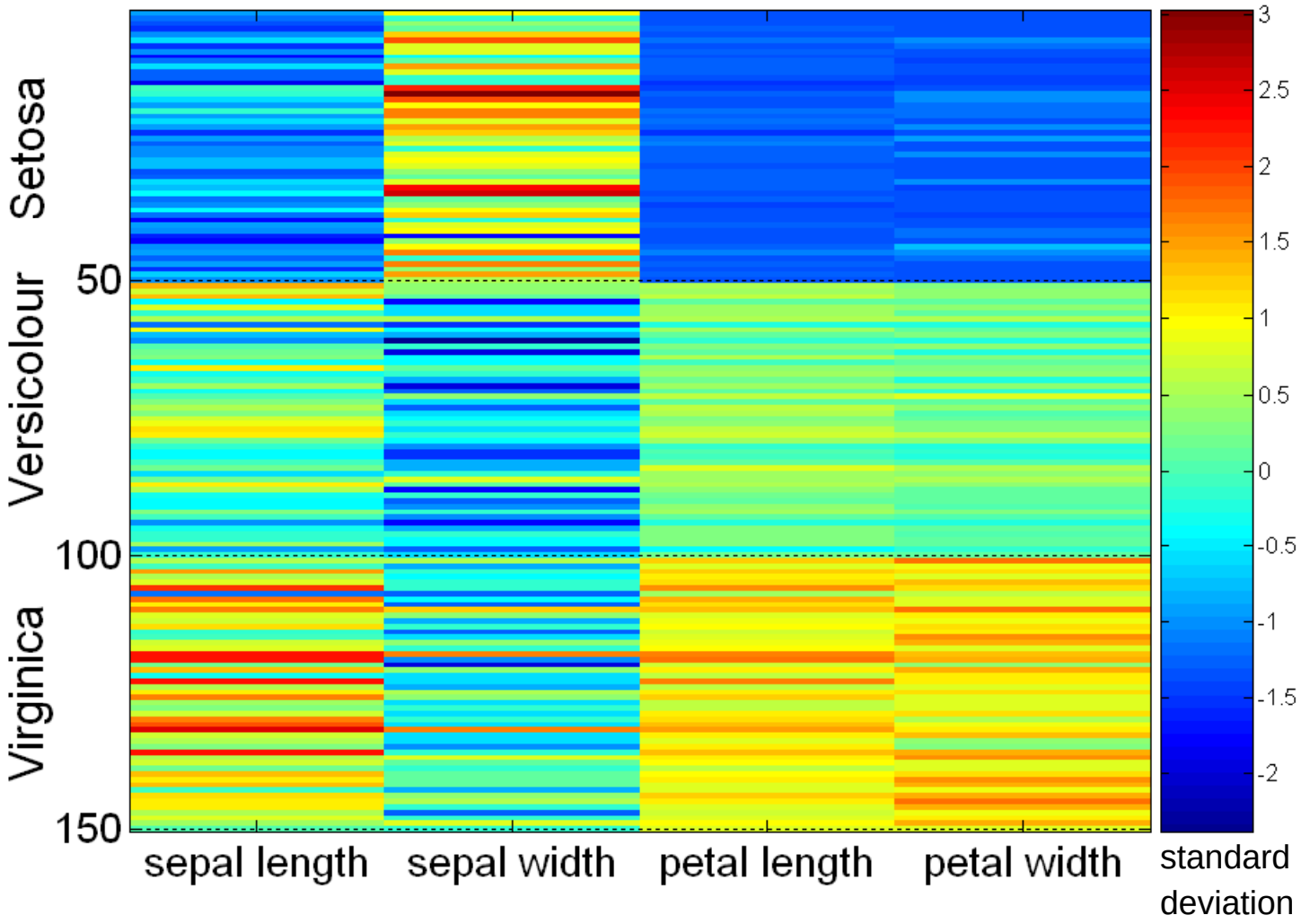
Scatter Plot di Iris



Matrix Plots

- **Per visualizzare la matrice**
  - **Utile quando è possibile definire un ordinamento nei dati**
    - **Molto utile per visualizzare relazioni tra dati**
      - Matrici di similarità
  - **Necessita normalizzazione**

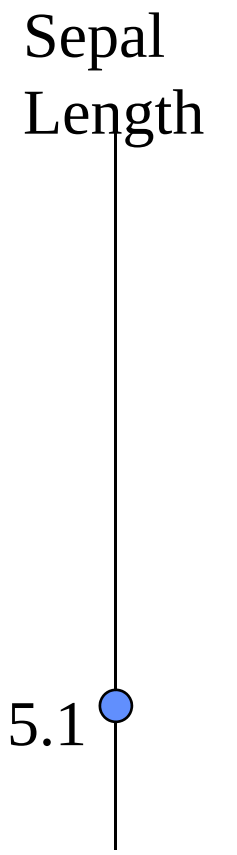
Iris Data


Setosa
Versicolour
Virginica
50
100
150
sepal length
sepal width
petal length
petal width
3
2.5
2
1.5
1
0.5
0
-0.5
-1
-1.5
-2
standard deviation

Coordinate parallele

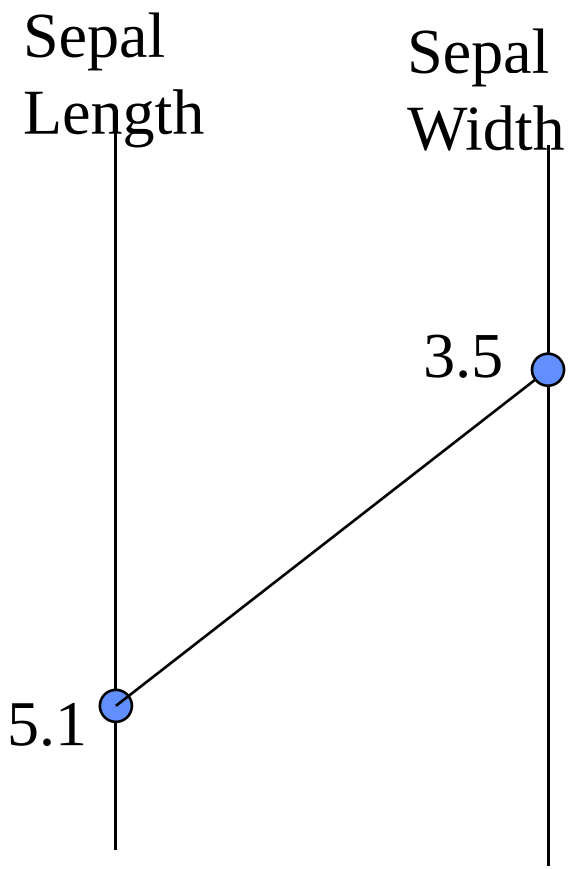
- **Per rappresentare relazioni con dati ad alta dimensionalità**
  - **Un asse per ogni attributo**
  - **I valori sugli assi**
- **Un oggetto è rappresentato come una linea**

Coordinate Parallele



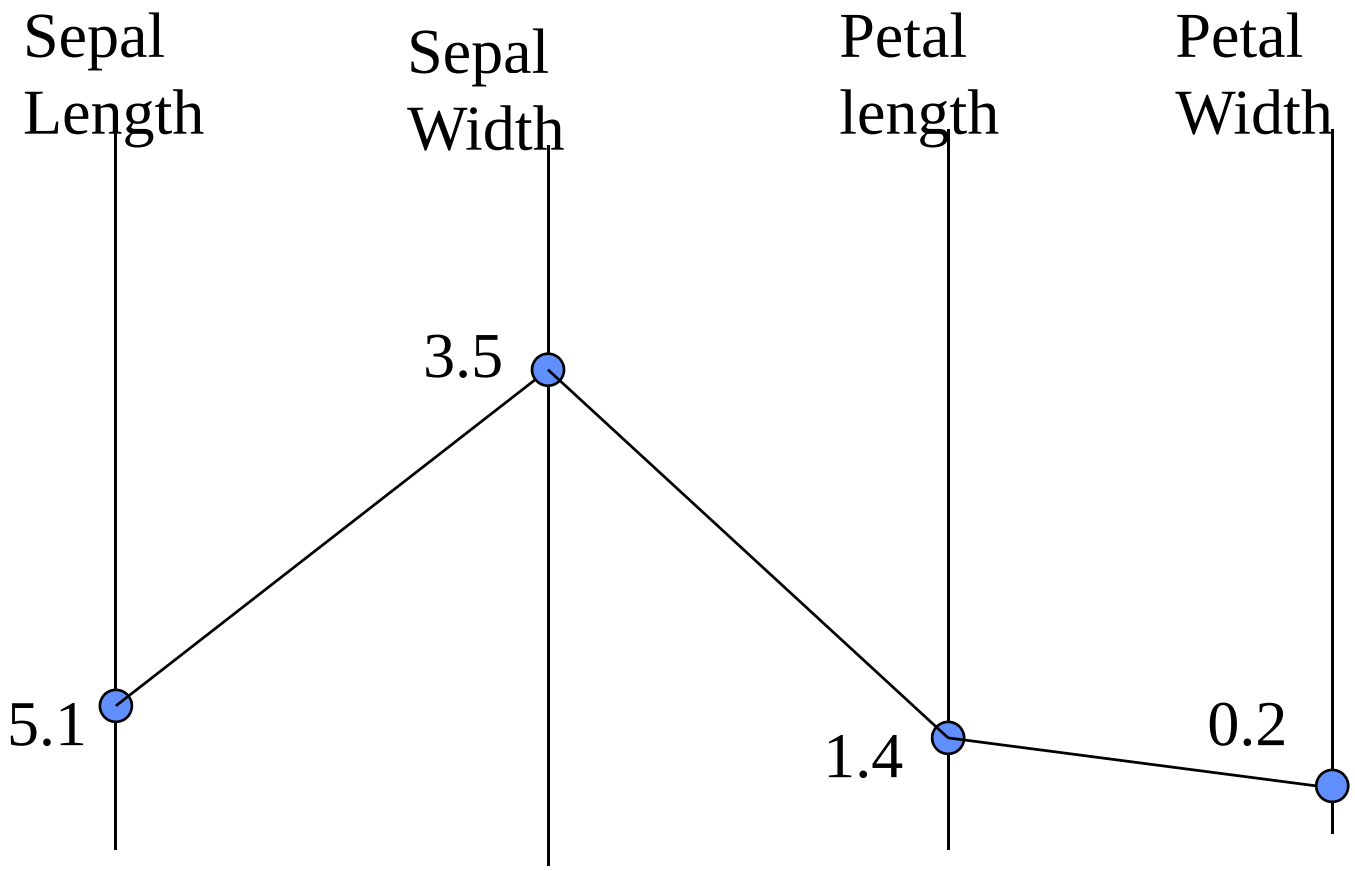
sepal length	sepal width	petal length	petal width
5.1	3.5	1.4	0.2

Parallel Coordinates: 2 D



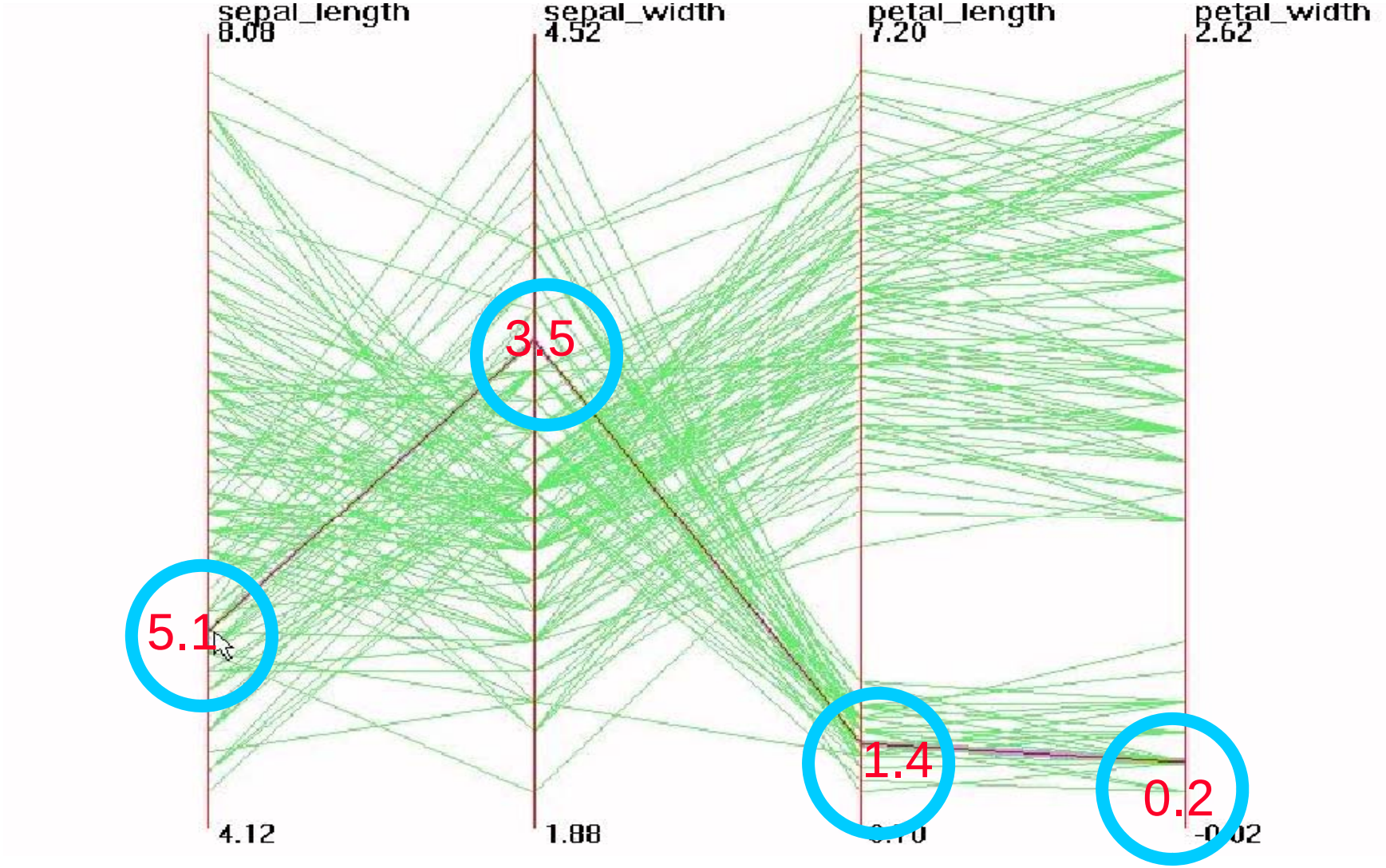
sepal length	sepal width	petal length	petal width
5.1	3.5	1.4	0.2

Parallel Coordinates: 4 D



sepal length	sepal width	petal length	petal width
5.1	3.5	1.4	0.2

Visualization di Iris data



Misure descrittive dei dati

- **Tendenza centrale o posizione**
  - **per individuare il valore intorno al quale i dati sono raggruppati**
- **dispersione o variabilita`**
  - **per definire la forma più o meno raccolta della distribuzione**
- **forma**
  - **simmetria, curtosi**

Media Aritmetica

- **Per effettuare la correzione di errori accidentali**
  - **permette di sostituire i valori di ogni elemento senza cambiare il totale**
    - **Sostituzione di valori NULL**
- **Monotona crescente**

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

$$\frac{1}{n+k}\left(\sum_{i=1}^{n} x_i + k\bar{x}\right) = \bar{x}$$

Media Geometrica

$$x_g = \sqrt[n]{\prod_{i=1}^{n} x_i}$$

- **Per bilanciare proporzioni**
- **dati moltiplicativi**

Prodotto	***Variazioni Prezzi***	
	1996	*1997*
A	100	200
B	100	50
Media	100	125

$$x_g = 100$$

- **La media aritmetica dei logaritmi è il logaritmo della media geometrica**
- **Monotona crescente**

$$\log x_g = \frac{1}{n}\sum_{i=1}^{n} \log x_i$$

Media Armonica

- **Monotona decrescente**
- **Per misure su dimensioni fisiche**
- **E.g., serie temporali**

$$x_a = \frac{n}{\sum_{i=1}^{n} \frac{1}{x_i}}$$

Mediana

- **Il valore centrale in un insieme ordinato di dati**
- **Robusta**
  - **poco influenzata dalla presenza di dati anomali**

1 7 12 23 34 54 68

$$\bar{x} = 21.3$$

$$M = 23$$

Mediana e Quartili

- Divide un insieme di dati a meta`
  - statistica robusta (non influenzata da valori con rilevanti differenze)
  - ulteriori punti di divisione
- interquartili
  - mediane degli intervalli dei dati superiore e inferiore
  - Un quarto dei dati osservati è sopra/sotto il quartile
- percentili
  - di grado p: il p% dei dati osservati è sopra/sotto il percentile
  - mediana: 50-esimo percentile
  - primo quartile: 25-esimo percentile
  - secondo quartile: 75-esimo percentile
- max, min
  - range = max-min

Percentili

- **Rappresentati con x_p**
- **Utilizziamo le lettere per esprimerli**

Etichetta	*P*
M	½= 0.5
F	¼=0.25
E	1/8=.125
D	1/16=0.625
C	1/32=0.3125
B	1/64
A	1/128
Z	1/256
Y	1/512
X	1/1024

Moda

- **Misura della frequenza dei dati**

$$a\ a\ b\ b\ c\ c\ a\ d\ b\ c\ a\ e\ c\ b\ a\ a$$

moda = a ($f = 6$)

- **Significativo per dati categorici**
- **Non risente di picchi**
- **Molto instabile**

Range, Deviazione media

- **Intervallo di variazione**

$$r = max\text{-}min$$

- **Scarti interquantili**

$$r_p = x_{100-p} - x_p$$

- **Scarto medio assoluto**

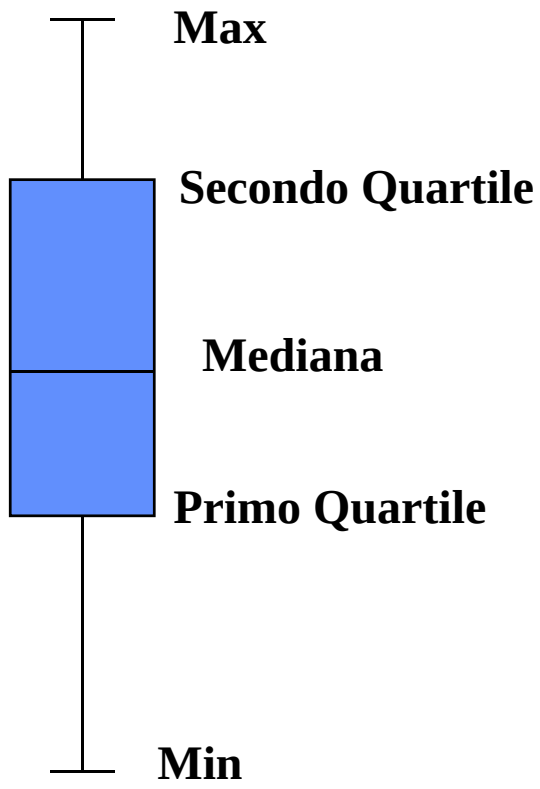
$$S_n = \frac{1}{n}\sum_{i=1}^{n}\left|x_i - \bar{x}\right|$$

- **Scarto medio assoluto dalla mediana**
  - ***In generale,*** $S_{.5} \leq S_n$

$$S_M = \frac{1}{n}\sum_{i=1}^{n}\left|x_i - M\right|$$

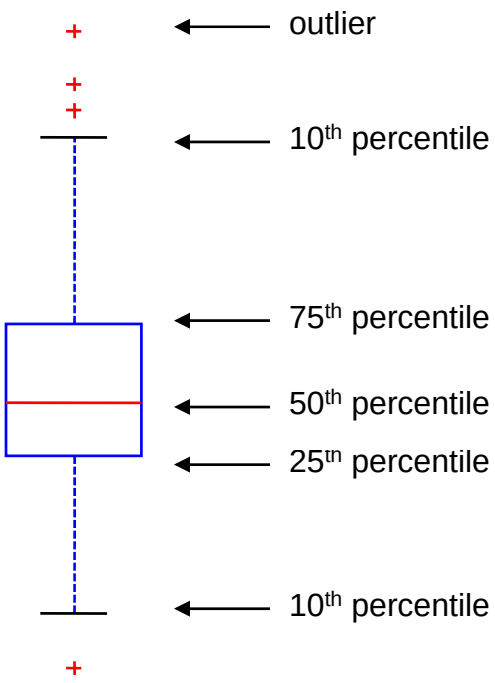
Rappresentazioni Boxplot

- **Rappresentano**
  - **il grado di dispersione o variabilità dei dati (w.r.t. mediana e/o media)**
  - **la simmetria**
  - **la presenza di valori anomali**
- **Le distanze tra i quartili definiscono la dispersione dei dati**



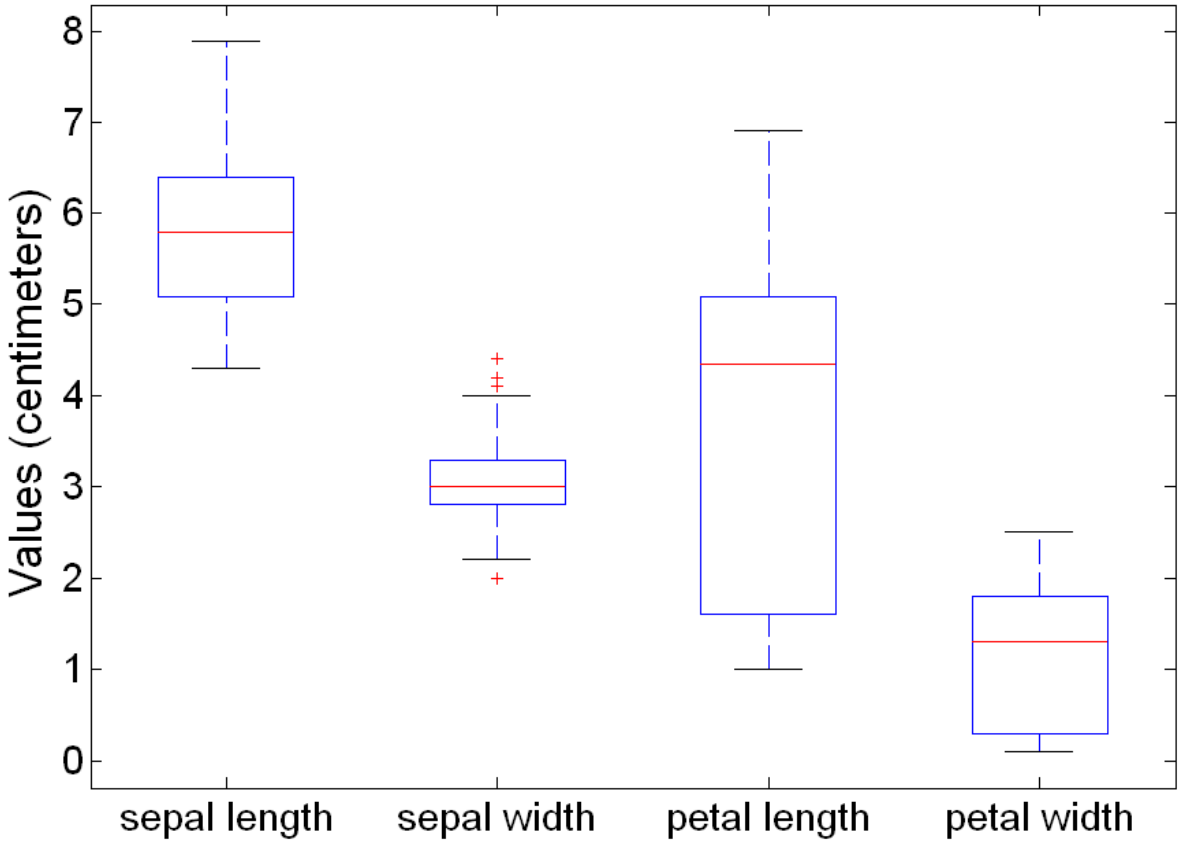
Box Plots

- **Inventati J. Tukey**
- **Permettono di riassumere la distribuzione dei dati**



Esempio

- **Utile per confrontare attributi**



Varianza, deviazione standard

- **misure di mutua variabilità tra i dati di una serie**
- **Devianza empirica**

$$dev = \sum_{i=1}^{n} (x_i - \bar{x})^2$$

- **Varianza**

$$s^2 = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2$$

- **Coefficiente di variazione**
  - **misura relativa**

$$V = \frac{s}{\bar{x}}$$

Simmetria

- **Si ha simmetria quando media, moda e mediana coincidono**
  - **condizione necessaria, non sufficiente**
  - **Asimmetria sinistra: moda, mediana, media**
  - **Asimmetria destra: media, mediana, moda**

Simmetria (Cont.)

- **Indici di asimmetria**
  - **medie interquartili**

$$\bar{x}_p = (x_{1-p} + x_p)/2$$

  - **Momenti centrali**

$$m_k = \frac{1}{n-1}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^k$$

- **indice di Fisher**
  - γ **nullo per distribuzioni simmetriche**
  - $\gamma > 0$**: sbilanciamenti a destra**
  - $\gamma < 0$**: sbilanciamento a sinistra**

$$\gamma = \frac{m_3}{\hat{s}^3}$$

Misure di correlazione

- **Covarianza**

$$\sigma(x,y) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$$

- **In D dimensioni**

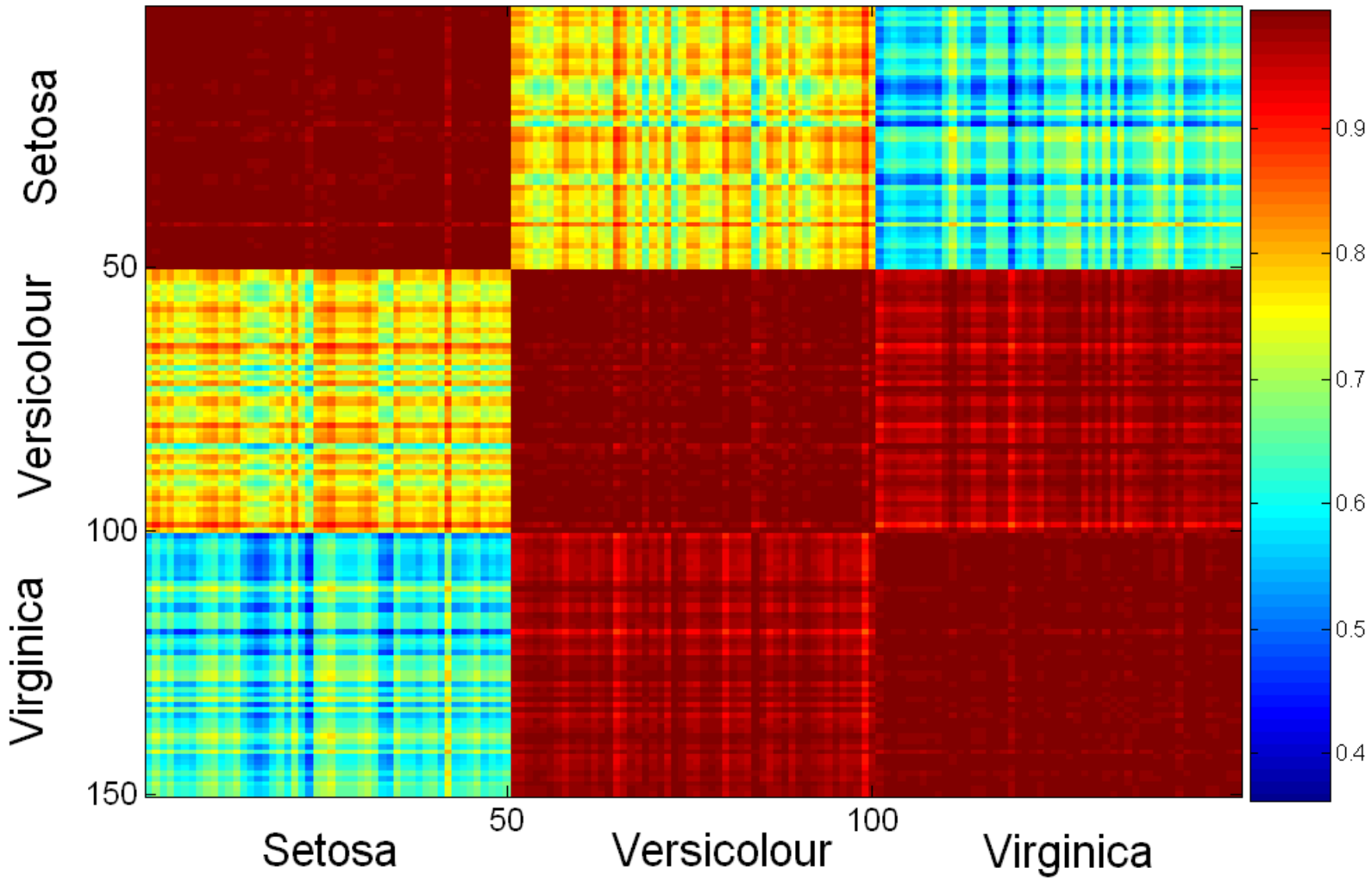
		A_1	A_2	A_3	A_4	A_5
	A_1	$\sigma^2(A_1)$	$\sigma(A_1,A_2)$	$\sigma(A_1,A_3)$	$\sigma(A_1,A_4)$	$\sigma(A_1,A_5)$
	A_2	$\sigma(A_2,A_1)$	$\sigma^2(A_2)$	$\sigma(A_2,A_3)$	$\sigma(A_2,A_4)$	$\sigma(A_2,A_5)$
	A_3	$\sigma(A_3,A_1)$	$\sigma(A_3,A_2)$	$\sigma^2(A_3)$	$\sigma(A_3,A_4)$	$\sigma(A_3,A_5)$
	A_4	$\sigma(A_4,A_1)$	$\sigma(A_4,A_2)$	$\sigma(A_4,A_3)$	$\sigma^2(A_4)$	$\sigma(A_4,A_5)$
	A_5	$\sigma(A_5,A_1)$	$\sigma(A_5,A_2)$	$\sigma(A_5,A_3)$	$\sigma(A_5,A_4)$	$\sigma^2(A_5)$

Coefficienti di Correlazione

- **Coefficiente di Pearson**

$$r_{xy} = \frac{\sigma(x, y)}{s_x s_y}$$

Matrice di correlazione per Iris



Outline del Modulo

- Introduzione e Concetti di Base
- Data Selection
- Information Gathering
- Data cleaning
- Data reduction
- Data transformation

Data Cleaning

- **Trattamento di valori anomali**
- **Trattamento di outliers**
- **Trattamento di tipi impropri**

Valori Anomali

- **Valori mancanti**
  - **NULL**
- **Valori sconosciuti**
  - **Privi di significato**
- **Valori non validi**
  - **Con valore noto ma non significativo**

Valori NULL

- I valori mancanti possono apparire in molte forme:
  - <empty field> “0” “.” “999” “NA” …
- I valori vanno standardizzati (e.g., utilizzando il simbolo NULL)
- Trattamento di valori nulli:
  - Ignorare I record con valori nulli
  - Trattare il valore null come un valore separato
  - Imputare: sostituire il valore null con altri valori

Valori nulli: esempio

- **Un valore può essere mancante perché non registrato o perché è inapplicabile**
- **Per** Jane **non è registrato, mentre per** Joe **o** Anna **dovrebbe essere considerato** Non applicabile
- **I valori null possono essere inferiti**

Pronto soccorso Ospedale

Nome	Età	Sesso	Incinta	..
Mary	25	F	N	
Jane	27	F	-	
Joe	30	M	-	
Anna	2	F	-	

Trattamento di valori nulli

- Utilizzando media/mediana/moda
- Predicendo i valori mancanti utilizzando la distribuzione dei valori non nulli
- Segmentando i dati (tramite le distribuzioni di altre variabili) e utilizzando misure statistiche (media/moda/mediana) di ogni segmento
- Segmentando i dati e utilizzando le distribuzioni di probabilità all'interno dei segmenti
- Costruendo un modello di classificazione/regressione e utilizzando il modello per calcolare i valori nulli
  - In dati numerici il trattamento può influenzare la distribuzione

Un caso particolare: le date

- Vogliamo trasformare tutte le date in uno stesso formato
- Problema molto sentito
  - e.g. “Sep 24, 2003” , 9/24/03, 24.09.03, etc
- Rappresentazioni categoriche: YYYYMM / YYYYMMDD
  - YYYYMMDD non preserva gli intervalli:
  - 20040201 - 20040131 /= 20040131 – 20040130
  - Ciò può introdurre bias nei modelli

Opzioni possibili

- **Possiamo usare:**
  - **Unix system date: numero di secondi dal 1970**
  - **Numero di giorni dal 1 gennaio 1960 (SAS)**
- **Problemi:**
  - **I valori non sono ovvii**
  - **Non aiuta l'interpretazione**
  - **Facile commettere errori**

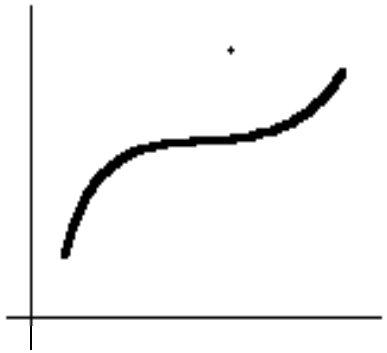
Un Formato possibile

$$\text{Date} = \text{YYYY} + \frac{\text{giorni\_dal\_1\_gennaio} - 0.5}{365 + 1\_\text{se\_bisestile}}$$

- Preserva gli intervalli
- L'anno e il quadrimestre sono facili da estrapolare
  - Sep 24, 2003 is 2003 + (267-0.5)/365= 2003.7301 (round to 4 digits)
- Può essere esteso al tempo

Rimozione di Outlier

- Outliers = Valori inconsistenti con la maggioranza dei dati
- Differente significato per gli outliers
  - Valido: il salario di un amministratore delegato
  - Rumore: Età = 200
- Rimozione
  - Clustering
  - Curve-fitting
  - Test di ipotesi con un modello precalcolato



Conversione: da Nominali a Numerici

- Alcuni algoritmi possono lavorare con valori nominali
- Altri metodi (reti neurali, regressione) lavorano solo con valori numerici
  - Conseguenza: trasformazione
- Strategie differenti

Da Binari a Numerici

- **Campi binari**
  - **E.g. Sesso=M, F**
- **Convertito in Campo_0_1 con valori 0, 1**
  - **e.g. Sesso= M → Sesso_0_1 = 0**
  - **Sesso = F → Sesso_0_1 = 1**

Da Ordinali a Numerici

- **Attributi ordinati (ad esempio, Giudizio) possono essere convertiti in numeri che preservano l'ordine naturale**
  - **Ottimo → 10.0**
  - **Discreto → 8**
  - **Sufficiente → 6**
  - **Insufficente → 4**
  - **Scarso → 2**
- **D: Perché è importante preservare l'ordine?**
- **R: Per permettere confronti significativi: Giudizio > 6**

Caso particolare: da ordinale a booleano

- **Un ordinale con n valori può essere codificato utilizzando *n–1* attributi booleani**
- **Esempio: l'attributo "temperature"**

Dati originali

Temperature
Cold
Medium
Hot

⇨

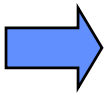
Dati trasformati

Temperature > cold	Temperature > medium
False	False
True	False
True	True

Nominali con pochi valori

- **Attributi numinali con pochi (*regola pratica < 20*) valori**
  - **e.g. Colore=Rosso, Arancio, Giallo, …, Viola**
  - **Per ogni valore *v creiamo una variabile* binaria C_*v* , che assumerà valore 1 if Colore=*v*, 0 altrimenti**

ID	Colore	…
371	rosso	
433	giallo	



ID	C_rosso	C_arancio	C_giallo	…
371	1	0	0	
433	0	0	1	

Categorici

- **Esempi:**
  - **Codici Postali (~10.000 valori)**
  - **Codici professionali (7,000 valori)**
- **D: Come gestirli ?**
- **R: Ignoriamo gli attributi che si comportano come chiavi (= con valori unici per ogni record)**
- **Gli altri attributi dovrebbero essere raggruppati in gruppi "naturali" :**
  - **Esempio: Codici postali→ regioni**
  - **Professioni – selezionare le più frequenti, raggruppare le altre**
- **Trattare le nuove categorie come attributi nominali**

Outline del Modulo

- Introduzione e Concetti di Base
- Data Selection
- Information Gathering
- Data cleaning
- Data reduction
- Data transformation

Data Reduction

- **Riduzione del volume dei dati**
  - **Orizzontale: eliminazione di colonne**
    - **Factor Analysis**
    - **Principal Component Analysis**
  - **Verticale: eliminazione di tuple**
    - **Data Sampling**
    - **Clustering**

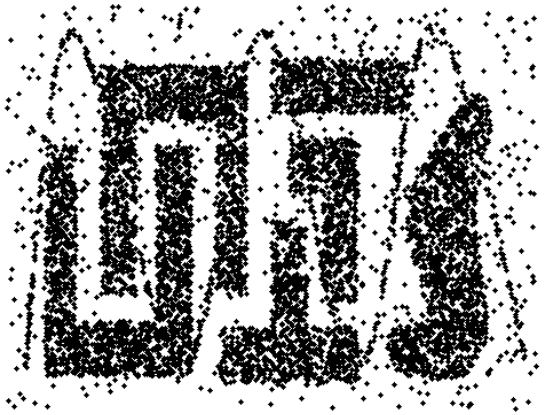
Sampling

- **Permette ad un algoritmo di mining di essere eseguito con una complessità minore (su una porzione sublineare della dimensione dei dati)**
- **Problema: scegliere un sottoinsieme rappresentativo dei dati**
  - **Un campione è rappresentativo se ha le stesse proprietà (di interesse) del dataset originale**
  - **Schemi semplici possono risultare inadeguati in presenza di picchi/sbilanciamenti**

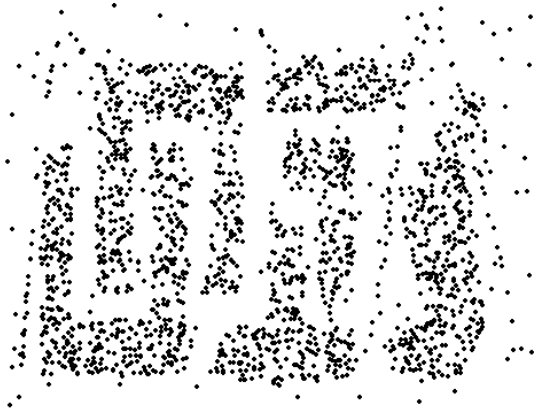
Sampling

- **Simple Random Sampling**
  - **Stessa probabilità di selezionare un oggetto**
- **Sampling senza rimpiazzamento**
  - **Gli oggetti selezionati sono rimossi dal dataset originale**
- **Sampling con rimpiazzamento**
  - **Gli oggetti selezionati non sono rimossi**
    - **Lo stesso oggetto può essere scelto più volte**
- **Stratified sampling**
  - **Dividi i dati in più partizioni; campiona da ogni partizione**

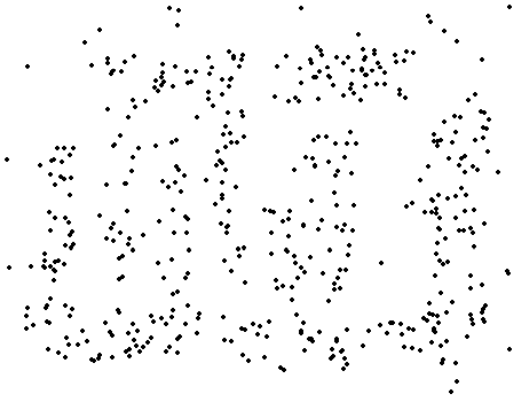
Sampling



8000 oggetti



2000 oggetti

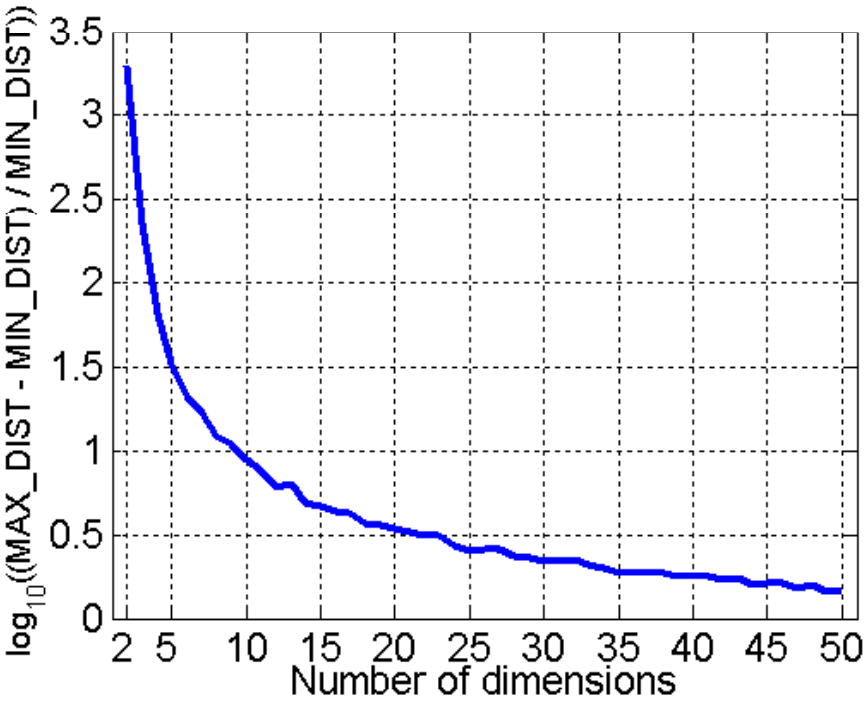


500 oggetti

“Curse of Dimensionality”

- **Quando il numero di attributi cresce, i dati diventano progressivamente sparsi**

- **Densità e distanza perdono di significatività**



- **La differenza tra max e min diminuisce progressivamente**

Riduzione della dimensionalità

- Evita il problema descritto precedentemente
- Migliora le performances degli algoritmi
- Permette una migliore visualizzazione
- Può eliminare attributi irrilevanti e ridurre il rumore

- Tecniche
  - Principle Component Analysis
  - Singular Value Decomposition
  - Altri metodi (più avanti nel corso)

Principal Component Analysis

sepallength	*sepalwidth*	*petallength*	*petalwidth*	*class*
7.2	3.6	6.1	2.5	Iris-virginica
5.9	3	4.2	1.5	Iris-versicolor
5.4	3.4	1.5	0.4	Iris-setosa
5	3.3	1.4	0.2	Iris-setosa
6.7	3	5.2	2.3	Iris-virginica
5.1	3.5	1.4	0.2	Iris-setosa
6.7	3.3	5.7	2.5	Iris-virginica
6.7	3.1	5.6	2.4	Iris-virginica
5.8	2.7	5.1	1.9	Iris-virginica
6.7	3.1	5.6	2.4	Iris-virginica
5.2	2.7	3.9	1.4	Iris-versicolor
6.9	3.1	4.9	1.5	Iris-versicolor
5.1	3.8	1.9	0.4	Iris-setosa
6.6	3	4.4	1.4	Iris-versicolor
5.1	3.5	1.4	0.3	Iris-setosa

$$\mu = [6.0067\ 3.2067\ 3.8867\ 1.4200]$$
$$\sigma = [0.8040\ 0.3173\ 1.8291\ 0.9096]$$

Principal Component Analysis

$$\mathbf{X} = \begin{bmatrix} 1.4843 & 1.2397 & 1.2101 & 1.1873 \\ -0.1327 & -0.6514 & 0.1713 & 0.0879 \\ -0.7546 & 0.6093 & -1.3048 & -1.1213 \\ -1.2521 & 0.2942 & -1.3595 & -1.3412 \\ 0.8624 & -0.6514 & 0.7180 & 0.9674 \\ -1.1277 & 0.9245 & -1.3595 & -1.3412 \\ 0.8624 & 0.2942 & 0.9914 & 1.1873 \\ 0.8624 & -0.3362 & 0.9367 & 1.0774 \\ -0.2571 & -1.5969 & 0.6634 & 0.5277 \\ 0.8624 & -0.3362 & 0.9367 & 1.0774 \\ -1.0033 & -1.5969 & 0.0073 & -0.0220 \\ 1.1111 & -0.3362 & 0.5540 & 0.0879 \\ -1.1277 & 1.8701 & -1.0862 & -1.1213 \\ 0.7380 & -0.6514 & 0.2807 & -0.0220 \\ -1.1277 & 0.9245 & -1.3595 & -1.2313 \end{bmatrix}$$

Principal Component Analysis

$$\mathbf{S} = \begin{bmatrix} 1 & -0.2 & 0.89 & 0.87 \\ -0.2 & 1 & -0.47 & -0.43 \\ 0.89 & -0.47 & 1 & 0.98 \\ 0.87 & -0.43 & 0.98 & 1 \end{bmatrix}$$

$$\lambda_1 = 3.0291, \lambda_2 = 0.8472, \lambda_3 = 0.1108, \lambda_4 = 0.013$$

$$\sum_{i=1}^{4} = 4 \qquad \lambda_1 + \lambda_2 = 3.8763$$

$$\mathbf{e}_1 = \begin{bmatrix} 0.5197 \\ -0.3017 \\ 0.5696 \\ 0.5607 \end{bmatrix} \qquad \mathbf{e}_2 = \begin{bmatrix} 0.3754 \\ 0.9207 \\ 0.0541 \\ 0.0926 \end{bmatrix}$$

Principal Component Analysis

$0.52sl$-$0.302sw$+$0.57pl$+$0.561pw$	$0.375sl$+$0.921sw$+$0.054pl$+$0.093pw$	*class*
1.752354	1.873861	Iris-virginica
0.274451	-0.632121	Iris-versicolor
-1.947978	0.10332	Iris-setosa
-2.265912	-0.396837	Iris-setosa
1.596159	-0.14761	Iris-virginica
-2.391448	0.230147	Iris-setosa
1.589859	0.758106	Iris-virginica
1.687302	0.164595	Iris-virginica
1.021998	-1.481964	Iris-virginica
1.687302	0.164595	Iris-virginica
-0.047758	-1.848442	Iris-versicolor
1.043771	0.145663	Iris-versicolor
-2.397804	1.135858	Iris-setosa
0.727673	-0.309538	Iris-versicolor
-2.32983	0.240319	Iris-setosa

Outline del Modulo

- Introduzione e Concetti di Base
- Data Selection
- Information Gathering
- Data cleaning
- Data reduction
- Data transformation

Data Transformation: Motivazioni

- Errori nei dati
- Dati incompleti
- forte asimmetria nei dati
  - diversi raggruppamenti esprimono comportamenti differenti
- molti picchi
  - residui larghi e sistematici nella definizione di un modello
- La modifica della forma dei dati può alleviare questi problemi

Obiettivi

- In una matrice X
  - X_{ik} rappresenta un elemento della matrice
    - ($i = 1..n$), n numero di righe
    - ($k = 1…l$) l numero di attributi
- Vogliamo definire una trasformazione T t.c.

$$Y_{ij} = T(X_{ik})$$

  - ($j = 1…m$), m numero di attributi dopo la trasformazione
  - Y_{ij} preserva l'informazione "rilevante" di X_{ik}
  - Y_{ij} elimina almeno uno dei problemi di X_{ik}
  - Y_{ij} è piu` utile di X_{ik}
- In generale, $m \neq l$

Obiettivi

- **scopi principali:**
  - **stabilizzare le varianze**
  - **linealizzare le relazioni tra variabili**
  - **normalizzare le distribuzioni**
- **scopi secondari:**
  - **semplificare l'elaborazione di dati che presentano caratteristiche non gradite**
  - **rappresentare i dati in una scala ritenuta più adatta.**

Similarita’ e Differenze

- Molte metodologie statistiche richiedono correlazioni lineari, distribuzioni normali, assenza di outliers
- Molti algoritmi di Data Mining hanno la capacita` di trattare automaticamente nonlinearita’ e non normalita’
  - Gli algoritmi lavorano comunque meglio se tali problemi sono trattati

Metodi

- **Trasformazioni esponenziali**

$$T_p(x) = \begin{cases} ax^p + b & (p \neq 0) \\ c \log x + d & (p = 0) \end{cases}$$

- **con a,b,c,d e p valori reali**
  - **Preservano l'ordine**
  - **Preservano alcune statistiche di base**
  - **sono funzioni continue**
  - **ammettono derivate**
  - **sono specificate tramite funzioni semplici**

Migliorare l'interpretabilita`

- **Trasformazioni lineari**

1€ = 1936.27 Lit.

– *p=1, a= 1936.27 ,b =0*

ºC= 5/9(ºF -32)

– *p = 1, a = 5/9, b = -160/9*

Normalizzazioni

- **min-max normalization**

$$v' = \frac{v - min_A}{max_A - min_A}(new\_max_A - new\_min_A) + new\_min_A$$

- **z-score normalization**

$$v' = \frac{v - mean_A}{stand\_dev_A}$$

- **normalization tramite decimal scaling**

$$v' = \frac{v}{10^j}$$ dove j è il più piccolo intero tale che Max($|v'|$)<1

Stabilizzare varianze

- **Trasformazione logaritmica**

$$T(x) = c \log x + d$$

  - **Si applica a valori positivi**
  - **omogeneizza varianze di distribuzioni lognormali**
  - **E.g.: normalizza picchi stagionali**

Trasformazione logaritmica: esempio

Bar	*Birra*	*Ricavo*
A	Bud	20
A	Becks	10000
C	Bud	300
D	Bud	400
D	Becks	5
E	Becks	120
E	Bud	120
F	Bud	11000
G	Bud	1300
H	Bud	3200
H	Becks	1000
I	Bud	135

2300	Media
2883,3333	Scarto medio assoluto
3939,8598	Deviazione standard
5	Min
120	Primo Quartile
350	Media
1775	Secondo Quartile
11000	Max

Dati troppo dispersi!!!

Trasformazione Logaritmica: esempio

Bar	*Birra*	*Ricavo (log)*
A	Bud	1,301029996
A	Becks	4
C	Bud	2,477121255
D	Bud	2,602059991
D	Becks	0,698970004
E	Becks	2,079181246
E	Bud	2,079181246
F	Bud	4,041392685
G	Bud	3,113943352
H	Bud	3,505149978
H	Becks	3
I	Bud	2,130333768

Media	2,585697
Scarto medio assoluto	0,791394
Deviazione standard	1,016144
Min	0,69897
Primo Quartile	2,079181
Media	2,539591
Secondo Quartile	3,211745
Max	4,041393

Stabilizzare varianze

$$T(x) = ax^p + b$$

- **Trasformazione in radice**
  - $p = 1/c$, c **numero intero**
  - **per omogeneizzare varianze di distribuzioni particolari, e.g., di Poisson**
- **Trasformazione reciproca**
  - $p < 0$
  - **Per l'analisi di serie temporali, quando la varianza aumenta in modo molto pronunciato rispetto alla media**

Creare simmetria nei dati

- **Aggiustiamo la media interpercentile**
  - **In generale,**
  - **Se la media interpercentile è sbilanciata, allora la distribuzione dei dati è asimmetrica**
    - **sbilanciata a destra**
    - **sbilanciata a sinistra**

$$M - x_p = x_{1-p} - M \Leftrightarrow \frac{x_{1-p} + x_p}{2} = M$$

$$\bar{x}_p > M$$

$$\bar{x}_p < M$$

Creare simmetria nei dati: esempio

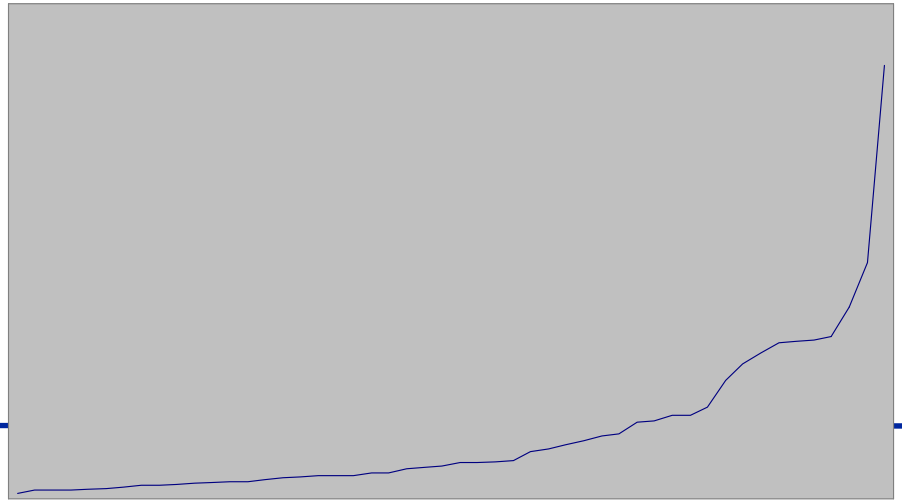
- **Verifichiamo la simmetria**

2.808	14.001	4.227	5.913	6.719
3.072	29.508	26.463	1.583	78.811
1.803	3.848	1.643	15.147	8.528
43.003	11.768	28.336	4.191	2.472
24.487	1.892	2.082	5.419	2.487
3.116	2.613	14.211	1.620	21.567
4.201	15.241	6.583	9.853	6.655
2.949	11.440	34.867	4.740	10.563
7.012	9.112	5.732	4.030	28.840
16.723	4.731	3.440	28.608	995

Creare simmetria : esempio

- I valori della media interpercentile crescono col percentile considerato
- Distribuzione sbilanciata a destra

Percentile	Media	Low	High
M	6158	6158	6158
F	9002	3278	14726
E	12499	2335	22662
D	15420	2117	28724
C	16722	2155	31288
1	39903	995	78811



Trasformation plot

- **Consideriamo i percentili x_U e x_L**
- **i valori c ottenuti tramite la formula**

$$\frac{x_U + x_L}{2} - M = (1-c)\frac{(x_U - M)^2 + (M - x_L)^2}{4M}$$

permettono di trovare il valore adeguato per p

- **Intuitivamente, compariamo la differenza assoluta e relativa tra mediana e medie interpercentili**
- **il valore medio (mediano) dei valori di c è il valore della trasformazione**

Trasformation plot: esempio

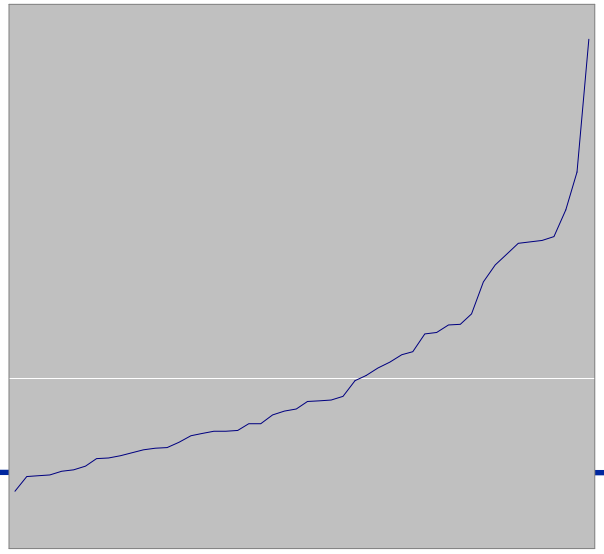
$(x_L-x_U)/2-M$	$((M- x_L)^2+(x_U -M)^2)/4M$	c
2844.5	3317.5	0.14258
6341	11652.8	0.45583
9262.7	21338.8	0.56592
10564.3	26292.5	0.59820

- **Calcolando la mediana dei valori *c* otteniamo *p=0.5188***
- **Proviamo le possibili approssimazioni razionali...**

Approssimazione 1: radice quatrata

$$T(x) = \sqrt{x}$$

Percentile	Media	Low	High	
M	78,42283	78,42283	78,42283	0,50000
F	89,28425	57,23633	121,33217	0,25000
E	99,37319	48,27950	150,46688	0,12500
D	107,58229	45,68337	169,48122	0,06250
C	110,87427	45,05801	176,69054	0,03125
1	156,13829	31,54362	280,73297	

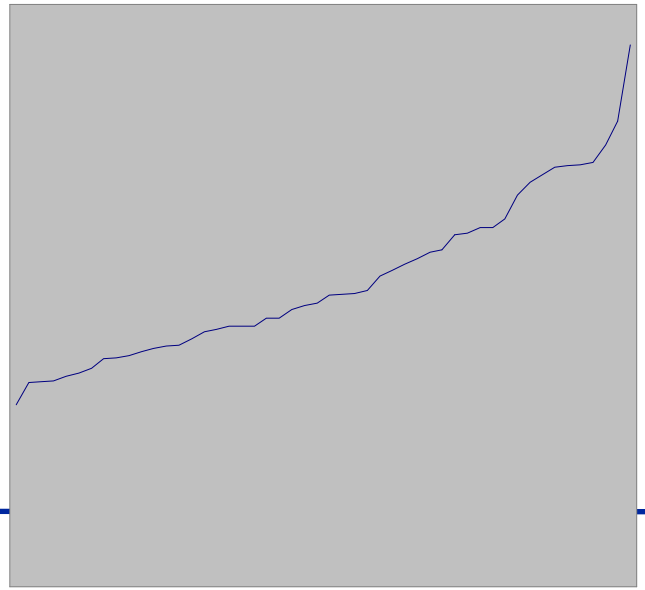


- La curva si tempera,ma i valori alti continuano a produrre differenze notevoli
- Proviamo a diminuire p...

Trasformazione 2: radice quarta

$$T(x) = \sqrt[4]{x}$$

Percentile	Media	Low	High	
M	8,85434	8,85434	8,85434	0,50000
F	9,28978	7,56489	11,01467	0,25000
E	9,60590	6,94676	12,26503	0,12500
D	9,88271	6,74694	13,01849	0,06250
C	9,97298	6,65710	13,28886	0,03125
1	11,18573	5,61637	16,75509	

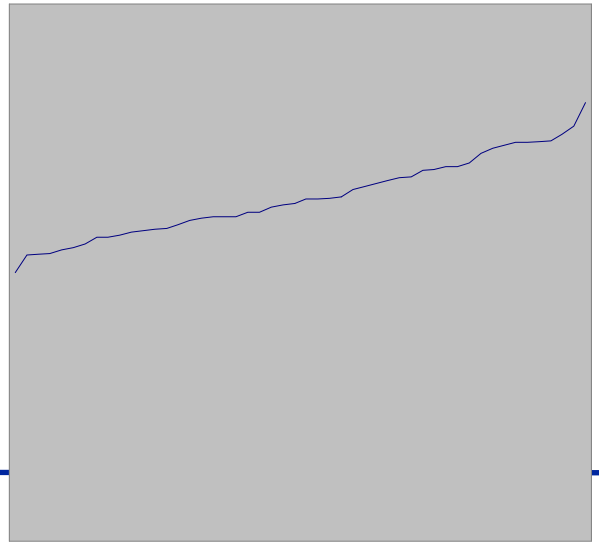


- I valori alti continuano ad influenzare
- Proviamo con il logaritmo...

Approssimazione 3: logaritmo

$$T(x) = \log x$$

Percentile	Media	Low	High	
M	3,78836502	3,78836502	3,78836502	0,50000
F	3,84144850	3,51507795	4,16781905	0,25000
E	3,86059853	3,36672764	4,35446943	0,12500
D	3,88578429	3,31332721	4,45824138	0,06250
C	3,88573156	3,27798502	4,49347811	0,03125
1	3,94720496	2,99782308	4,89658684	



- **Abbiamo ottenuto simmetria!**

Semplificare le relazioni tra piu` attributi

- **E.g., nel caso della regressione**
  - **La formula**

$$y = \alpha x^{p}$$

puo' essere individuata studiando la relazione

$$z = \log \alpha + pw$$

dove $z = \log y$ e $w = \log x$

Discretizzazione

- **Unsupervised vs. Supervised**
- **Globale vs. Locale**
- **Statica vs. Dinamica**
- **Task difficile**
  - **Difficile capire a priori qual'è la discretizzazione ottimale**
    - **bisognerebbe conoscere la distribuzione reale dei dati**

Discretizzazione: Vantaggi

- I dati originali possono avere valori continui estremamente sparsi
- I dati originali possono avere variabili multimodali
- I dati discretizzati possono essere più semplici da interpretare
- Le distribuzioni dei dati discretizzate possono avere una forma “Normale”
- I dati discretizzati possono essere ancora estremamente sparsi
  - Eliminazione della variabile in oggetto

Unsupervised Discretization

- Non etichetta le istanze
- Il numero di classi è noto a priori
- Natural binning
  - intervalli di identica ampiezza
- Equal Frequency binning
  - intervalli di identica frequenza
- Statistical binning
  - Utilizzando informazioni statistiche
    - media e varianza
    - Quartili

Quante classi?

- **Troppo poche $\Rightarrow$ perdita di informazione sulla distribuzione**
- **troppe $\Rightarrow$ disperde i valori e non manifesta la foma della distribuzione**
- **Il numero ottimale C di classi è funzione del numero N di elementi (Sturges, 1929)**

$$C = 1 + \frac{10}{3}\log_{10}(N)$$

- **L'ampiezza ottimale delle classi dipende dalla varianza e dal numero dei dati (Scott, 1979)**

$$h = \frac{3{,}5 \cdot s}{\sqrt{N}}$$

Natural Binning

- **Semplice**
- **Ordino i valori, quindi divido il range di valori in *k* parti della stessa dimensione**

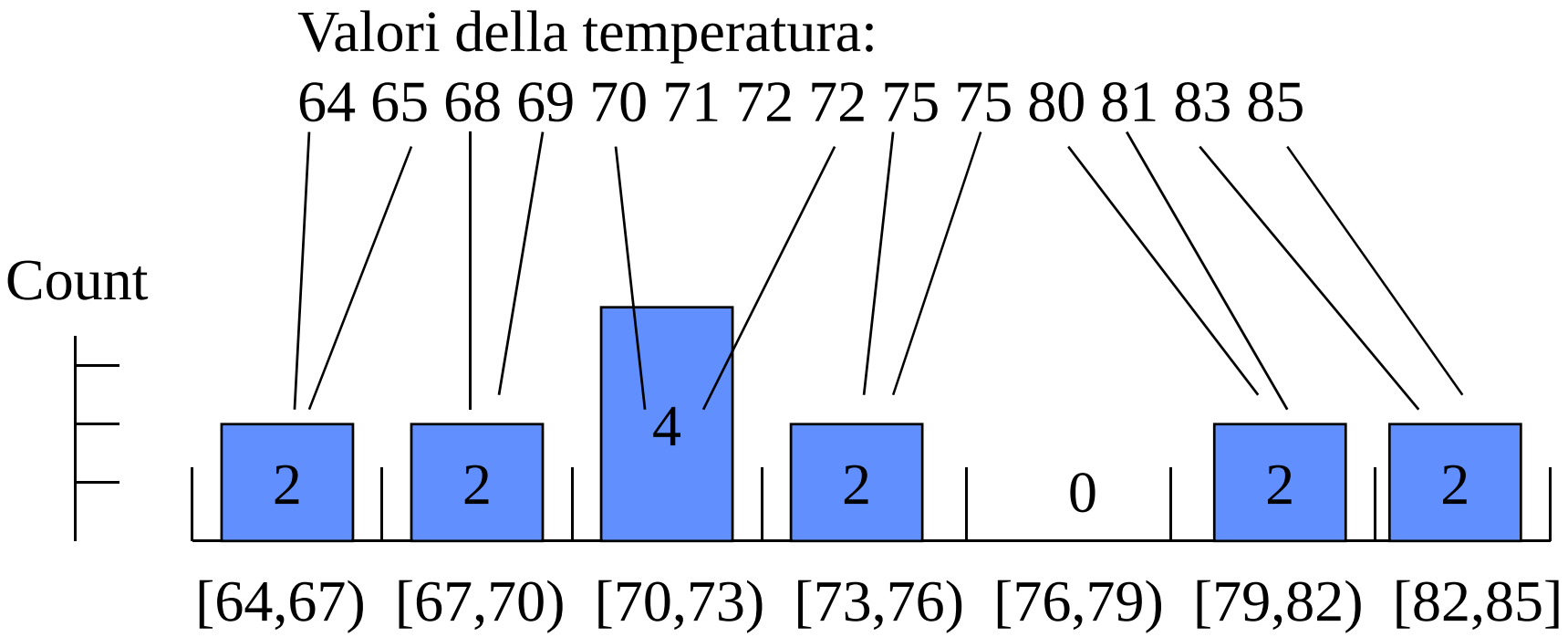
$$\delta = \frac{x_{\max} - x_{\min}}{k}$$

- **l'elemento x_j appartiene alla classe i se**

$$x_j \in [x_{min} + i\delta,\ x_{min} + (i+1)\delta)$$

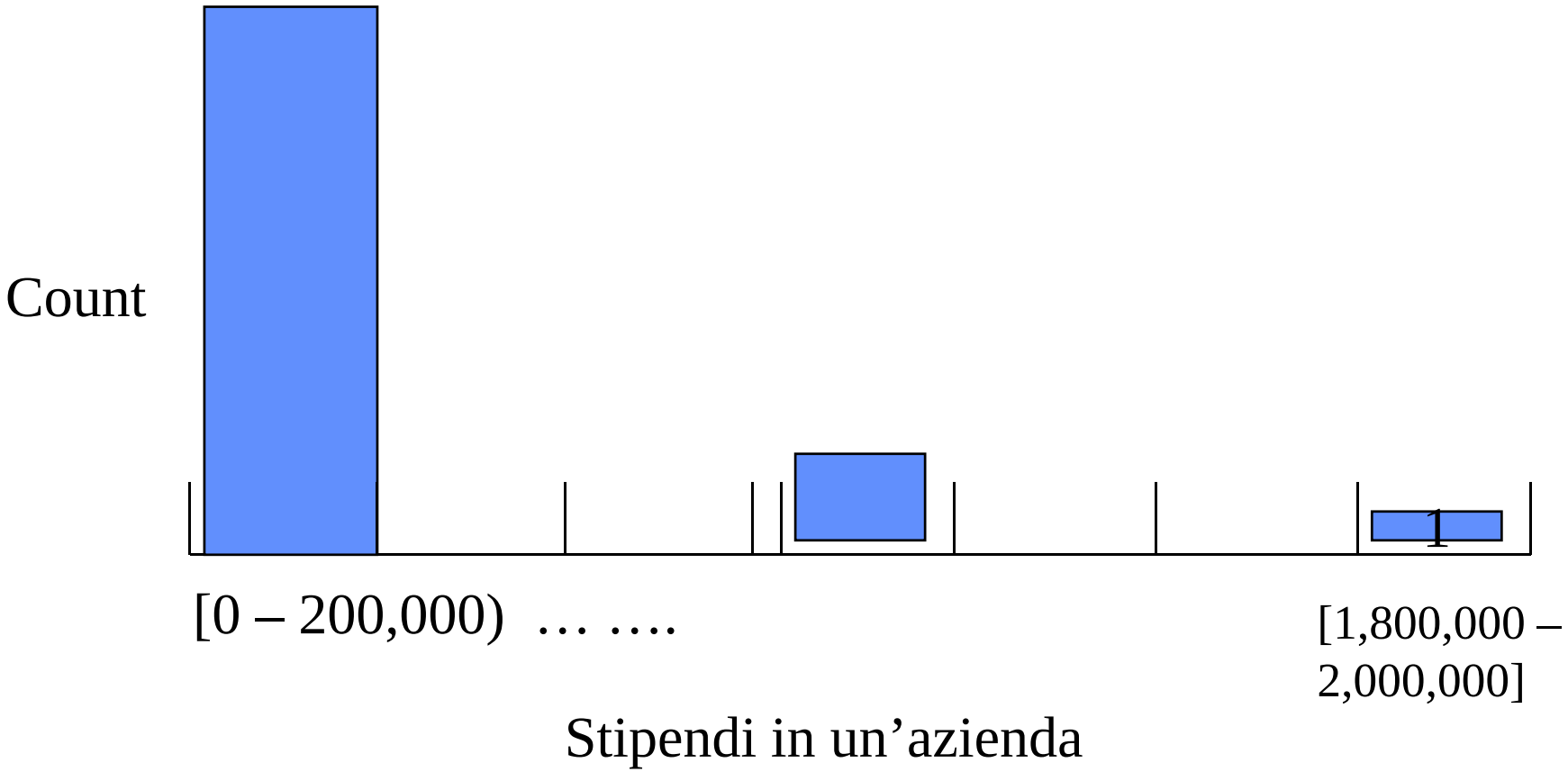
- **Puo` produrre distribuzioni molto sbilanciate**

Natural binning



Ampiezza dell'intervallo prefissata

Il natural binning può produrre raggruppamenti



Esempio

Bar	Beer	Price
A	Bud	100
A	Becks	120
C	Bud	110
D	Bud	130
D	Becks	150
E	Becks	140
E	Bud	120
F	Bud	110
G	Bud	130
H	Bud	125
H	Becks	160
I	Bud	135

- δ =(160-100)/4 = 15
- **classe 1: [100,115)**
- **classe 2: [115,130)**
- **classe 3: [130,145)**
- **classe 4: [145, 160]**

- **Caratterizza il prezzo di Bud**
- **Non caratterizza il prezzo di Becks**

Equal Frequency Binning

- **Ordino e Conto gli elementi, quindi definisco il numero di intervalli calcolando**

$$f = \frac{N}{k}$$

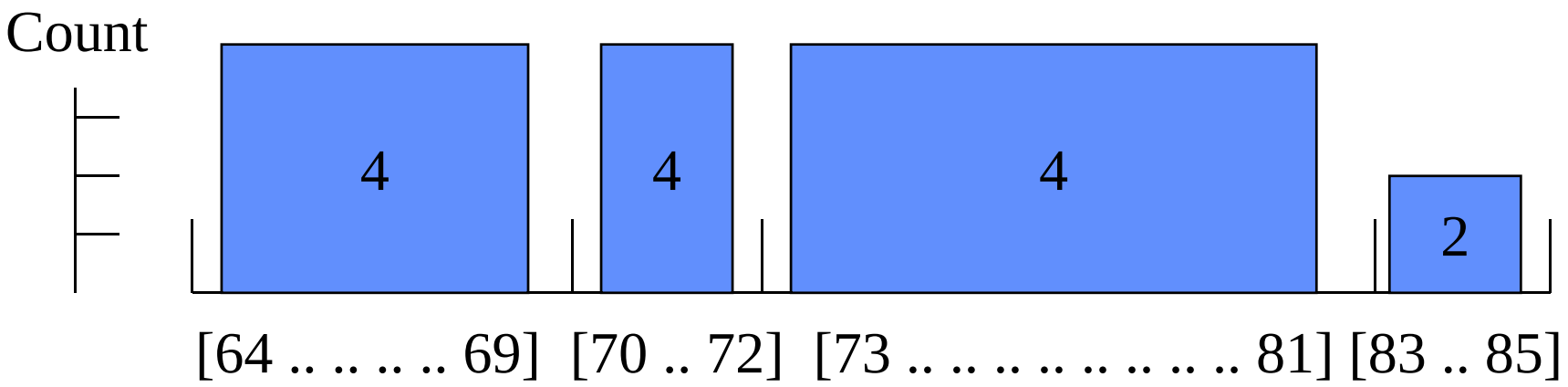
- **Dove *N* è il numero di elementi del campione**
- **l'elemento x_i appartiene alla classe j se**

$$j \times f \leq i < (j+1) \times f$$

- **Non sempre adatta ad evidenziare correlazioni interessanti**

Frequency binning

Valori di temperatura:
64 65 68 69 70 71 72 72 75 75 80 81 83 85



Altezza identica= 4 (tranne che per l'ultimo intervallo)

Vantaggi

- **Preferita perché evita i raggruppamenti**
- **In aggiunta:**
  - **Non separa valori frequenti ai bordi degli intervalli**
  - **Crea intervalli specifici per valori speciali (e.g. 0)**

Esempio

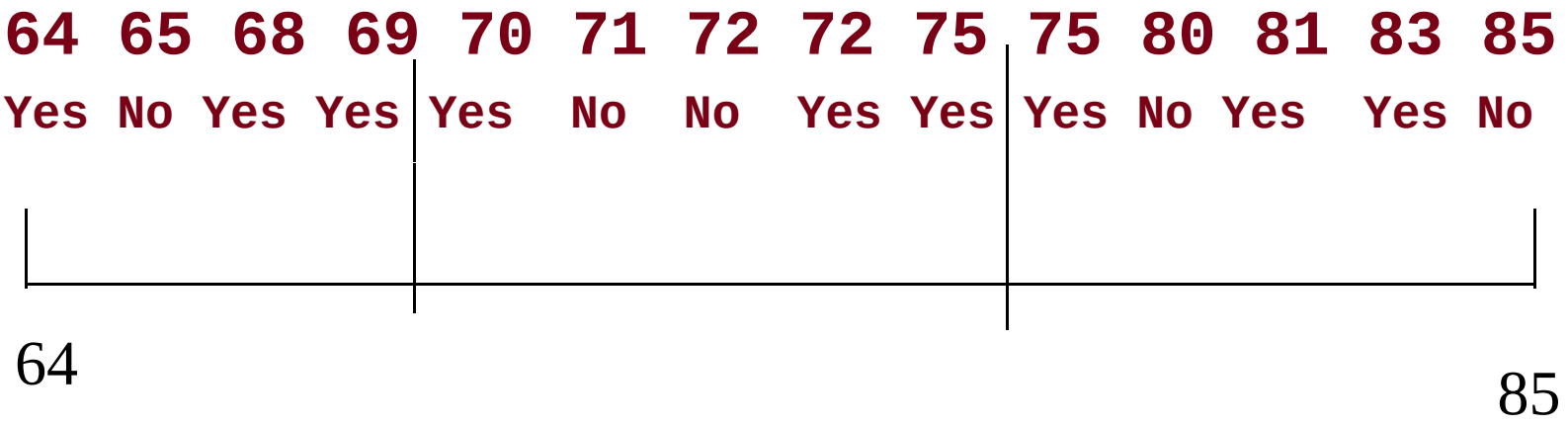
Bar	Beer	Price
A	Bud	100
A	Becks	120
C	Bud	110
D	Bud	130
D	Becks	150
E	Becks	140
E	Bud	120
F	Bud	110
G	Bud	130
H	Bud	125
H	Becks	160
I	Bud	135

- $f = 12/4 = 3$
- **classe 1: {100,110,110}**
- **classe 2: {120,120,125}**
- **classe 3: {130,130,135}**
- **classe 4: {140,150,160}**

- **Non caratterizza il prezzo di Becks**

Supervised Discretization

- **La discretizzazione ha un obiettivo quantificabile**
- **Il numero di intervalli non è noto a priori**
- **Esempio: voglio che in ogni intervallo di siano almeno tre valori identici per un altro attributo**



Supervised Discretization: ChiMerge

- **Bottom-up**
- **Inizialmente, ogni valore è un intervallo a se'**
- **Intervalli adiacenti sono iterativamente uniti se sono simili**
- **La similitudine è misurata sulla base dell'attributo target, contando quanto i due intervalli sono "diversi"**

ChiMerge: criterio di similitudine

- **Basato sul test del Chi quadro**
- **k numero di valori differenti dell'attributo target**
- **A_{ij} numero di casi della j-esima classe nell'i-esimo intervallo**
- **R_i numero di casi nell'i-esimo intervallo ($\sum_{j=1}^{k} A_{ij}$)**
- **C_j numero di casi nella j-esima classe ($\sum_{i=1}^{2} A_{ij}$)**
- **E_{ij} frequenza attesa di A_{ij} $(R_i * C_j / N)$**

Reminder: test del Chi Quadro

- **Obiettivo: data una tabella di contingenza, verificare se righe e colonne sono indipendenti**
  - **Per un dato elemento in classe i,j la sua probabilità è p_{ij}**
  - **Se righe e colonne sono indipendenti, allora**
    - **$p_{ij} = u_i v_j$**

		Columns			
		1	2	…	c
Rows	1	O_{11}	O_{12}	…	O_{1c}
	2	O_{21}	O_{22}	…	O_{2c}
	⋮	⋮	⋮	⋮	⋮
	r	O_{r1}	O_{r2}	…	O_{rc}

$$\hat{u}_i = \frac{1}{n} \sum_{j=1}^{c} O_{ij}$$

$$\hat{v}_j = \frac{1}{n} \sum_{i=1}^{r} O_{ij}$$

Test dell'indipendenza

- **Se l'indipendenza vale, allora**

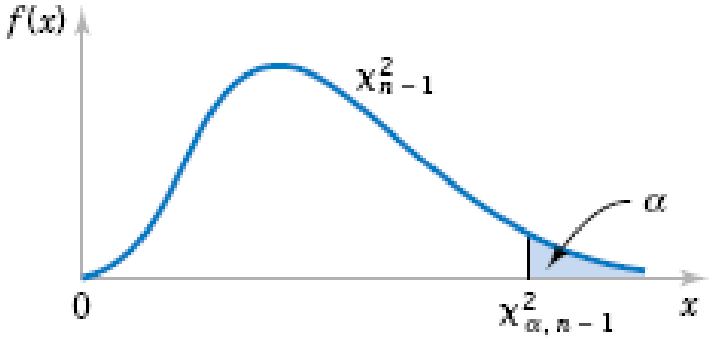
$$E_{ij} = n\hat{u}_i\hat{v}_j = \frac{1}{n}\sum_{j=1}^{c} O_{ij} \sum_{i=1}^{r} O_{ij}$$

- **La statistica**

  Ha una distribuzione del Chi quadro con (r-1)(c-1) gradi di libertà

$$\chi_0^2 = \sum_{i=1}^{r}\sum_{j=1}^{c} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

$$\Pr(\chi_0^2 \leq \chi^2_{\alpha,(r-1)(c-1)}) = 1-\alpha$$



Conseguenze

- **Dato un valore per α (tipicamente, 0.05)**
  - **Se la statistica associata ha un valore maggiore a $\chi^2_{\alpha,(r-1)(c-1)}$**
  - **Il test viene rigettato e le colonne non sono indipendenti**

Esempio

Valori attuali

Job Classification	Pension Plan 1	2	3	Totals
Salaried workers	160	140	40	340
Hourly workers	40	60	60	160
Totals	200	200	100	500

Valori attesi

Job Classification	Pension Plan 1	2	3	Totals
Salaried workers	136	136	68	340
Hourly workers	64	64	32	160
Totals	200	200	100	500

$$\chi_0^2 > \chi_{0.05,2}^2 = 5.99$$

$$\chi_0^2 = \sum_{i=1}^{2}\sum_{j=1}^{3}\frac{(o_{ij} - E_{ij})^2}{E_{ij}}$$

$$= \frac{(160 - 136)^2}{136} + \frac{(140 - 136)^2}{136} + \frac{(40 - 68)^2}{68} + \frac{(40 - 64)^2}{64}$$

$$+ \frac{(60 - 64)^2}{64} + \frac{(60 - 32)^2}{32} = 49.63$$

Test del Chi Quadro per la discretizzazione

	1	2	…	K	Total
1	$\mathbf{A_{11}}$	$\mathbf{A_{12}}$	…	$\mathbf{A_{1k}}$	$\mathbf{R_1}$
2	$\mathbf{A_{21}}$	$\mathbf{A_{22}}$	…	$\mathbf{A_{2k}}$	$\mathbf{R_2}$
Total	$\mathbf{C_1}$	$\mathbf{C_2}$	…	$\mathbf{C_k}$	**N**

$$\chi^2 = \sum_{i=1}^{2}\sum_{j=1}^{k}\frac{(A_{ij} - E_{ij})^2}{E_{ij}}$$

- **Si individua quanto “distinti” sono due intervalli**
- **k-1 gradi di liberta`**
- **La significativita` del test è data da un threshold α**
  - **Probabilita` che l’intervallo in questione e la classe siano indipendenti**

Esempio

Bar	Beer	Price
A	Bud	100
A	Becks	120
C	Bud	110
D	Bud	130
D	Becks	150
E	Becks	140
E	Bud	120
F	Bud	110
G	Bud	130
H	Bud	125
H	Becks	160
I	Bud	135

- **Discretizzazione w.r.t. Beer**
- **threshold 50% confidenza**
- **Vogliamo ottenere una discretizzazione del prezzo che permetta di mantenere omogeneita` w.r.t. Beer**

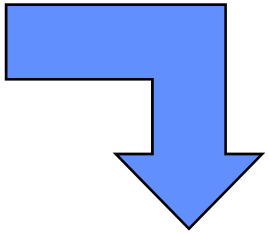
Esempio: Chi Values

	Bud	*Becks*
100	1	0
110	2	0
120	1	1
125	1	0
130	2	0
135	1	0
140	0	1
150	0	1
160	0	1

Scegliamo gli elementi adiacenti con Chi-Value minimo

Esempio: passo 1

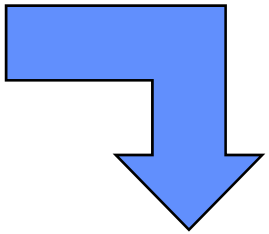
	Bud	Becks	Chi Value
100	1	0	0
110	2	0	1.33333
120	1	1	0.75
125	1	0	0
130	2	0	0
135	1	0	2
140	0	1	0
150	0	1	0
160	0	1	1.38629



	Bud	Becks	Chi Value
100	1	0	0
110	2	0	1.33333
120	1	1	0.75
125	1	0	0
130	2	0	0
135	1	0	2
140	0	1	0
150-160	0	2	1.38629

Esempio: passo 2

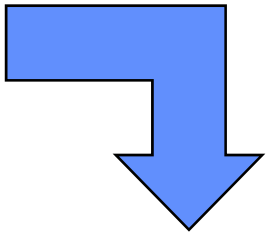
	Bud	Becks	Chi Value
100	1	0	0
110	2	0	1.33333
120	1	1	0.75
125	1	0	0
130	2	0	0
135	1	0	2
140	0	1	0
150-160	0	2	1.38629



	Bud	Becks	Chi Value
100	1	0	0
110	2	0	1.33333
120	1	1	0.75
125	1	0	0
130	2	0	0
135	1	0	4
140-150-160	0	3	1.38629

Esempio: passo 3

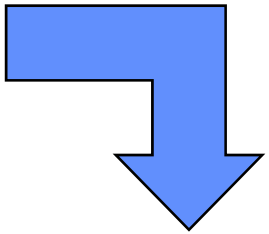
	Bud	*Becks*	*Chi Value*
100	1	0	0
110	2	0	1.33333
120	1	1	0.75
125	1	0	0
130	2	0	0
135	1	0	4
140-150-160	0	3	1.38629



	Bud	*Becks*	*Chi Value*
100	1	0	0
110	2	0	1.33333
120	1	1	0.75
125	1	0	0
130-135	3	0	6
140-150-160	0	3	1.38629

Esempio: passo 4

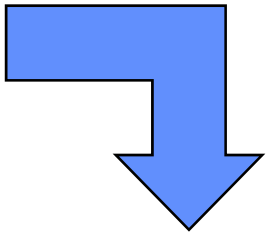
	Bud	*Becks*	*Chi Value*
100	1	0	0
110	2	0	1.33333
120	1	1	0.75
125	1	0	0
130-135	3	0	6
140-150-160	0	3	1.38629



	Bud	*Becks*	*Chi Value*
100	1	0	0
110	2	0	1.33333
120	1	1	2.4
125-130-135	4	0	7
140-150-160	0	3	1.38629

Esempio: passo 5

	Bud	Becks	Chi Value
100	1	0	0
110	2	0	1.33333
120	1	1	2.4
125-130-135	4	0	7
140-150-160	0	3	1.38629



Tutti i valori sono oltre il 50% di confidenza (1.38)

	Bud	Becks	Chi Value
100-110	3	0	1.875
120	1	1	2.4
125-130-135	4	0	7
140-150-160	0	3	1.38629

Considerazioni finali

- Natural binning è il metodo più semplice (e va bene nella maggioranza dei casi)
  - Fallisce miseramente con distribuzioni sbilanciate)
- Frequency binning può dare risultati migliori
  - Ma non può essere utilizzato con tutte le tecniche
- La discretizzazione supervisionata è particolarmente importante per l'analisi delle dipendenze
  - Alcuni algoritmi (Naïve Bayes) richiedono la discretizzazione