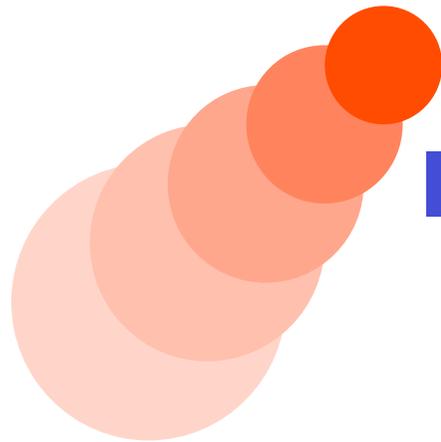


Data Warehousing, Data Mining & Business Intelligence



Data Preparation

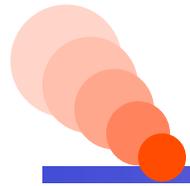
Paolo G. Franciosa

Dipartimento di Statistica, Probabilità e Statistiche Applicate

Università "La Sapienza"

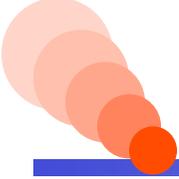
paolo.franciosa@uniroma1.it

Questo materiale deriva dalla traduzione e adattamento delle presentazioni
pubblicate dal prof. Jiawei Han



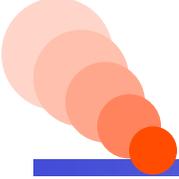
Data Preprocessing

- Motivazioni
- Data cleaning
- Integrazione e trasformazione
- Riduzione
- Discretizzazione e gerarchie di concetti



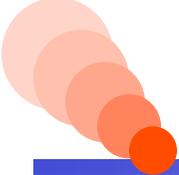
Motivazioni

- Dati reali:
 - **incompleti**: valori mancanti, attributi di interesse mancanti, dati aggregati (mancano microdati)
 - **rumorosi**: errori o anomalie
 - **inconsistenti**: discrepanze in codifiche o nomi
- La qualità delle decisioni dipende dalla qualità dei dati
- L'azione di integrazione e pulizia è fondamentale



Misurare la qualità

- Accuratezza
- Completezza
- Consistenza
- Aggiornamento
- Credibilità
- Valore aggiunto
- Interpretabilità
- Accessibilità



Processi principali

➤ Data cleaning

- Riempire campi mancanti, correggere valori affetti da rumore, individuare o eliminare anomalie, risolvere inconsistenze

➤ Integrazione

- Integrazione di database, data cubes o file eterogenei

➤ Trasformazione

- Normalizzazione e aggregazione

➤ Riduzione

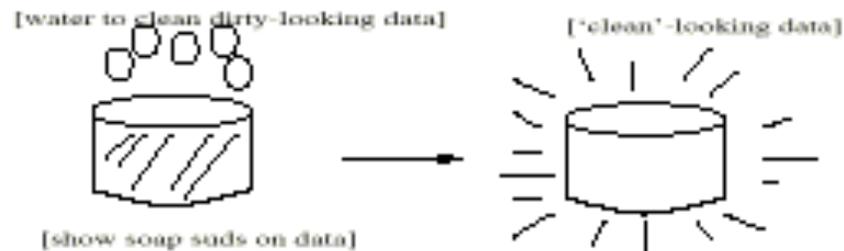
- Limitazione del volume dei dati per ottenere risultati analitici simili

➤ Discretizzazione

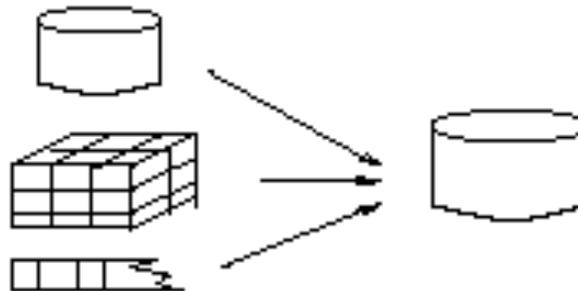
- Un tipo di riduzione per dati numerici

Aspetti del preprocessing

Data Cleaning



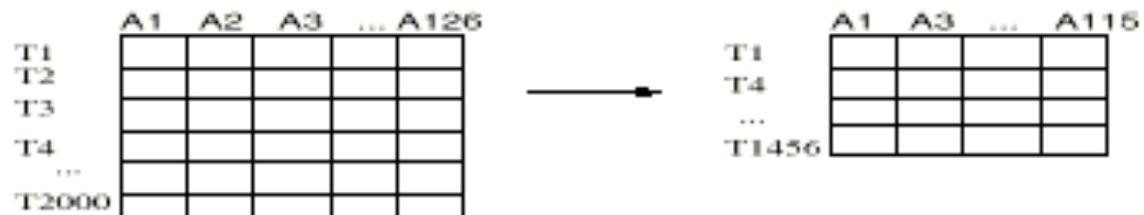
Data Integration

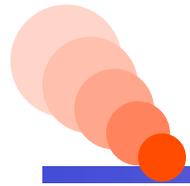


Data Transformation

-2, 32, 100, 59, 48 → -0.02, 0.32, 1.00, 0.59, 0.48

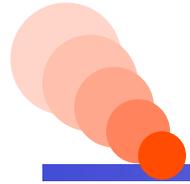
Data Reduction





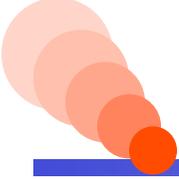
Data Preprocessing

- Motivazioni
- Data cleaning
- Integrazione e trasformazione
- Riduzione
- Discretizzazione e gerarchie di concetti



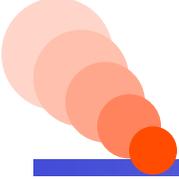
Data Cleaning

- Riempire campi mancanti
- Correggere rumore, individuare o eliminare anomalie
- Risolvere inconsistenze



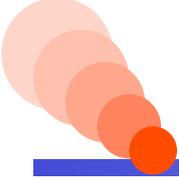
Missing Data

- Non sempre i dati sono disponibili
- Cause:
 - malfunzionamenti
 - cancellati a causa di inconsistenze
 - non immessi per incomprensioni
 - alcuni dati potrebbero essere considerati irrilevanti al momento del data entry
 - mancanza di dati storici
- I dati mancanti possono essere inferiti?



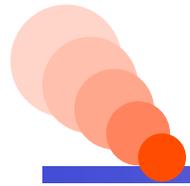
Come gestire i dati mancanti?

- Ignorare la tupla: generalmente quando manca il valore target
- Riempimento manuale (non sempre possibile)
- Rimpiazzare con un valore predefinito (null?). Non consigliabile per attributi target
- Rimpiazzare con la media dei valori per l'attributo
- Rimpiazzare con la media dei valori per l'attributo, calcolata per ciascuna classe
- Rimpiazzare con il valore più "probabile" (richiede attività di analisi)



Dati rumorosi (1)

- Rumore: errore random oppure varianza in dati misurati
- Valori errati dovuti a:
 - dispositivi di raccolta affetti da errore
 - problemi nel data entry
 - problemi di trasmissione
 - limitazioni tecnologiche
 - inconsistenze di varia natura
- Altri motivi per il data cleaning
 - record duplicati
 - dati incompleti
 - dati inconsistenti



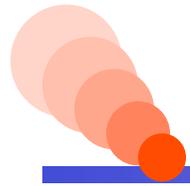
Dati rumorosi (2)

- Binning:
 - partizionare i dati ordinati in “bins” della stessa cardinalità
 - smooth by bin means, smooth by bin median, smooth by bin boundaries, etc.

- Clustering
 - eliminare outliers

- Analisi (semi)manuale
 - scoprire ed eliminare valori sospetti

- Regressione



Discretizzazione: Binning

- Partizionamento basato su **ampiezza** (distanza):
 - partiziona secondo una griglia regolare (equispaziata), che divide in N intervalli la distanza tra minimo e massimo per l'attributo
 - molto diretta
 - affetta da outliers
 - non adatta a dati "concentrati"
 - richiede la definizione di una metrica
- Partizionamento basato su **frequenza**:
 - partiziona in N intervalli che contengono lo stesso numero di osservazioni
 - è sufficiente definire un ordine totale sull'attributo

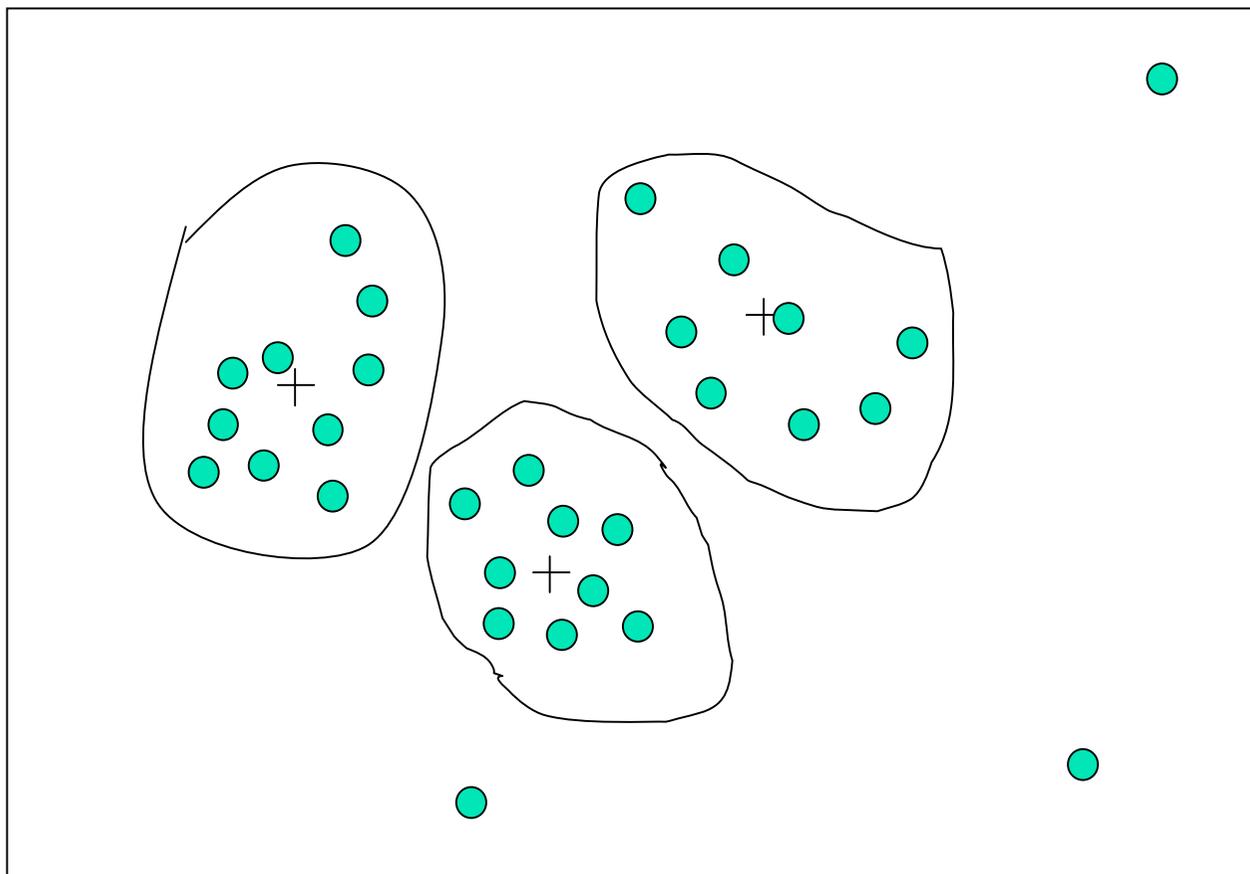
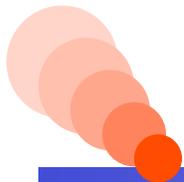


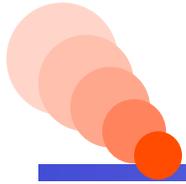
Smoothing

Prezzi: 4, 8, 9, 15, 21, 21, 24, 25, 26, 28, 29, 34

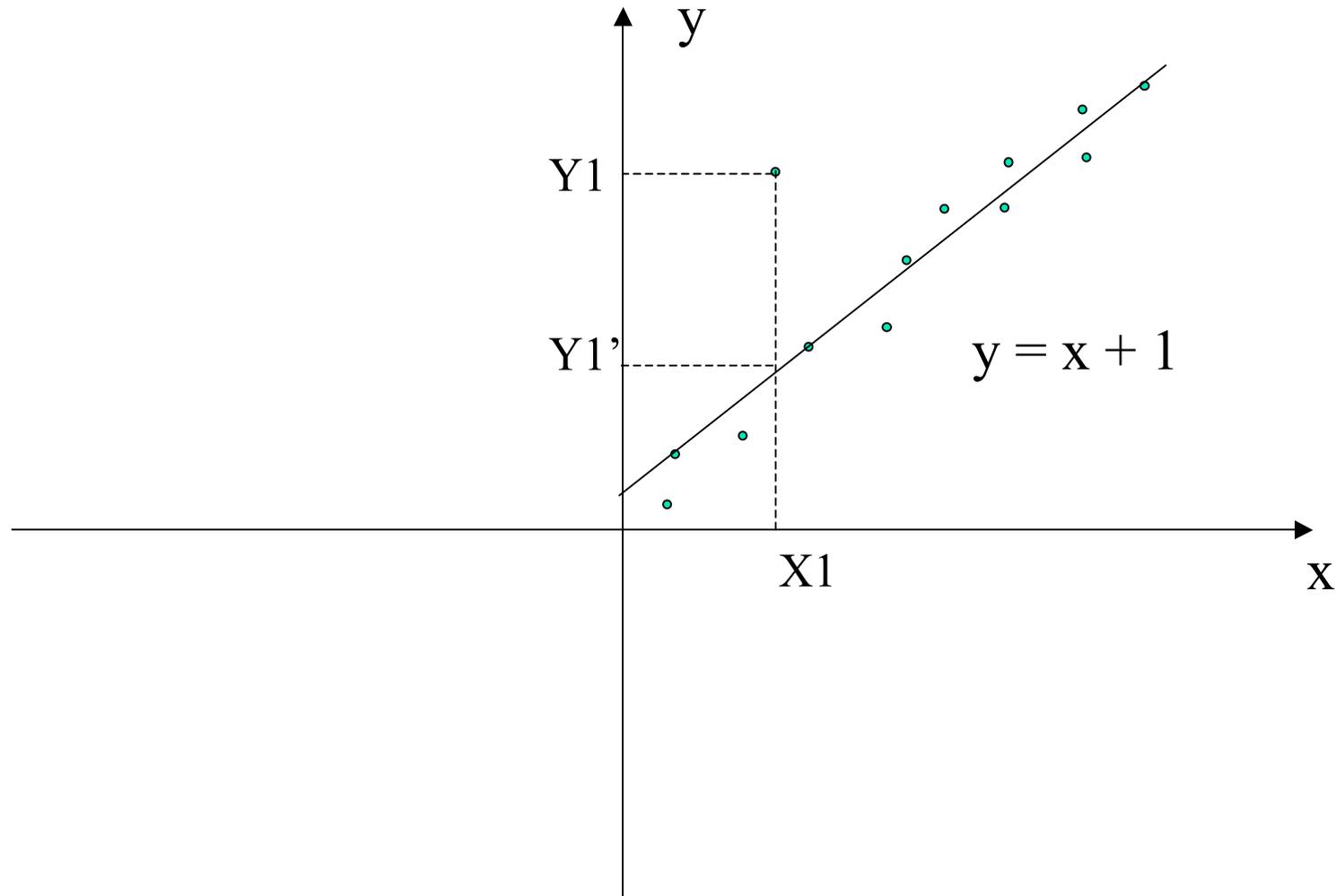
- Partiziono sulla frequenza:
 - Bin 1: 4, 8, 9, 15
 - Bin 2: 21, 21, 24, 25
 - Bin 3: 26, 28, 29, 34
- Smoothing by bin means:
 - Bin 1: 9, 9, 9, 9
 - Bin 2: 23, 23, 23, 23
 - Bin 3: 29, 29, 29, 29
- Smoothing by bin boundaries:
 - Bin 1: 4, 4, 4, 15
 - Bin 2: 21, 21, 25, 25
 - Bin 3: 26, 26, 26, 34

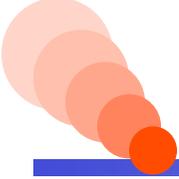
Cluster Analysis





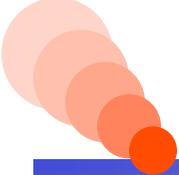
Regressione





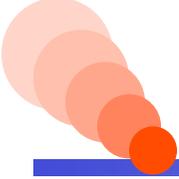
Data Preprocessing

- Motivazioni
- Data cleaning
- **Integrazione e trasformazione**
- Riduzione
- Discretizzazione e gerarchie di concetti



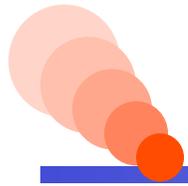
Integrazione

- Integrazione dei dati:
 - associare coerentemente dati da sorgenti eterogenee
- Integrazione degli schemi
 - integrare metadati
 - identificare/distinguere concetti e attributi
- Risolvere conflitti tra i valori
 - dovuti a unità di misura, precisione, etc.



Trasformazione dei dati

- Smoothing: eliminare rumore
- Aggregazione
- Generalizzazione: variare livello di dettaglio (roll-up)
- Normalizzazione:
 - min-max
 - z-score
 - decimal scaling
- Generazione di nuovi attributi (derivati)



Normalizzazione

- min-max

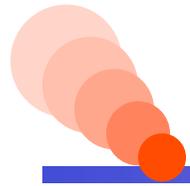
$$v' = \frac{v - \mathit{min}_A}{\mathit{max}_A - \mathit{min}_A} (\mathit{new_max}_A - \mathit{new_min}_A) + \mathit{new_min}_A$$

- z-score

$$v' = \frac{v - \mathit{mean}_A}{\mathit{stddev}_A}$$

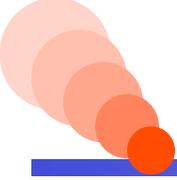
- decimal scaling

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



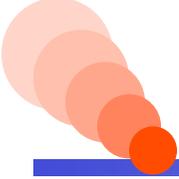
Data Preprocessing

- Motivazioni
- Data cleaning
- Integrazione e trasformazione
- **Riduzione**
- Discretizzazione e gerarchie di concetti



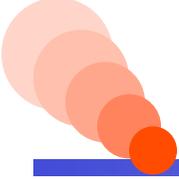
Riduzione: generalità

- Ridurre il volume dei dati per accelerare i processi di analisi/mining
 - ottenere una rappresentazione molto più compatta sulla quale sia possibile ottenere gli stessi risultati
- **Strategie**
 - Aggregazione
 - Riduzione dimensionale
 - Riduzione di cardinalità
 - Discretizzazione nella gerarchia dei concetti



Aggregazione

- Livello minimo nel data cube (base data cuboid)
 - dati relativi a entità individuali (p.es. un cliente nel data warehouse di un gestore telefonico)
- Livelli superiori di aggregazione nel datacube
 - riducono molto il volume dei dati
- Scegliere livelli appropriati
 - individuare la rappresentazione minima che permette di risolvere il problema



Riduzione dimensionale

- Scelta degli attributi rilevanti:
 - insieme di attributi che spiegano il fenomeno
 - permette di applicare algoritmi più efficienti
 - permette di ottenere modelli più comprensibili

- Metodi euristici (spazio di ricerca esponenziale):
 - step-wise forward selection
 - step-wise backward elimination
 - studio degli alberi di decisione: conservo gli attributi usati come discriminatori nei nodi interni dell'albero