

Preparazione e caratteristiche dei Dati per Data Mining

Fosca Giannotti
f.giannotti@isti.cnr.it

Dino Pedreschi
dino.pedreschi@unipi.it

ISTI-CNR Pisa & Università di Pisa



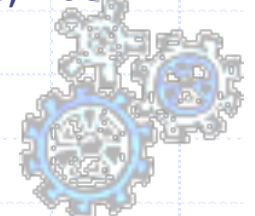
Materiale

◆ **Lucidi delle lezioni** (Slides PowerPoint):

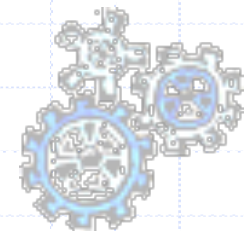
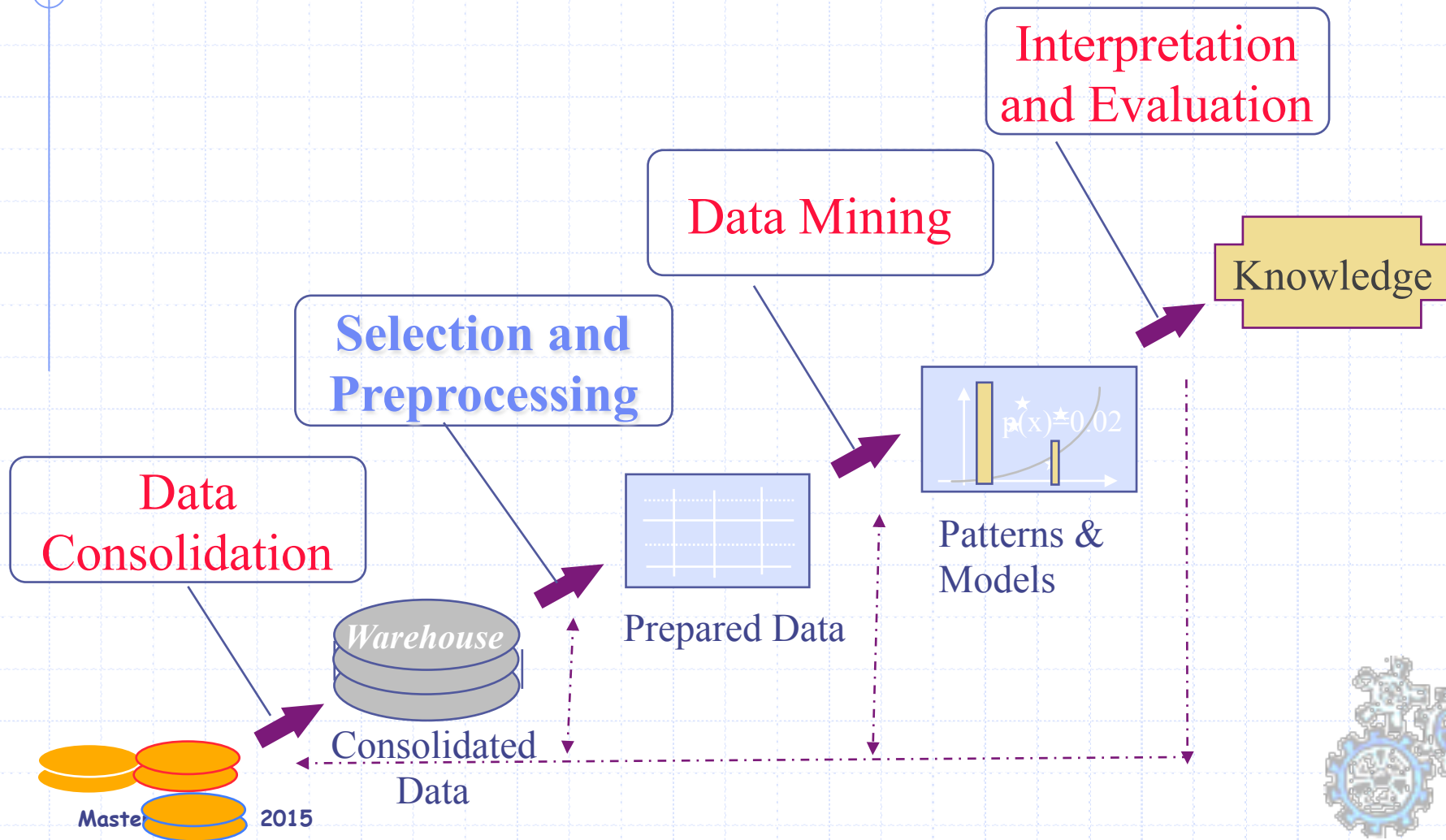
- Primo autore: G. Manco Revisione: M. Nanni
- Versione attuale: In distribuzione

◆ **Testi di Riferimento**

- J. Han, M. Kamber. ***Data Mining: Concepts and Techniques***. Morgan Kaufmann, 2000.
- Dorian Pyle. ***Data Preparation for Data Mining***. Morgan Kaufmann, 1999.
- D. Hand, H. Mannila, P. Smyth. ***Principles of Data Mining***. MIT Press, 2001.



Il Processo di KDD



Data Sources

Preparazione di Dati per Data Mining

I Contenuti

◆ 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

◆ Data Similarity

- Similarity and Dissimilarity (on Single attribute)
- Distance (Many attributes)
- Distance on Binary data (Simple matching; Jaccard)
- Distance on Document Data

- Data Exploration (multidimensional array)



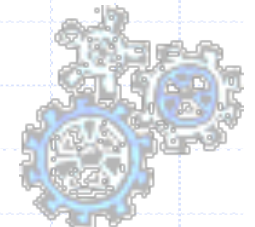
Problemi tipici

◆ Troppi dati

- dati sbagliati, rumorosi
- dati non rilevanti
- dimensione intrattabile
- mix di dati numerici/simbolici

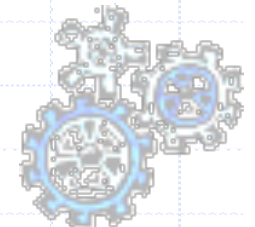
◆ Pochi dati

- attributi mancanti
- valori mancanti
- dimensione insufficiente



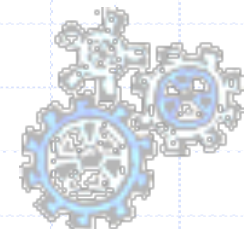
Il Data Preprocessing è un Processo

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



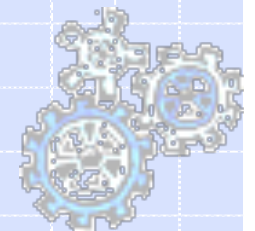
Il Data Preprocessing dipende (ma non sempre) dall' Obiettivo

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



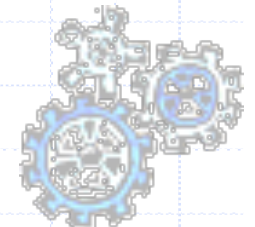
Outline del Modulo

- ◆ Introduzione e Concetti di Base
- ◆ Data Selection ←



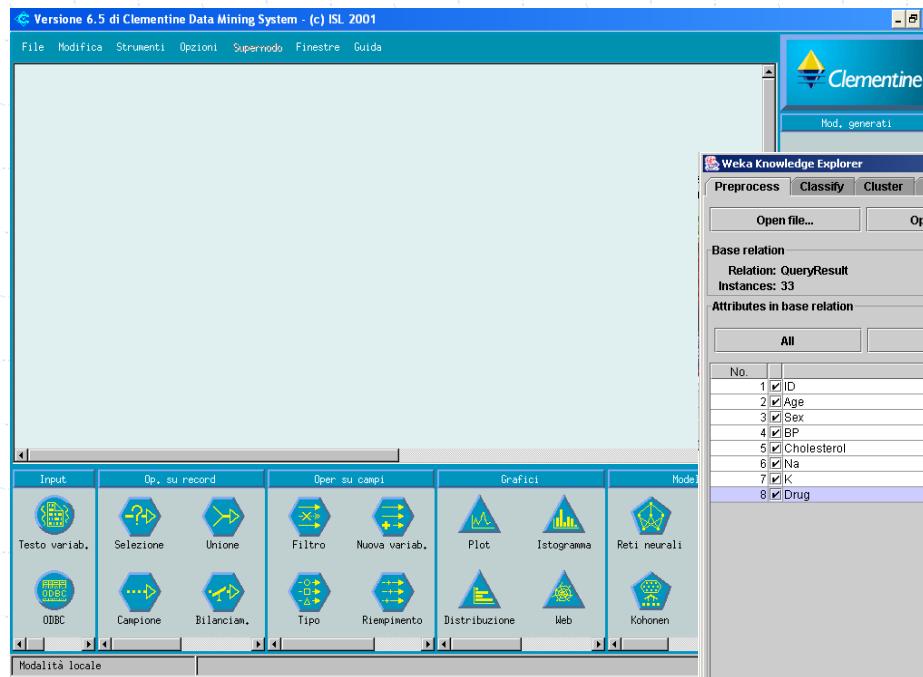
E' sempre necessario SQL?

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

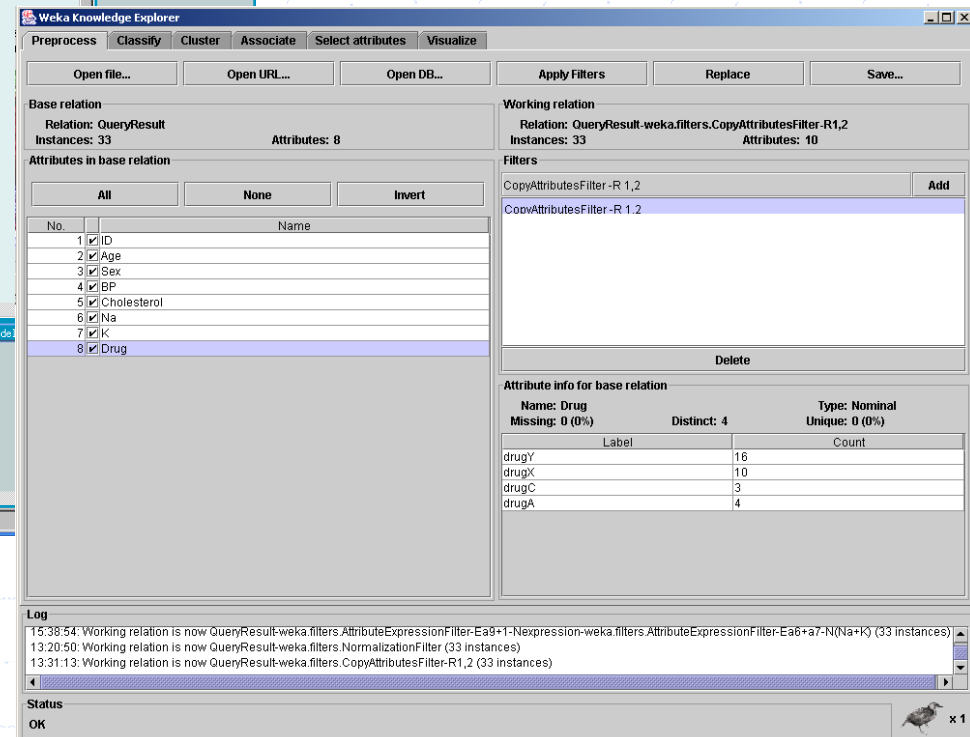


Es. due piattaforme per DM

Clementine



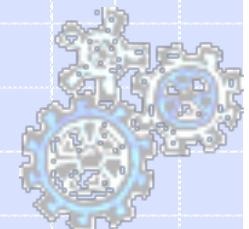
Weka



Master MAINS, 2015

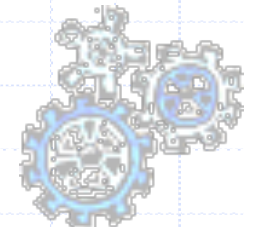
Outline del Modulo

- ◆ Introduzione e Concetti di Base
- ◆ Data Selection
- ◆ **Information Gathering** ←



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**

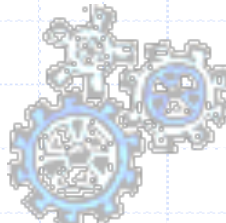
Proprietà (feature)

Name	Age	Height
John	21	181
Carl		169
Max	31	
Tom		
Louis	42	176
Edna	14	171

Oggetti da studiare

Variabile

Misurazione



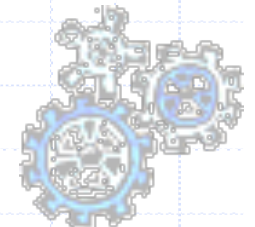
Tipi di misure

◆ Misure Discrete (simboliche, categoriche, qualitative)

- Nominali → identificatori univoci (Cod. Fiscale)
- Ordinali → è definito un ordine (low < high)
- Binarie → due soli valori (T/F, 1/0,...)

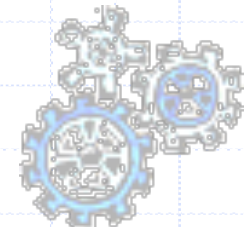
◆ Misure Continue

- Interval-Based → Scalabili di fattore costante (es.: misure in MKS e CGS)
- Ratio-Scaled → Scalabili linearmente ($ax+b$) (es.: temperature °C e °F)



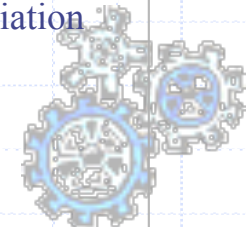
Properties of Attribute Values

- ◆ The type of an attribute depends on which of the following properties it possesses:
 - Distinctness: = ≠
 - Order: < >
 - Addition: + -
 - Multiplication: * /
- Nominal attribute: distinctness
- Ordinal attribute: distinctness & order
- Interval attribute: distinctness, order & addition
- Ratio attribute: all 4 properties



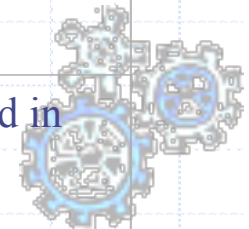
•Attribute Type	•Description	•Examples	•Operations
•Nominal	•The values of a nominal attribute are just different names, i.e., nominal attributes provide only enough information to distinguish one object from another. (=, ≠)	•zip codes, employee ID numbers, eye color, sex: { <i>male, female</i> }	•mode, entropy, contingency correlation, χ^2 test
•Ordinal	•The values of an ordinal attribute provide enough information to order objects. (<, >)	•hardness of minerals, { <i>good, better, best</i> }, grades, street numbers	•median, percentiles, rank correlation, run tests, sign tests
•Interval	•For interval attributes, the differences between values are meaningful, i.e., a unit of measurement exists. (+, -)	•calendar dates, temperature in Celsius or Fahrenheit	•mean, standard deviation, Pearson's correlation, <i>t</i> and <i>F</i> tests
•Ratio	•For ratio variables, both differences and ratios are meaningful. (*, /)	•temperature in Kelvin, monetary quantities, counts, age, mass, length, electrical current	•geometric mean, harmonic mean, percent variation

Master MAINS, 2015



•Attribute Level	•Transformation	•Comments
•Nominal	•Any permutation of values	•If all employee ID numbers were reassigned, would it make any difference?
•Ordinal	•An order preserving change of values, i.e., $new_value = f(old_value)$ where f is a monotonic function.	•An attribute encompassing the notion of good, better best can be represented equally well by the values {1, 2, 3} or by {0.5, 1, 10}.
•Interval	• $new_value = a * old_value + b$ where a and b are constants	•Thus, the Fahrenheit and Celsius temperature scales differ in terms of where their zero value is and the size of a unit (degree).
•Ratio	• $new_value = a * old_value$	•Length can be measured in meters or feet.

Master MAINS, 2015



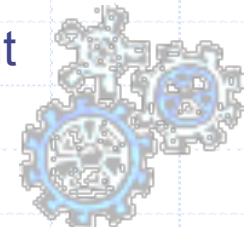
Discrete and Continuous Attributes

◆ Discrete Attribute

- Has only a finite or countably infinite set of values
- Examples: zip codes, counts, or the set of words in a collection of documents
- Often represented as integer variables.
- Note: binary attributes are a special case of discrete attributes

◆ Continuous Attribute

- Has real numbers as attribute values
- Examples: temperature, height, or weight.
- Practically, real values can only be measured and represented using a finite number of digits.
- Continuous attributes are typically represented as floating-point variables.



Types of data sets

◆ Record

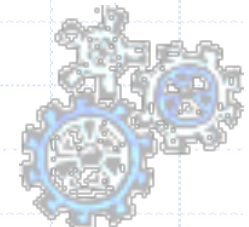
- Data Matrix
- Document Data
- Transaction Data

◆ Graph

- World Wide Web
- Molecular Structures

◆ Ordered

- Spatial Data
- Temporal Data
- Sequential Data
- Genetic Sequence Data

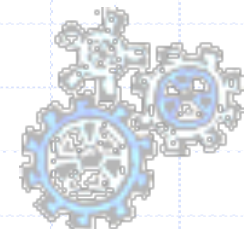


Record Data

- ◆ Data that consists of a collection of records, each of which consists of a fixed set of attributes

<i>Tid</i>	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

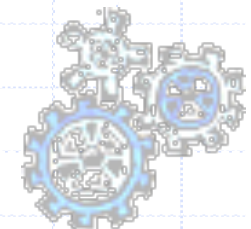
Master MAINS, 2015



Data Matrix

- ◆ If data objects have the same fixed set of numeric attributes, then the data objects can be thought of as points in a multi-dimensional space, where each dimension represents a distinct attribute
- ◆ Such data set can be represented by an m by n matrix, where there are m rows, one for each object, and n columns, one for each attribute

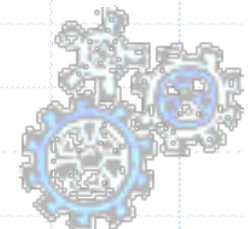
Projection of x Load	Projection of y load	Distance	Load	Thickness
10.23	5.27	15.22	2.7	1.2
12.65	6.25	16.22	2.2	1.1



Document Data

- ◆ Each document becomes a `term' vector,
 - each term is a component (attribute) of the vector,
 - the value of each component is the number of times the corresponding term occurs in the document.

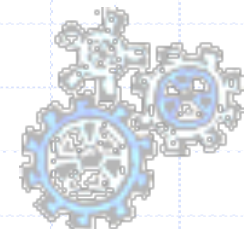
	team	coach	play	ball	score	game	win	lost	timeout	season
Document 1	3	0	5	0	2	6	0	2	0	2
Document 2	0	7	0	2	1	0	0	3	0	0
Document 3	0	1	0	0	1	2	2	0	3	0



Transaction Data

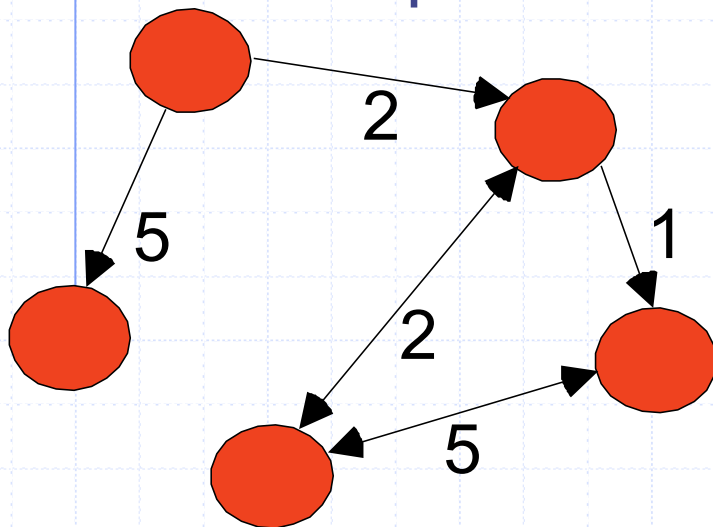
- ◆ A special type of record data, where
 - each record (transaction) involves a set of items.
 - For example, consider a grocery store. The set of products purchased by a customer during one shopping trip constitute a transaction, while the individual products that were purchased are the items.

<i>TID</i>	<i>Items</i>
1	Bread, Coke, Milk
2	Beer, Bread
3	Beer, Coke, Diaper, Milk
4	Beer, Bread, Diaper, Milk
5	Coke, Diaper, Milk

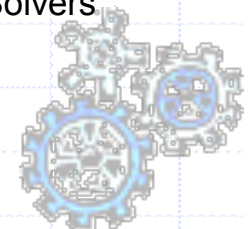


Graph Data

◆ Examples: Generic graph and HTML Links

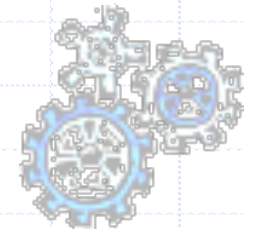
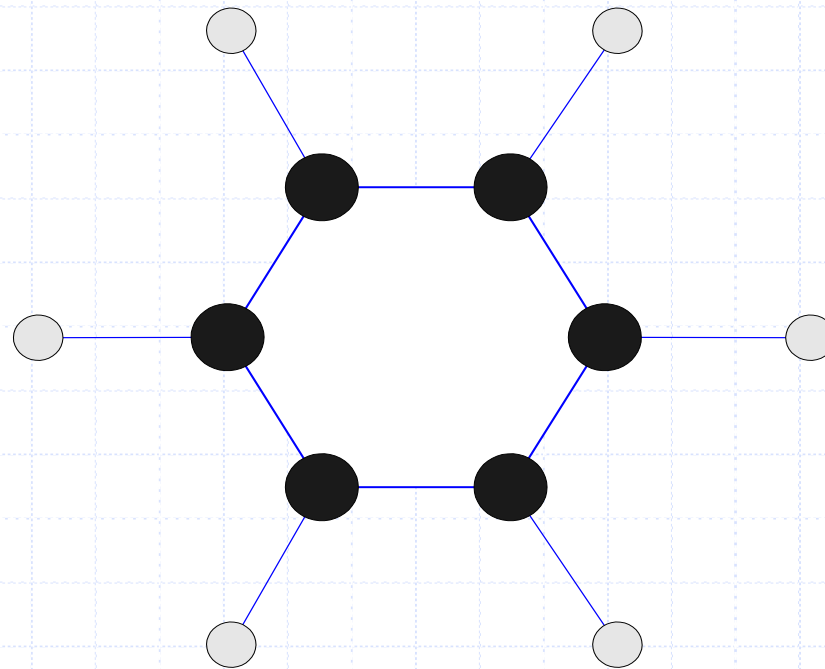


```
<a href="papers/papers.html#bbbb">  
Data Mining </a>  
<li>  
<a href="papers/papers.html#aaaa">  
Graph Partitioning </a>  
<li>  
<a href="papers/papers.html#aaaa">  
Parallel Solution of Sparse Linear System of Equations </a>  
<li>  
<a href="papers/papers.html#ffff">  
N-Body Computation and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



Chemical Data

◆ Benzene Molecule: C_6H_6



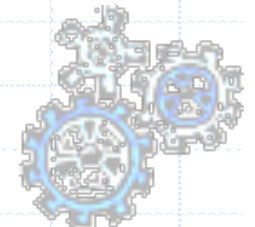
Ordered Data

- ◆ Sequences of transactions
 - Items/Events

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



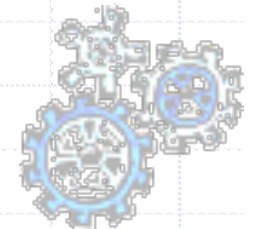
- An element of the sequence



Ordered Data

◆ Genomic sequence data

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

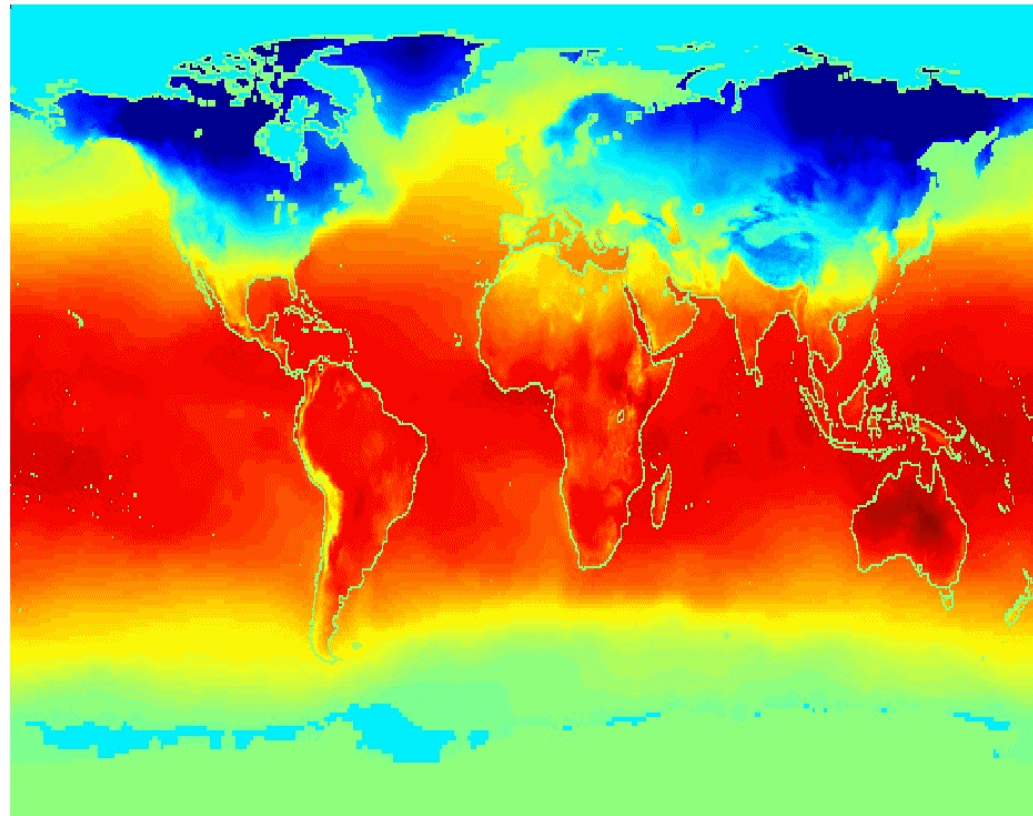


Ordered Data

◆ Spatio-Temporal Data

**Average Monthly
Temperature of
land and ocean**

Jan



Caratteristiche delle Variabili (dei data sets)

◆ 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 $[-\infty, \infty]$ (o simili)
- Non ha senso considerare l'intero intervallo

◆ Outliers

- Valori singoli o con frequenza estremamente bassa
- Possono distorcere le informazioni sui dati

◆ Dimensionalità delle variabili

- Il numero di valori che una variabile può assumere può essere estremamente alto
 - ◆ Tipicamente riguarda valori categorici

◆ Dimensionalità degli oggetti

- Il numero di attributi che un oggetto ha può essere estremamente alto
 - ◆ Es. prodotti di un market basket

◆ Anacronismo

- Una variabile può essere contingente: abbiamo i valori in una sola porzione dei dati

Master MAINS, 2015



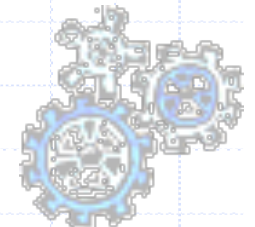
Descrizione dei dati

◆ Grafici

- Distribuzione frequenze
- Correlazione
- Dispersione

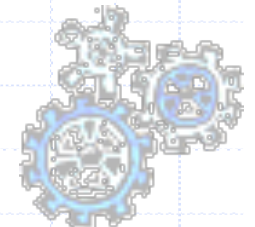
◆ Misure

- Media, mediana, quartili
- Varianza, deviazione standard
- Forma, simmetria, curtosi



Visualizzazione dati qualitativi

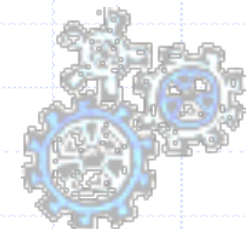
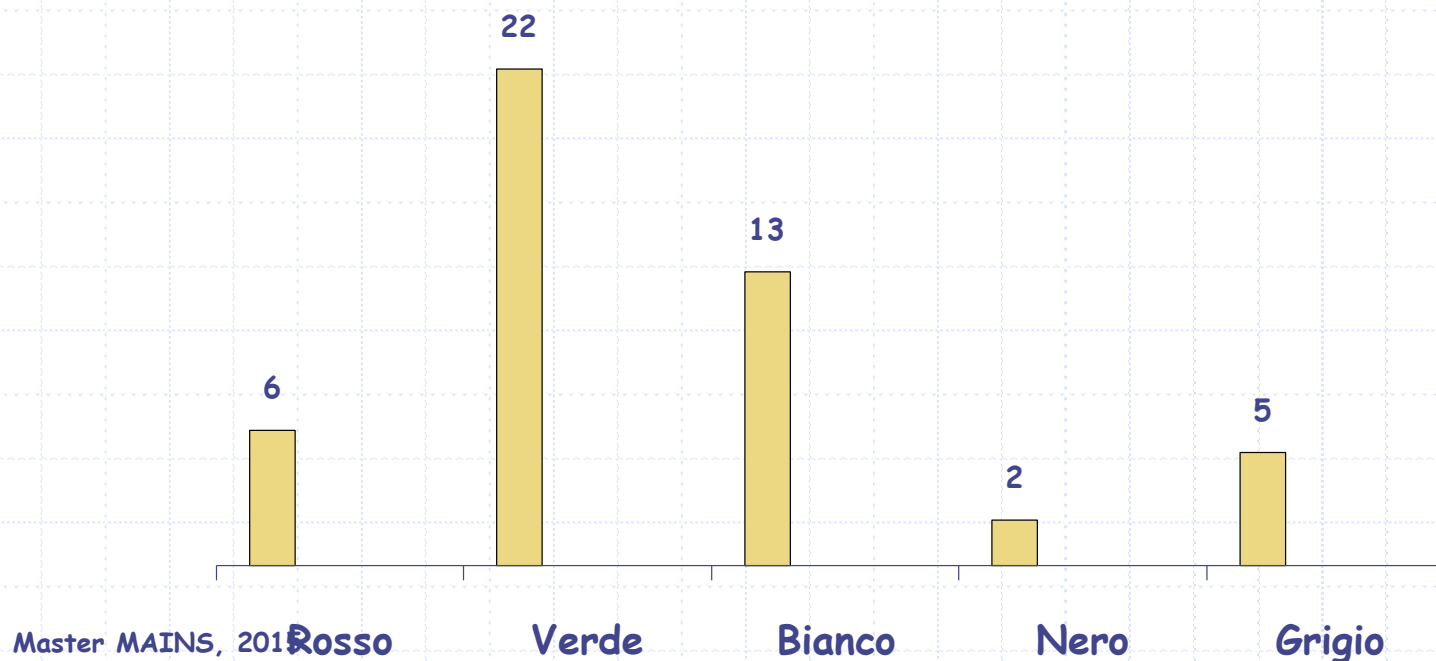
- ◆ Rappresentazione delle frequenze
 - Diagrammi a barre
 - Ortogrammi
 - Aerogrammi
- ◆ Correlazione
 - Web diagrams
- ◆ Ciclicità
 - Diagrammi polari



Diagrammi di Pareto



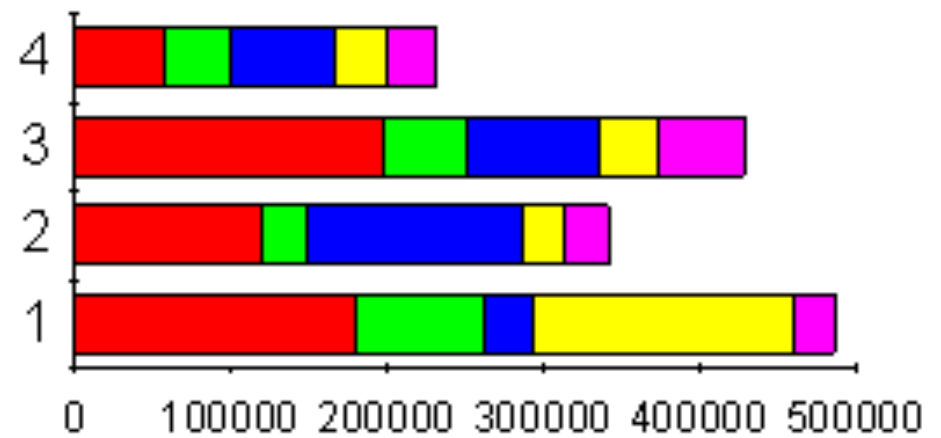
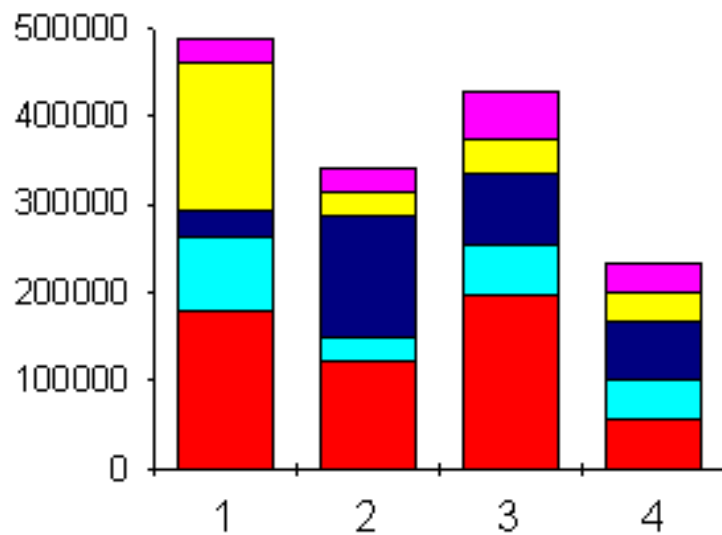
- ◆ Diagrammi a barre distanziate
- ◆ Un assortimento di eventi presenta pochi picchi e molti elementi comuni



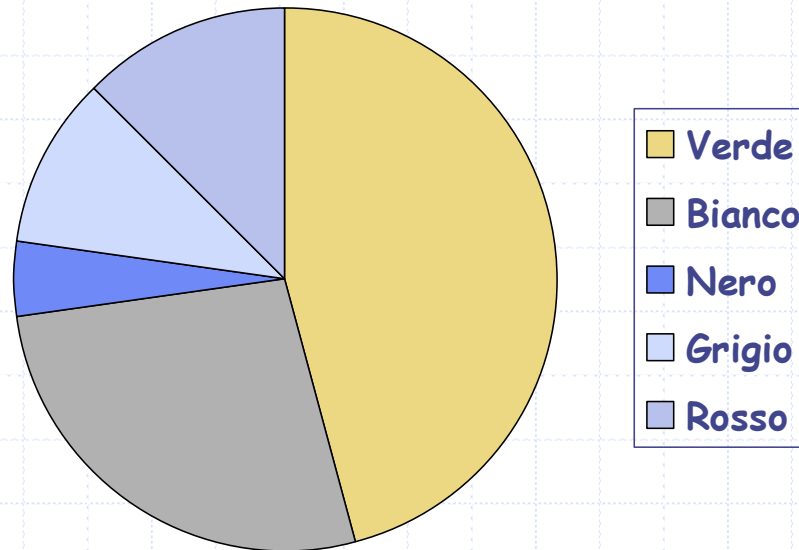
Ortogrammi



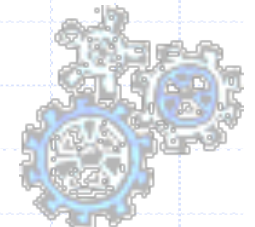
- ◆ Ogni colonna indica la la distribuzione interna per un dato valore e la frequenza



Aerogrammi



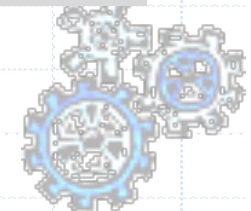
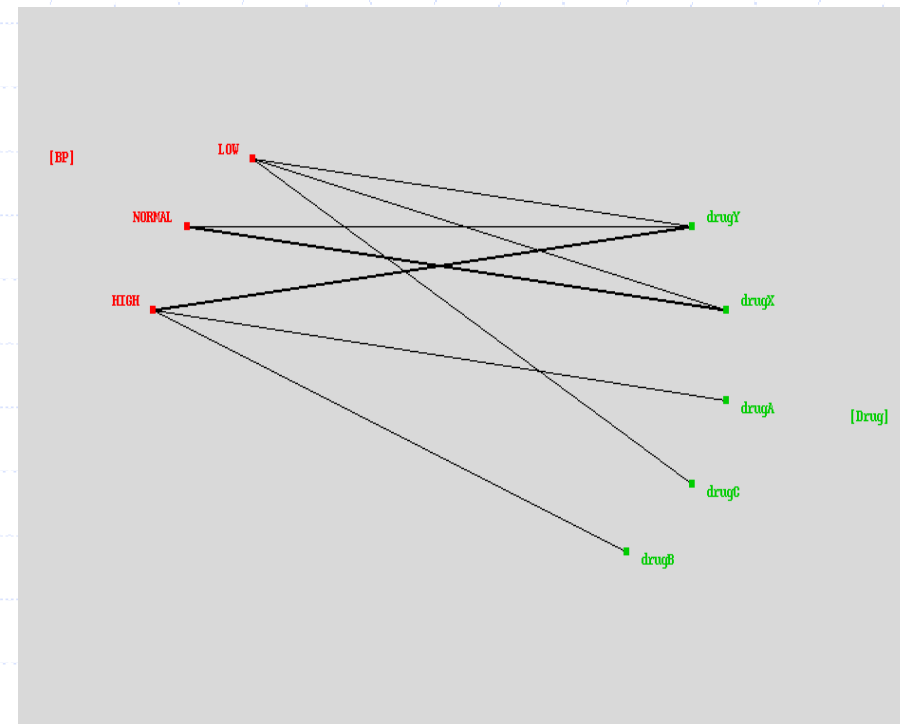
- ◆ Rappresentazioni a torta
- ◆ frequenza della distribuzioni



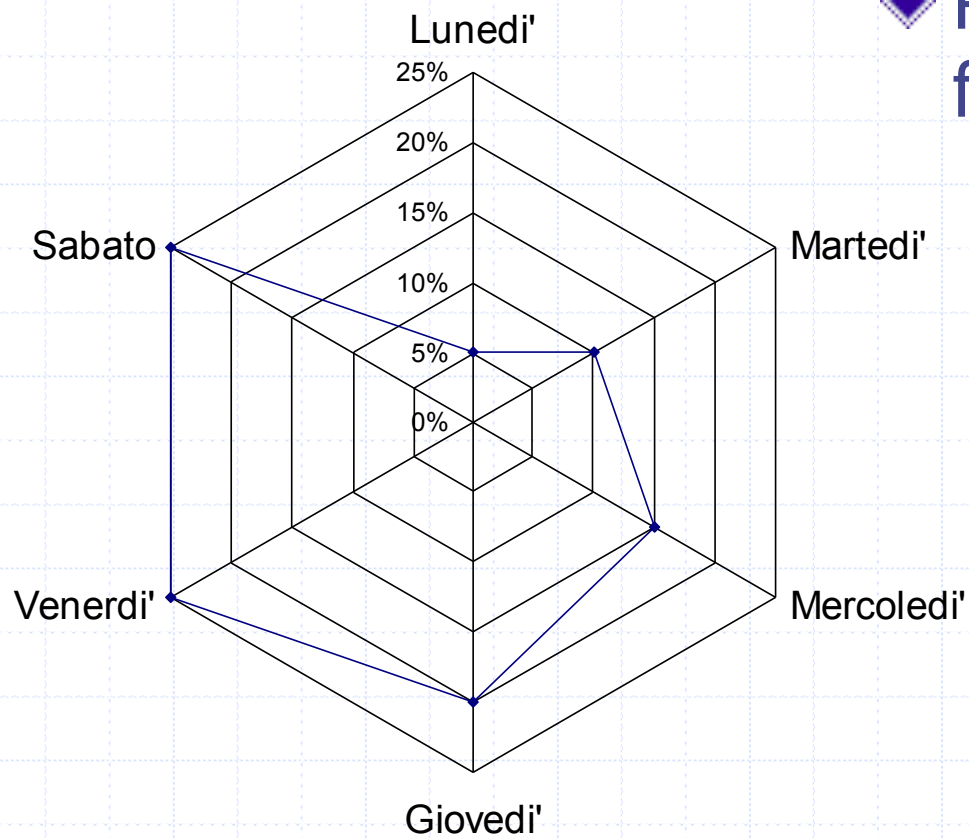
Web



◆ Visualizzano correlazioni tra valori simbolici

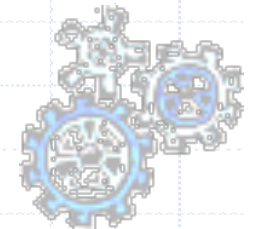


Diagrammi polari



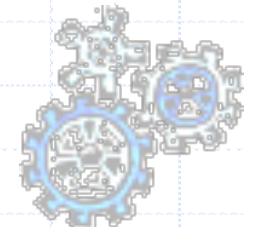
◆ Rappresentano fenomeni ciclici

- E.g., concentrazione delle vendite nell'arco settimanale

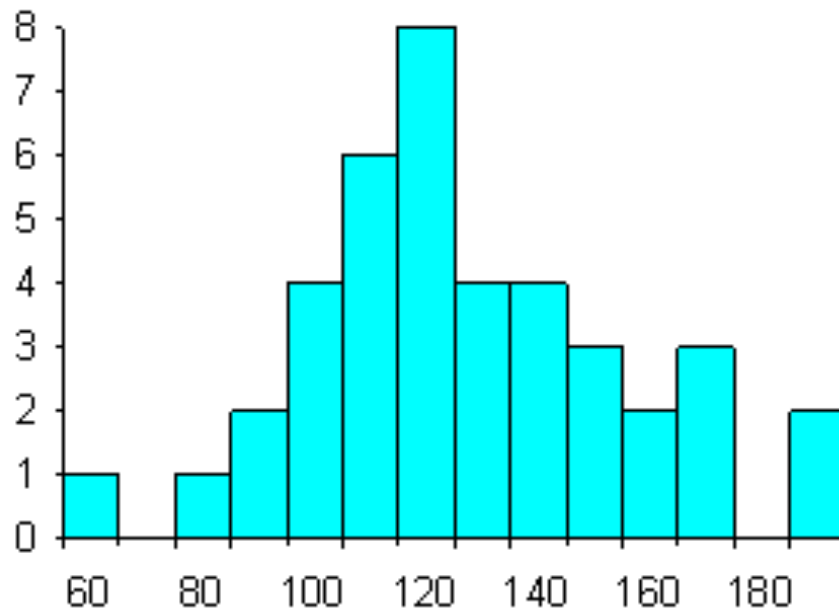
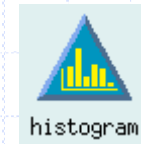


Dati Quantitativi

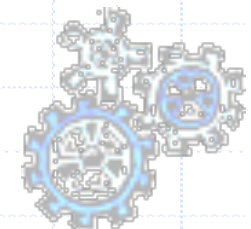
- ◆ Istogrammi
- ◆ Poligoni
- ◆ Stem and leaf
- ◆ Dot Diagrams
- ◆ Diagrammi quantili



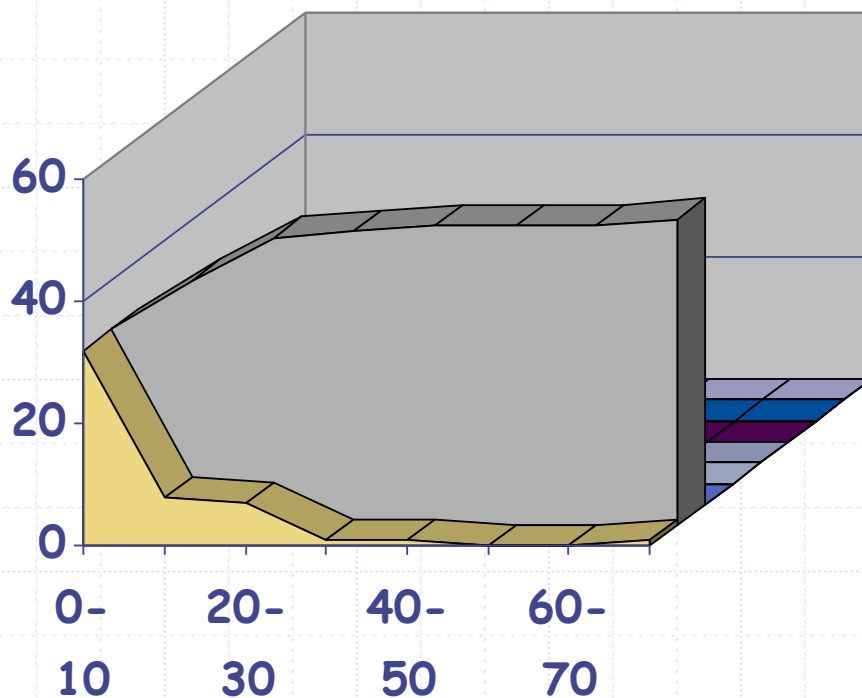
Istogrammi



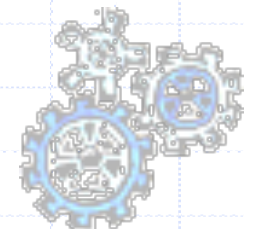
- ◆ Rappresentazioni a barre
- ◆ Evidenziano la frequenza su intervalli adiacenti
 - La larghezza di ogni rettangolo misura l'ampiezza degli intervalli
 - Quale larghezza?



Poligoni



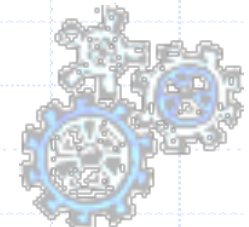
- ◆ Per la descrizione di frequenze cumulative
- ◆ I punti sono uniti tramite linee



Rappresentazione “Stem & Leaf”

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

- ◆ Simile a istogrammi
- ◆ Evita la perdita di informazione
- ◆ Utile per pochi dati

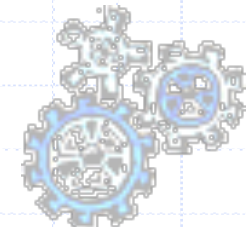
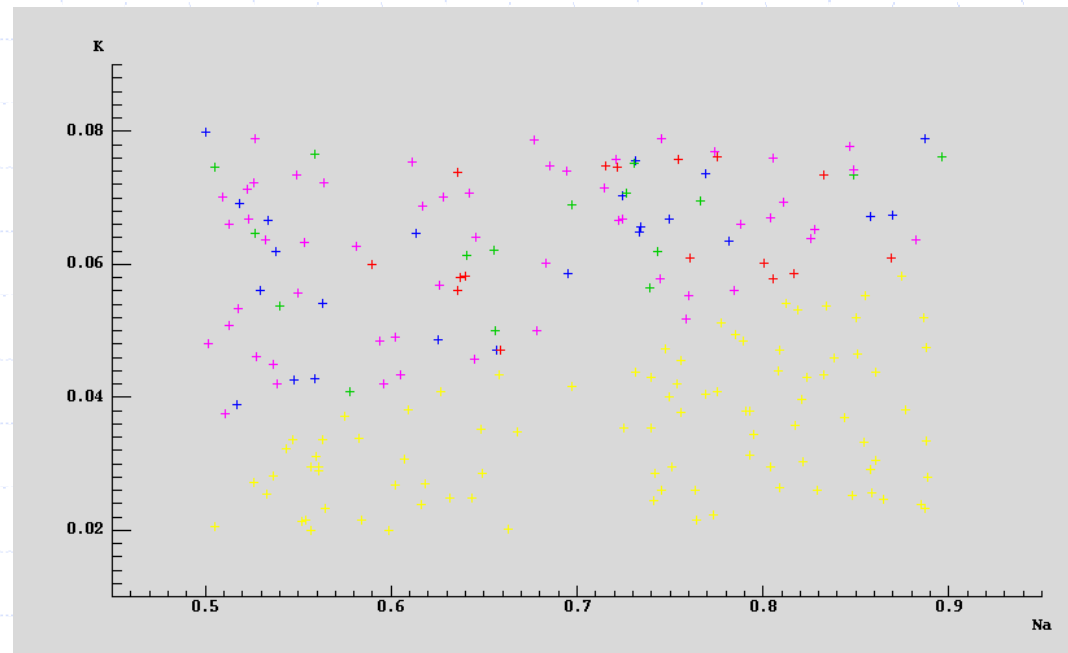


Dot Diagrams, Scatters



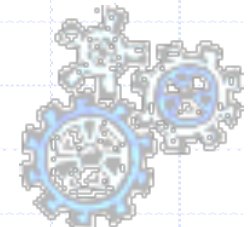
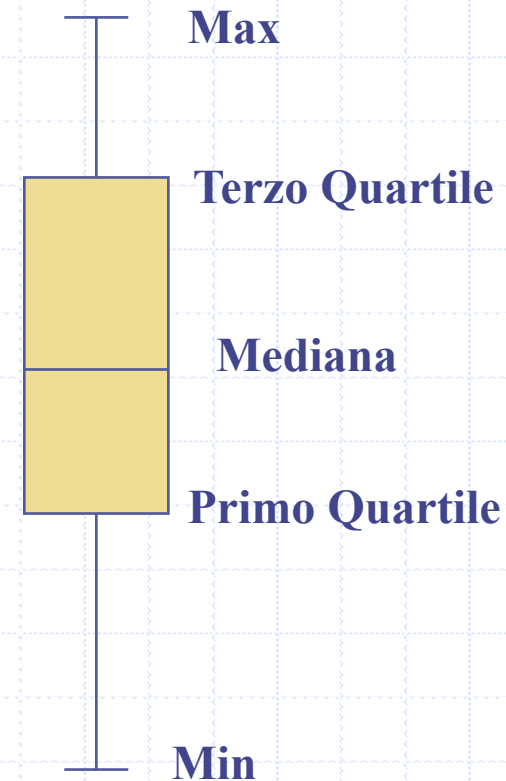
Weka

- ◆ Visualizza la Dispersione plot dei dat



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**

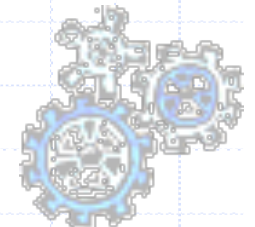
- Media aritmetica, geometrica e armonica, mediana, quartili, percentili, moda

◆ **Dispersione o variabilità**

- Range, scarto medio, varianza, deviazione standard

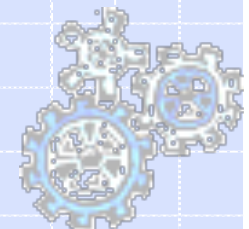
◆ **Forma della distribuzione**

- Simmetria (medie interquartili, momenti centrali, indice di Fisher) e curtosi (indice di Pearson, coefficiente di curtosi)



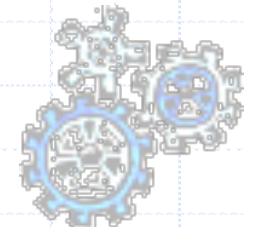
Outline del Modulo

- ◆ Introduzione e Concetti di Base
- ◆ Data Selection
- ◆ Information Gathering
- ◆ Data cleaning ←



Data Cleaning

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



Valori Anomali

◆ Valori mancanti

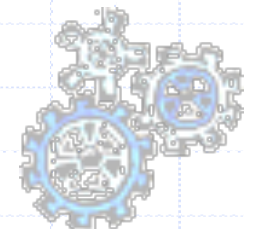
- NULL

◆ Valori sconosciuti

- Privi di significato

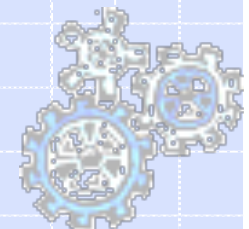
◆ Valori non validi

- Con valore noto ma non significativo

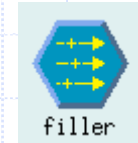


Outline del Modulo

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



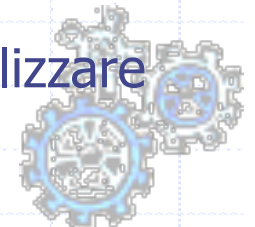
Trattamento di valori nulli



1. Eliminazione delle tuple
2. Sostituzione dei valori nulli

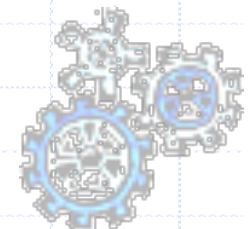
N.B.: può influenzare la distribuzione dei dati numerici

- Utilizzare media/mediana/moda
- Predirre 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



Data Reduction

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

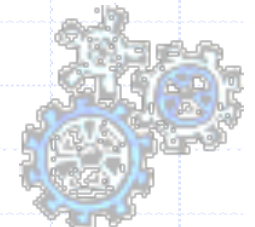


Sampling

(Riduzione verticale)

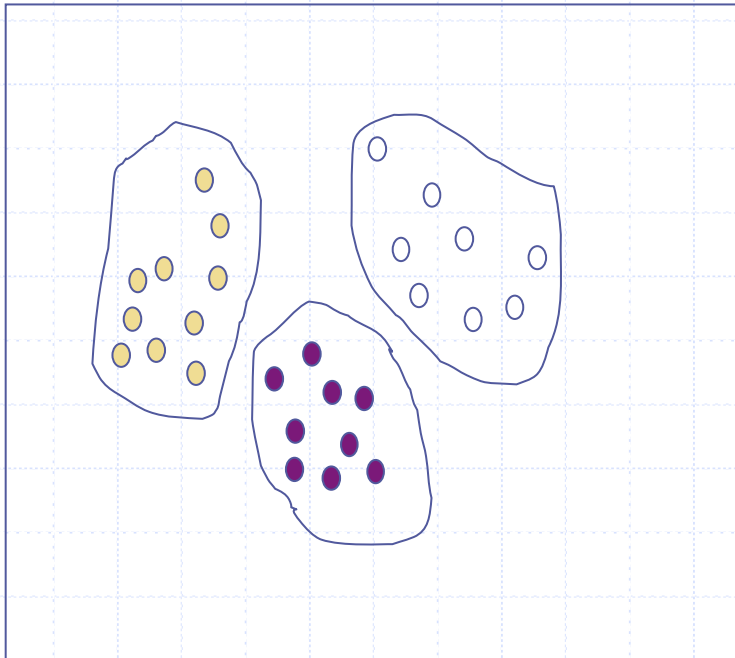


- ◆ Riduce la complessità di esecuzione degli algoritmi di Mining
- ◆ Problema: scegliere un sottoinsieme **rappresentativo** dei dati
 - La scelta di un campionamento casuale può essere problematica per la presenza di picchi
- ◆ Alternative: Schemi adattativi
 - **Stratified sampling:**
 - ◆ Approssimiamo 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).

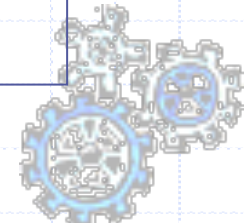
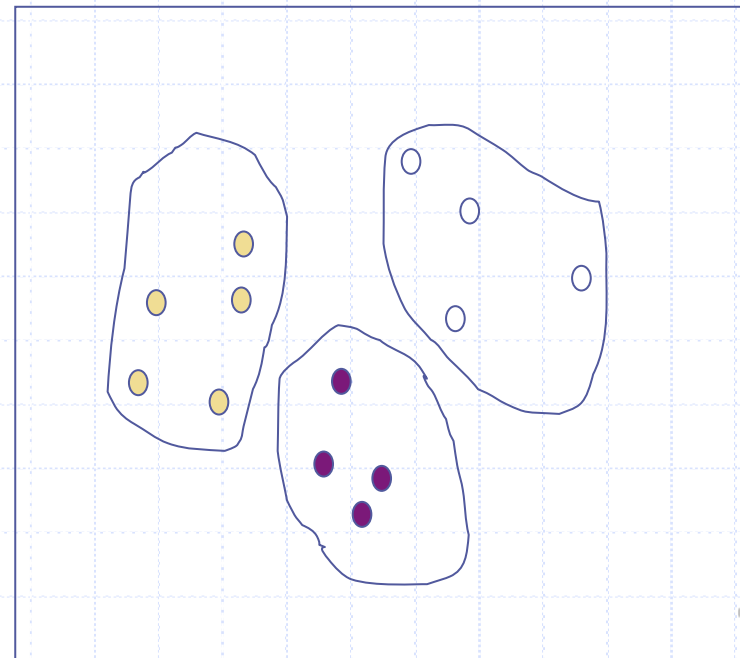


Sampling

Raw Data



Cluster/Stratified Sample



Riduzione Dimensionalità

(Riduzione orizzontale)

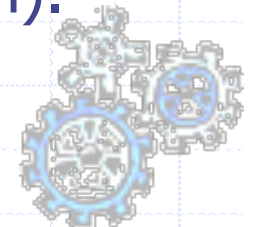
◆ Selezione di un sotto-insieme di attributi

■ 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).

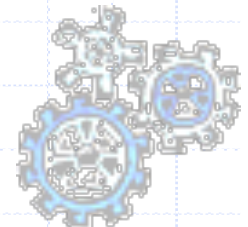


Riduzione Dimensionalità

(Riduzione orizzontale)

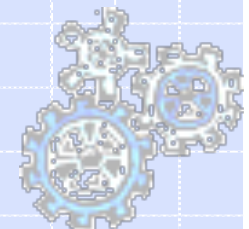


- ◆ Creazione di nuovi attributi con i quali rappresentare le tuple
 - Principal components analysis (PCA)
 - ◆ Trova le combinazioni lineari degli attributi nei k vettori ortonormali più significativi
 - ◆ Proietta le vecchie tuple sui nuovi attributi
 - Altri metodi
 - ◆ Factor Analysis
 - ◆ Decomposizione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



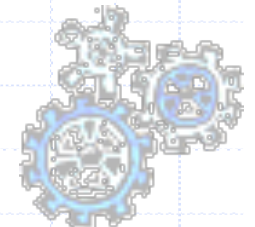
Outline del Modulo

- ◆ 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



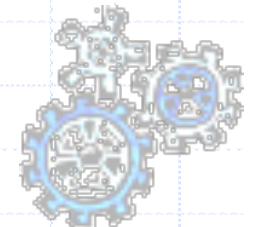
Obiettivi

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

$$Y = T(X)$$

tale che:

- Y preserva l' informazione “rilevante” di X
- Y elimina almeno uno dei problemi di X
- Y è più “utile” di X



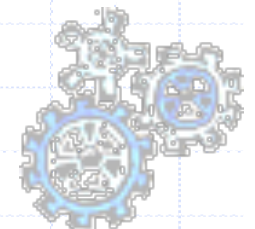
Obiettivi

◆ Scopi principali:

- stabilizzare le varianze
- normalizzare le distribuzioni
- linearizzare le relazioni tra variabili

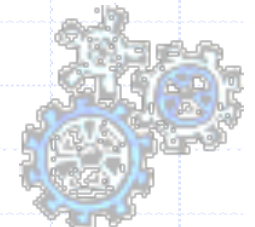
◆ Scopi secondari:

- 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 outliers
- ◆ 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 trattati



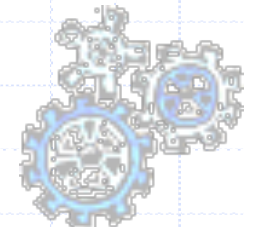
Metodi

◆ 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`

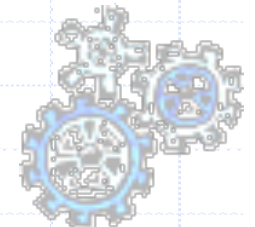
◆ Trasformazioni lineari

$$1\text{€} = 1936.27 \text{ Lit.}$$

- $p=1, a=1936.27, b=0$

$$^{\circ}\text{C} = 5/9(^{\circ}\text{F} - 32)$$

- $p = 1, a = 5/9, b = -160/9$



Normalizzazioni

◆ min-max normalization

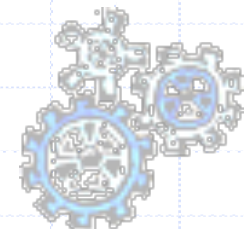
$$v' = \frac{v - \min_A}{\max_A - \min_A} (\text{new_max}_A - \text{new_min}_A) + \text{new_min}_A$$

◆ z-score normalization

$$v' = \frac{v - \text{mean}_A}{\text{stand_dev}_A}$$

◆ normalization tramite decimal scaling

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

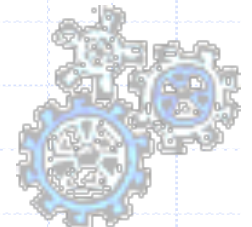


Stabilizzare varianze

◆ Trasformazione logaritmica

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

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

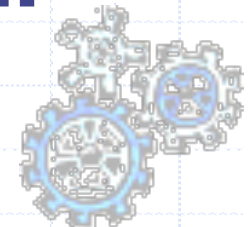


Trasformazione logaritmica: esempio

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	Mediana
1775	Secondo Quartile
11000	Max

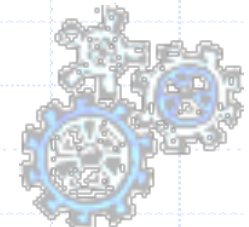
Dati troppo dispersi!!!



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

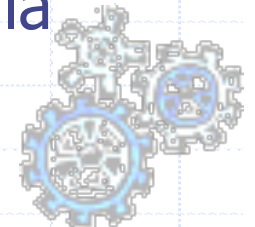
$$T(x) = ax^p + b$$

◆ Trasformazione in radice

- $p = 1/c$, c numero intero
- per omogeneizzare varianze di distribuzioni particolari, e.g., di Poisson

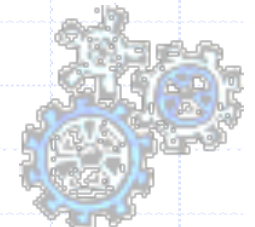
◆ Trasformazione reciproca

- $p < 0$
- per l'analisi di serie temporali, quando la varianza aumenta in modo molto pronunciato rispetto alla media



Simmetria

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



Asimmetria dei dati

◆ Simmetria e Media interpercentile

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

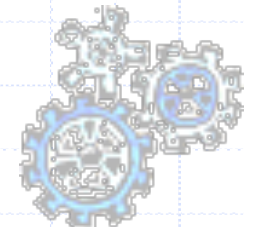
◆ Se la media interpercentile è sbilanciata, allora la distribuzione dei dati è asimmetrica

- ◆ sbilanciata a destra

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

- ◆ sbilanciata a sinistra

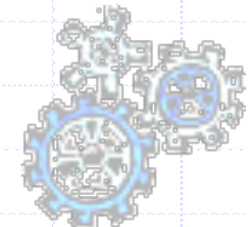
$$\bar{x}_p < M$$



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



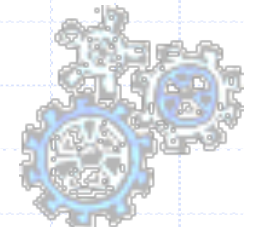
Asimmetria : esempio

- ◆ 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



Master MAINS, 2015



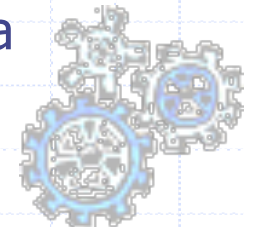
Creare simmetria nei dati: Transformation plot

- ◆ 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

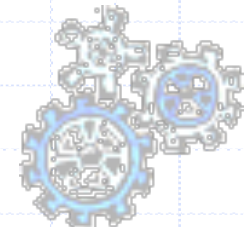
- ◆ Intuitivamente, confrontiamo la differenza assoluta e relativa tra mediana e medie interpercentili
- ◆ il valore medio (mediano) dei valori di c è il valore della trasformazione



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

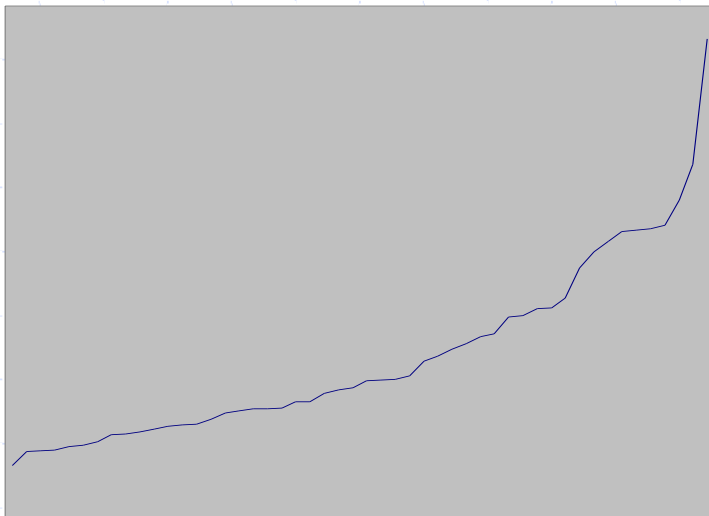
- ◆ Calcolando la mediana dei valori c otteniamo $p=0.5188$
- ◆ Proviamo con $p=1/2...$



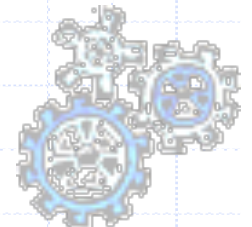
Trasformazione 1: radice quadrata

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

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	



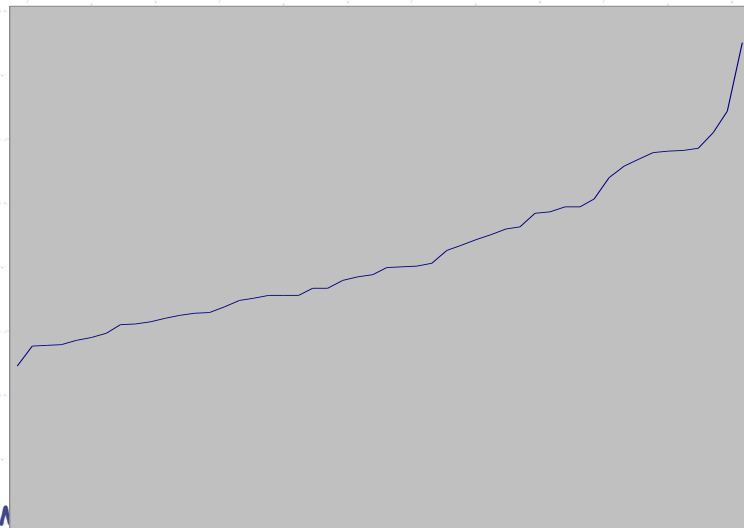
- La curva si tempera, ma i valori alti continuano a produrre differenze notevoli
- Proviamo a diminuire $p...$



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	



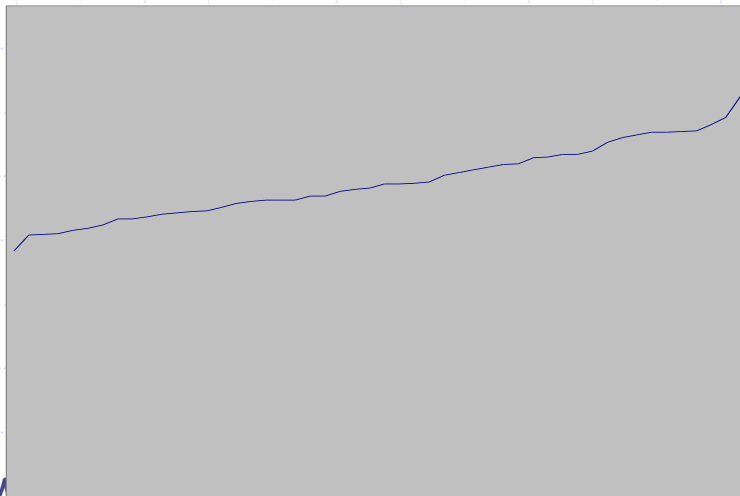
- ◆ I valori alti continuano ad influenzare
- ◆ Proviamo con il logaritmo...



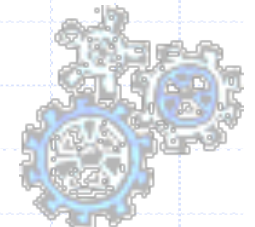
Trasformazione 3: logaritmo

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

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

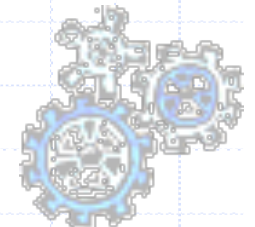
- La formula

$$y = \alpha x^p$$

puo' essere individuata studiando la relazione

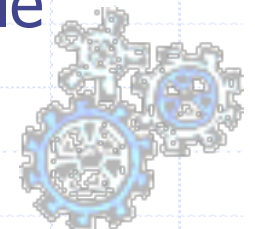
$$z = \log \alpha + pw$$

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



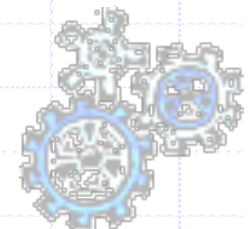
Discretizzazione

- ◆ 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 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



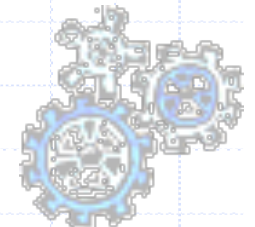
Unsupervised Discretization

◆ 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)



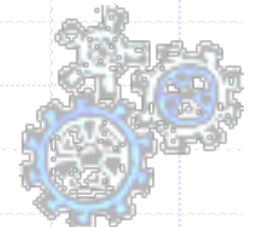
Discretization of quantitative attributes

• **Solution:** each value is replaced by the interval to which it belongs.

- **height:** 0-150cm, 151-170cm, 171-180cm, >180cm
- **weight:** 0-40kg, 41-60kg, 60-80kg, >80kg
- **income:** 0-10ML, 11-20ML, 20-25ML, 25-30ML, >30ML

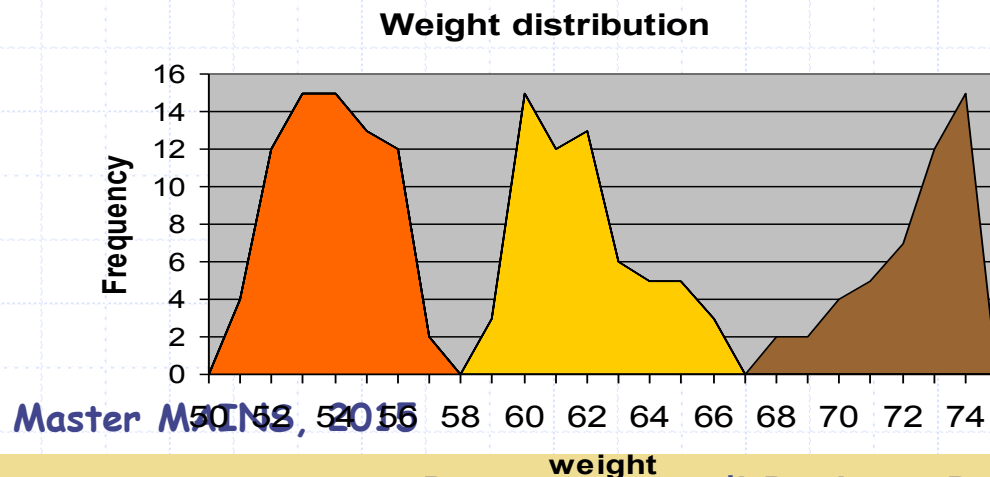
CID	height	weight	income
1	151-171	60-80	>30
2	171-180	60-80	20-25
3	171-180	60-80	25-30
4	151-170	60-80	25-30

• **Problem:** the discretization may be useless (see **weight**).

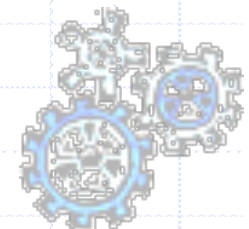


How to choose intervals?

1. Interval with a fixed "reasonable" granularity
Ex. intervals of 10 cm for height.
2. Interval size is defined by some domain dependent criterion
Ex.: 0-20ML, 21-22ML, 23-24ML, 25-26ML, >26ML
3. Interval size determined by analyzing data, studying the distribution or using clustering



- 50 - 58 kg
- 59-67 kg
- > 68 kg



Natural Binning

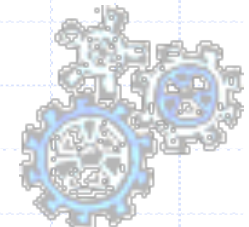
- ◆ Semplice
- ◆ Ordino i valori, quindi divido il range di valori in k parti della stessa dimensione

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

- ◆ l'elemento x_j appartiene alla classe i se

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

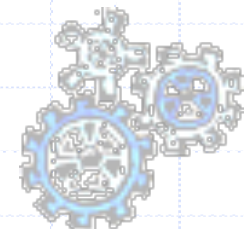
- ◆ Può produrre distribuzioni molto sbilanciate



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

- ◆ $\delta = (160-100)/4 = 15$
- ◆ 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:

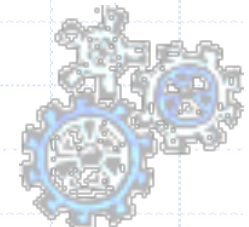
$$f = \frac{N}{k}$$

(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



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

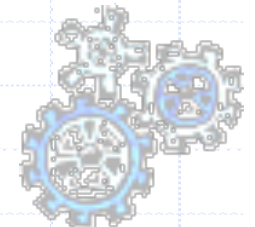
◆ $f = 12/4 = 3$

◆ classe 1: {100,110,110}

◆ classe 2: {120,120,125}

◆ classe 3: {130,130,135}

◆ classe 4: {140,150,160}



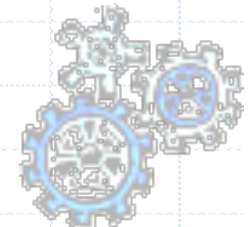
Quante classi?

- ◆ 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)

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



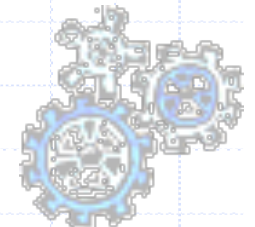
Supervised Discretization

◆ Caratteristiche:

- La discretizzazione ha un obiettivo quantificabile
- Il numero di classi non è noto a priori

◆ Tecniche:

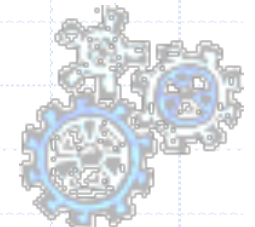
- ChiMerge
- Discretizzazione basata sull' Entropia
- Discretizzazione basata sui percentili



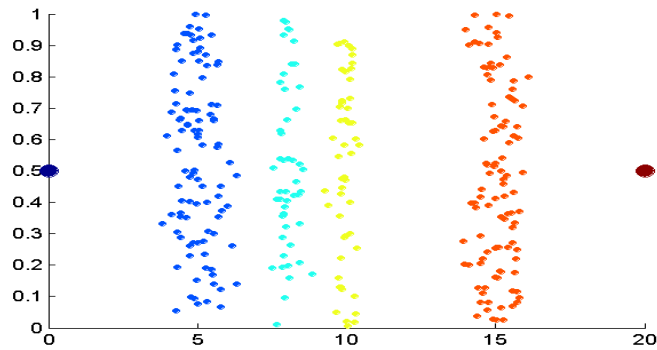
Supervised Discretization: ChiMerge

◆ Procedimento Bottom-up:

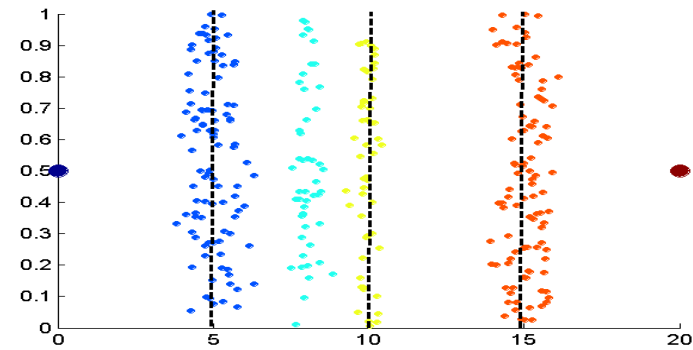
- Inizialmente, ogni valore è un intervallo a se'
- Intervalli adiacenti sono iterativamente uniti se sono simili
- La similitudine è misurata sulla base dell' attributo target, contando quanto i due intervalli sono "diversi"



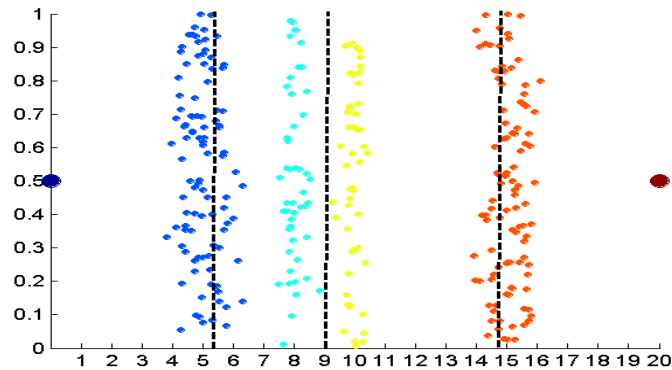
Labels



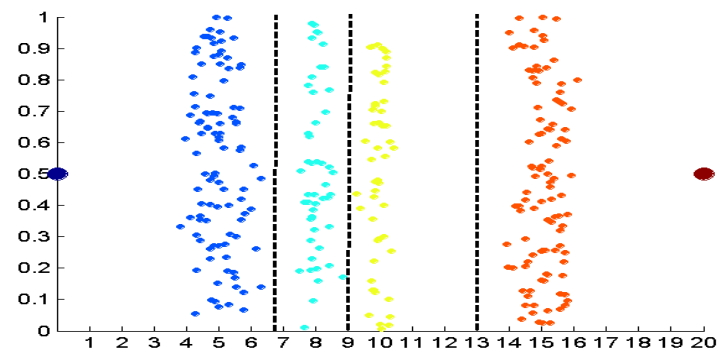
Data



Equal interval width

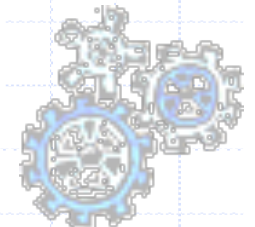


Equal frequency



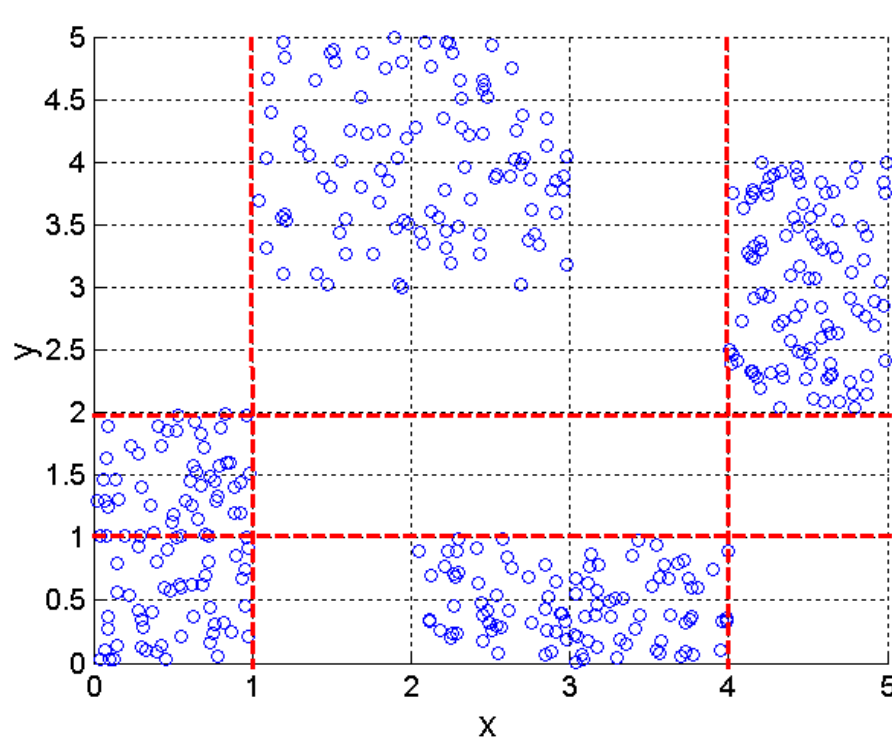
K-means

Master MAINS, 2015

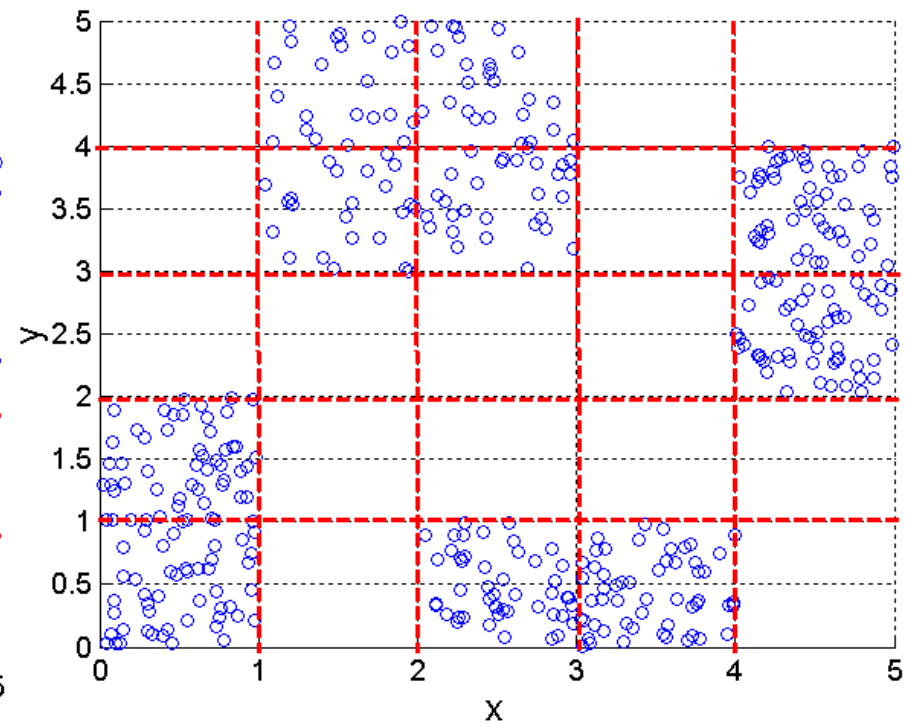


Discretization Using Class Labels

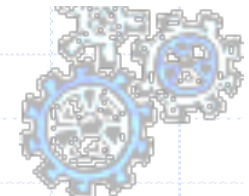
◆ Entropy based approach



3 categories for both x and y

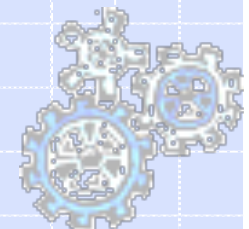


5 categories for both x and y



Outline del Modulo

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



Similarity and Dissimilarity

◆ Similarity

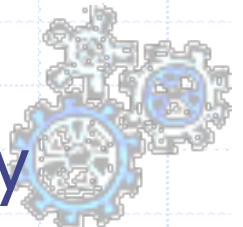
- Numerical measure of how alike two data objects are.
- Is higher when objects are more alike.
- Often falls in the range $[0,1]$

◆ Dissimilarity

- Numerical measure of how different are two data objects
- Lower when objects are more alike
- Minimum dissimilarity is often 0
- Upper limit varies

◆ Proximity refers to a similarity or dissimilarity

Master MAINS, 2015



Similarity/Dissimilarity for ONE Attribute

p and q are the attribute values for two data objects.

Attribute Type	Dissimilarity	Similarity
Nominal	$d = \begin{cases} 0 & \text{if } p = q \\ 1 & \text{if } p \neq q \end{cases}$	$s = \begin{cases} 1 & \text{if } p = q \\ 0 & \text{if } p \neq q \end{cases}$
Ordinal	$d = \frac{ p-q }{n-1}$ (values mapped to integers 0 to $n-1$, where n is the number of values)	$s = 1 - \frac{ p-q }{n-1}$
Interval or Ratio	$d = p - q $	$s = -d, s = \frac{1}{1+d} \text{ or } s = 1 - \frac{d - \min_d}{\max_d - \min_d}$

Table 5.1. Similarity and dissimilarity for simple attributes



Many attributes: Euclidean Distance

◆ Euclidean Distance

$$\mathit{dist} = \sqrt{\sum_{k=1}^n (p_k - q_k)^2}$$

Where n is the number of dimensions (attributes) and p_k and q_k are, respectively, the value of k^{th} attributes (components) or data objects p and q .

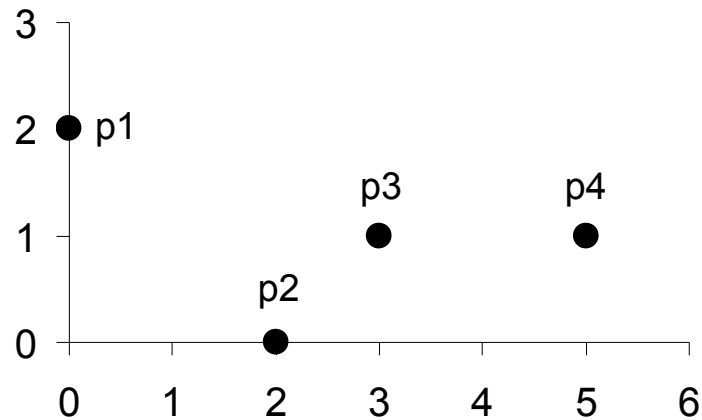


Master MAINS, 2015

Standardization is necessary, if scales differ.



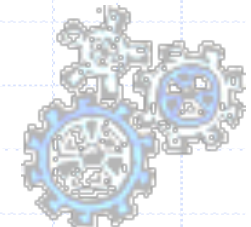
Euclidean Distance



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

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

Distance Matrix

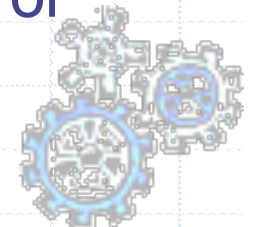


Minkowski Distance

- ◆ Minkowski Distance is a generalization of Euclidean Distance

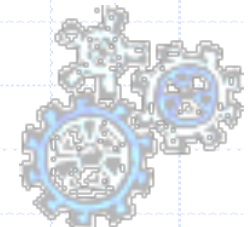
$$\mathit{dist} = \left(\sum_{k=1}^n |p_k - q_k|^r \right)^{\frac{1}{r}}$$

Where r is a parameter, n is the number of dimensions (attributes) and p_k and q_k are, respectively, the k th attributes (components) or data objects p and q .

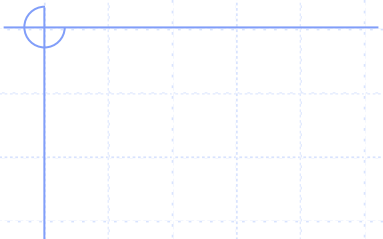


Minkowski Distance: Examples

- ◆ $r = 1$. City block (Manhattan, taxicab, L_1 norm) distance.
 - A common example of this is the Hamming distance, which is just the number of bits that are different between two binary vectors
- ◆ $r = 2$. Euclidean distance
- ◆ $r \rightarrow \infty$. "supremum" (L_{\max} norm, L_{∞} norm) distance.
 - This is the maximum difference between any component of the vectors
- ◆ Do not confuse r with n , i.e., all these distances are defined for all numbers of dimensions.



Minkowski Distance



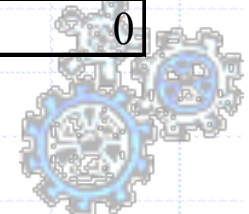
point	x	y
p1	0	2
p2	2	0
p3	3	1
p4	5	1

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

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

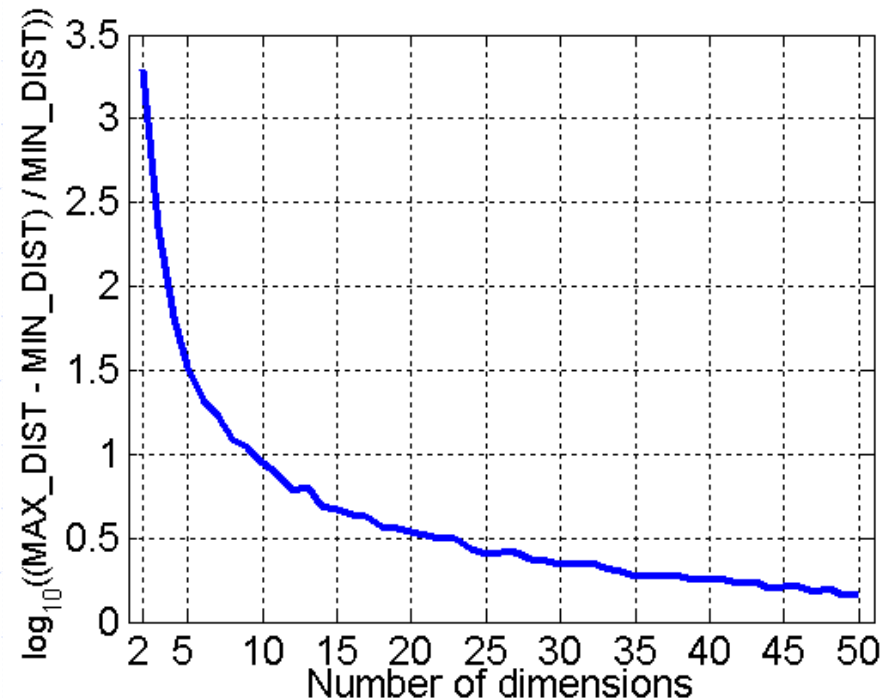
L_∞	p1	p2	p3	p4
p1	0	2	3	5
p2	2	0	1	3
p3	3	1	0	2
p4	5	3	2	0

Distance Matrix

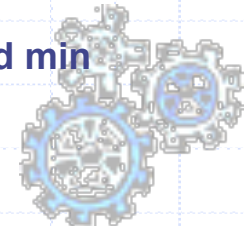


Curse of Dimensionality

- ◆ When dimensionality increases, data becomes increasingly sparse in the space that it occupies
- ◆ Definitions of density and distance between points, which is critical for clustering and outlier detection, become less meaningful



- Randomly generate 500 points
- Compute difference between max and min distance between any pair of points



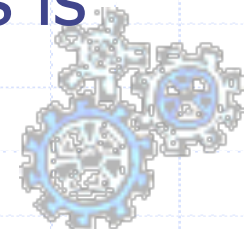
Common Properties of a Distance

◆ Distances, such as the Euclidean distance, have some well known properties.

1. $d(p, q) \geq 0$ for all p and q and $d(p, q) = 0$ only if $p = q$. (Positive definiteness)
2. $d(p, q) = d(q, p)$ for all p and q . (Symmetry)
3. $d(p, r) \leq d(p, q) + d(q, r)$ for all points $p, q,$ and r . (Triangle Inequality)

where $d(p, q)$ is the distance (dissimilarity) between points (data objects), p and q .

◆ A distance that satisfies these properties is a **metric**

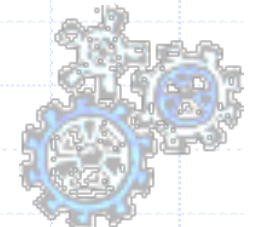


Common Properties of a Similarity

◆ Similarities, also have some well known properties.

1. $s(p, q) = 1$ (or maximum similarity) only if $p = q$.
2. $s(p, q) = s(q, p)$ for all p and q . (Symmetry)

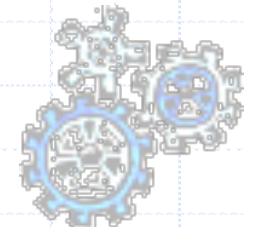
where $s(p, q)$ is the similarity between points (data objects), p and q .



Binary Data

Categorical	insufficient	sufficient	good	very good	excellent
p1	0	0	1	0	0
p2	0	0	1	0	0
p3	1	0	0	0	0
p4	0	1	0	0	0

item	bread	butter	milk	apple	tooth-past
p1	1	1	0	1	0
p2	0	0	1	1	1
p3	1	1	1	0	0
p4	1	0	1	1	0



Similarity Between Binary Vectors

- ◆ Common situation is that objects, p and q , have only binary attributes
- ◆ Compute similarities using the following quantities

M_{01} = the number of attributes where p was 0 and q was 1

M_{10} = the number of attributes where p was 1 and q was 0

M_{00} = the number of attributes where p was 0 and q was 0

M_{11} = the number of attributes where p was 1 and q was 1

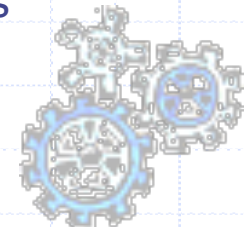
- ◆ Simple Matching and Jaccard Coefficients

SMC = number of matches / number of attributes

$$= (M_{11} + M_{00}) / (M_{01} + M_{10} + M_{11} + M_{00})$$

J = number of 11 matches / number of not-both-zero attributes values

$$= (M_{11}) / (M_{01} + M_{10} + M_{11})$$



SMC versus Jaccard: Example

$$p = 1000000000$$

$$q = 0000001001$$

$M_{01} = 2$ (the number of attributes where p was 0 and q was 1)

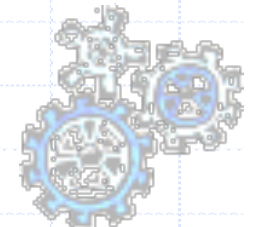
$M_{10} = 1$ (the number of attributes where p was 1 and q was 0)

$M_{00} = 7$ (the number of attributes where p was 0 and q was 0)

$M_{11} = 0$ (the number of attributes where p was 1 and q was 1)

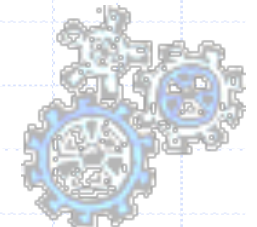
$$\text{SMC} = (M_{11} + M_{00}) / (M_{01} + M_{10} + M_{11} + M_{00}) = (0+7) / (2+1+0+7) = 0.7$$

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



Document Data

	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



Cosine Similarity

◆ If d_1 and d_2 are two document vectors, then

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

where \cdot indicates vector dot product and $||d||$ is the length of vector d .

◆ Example:

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

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

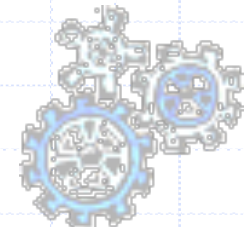
$$d_1 \cdot d_2 = 3*1 + 2*0 + 0*0 + 5*0 + 0*0 + 0*0 + 0*0 + 2*1 + 0*0 + 0*2 = 5$$

$$||d_1|| = (3*3 + 2*2 + 0*0 + 5*5 + 0*0 + 0*0 + 0*0 + 2*2 + 0*0 + 0*0)^{0.5} = (42)^{0.5} = 6.481$$

$$||d_2|| = (1*1 + 0*0 + 0*0 + 0*0 + 0*0 + 0*0 + 0*0 + 1*1 + 0*0 + 2*2)^{0.5} = (6)^{0.5} = 2.245$$

$$\cos(d_1, d_2) = .3150$$

Master MAINS, 2015



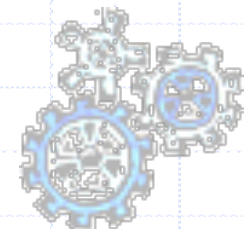
Correlation

- ◆ Correlation measures the linear relationship between objects (binary or continuous)
- ◆ To compute correlation, we standardize data objects, p and q , and then take their dot product (covariance/standard deviation)

$$p'_k = (p_k - \text{mean}(p)) / \text{std}(p)$$

$$q'_k = (q_k - \text{mean}(q)) / \text{std}(q)$$

$$\text{correlation}(p, q) = p' \cdot q'$$



General Approach for Combining Similarities

◆ Sometimes attributes are of many different types, but an overall similarity is needed.

1. For the k^{th} attribute, compute a similarity, s_k , in the range $[0, 1]$.
2. Define an indicator variable, δ_k , for the k^{th} attribute as follows:

$$\delta_k = \begin{cases} 0 & \text{if the } k^{th} \text{ attribute is a binary asymmetric attribute and both objects have} \\ & \text{a value of 0, or if one of the objects has a missing values for the } k^{th} \text{ attribute} \\ 1 & \text{otherwise} \end{cases}$$

3. Compute the overall similarity between the two objects using the following formula:

$$\text{similarity}(p, q) = \frac{\sum_{k=1}^n \delta_k s_k}{\sum_{k=1}^n \delta_k}$$

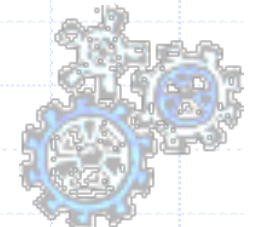


Using Weights to Combine Similarities

- ◆ May not want to treat all attributes the same.
 - Use weights w_k which are between 0 and 1 and sum to 1.

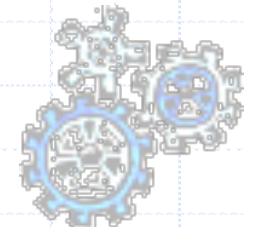
$$\text{similarity}(p, q) = \frac{\sum_{k=1}^n w_k \delta_k s_k}{\sum_{k=1}^n \delta_k}$$

$$\text{distance}(p, q) = \left(\sum_{k=1}^n w_k |p_k - q_k|^r \right)^{1/r}$$



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} $(R_i * C_j / N)$

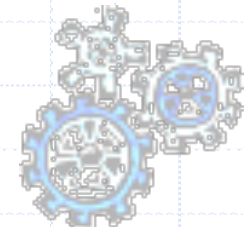


Test del Chi Quadro per la discretizzazione

	1	2	...	K	Total
1	A_{11}	A_{12}	...	A_{1k}	R_1
2	A_{21}	A_{22}	...	A_{2k}	R_2
Total	C_1	C_2	...	C_k	N

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

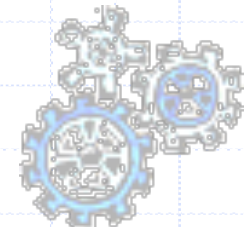
- ◆ Si individua quanto “distinti” sono due intervalli
- ◆ $k-1$ gradi di liberta`
- ◆ La significativita` del test è data da un threshold δ
 - Probabilita` che l’intervallo in questione e la classe siano indipendenti



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

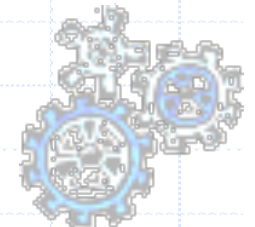
- ◆ Discretizzazione w.r.t. Beer
- ◆ threshold 50% confidenza
- ◆ Vogliamo ottenere una discretizzazione del prezzo che permetta di mantenere omogeneità w.r.t. Beer



Esempio: Chi Values

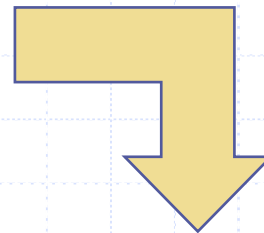
	<i>Bud</i>	<i>Becks</i>
100	1	0
110	2	0
120	1	1
125	1	0
130	2	0
135	1	0
140	0	1
150	0	1
160	0	1

Scegliamo gli elementi adiacenti
con Chi-Value minimo



Esempio: passo 1

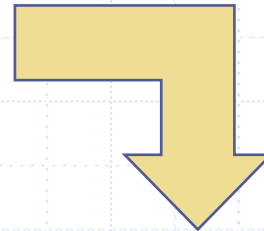
	<i>Bud</i>	<i>Becks</i>	<i>Chi Value</i>
100	1	0	0
110	2	0	1.33333
120	1	1	0.75
125	1	0	0
130	2	0	0
135	1	0	2
140	0	1	0
150	0	1	0
160	0	1	1.38629



	<i>Bud</i>	<i>Becks</i>	<i>Chi Value</i>
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

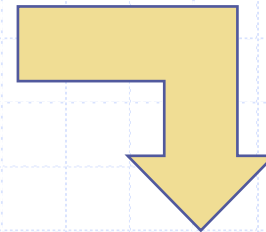
	<i>Bud</i>	<i>Becks</i>	<i>Chi Value</i>
100	1	0	0
110	2	0	1.33333
120	1	1	0.75
125	1	0	0
130	2	0	0
135	1	0	2
140	0	1	0
150-160	0	2	1.38629



	<i>Bud</i>	<i>Becks</i>	<i>Chi Value</i>
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

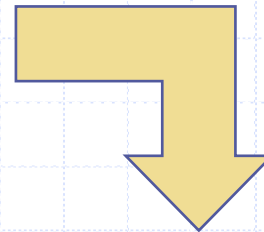
	<i>Bud</i>	<i>Becks</i>	<i>Chi Value</i>
100	1	0	0
110	2	0	1.33333
120	1	1	0.75
125	1	0	0
130	2	0	0
135	1	0	4
140-150-160	0	3	1.38629



	<i>Bud</i>	<i>Becks</i>	<i>Chi Value</i>
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

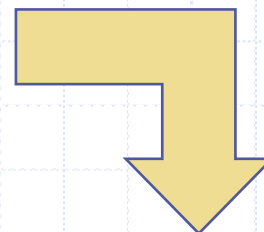
	<i>Bud</i>	<i>Becks</i>	<i>Chi Value</i>
100	1	0	0
110	2	0	1.33333
120	1	1	0.75
125	1	0	0
130-135	3	0	6
140-150-160	0	3	1.38629



	<i>Bud</i>	<i>Becks</i>	<i>Chi Value</i>
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

	<i>Bud</i>	<i>Becks</i>	<i>Chi Value</i>
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)

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

Esercitazione KNIME



Appendice

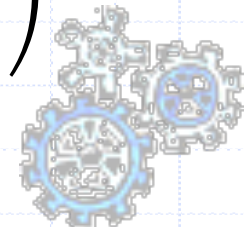
Misure descrittive dei dati

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

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

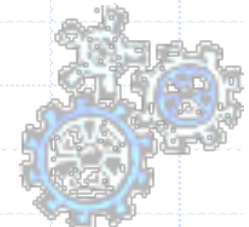
- ◆ Per bilanciare proporzioni
- ◆ dati moltiplicativi

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

<i>Prodotto</i>	<i>Variazioni Prezzi</i>	
	1996	1997
A	100	200
B	100	50
<i>Media</i>	100	125

$$x_g = 100$$

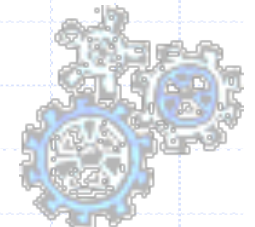
$$\log x_g = \frac{1}{n} \sum_{i=1}^n \log x_i$$



Media Armonica

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

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



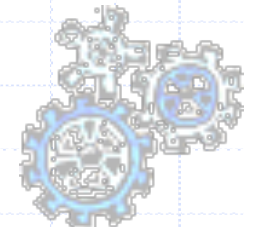
Mediana

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

1 7 12 18 23 34 54

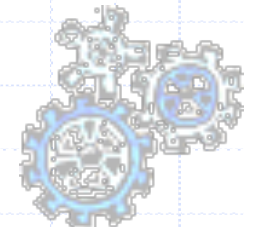
$$\bar{x} = 21.3$$

$$M = 23$$



Mediana e Quartili

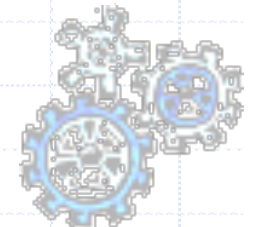
- ◆ Divide un insieme di dati a metà
 - statistica robusta (non influenzata da valori con rilevanti differenze)
 - ulteriori punti di divisione
- ◆ interquartili
 - mediane degli intervalli dei dati superiore e inferiore
 - Un quarto dei dati osservati è sopra/sotto il quartile
- ◆ percentili
 - di grado 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

<i>Etichetta</i>	<i>P</i>
M	$\frac{1}{2}=0.5$
F	$\frac{1}{4}=0.25$
E	$\frac{1}{8}=0.125$
D	$\frac{1}{16}=0.0625$
C	$\frac{1}{32}=0.03125$
B	$\frac{1}{64}$
A	$\frac{1}{128}$
Z	$\frac{1}{256}$
Y	$\frac{1}{512}$
X	$\frac{1}{1024}$



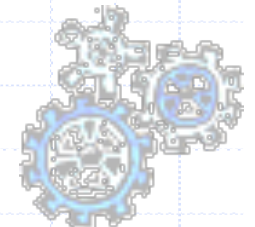
Moda

- ◆ 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

$$r = \max - \min$$

- ◆ Scarti interquantili

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

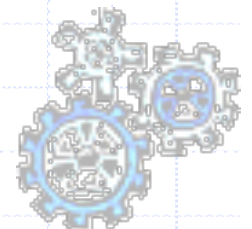
- ◆ Scarto medio assoluto

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

- ◆ Scarto medio assoluto dalla mediana

- *In generale, $S_{.5} \leq S_n$*

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



Varianza, deviazione standard

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

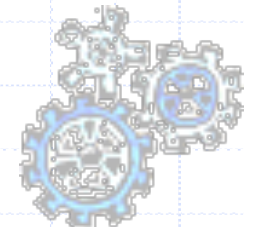
$$dev = \sum_{i=1}^n (x_i - \bar{x})^2$$

- ◆ Varianza

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

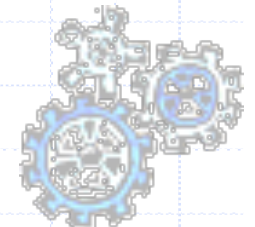
- ◆ Coefficiente di variazione
 - misura relativa

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



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

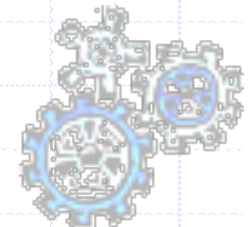
$$\bar{x}_p = (x_{1-p} + x_p) / 2$$

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

◆ indice di Fisher

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

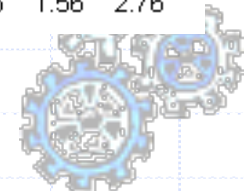
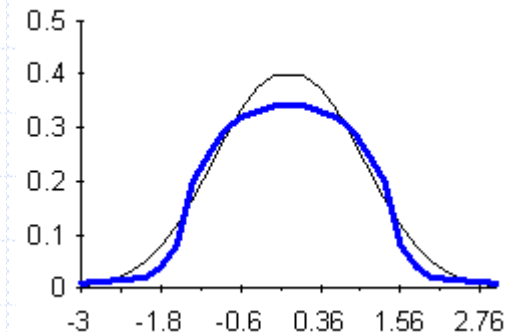
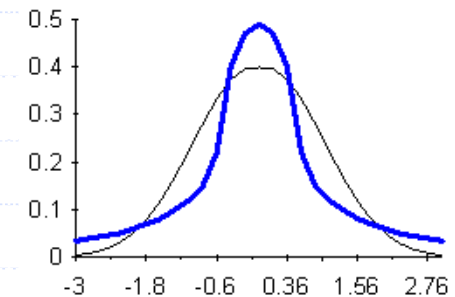
$$\gamma = \frac{m_3}{\hat{s}^3}$$



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

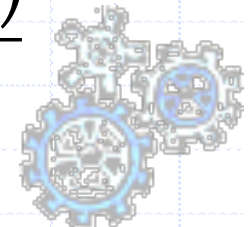
- $\beta=3$: distribuzione mesocurtica
- $\beta > 3$: distribuzione leptocurtica
- $\beta < 3$: distribuzione platicurtica

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

◆ Coefficiente di curtosi

- Una distribuzione leptocurtica ha $K \sim 1/2$
- platicurtosi: $k \sim 0$

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



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

$$r_{xy} = \frac{Cov(x, y)}{S_x S_y}$$

