Data preparation

Sistemi informativi per le Decisioni

Slide a cura di prof. Claudio Sartori

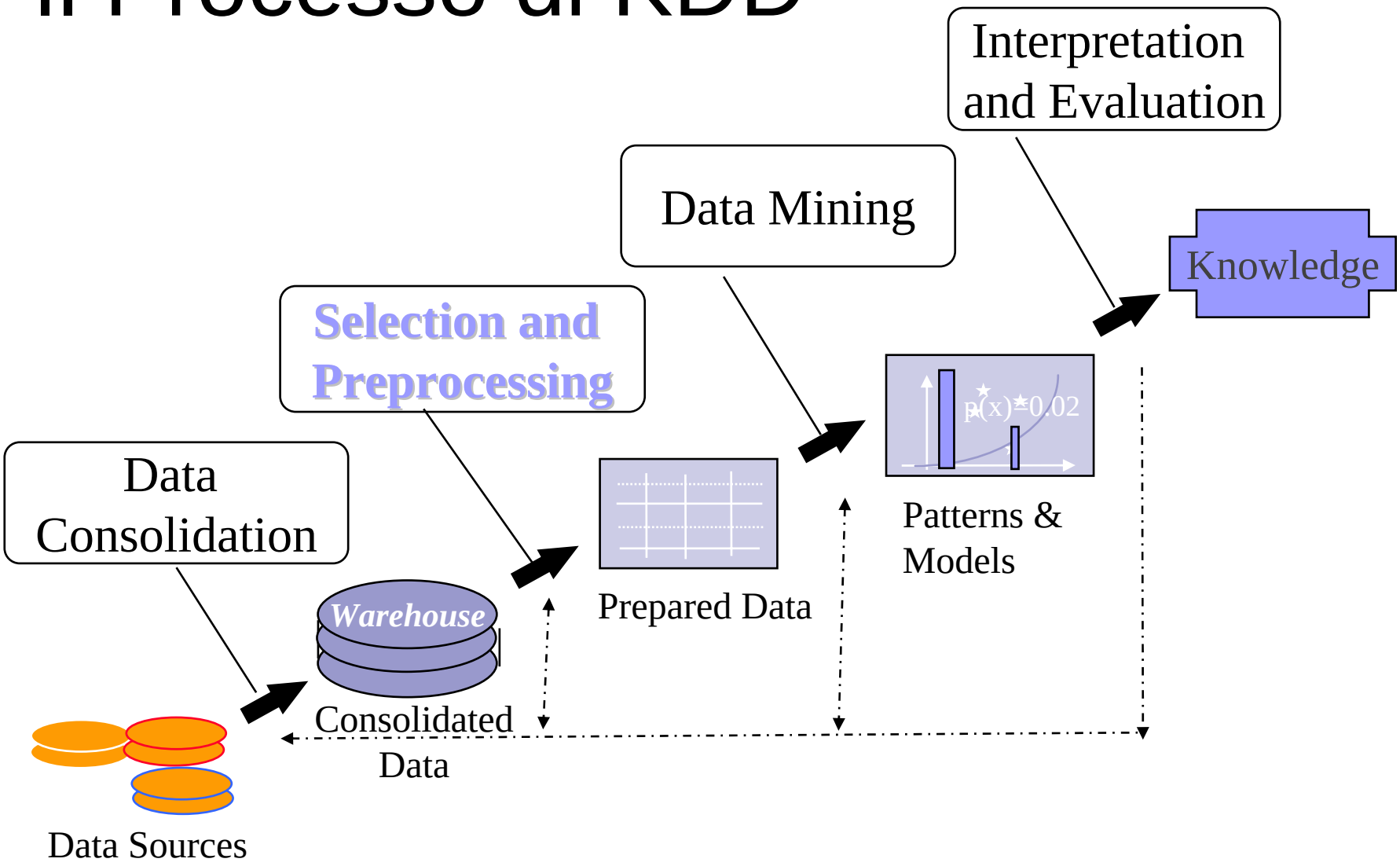
Preparazione dati

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

Outline

- 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

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

Un tool Fondamentale: le query

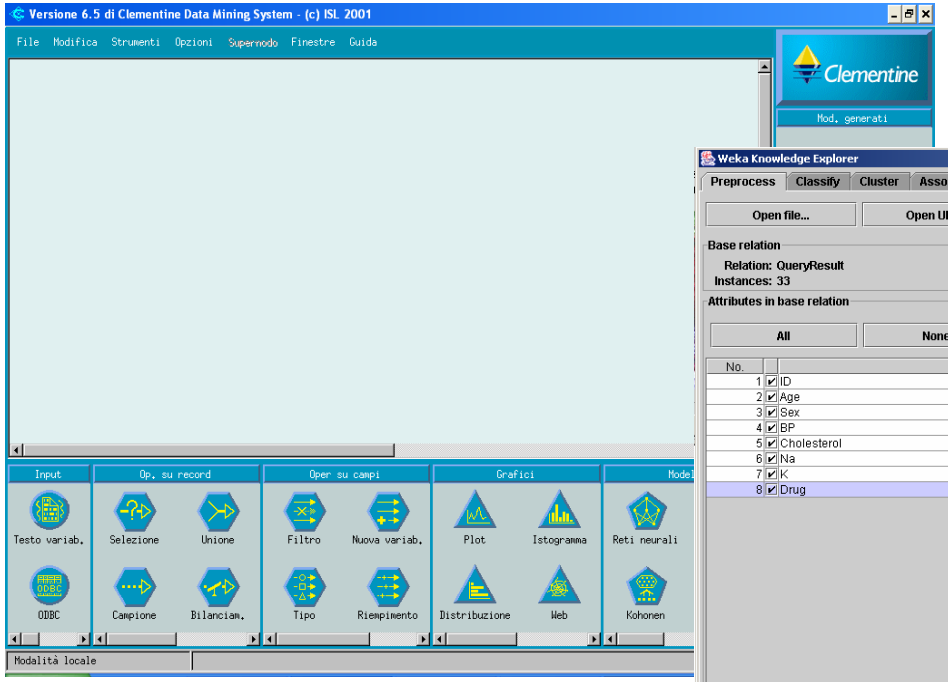
- Base di partenza: un data-mart
- Dal data-mart estraiamo una tabella
- Le informazioni sulla tabella permettono di effettuare data preprocessing
  - Selezione dati: SELECT
  - Aggiornamento dati: UPDATE e DELETE

È sempre necessario SQL?

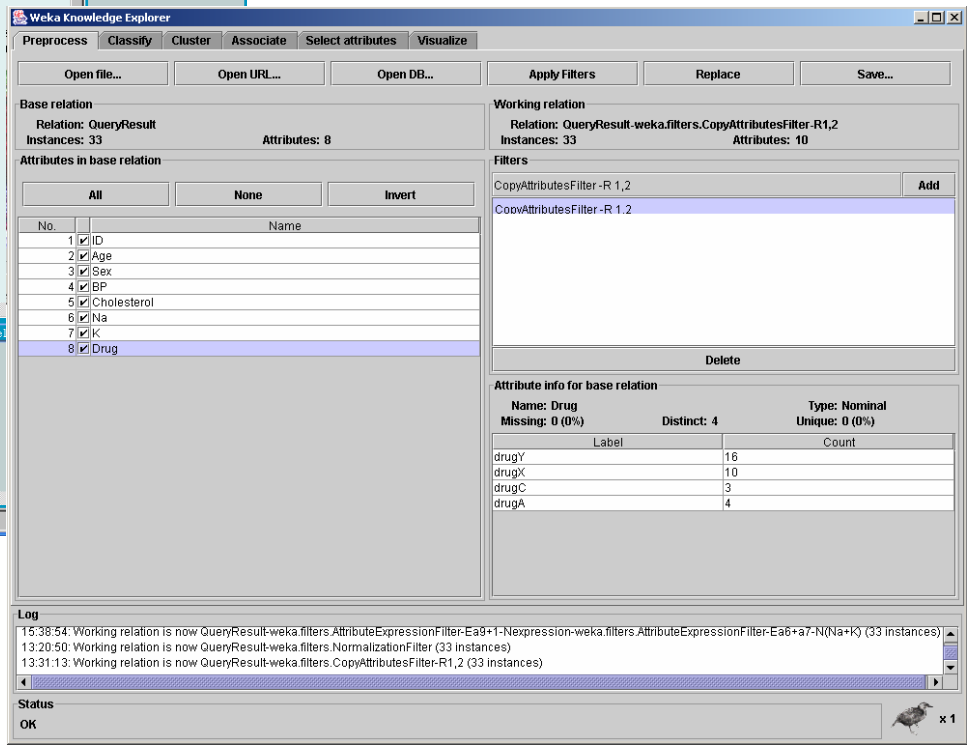
- I moderni tool raggruppano una serie di operazioni in maniera uniforme
- La metafora di interazione è visuale
  - Esempi:
    - Clementine
    - Weka
- SQL è più generico
  - Ma anche più difficile da usare

Overview di due strumenti

Clementine

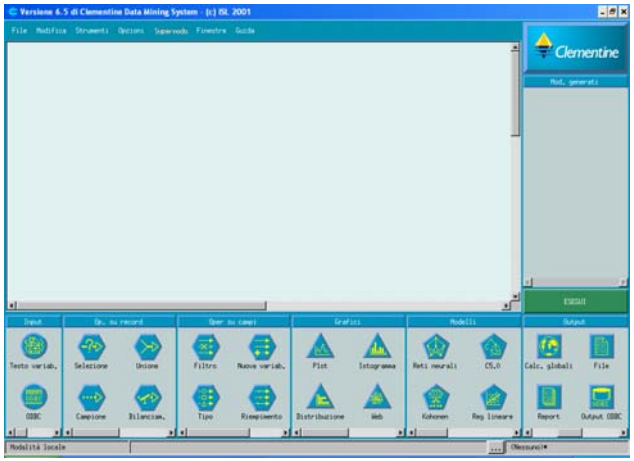


Weka



Gli strumenti: Clementine

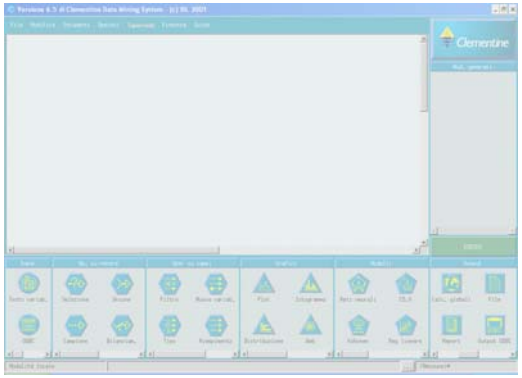
- Ambiente grafico intuitivo



- Processo = flusso di dati (stream):
  - Parte da nodi sorgente → 
  - Attraversa nodi di trasformazione → 
  - Arriva a nodi terminali →  

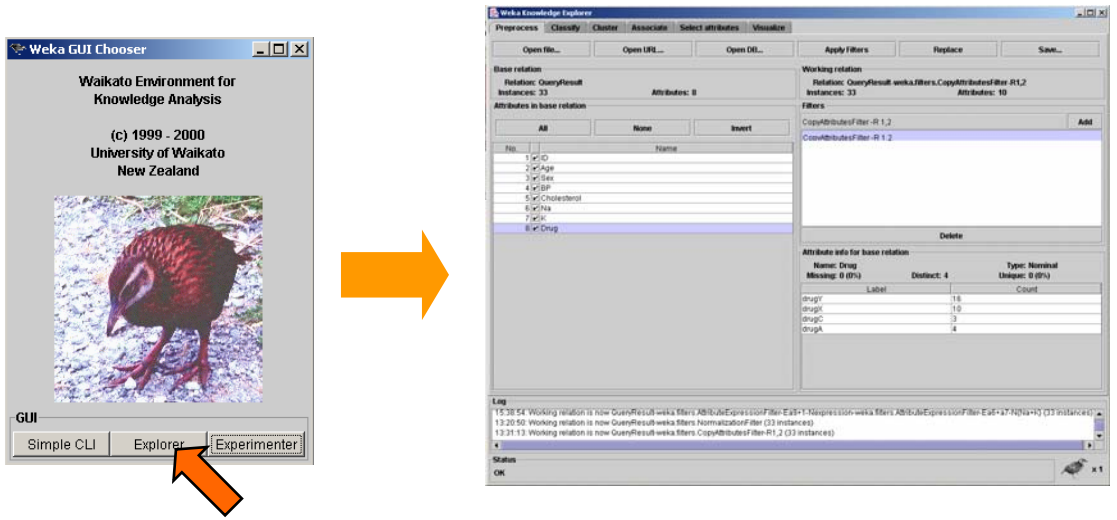
Clementine

- Tool di Data Mining
  - Nodi per la generazione di modelli → 
  - Nodi per i modelli scoperti → 



Gli Strumenti: Weka

- Libreria Java Open Source ricca di tool per il preprocessing e il Data Mining
- Interfaccia grafica semplificata: Explorer

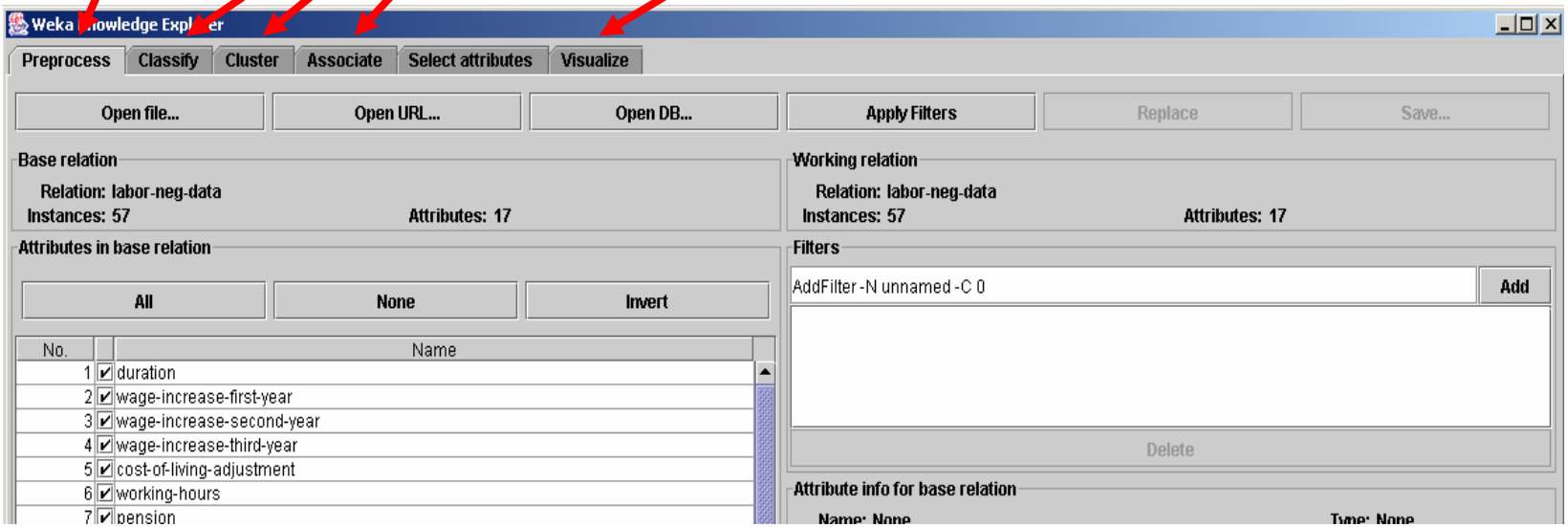


Weka: le 3 fasi del processo

1. Pannello per caricamento dati e preprocessing

2. Pannelli per data mining

3. Pannello per visualizzazione (dot diagrams)



SQL: Selezione tuple

- 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: Selezione attributi

- Consente anche la rinomina delle colonne
- Tabella coinvolta:

  Beers(name, manf)
- Query:

```
SELECT name AS beer
FROM Beers
```

- Risposta:

beer
Bud
Bud Lite
Michelob

SQL: Attributi derivati

- 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
…	…	…

(Inner) Join

- Query che coinvolgono valori correlati in due tabelle diverse
- Tabelle coinvolte:

  Likes(drinker, beer)

  Frequents(drinker, bar)
- Query: SELECT drinker, beer, bar

  FROM Frequents, Likes

  WHERE Frequents.drinker = Likes.drinker

Query su più relazioni

- Esempio: selezione (join vincolata)

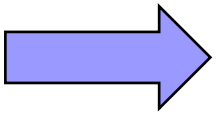
- Trova le birre che piacciono ai frequentatori del bar “Joe’s”
- Query: SELECT beer
  FROM Frequents, Likes
  WHERE bar = “Joe’s Bar” AND
  Frequents.drinker = Likes.drinker

Risposte multiple

- Le risposte sono “bags”

  SELECT beer

  FROM Sells

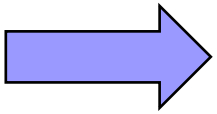


beer
Bud
Miller
Bud
…

- Possiamo comunque utilizzare la parola chiave DISTINCT

  SELECT DISTINCT beer

  FROM Sells



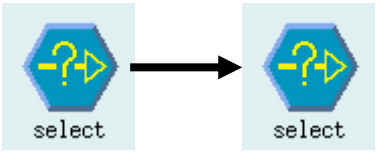
beer
Bud
Miller
…

Unioni di query

- 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)
```

Subquery



- I risultati possono essere annidati

```
SELECT *
FROM (
        SELECT beer
        FROM Likes
        WHERE drinker = ‘Fred’
      )
WHERE price < 100
```

Aggregati

- Trova il prezzo medio della “Bud”

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

- Possiamo aggiungere in fondo al costrutto la parola chiave GROUP BY e una lista di attributi

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

Ordinamento

- Ordina il risultato della query secondo un attributo:

```
SELECT beer
FROM Likes
ORDER BY Price
```

Outline

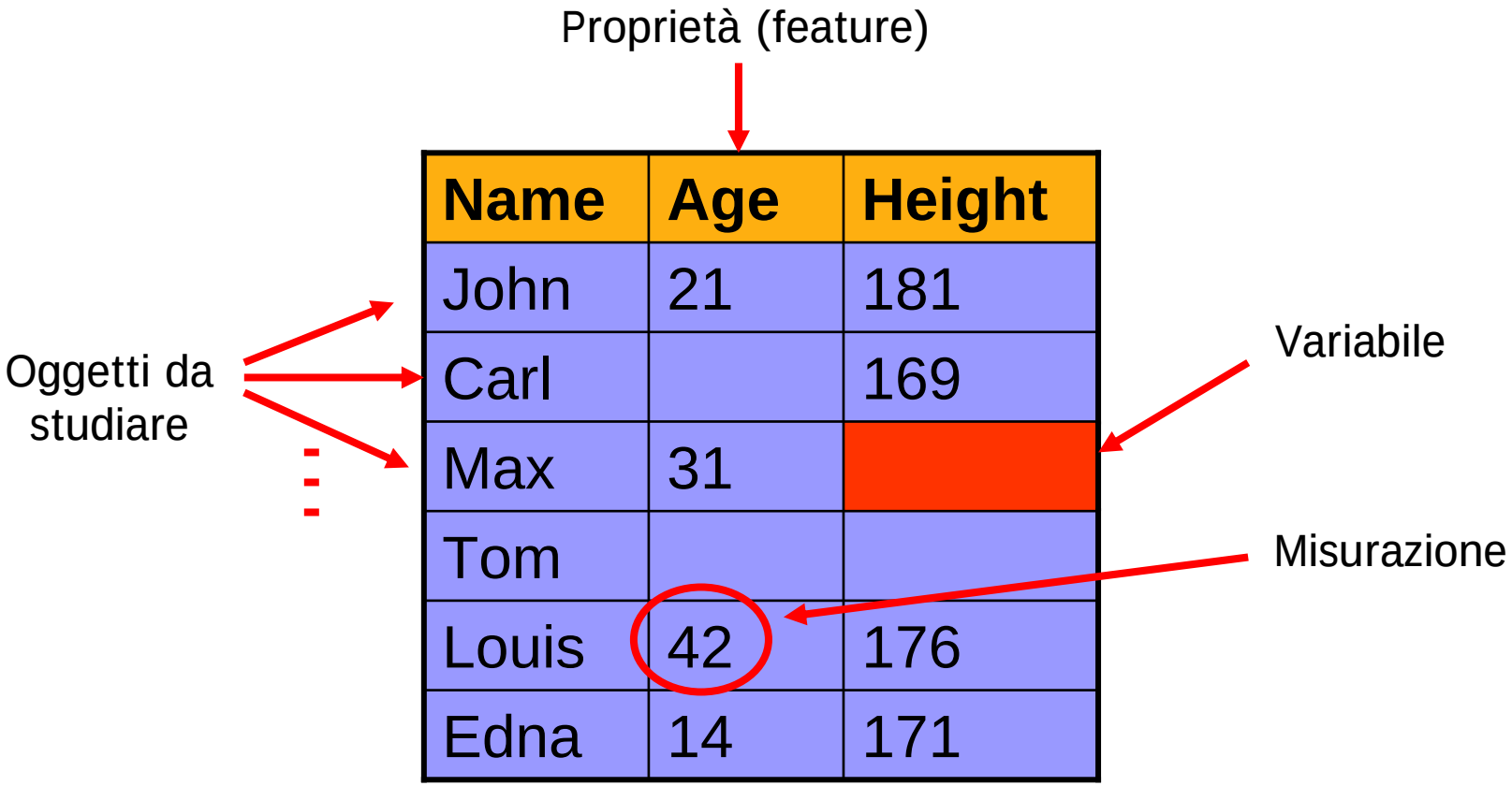
- Introduzione e Concetti di Base
- Data Selection
- Information Gathering
- Data cleaning
- Data reduction
- Data transformation

Oggetti, Proprietà, Misurazioni

- Il mondo reale consiste di oggetti
  - Automobili, Vigili, Norme, …
- Ad ogni oggetto 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 è descritta da una tabella



Tipi di misure

- Misure Discrete (simboliche)
  - Nominali → identificatori univoci (Cod. Fiscale)
  - Categoriche → “etichette” ripetibili (Città)
  - Ordinali → è definito un ordine (low < high)
  - Binarie → due soli valori (T/F, 1/0,...)
- Misure Continue
  - Interval-Based → Scalabili di fattore costante (es.: misure in MKS e CGS)
  - Ratio-Scaled → Scalabili linearmente (ax+b) (es.: temperature °C e °F)

Caratteristiche delle 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
- Outlier
  - 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

Bias

- Un fattore esterno significativo e rilevante nei dati
  - Comporta problemi (espliciti o impliciti) nei dati
  - Molti valori 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

Descrizione dei dati

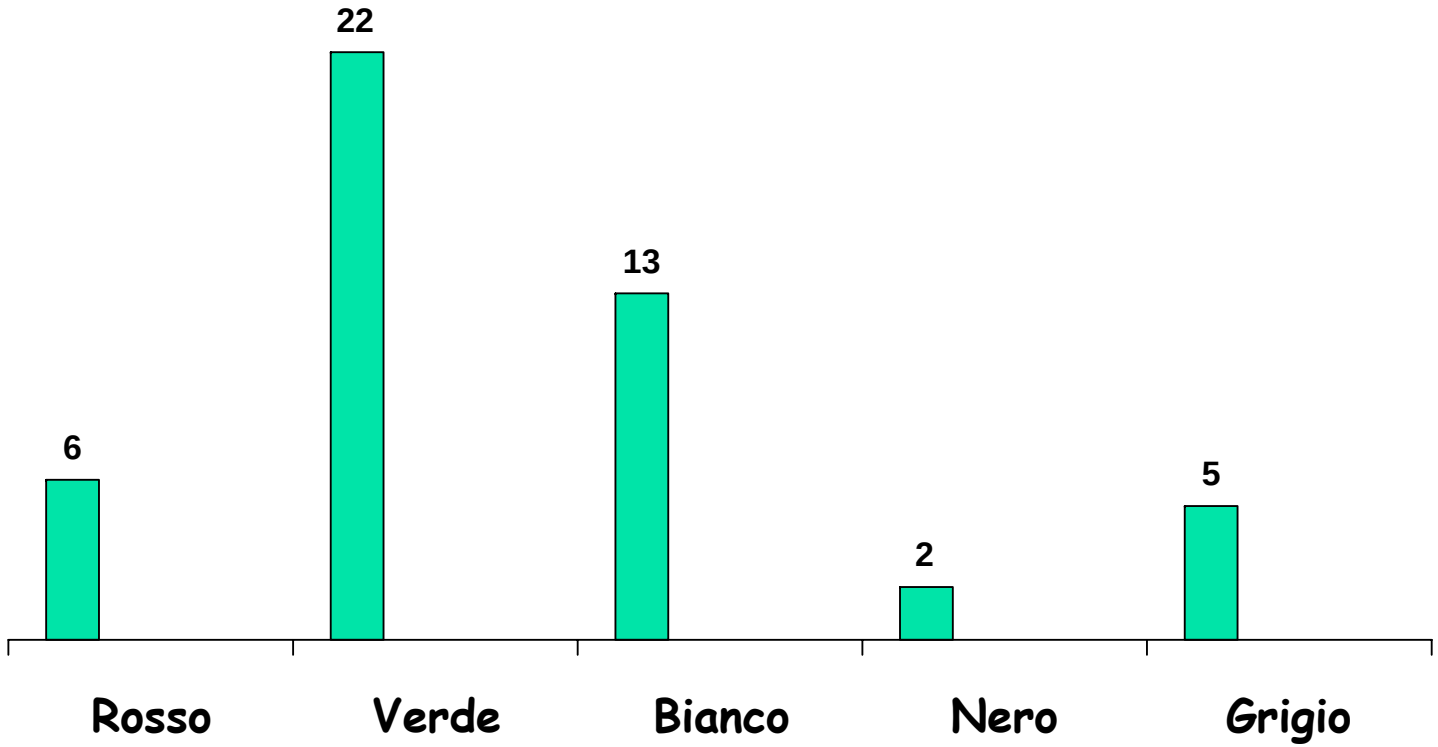
- Grafici
  - Distribuzione frequenze
  - Correlazione
  - Dispersione
- Misure
  - Media, mediana, quartili
  - Varianza, deviazione standard
  - Forma, simmetria, curtosi

Visualizzazione dati qualitativi

- Rappresentazione delle frequenze
  - Diagrammi a barre
  - Ortogrammi
  - Aerogrammi
- Correlazione
  - Web diagrams
- Ciclicità
  - Diagrammi polari

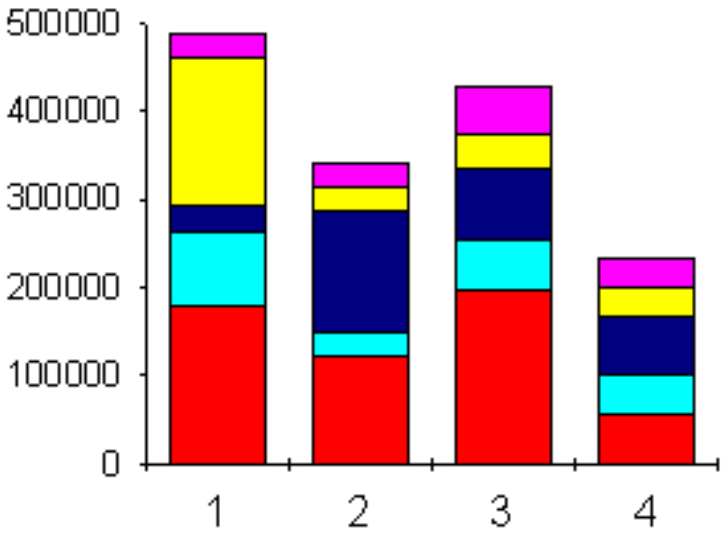
Diagrammi di Pareto

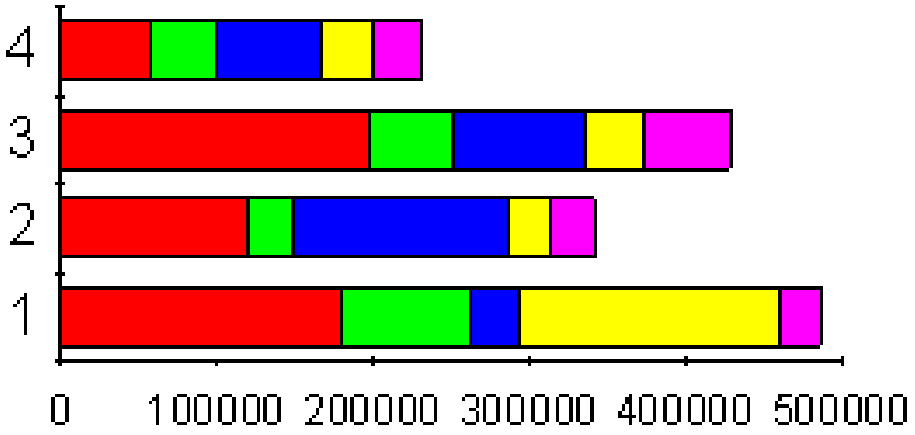
- Diagrammi a barre distanziate
- Un assortimento di eventi presenta pochi picchi e molti elementi comuni



Ortogrammi

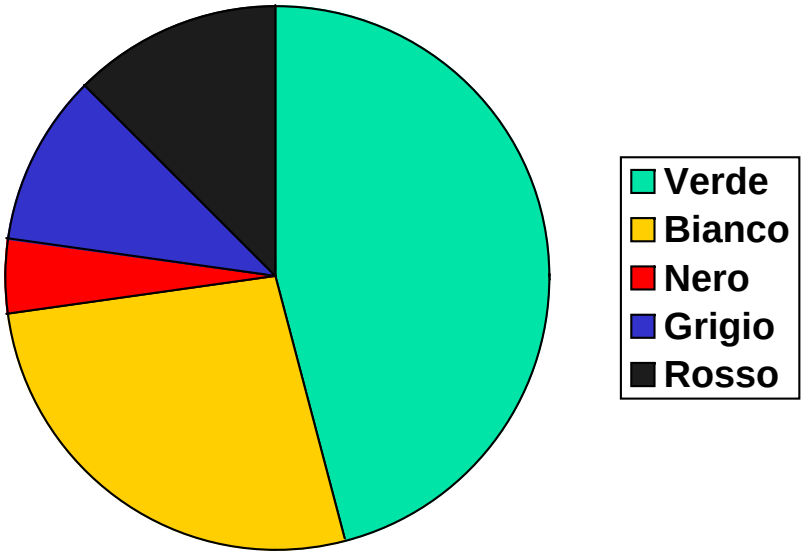
- Ogni colonna indica la la distribuzione interna per un dato valore e la frequenza





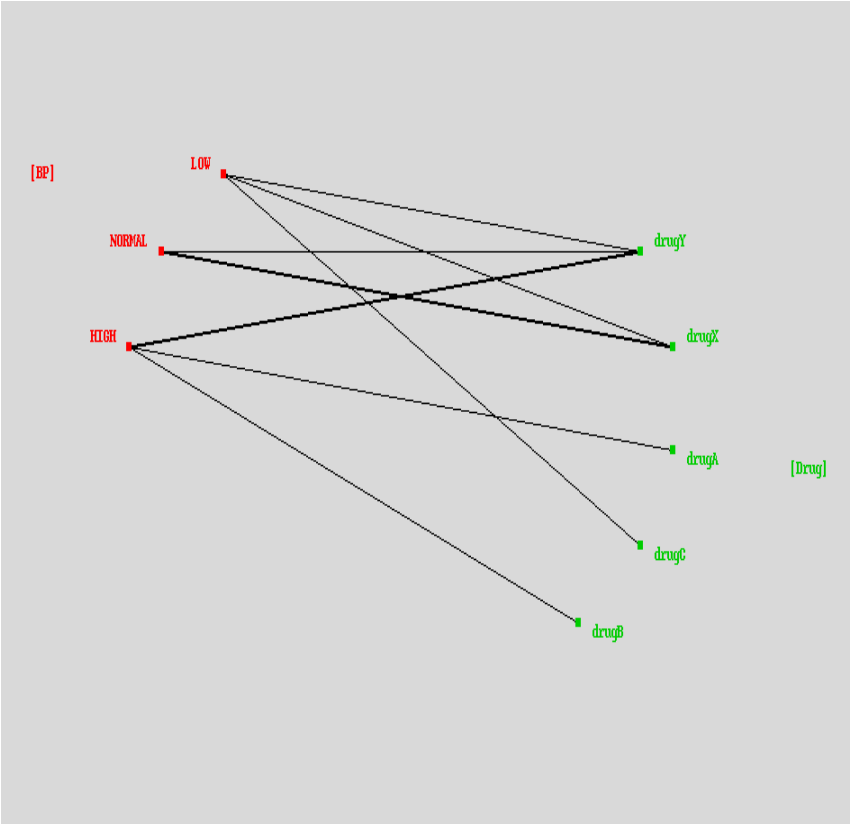
Aerogrammi

- Rappresentazioni a torta
- Frequenza della distribuzione



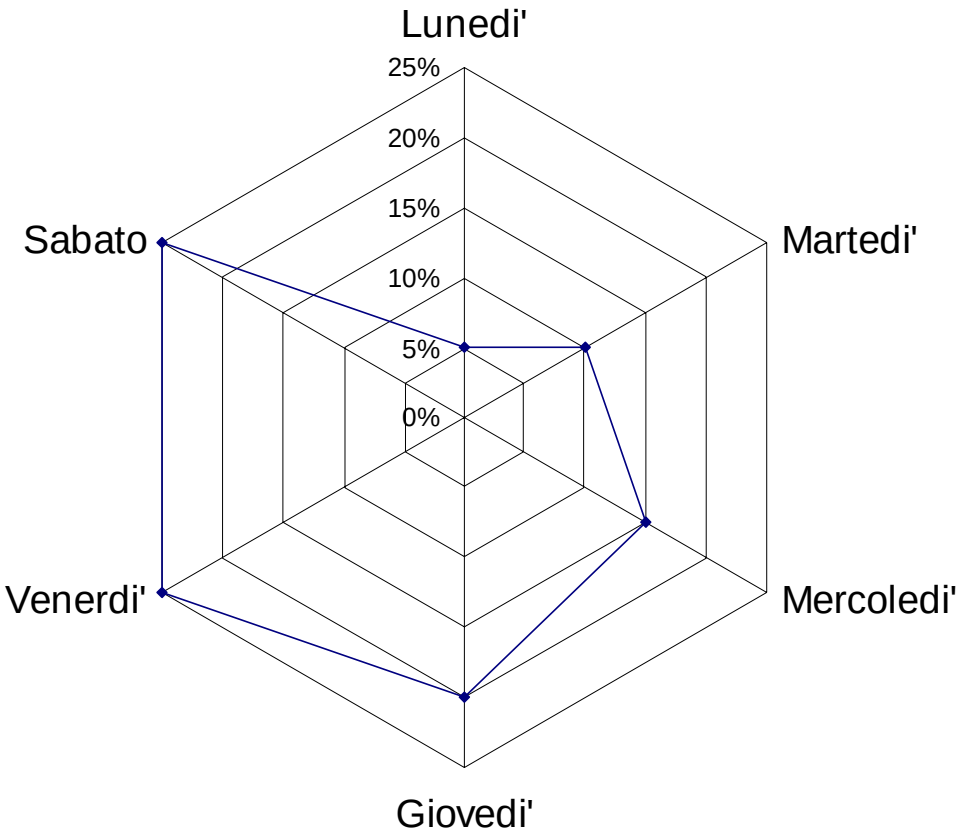
Web

- Visualizzano correlazioni tra valori simbolici



Diagrammi polari

- Rappresentano fenomeni ciclici
  - Es., concentrazione delle vendite nell’arco settimanale

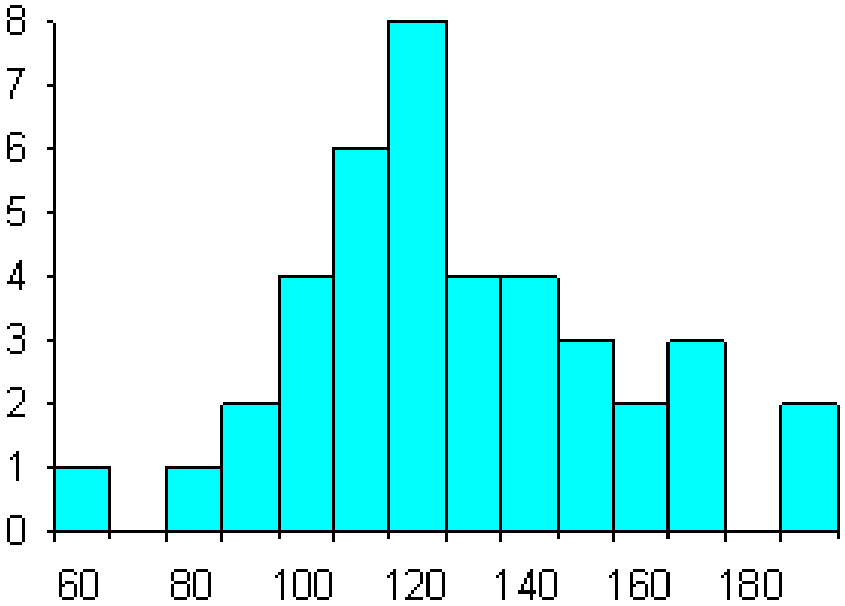


Dati Quantitativi

- Istogrammi
- Poligoni
- Stem and leaf
- Dot Diagrams
- Diagrammi quantili

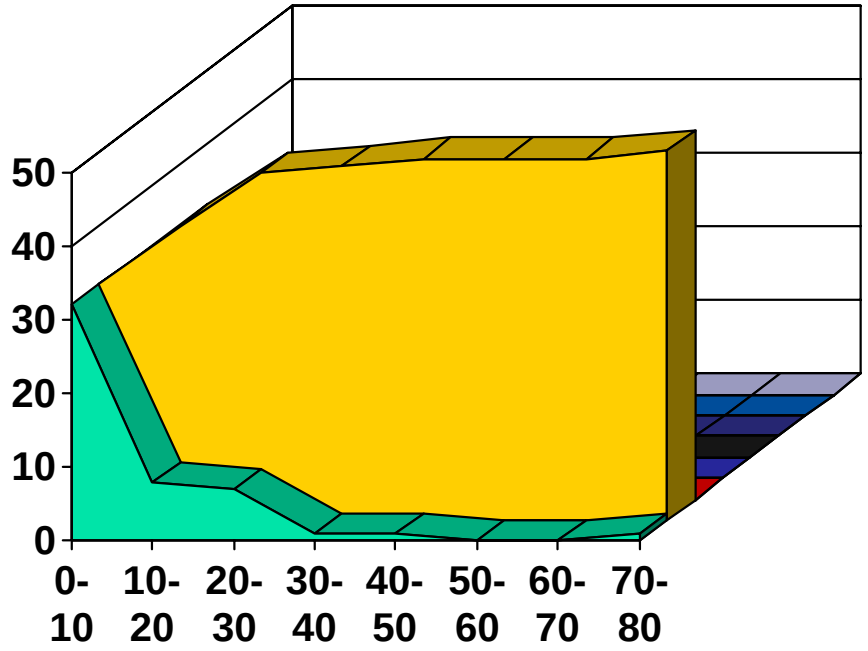
Istogrammi

- Rappresentazioni a barre
- Evidenziano la frequenza su intervalli adiacenti
  - La larghezza di ogni rettangolo misura l'ampiezza degli intervalli



Poligoni

- Per la descrizione di frequenze cumulative
- I punti sono uniti tramite linee



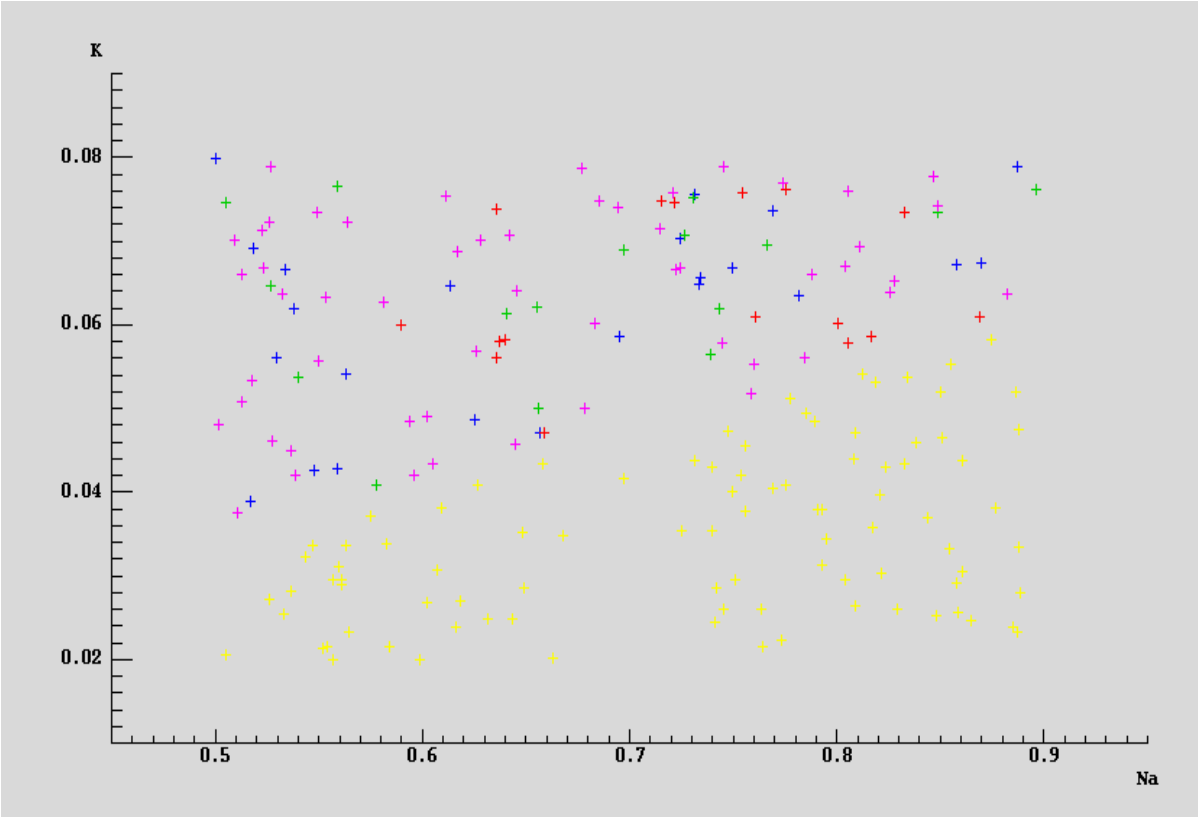
Rappresentazione “Stem & Leaf”

- Simile a istogrammi
- Evita la 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

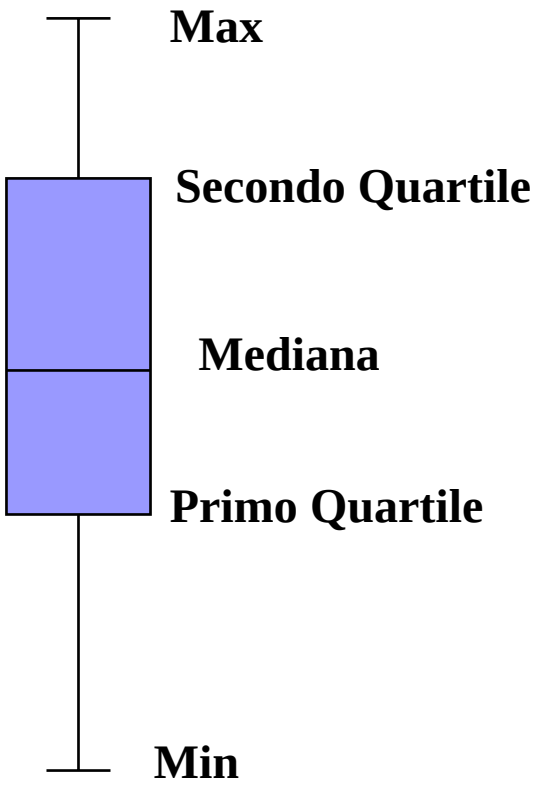
Dot Diagrams, Scatters

- Visualizza la Dispersione



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



Misure descrittive dei dati

- Tendenza centrale o posizione
  - Media aritmetica, geometrica e armonica, mediana, quartili, percentili, moda
- Dispersione o variabilità
  - Range, scarto medio, varianza, deviazione standard
- Forma della distribuzione
  - Simmetria (medie interquartili, momenti centrali, indice di Fisher)
  - Curtosi (indice di Pearson, coefficiente di curtosi)

Outline

- 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

Trattamento di valori nulli

- Eliminazione delle tuple
- Sostituzione dei valori nulli
  - N.B.: può influenzare la distribuzione dei dati numerici
  - Utilizzare media/mediana/moda
  - Predire i valori mancanti utilizzando la distribuzione dei valori non nulli
  - Segmentare i dati e utilizzare misure statistiche (media/moda/mediana) di ogni segmento
  - Segmentare i dati e utilizzare le distribuzioni di probabilità all'interno dei segmenti
  - Costruire un modello di classificazione/regressione e utilizzare il modello per calcolare i valori nulli

Outline

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

Data Reduction

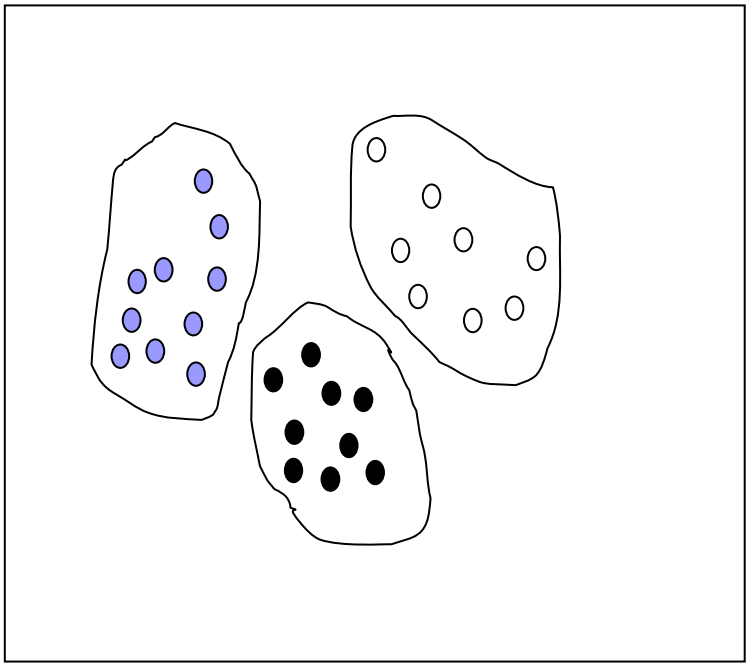
- Riduzione del volume dei dati
  - Verticale: riduzione numero di tuple
    - Data Sampling
    - Clustering
  - Orizzontale: riduzione numero di colonne
    - Seleziona un sottoinsieme di attributi
    - Crea un nuovo (e piccolo) insieme di attributi

Sampling (riduzione verticale)

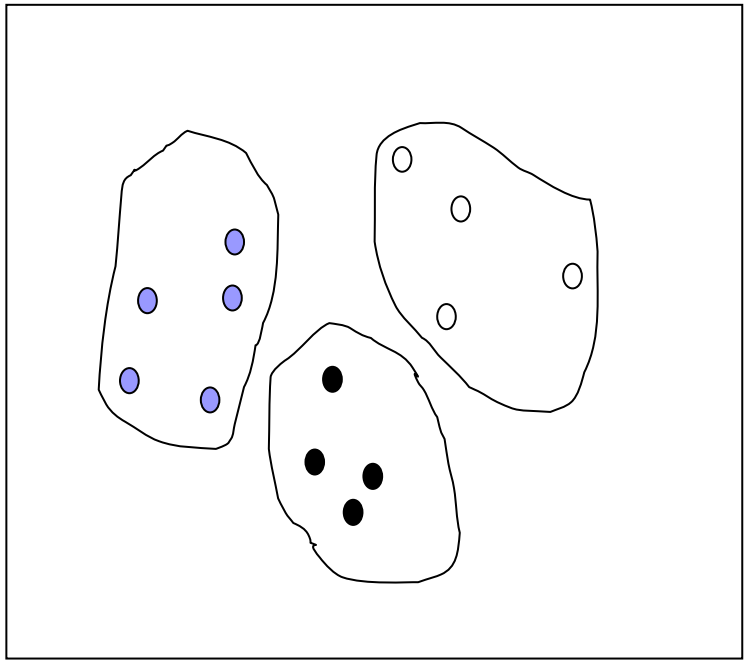
- Riduce la complessità di esecuzione degli algoritmi di Mining
- Problema: scegliere un sottoinsieme rappresentativo dei dati
  - La scelta di un campionamento casuale può essere problematica per la presenza di picchi
- Alternative: Schemi adattativi
  - *Stratified sampling*:
    - Approssimiamo la percentuale di ogni classe (o sottopopolazione di interesse rispetto all'intero database)
    - Adatto a distribuzioni con picchi: ogni picco è in uno strato
  - Possiamo combinare le tecniche random con la stratificazione
- N.B.: Il Sampling potrebbe non ridurre I tempi di risposta se i dati risiedono su disco (page at a time).

Sampling

Raw Data



Cluster/Stratified Sample



Riduzione Dimensionalità (Riduzione orizzontale)

- Selezione di un sottoinsieme di attributi
  - Manuale
    - In seguito a analisi di significatività e/o correlazione con altri attributi
  - Automatico
    - Selezione incrementale degli attributi “migliori”
    - “Migliore” = rispetto a qualche misura di significatività statistica (es.: information gain).

Riduzione Dimensionalità (Riduzione orizzontale)

- Creazione di nuovi attributi con i quali rappresentare le tuple
  - Principal components analysis (PCA)
    - Trova le combinazioni lineari degli attributi nei *k* vettori ortonormali più significativi
    - Proietta le vecchie tuple sui nuovi attributi
  - Altri metodi
    - Factor Analysis
    - Decomposizione SVD

Outline

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

Data Transformation: Motivazioni

- Dati con errori o incompleti
- Dati mal distribuiti
  - Forte asimmetria nei dati
  - Molti picchi
- La trasformazione dei dati può alleviare questi problemi

Obiettivi

- Vogliamo definire una trasformazione *T* sull'attributo *X*:

$$Y = T(X)$$

  tale che:

  - *Y* preservi l'informazione "rilevante" di *X*
  - *Y* elimini almeno uno dei problemi di *X*
  - *Y* sia più "utile" di *X*

Obiettivi

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

Perché normalità, linearità, ecc.?

- Molte metodologie statistiche richiedono correlazioni lineari, distribuzioni normali, assenza di outlier
- Molti algoritmi di Data Mining hanno la capacità di trattare automaticamente non-linearità e non-normalità
  - Gli algoritmi lavorano comunque meglio se tali problemi sono assenti

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’interpretabilità

- Trasformazioni lineari

$$1€ = 1936.27 \text{ Lit.}$$

  - 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 log-normali
  - es.: 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	Mediana
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
Mediana	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

Asimmetria dei dati

- Simmetria e Media interpercentile

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

- Se la media interpercentile è sbilanciata, allora la distribuzione dei dati è asimmetrica
  - sbilanciata a destra $\bar{x}_p > M$
  - sbilanciata a sinistra $\bar{x}_p < M$

Asimmetria nei dati: esempio

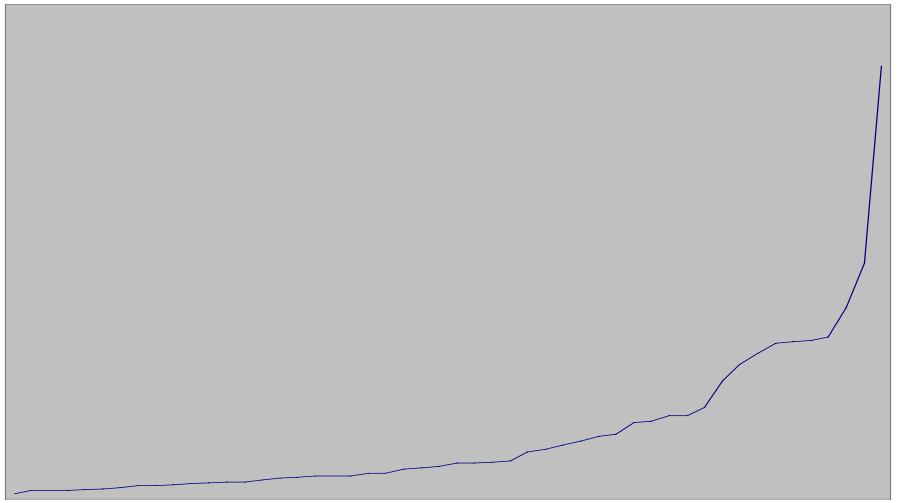
- Verifichiamo la simmetria (valori di un unico attributo)

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

Asimmetria: 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



Creare simmetria nei dati: Trasformation plot

- Trovare una trasformazione T_p che crei simmetria
  - Consideriamo i percentili x_U e x_L
  - I valori c ottenuti tramite la formula

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

suggeriscono dei valori adeguati 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

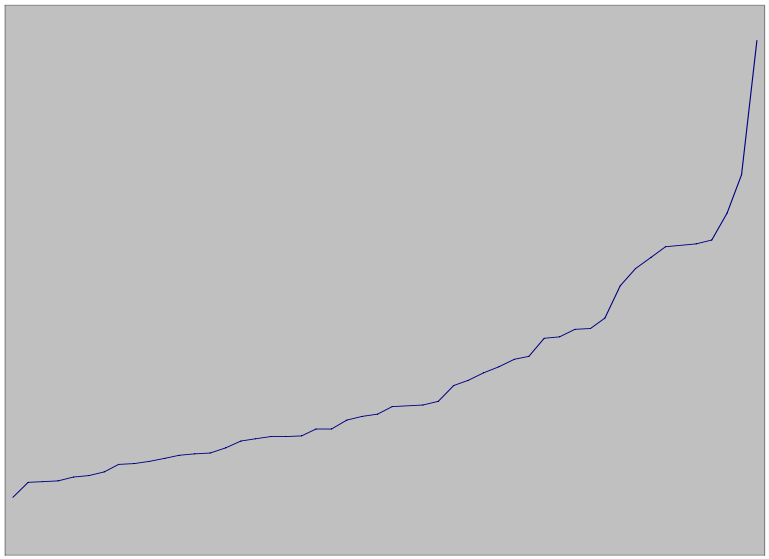
$(x_L-x_U)/2-M$	$((M- x_L)^2+(x_U -M)^2)/4M$	c
2844.5	3317.5	0.14258
6341	11652.8	0.45583
9262.7	21338.8	0.56592
10564.3	26292.5	0.59820

- Calcolando la mediana dei valori *c* otteniamo $p=0.5188$
- Proviamo con $p=1/2$...

Trasformazione 1: radice quadrata

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

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	

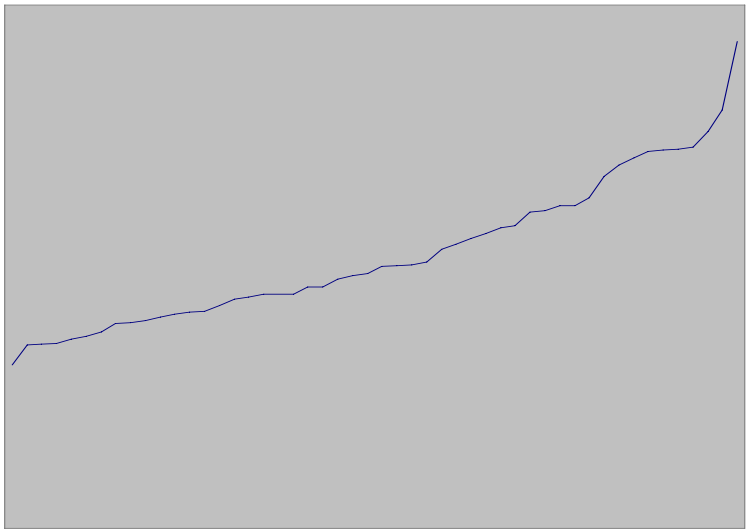


- La curva si tempera, ma i valori alti continuano a produrre differenze notevoli
- Proviamo a diminuire *p*…

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	

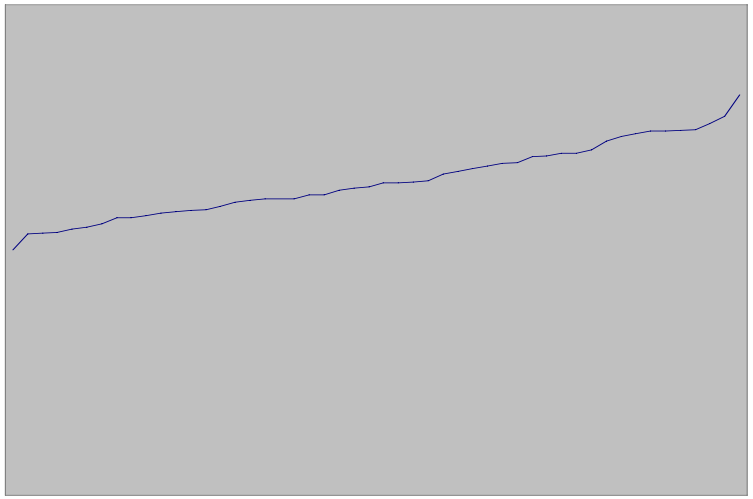


- I valori alti continuano ad influenzare
- Proviamo con il logaritmo…

Trasformazione 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 attributi

- Esempio: caso della regressione
  - La formula

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

può 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

- Caratteristiche:
  - Non etichetta le istanze
  - Il numero di classi è noto a priori
- Tecniche di binning:
  - Natural binning
    → Intervalli di identica ampiezza
  - Equal Frequency binning
    → Intervalli di identica frequenza
  - Statistical binning
    → Uso di informazioni statistiche
    (Media, varianza, Quartili)

Quante classi?

- Se troppo poche
  - perdita di informazione sulla distribuzione
- Se troppe
  - disperde i valori e non manifesta la “forma” 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

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

- Può produrre distribuzioni molto sbilanciate

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]

Equal Frequency Binning

- Ordino e conto gli elementi,
  quindi definisco *k* intervalli di *f* elementi, dove:

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

- L'elemento x_j appartiene alla classe *i* se:

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

- Non sempre adatta ad evidenziare correlazioni interessanti

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}

Supervised Discretization

- Caratteristiche:
  - La discretizzazione ha un obiettivo quantificabile
  - Il numero di classi non è noto a priori
- Tecniche:
  - ChiMerge
  - Discretizzazione basata sull’entropia
  - Discretizzazione basata sui percentili

Supervised Discretization: ChiMerge

- Procedimento Bottom-up:
  - Inizialmente, ogni valore è un intervallo a sé
  - 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} ($R_i * C_j / N$)

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 due intervalli sono “distinti”
- k-1 gradi di libertà
- La significatività del test è data da una soglia δ
  - Probabilità 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 rispetto a Beer
- soglia 50% confidenza
- Vogliamo ottenere una discretizzazione del prezzo che permetta di mantenere omogeneità su Beer

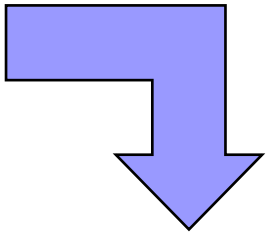
Esempio: valori di Chi

- Scegliamo i elementi adiacenti con Chi-Value minimo

	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

Esempio: passo 1

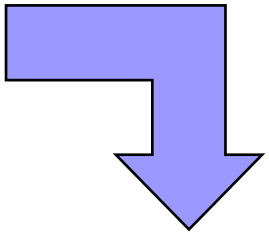
	Bud	Becks	Chi Value
100	1	0	0
110	2	0	1.33333
120	1	1	0.75
125	1	0	0
130	2	0	0
135	1	0	2
140	0	1	0
150	0	1	0
160	0	1	1.38629



	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

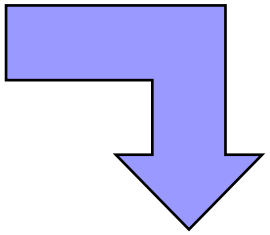
	Bud	*Becks*	*Chi Value*
100	1	0	0
110	2	0	1.33333
120	1	1	0.75
125	1	0	0
130	2	0	0
135	1	0	2
140	0	1	0
150-160	0	2	1.38629



	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

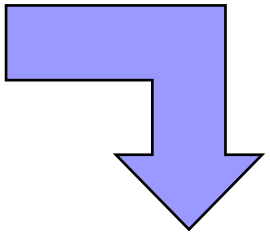
	Bud	*Becks*	*Chi Value*
100	1	0	0
110	2	0	1.33333
120	1	1	0.75
125	1	0	0
130	2	0	0
135	1	0	4
140-150-160	0	3	1.38629



	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

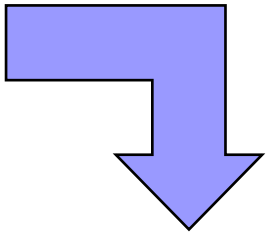
	Bud	Becks	Chi Value
100	1	0	0
110	2	0	1.33333
120	1	1	0.75
125	1	0	0
130-135	3	0	6
140-150-160	0	3	1.38629



	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
(min = 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

Appendice

Misure descrittive dei dati

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

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

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

- Per bilanciare proporzioni
- Dati moltiplicativi

$$x_g = 100$$

- 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
- Es., 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

$$1\ 7\ 12\ 18\ 23\ 34\ 54$$

$$\bar{x} = 21.3$$

$$M = 23$$

Mediana e Quartili

- Divide un insieme di dati a metà
  - 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

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

- 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, $S_{.5} \leq S_n$

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

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

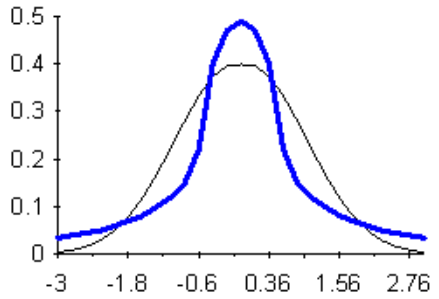
$$\bar{x}_p = (x_{1-p} + x_p)/2$$

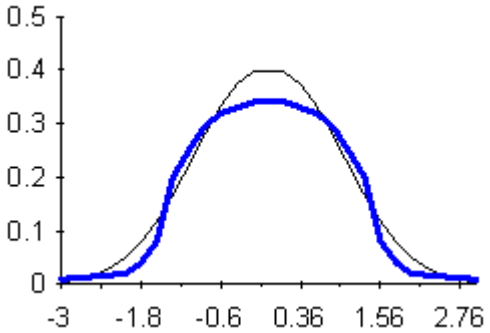
$$m_k = \frac{1}{n-1}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^k$$

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

Curtosi

- Grado di appiattimento della curva di distribuzione rispetto alla curva normale
  - mesocurtica: forma uguale alla distribuzione normale;
  - leptocurtica: una frequenza minore delle classi intermedie, frequenza maggiore delle classi estreme e dei valori centrali;
  - platicurtica: una frequenza minore delle classi centrali e di quelle estreme, con una frequenza maggiore di quelle intermedie
    - numero più ridotto di valori centrali.





Curtosi (cont.)

- Indice di Pearson
  - $\beta=3$: distribuzione mesocurtica
  - $\beta > 3$: distribuzione leptocurtica
  - $\beta < 3$: distribuzione platicurtica

$$\beta = \frac{m_4}{\hat{s}^4}$$

- Coefficiente di curtosi
  - Una distribuzione leptocurtica ha $K \sim 1/2$
  - platicurtosi: $K \sim 0$

$$K = \frac{\frac{1}{2}\left(x_{.75} - x_{.25}\right)}{\left(x_{.90} - x_{.10}\right)}$$

Coefficienti di Correlazione

- Covarianza

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

- Coefficiente di Pearson

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