Sistemi informativi aziendali
ERP e sistemi di data analysis

Data mining

Maurizio Pighin, Anna Marzona

Uso dei Data Warehouse

- Esistono tre tipi di applicazioni front-end per data warehouse
  - *Analisi statica (reporting)*
    - elaborazione di situazioni aziendali da lanciare con elevata periodicità, con modalità invarianti nel tempo
    - riflette informazione di base
  - *Analisi interattiva (OLAP)*
    - analisi interattiva basata su ipotesi
    - supporta operazioni OLAP di base: slice-dice, drill up-down, pivot
  - *Data mining*
    - fa emergere nuova conoscenza rilevando pattern nascosti
    - supporta modelli descrittivi e predittivi

Limiti dei sistemi di analisi OLAP

- Le informazioni non sono facilmente identificabili
  - *Quantità elevata dei dati*
  - *Complessità elevata delle relazioni esistenti tra i dati*
- In assenza di strumenti adeguati i dati raccolti corrono il rischio di restare sotto-utilizzati
  - *Il sistema è ricco di dati, ma povero di informazioni*
- Gli strumenti OLAP non sono sufficienti
  - *Operano a supporto di processi deduttivi dei decisori*
  - *Sviluppano percorsi di analisi da ipotesi formulate dall'utente, limitate dal suo bagaglio cognitivo*

Tipi di inferenza

Deductive, Inductive, and Abductive Syllogisms

Deductive	Inductive	Abductive
All men are mortal;	Socrates is a man;	All men are mortal;
Socrates is a man;	Socrates is mortal;	Socrates is mortal;
∴ Socrates is mortal.	∴ All men are mortal.	∴ Socrates is a man.

Adapted from: Hui, J., Cashman, T. and T. Deacon. 2008. Bateson's Method: Double Description. What is It? How Does It Work? What Do We Learn? in J. Hoffmeyer (ed.) A Legacy for Living Systems: Gregory Bateson As Precursor to Biosemiotics.

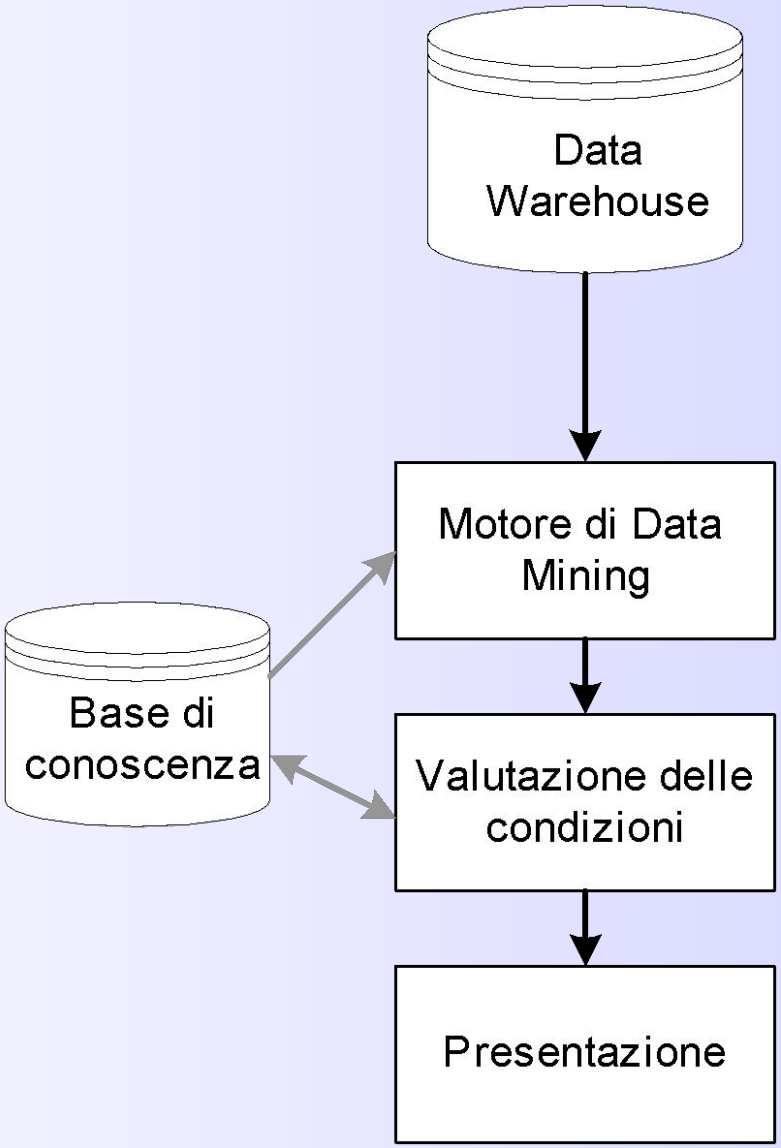
Data mining

- Attività volta a riconoscere ed estrarre automaticamente informazione da basi di dati di grandi dimensioni
- Passi del processo di mining
  - *Pulizia*
  - *Integrazione*
  - *Selezione*
  - *Trasformazione*
  - *Data mining*
  - *Valutazione dei pattern*
  - *Presentazione della conoscenza*
- Le prime fasi coincidono con quelle popolamento dei sistemi di data warehousing
- In ambito aziendale il data mining può essere considerato
  - *Ampliamento del sistema di data warehousing*
  - *Complemento dei sistemi OLAP di analisi dei dati*

Da OLAP a OLAM (On Line Analytical Mining)

- Partire dai Data warehouse garantisce l'accesso a dati ben strutturati, puliti, completi
- Il processo di mining non può essere completamente automatico
  - *I pattern rilevati potrebbero essere troppi e non interessanti*
- Il data mining deve essere un processo interattivo
  - *Gli utenti indicano la direzione in cui "scavare"*
- Lavorare con uno strumento interattivo consente l'affinamento iterativo delle ricerche

Architettura dei sistemi di Data mining



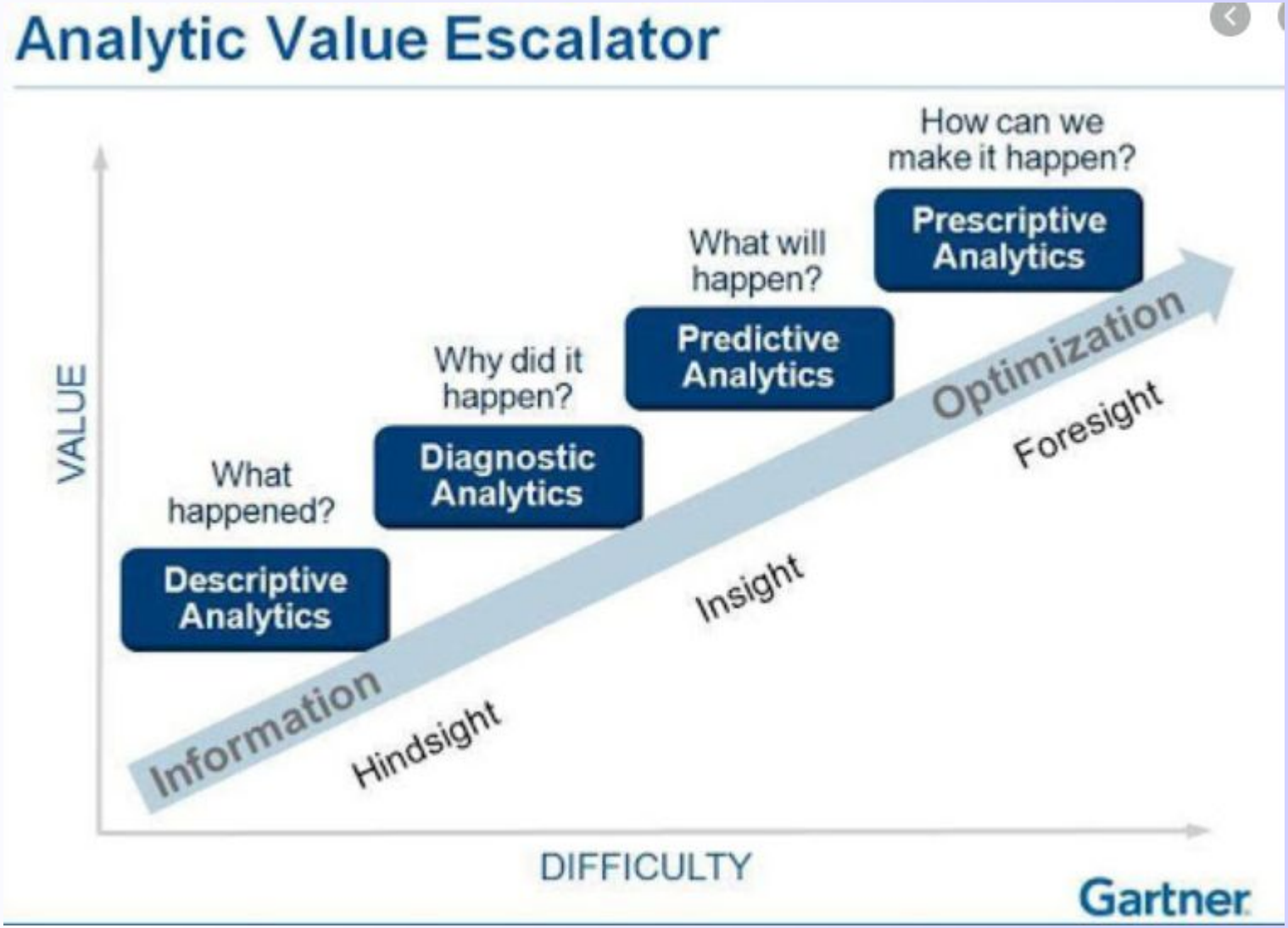
Architettura dei sistemi di Data Mining

- Data warehouse
  - *E' una base di dati pronta, di elevata qualità e multidimensionale*
  - *I dati da analizzare sono definiti da un'interrogazione OLAP*
- Base di conoscenza (Knowledge Base)
  - *Insieme di regole e conoscenze 'date per note' utilizzate per guidare la ricerca e per filtrare i risultati sulla base del loro effettivo interesse*
- Motore di data mining (Data Mining Engine)
  - *Insieme delle funzioni di analisi dei dati*
- Sistema di valutazione delle condizioni (Pattern Evaluation)
  - *Effettua un postprocessing delle informazioni estratte dal mining (pattern) mantenendo le sole condizioni interessanti*
- Sistema di presentazione
  - *Interfaccia utente per l'attivazione delle funzioni di mining e la visualizzazione dei pattern*

Macro classi del mining

- Le attività di mining possono essere ripartite in due macro classi
  - *Mining descrittivo*
    - estrae informazioni che descrivono le proprietà generali dei dati
  - *Mining predittivo*
    - determina regole generali e crea modelli per predire le tendenze nel futuro

I 4 principali tipi di analisi



I 4 principali tipi di analisi

https://lorenzogovoni.com/il-tipo-di-analisi-dei-dati-piu-semplice-analisi-descrittiva/

Funzioni di mining

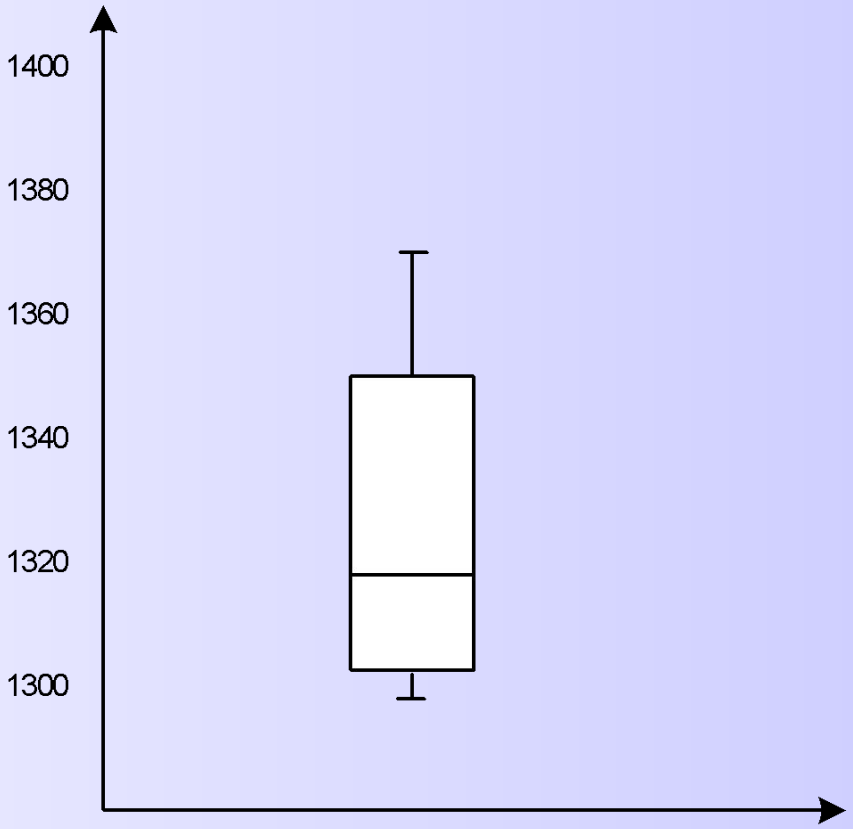
- I sistemi presenti sul mercato propongono diversi insiemi di funzioni di mining
- Ogni funzione permette di ricercare un certo tipo di informazione o costruire un particolare modello di predizione
- La stessa funzione può essere elaborata tramite algoritmi diversi
- Le funzioni sono riconducibili a cinque tipologie
  - *Caratterizzazione e discriminazione*
  - *Analisi associativa*
  - *Classificazione e predizione*
  - *Analisi dei cluster*
  - *Analisi degli outlier*

Caratterizzazione e discriminazione

- Strumenti che permettono di descrivere in modo sintetico ma preciso i dati contenuti nel database
- Operano tramite
  - *Generalizzazione*
    - classificazione dei dati elementari in gruppi (classi) caratterizzati da attributi comuni
    - opera tramite tecniche OLAP e funzioni di induzione sugli attributi
  - *Funzioni di descrizione delle classi*
    - caratterizzazione: descrivono le particolarità della classe
    - discriminazione: marcano le differenze tra classe e classe

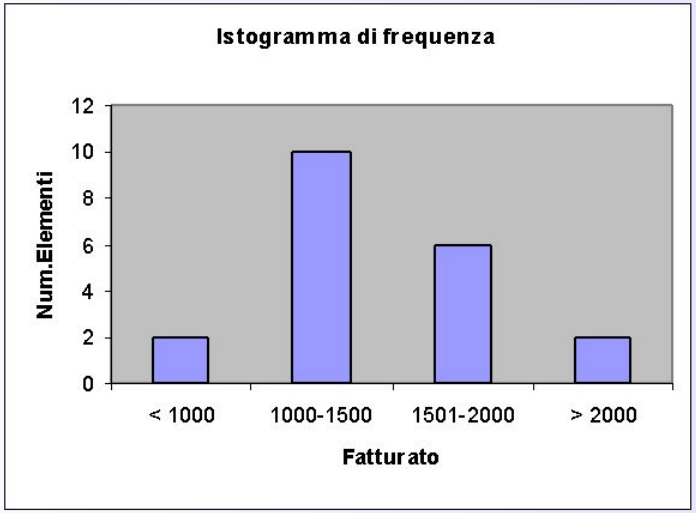
Caratterizzazione e discriminazione

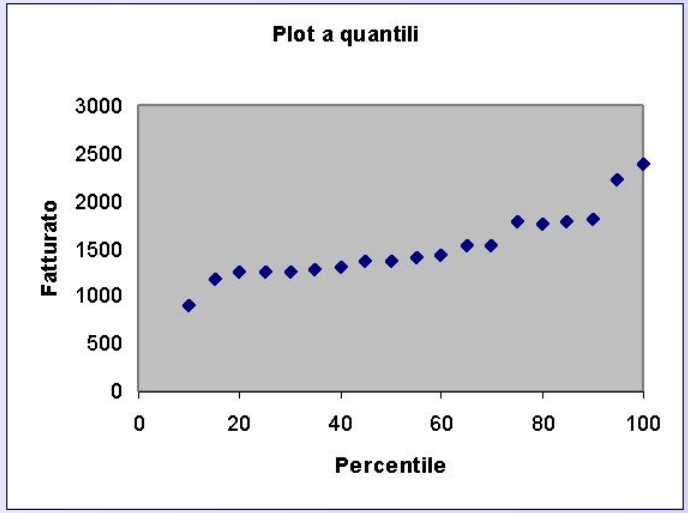
- Caratterizzazione
  - *Descrive le caratteristiche di una classe*
  - *Calcola misure*
    - di tendenza
    - di dispersione
  - *Rappresentazione dei dati*
    - Tabellare, Grafica, Boxplot
- Discriminazione
  - *Permette di rilevare le differenze tra una classe e classi diverse paragonabili tramite il confronto diretto dei dati su tabelle o su grafici*

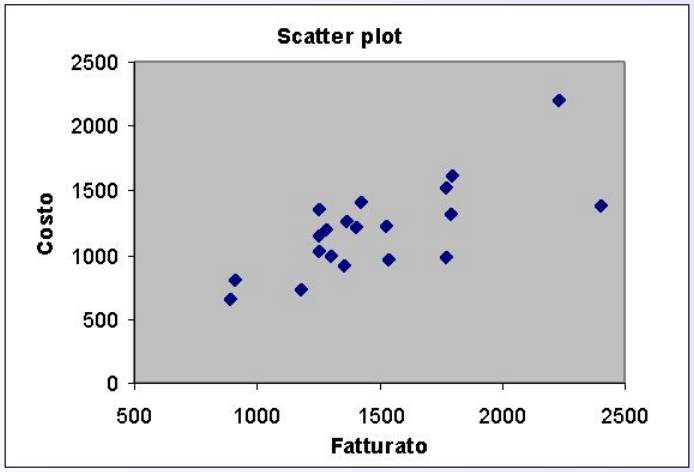


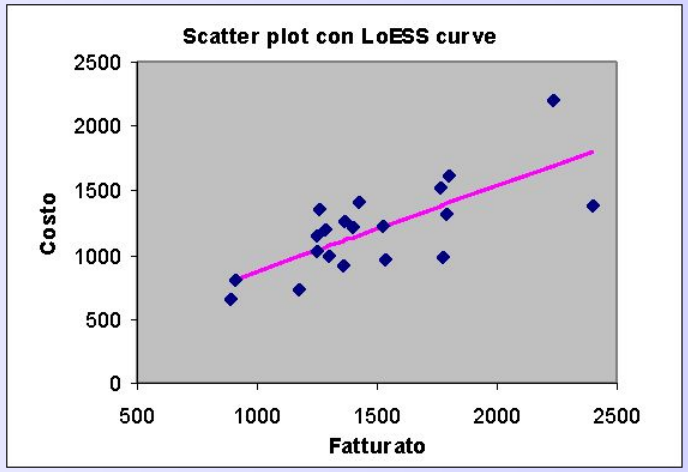
Esempio di boxplot sui dati di una classe

Caratterizzazione e discriminazione









Esempi di rappresentazioni grafiche delle caratteristiche dei dati appartenenti ad una classe

Analisi associativa

- Permette di identificare condizioni che si verificano contemporaneamente con elevata frequenza
- Rileva pattern che si ripetono su determinati attributi e ne deriva regole di implicazione del tipo

  $A \Rightarrow B$
- Esempi
  - *compra(X, "divano 2 posti") ⇒ compra (X, "poltrona")*
  - *fatturato (X, "> 100M") ∧ struttura(X, "Spa") ⇒ compra(X, "Jaguar")*
- Applicazioni
  - *market basket analysis*
  - *profili clienti (abitudini di acquisto)*
  - *ottimizzazione delle manutenzioni*
  - *...*

Significatività delle associazioni

- Viene valutata in base a
  - *Confidenza: misura la certezza del pattern*
  - *Supporto: misura la frequenza con cui il pattern è presente sulla base di dati*
  - *Esempio*

    Compra(X, "divano 2 posti") ⇒ Compra (X, "poltrona") [c.85%;s.30%]

    - L'85% di tutti coloro che comprano un divano 2 posti compra anche una poltrona
    - Nel 30% delle vendite il cliente ha comprato sia un divano a due posti che una poltrona

Classificazione e predizione

- Costruzione di modelli per
  - *Predire gli eventi futuri*
  - *Stimare il valore di elementi non noti*
- Classificazione
  - *Definizione di criteri che permettono di assegnare un soggetto ad una classe*
- Predizione
  - *Calcolo di funzioni di tendenza continue tramite l'interpolazione dei dati noti*

Classificazione e predizione

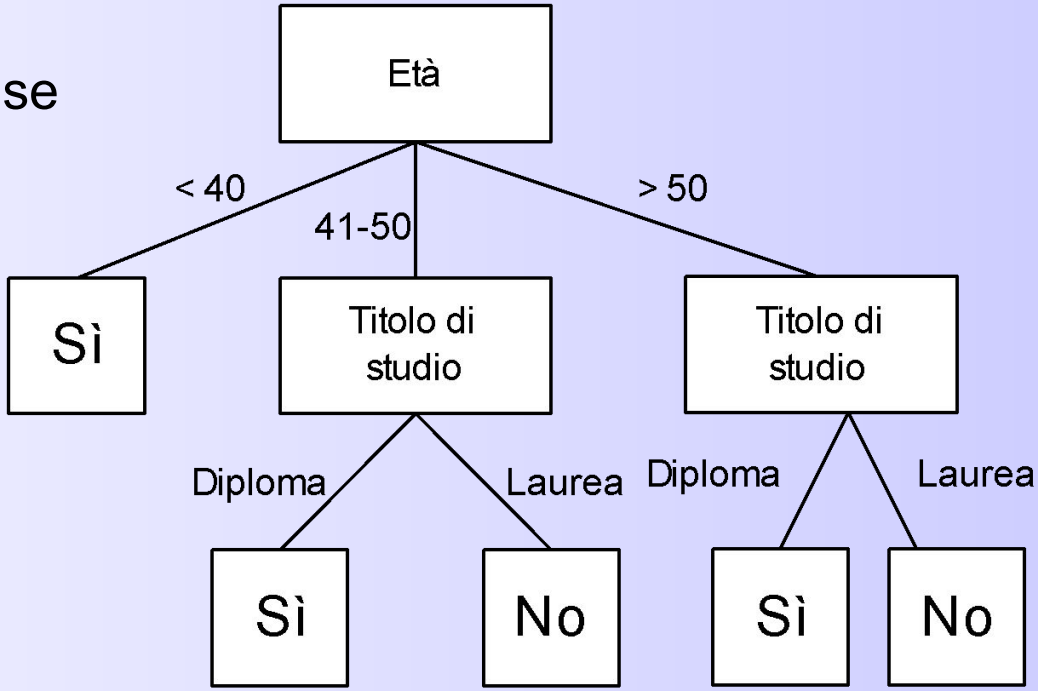
- Costruzione “basata su esempi”
  - *Il modello deriva da un sottoinsieme significativo dei dati esistenti*
  - *L’efficacia viene testata su un sottoinsieme diverso (disgiunto) dei dati*
  - *Se il modello si rivela efficace può essere usato come ‘predittore’*
- Applicazioni
  - *Propensione all’acquisto dei clienti*
  - *Qualità dei fornitori*
  - *Affidabilità dei prodotti*
  - *...*

Classificazione

- Permette di indicare l'appartenenza di un elemento ad una certa classe
- Diversi tipi di modelli
  - *Funzioni matematiche*
  - *Analisi statistiche*
  - *Regole associative*
  - *Alberi di decisione*
  - *Reti della verità bayesiane*
  - *Reti neurali*

Alberi di decisione

- Struttura di classificazione basata sulla valutazione di condizioni del tipo if-then-else
- Nodi interni
  - *Attributi del soggetto da classificare*
- Archi in uscita
  - *Etichettati con i valori che l'attributo può assumere*
- Nodi foglia
  - *Classi*
- La classificazione avviene seguendo un percorso guidato dai valori assunti dagli attributi dell'elemento da classificare

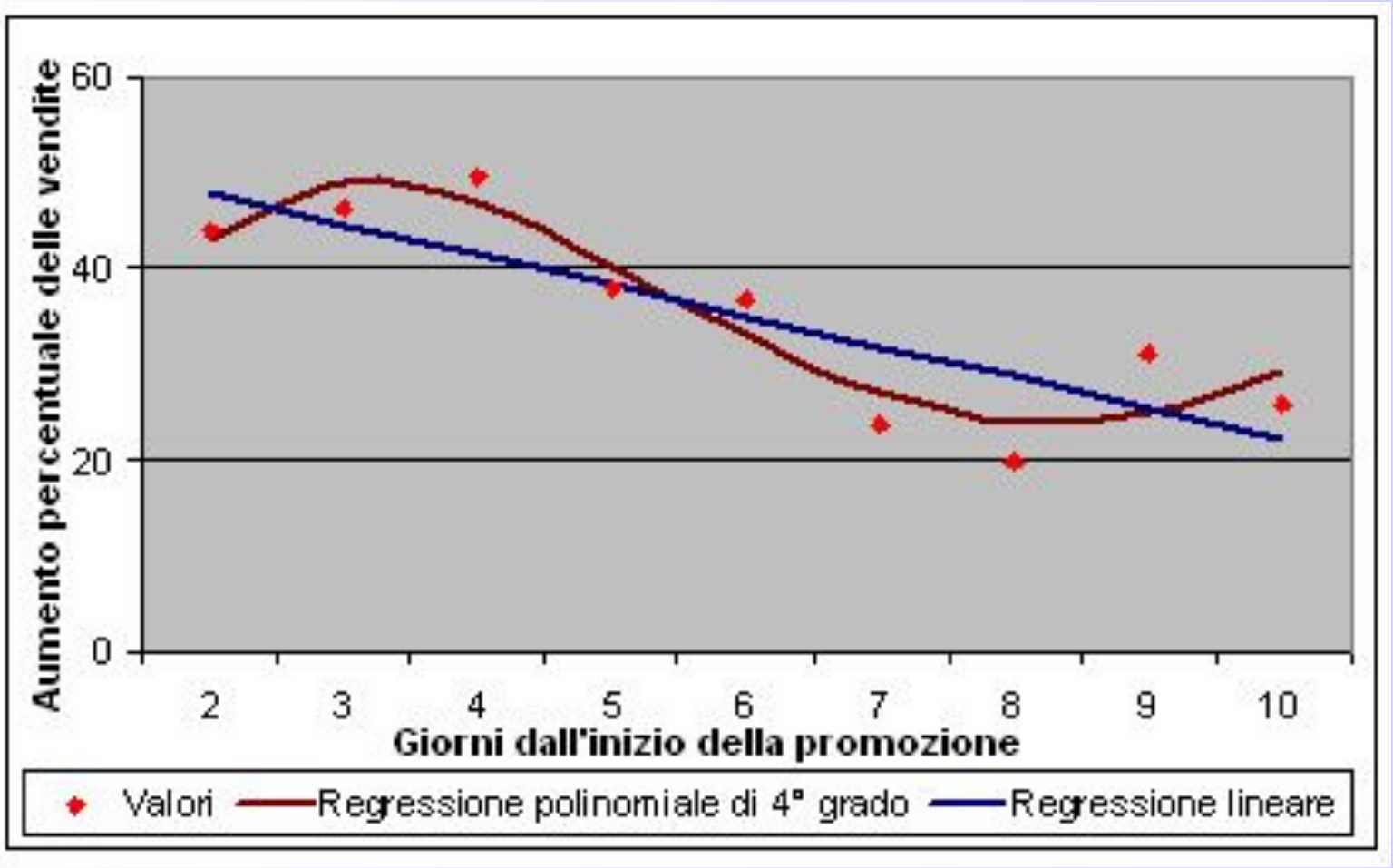


Esempio di albero di decisione

Predizione

- Permette di identificare valori non noti di elementi il cui dominio è continuo
- Costruzione delle funzioni di tendenza tramite interpolazione sui punti noti (regressione)
- Diversi modelli di regressione
  - *Lineare semplice (distribuzioni bivariate)*
    - $Y = q + m X$
  - *Multilineare (distribuzioni multivariate)*
    - $Y = q + m_1 X_1 + m_2 X_2 + m_3 X_3$
  - *Non-lineare (polinomiale, esponenziale, logaritmica, ...)*
    - $Y = q + m_1 X + m_2 X^2 + m_3 X^3$ (polinomiale di grado 3)

Predizione



Confronto tra regressione lineare e polinomiale di terzo grado sugli stessi punti

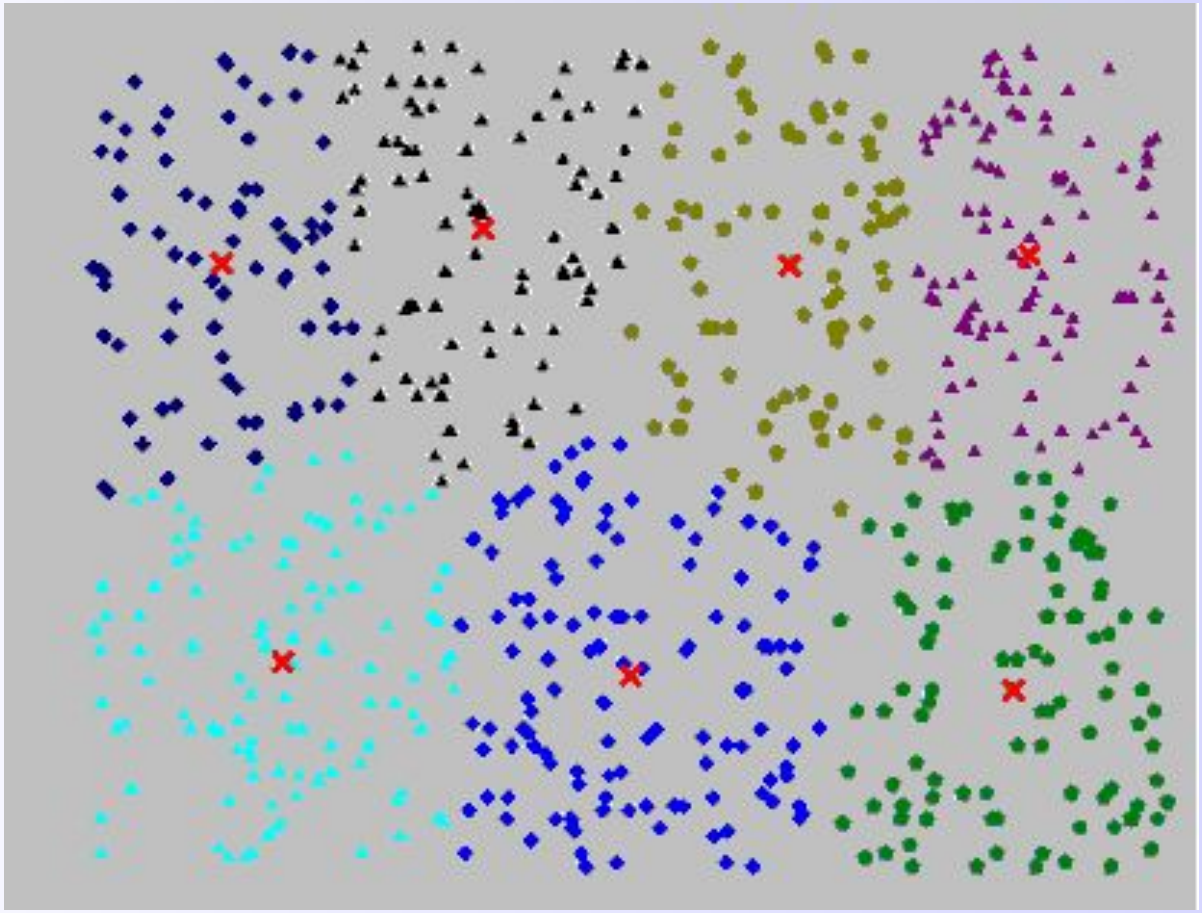
Clustering

- Ripartisce gli elementi in classi anonime sulla base delle affinità rilevate tramite l'osservazione dei dati
  - *Classi non definite a priori*
  - *Classi proposte all'utente come "agglomerati spontanei" di dati*
  - *Dall'analisi dei cluster l'utente può derivare informazioni e nuovi criteri su cui costruire modelli di classificazione*
- I cluster presentano
  - *La massima similarità tra gli elementi appartenenti ad una classe*
  - *La minima similarità tra gli elementi appartenenti classi diverse*

Clustering

- I metodi di clustering si fondano su diverse tecniche
  - *Partizionamento*
    - l'utente indica in quante classi ripartire i dati
    - l'algoritmo ripartisce gli elementi nel numero di classi indicato sulla base delle reciproche distanze
  - *Classificazione gerarchica*
    - Basata su aggregazione: costruisce le classi aggregando iterativamente gli elementi sulla base delle similitudini
    - Basata su divisione: ripartisce iterativamente l'insieme dei dati in sottoinsiemi di elementi simili
  - *Valutazione della densità*
    - i cluster sono identificati dalle zone topologicamente dense.

Clustering



Esempio di clustering con partizionamento su 7 classi

Ricerca degli outlier

- Outlier: eccezione, elemento fuori range
- La ricerca
  - *Si basa sugli stessi principi del clustering*
  - *Concentra gli sforzi sull'identificazione degli elementi che si discostano maggiormente dagli altri*
- Metodi per la ricerca degli outlier
  - *Statistici*
    - applicabili se sui dati è identificabile una distribuzione
  - *Basati sulla distanza*
    - ricercano gli elementi che massimizzano la distanza dai restanti elementi del set di analisi
  - *Basati sulla deviazione*
    - identificano gli outlier come elementi che 'deviano' dalle caratteristiche tendenziali del gruppo

Processo di mining dei dati

- L'utente effettua iterativamente interrogazioni di mining sul sistema
- Ogni analisi di mining dei dati si basa su
  - *Insieme dei dati di analisi*
    - query multidimensionale e condizioni di filtro
  - *Tipo di informazioni da ricercare*
    - funzione di mining che verrà attivata
  - *Misure di interesse*
    - criteri di interesse dei pattern
  - *Modalità di presentazione dei pattern*

Misure di interesse dei pattern

- Il mining può restituire insiemi molto numerosi di pattern
- E' necessario un passo di post-processing che permetta di identificare i pattern interessanti
- Caratteristiche che rendono un pattern interessante
  - *Novità*
    - riduzione tramite l'omissione di informazioni ridondanti
  - *Semplicità*
    - riduzione tramite valori di soglia (ad esempio sulla lunghezza delle regole associative, sul numero di livelli negli alberi di decisione)
  - *Certezza*
    - riduzione tramite valori di soglia (ad esempio su confidenza e supporto nell'analisi associativa)
  - *Utilità*
    - riduzione tramite valori di soglia (ad esempio, sul numero di elementi appartenenti ad un cluster)