DATA MINING: INTRODUZIONE E CLASSIFICAZIONE TECNICHE

Vedremo

- Cosa si intende per Data Mining
- Classificazione delle principali tecniche di Data Mining

Definizione di Data Mining

- **Processo iterativo** per l’analisi di grandi quantitativi di dati con l’obiettivo di estrarre informazione strategica
  - **accurata**
  - **utile ai fini del supporto decisionale**
  - **precedentemente sconosciuta**

- Le informazioni strategiche estratte rappresentano **nuova conoscenza**

Definizione di Data Mining (segue)

- Processo basato su **metodi di apprendimento induttivo** che permette di:
  - Trarre conclusioni da un un insieme di dati
  - Generalizzare le conclusioni individuate ad altri dati inizialmente non noti, nel modo più accurato possibile
- Due obiettivi:
  - **Metodi descrittivi/interpretativi: descrivere** i dati, rappresentandone in modo efficace le regolarità
  - **Metodi predittivi: predire** il valore che una variabile assumerà in futuro o stimare la probabilità di un evento futuro
- I dati sono generalmente memorizzati in un Data Warehouse
- Gli strumenti di front-end di un sistema di Data Warehousing includono gli strumenti di Data Mining

Confronto tra gli strumenti di front-end in un Data Warehouse

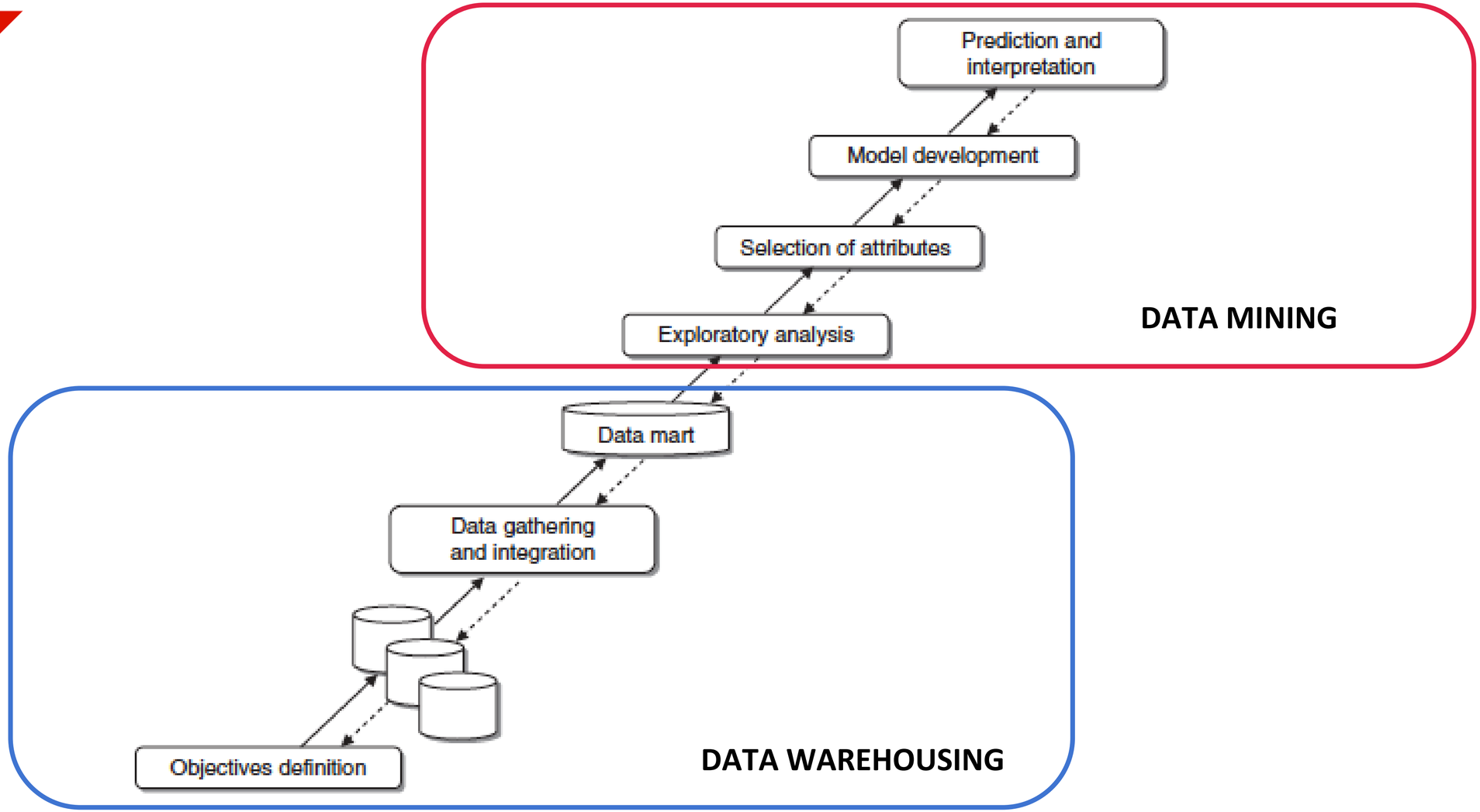
OLAP	Statistici	Data Mining
Estrazione di dati aggregati da informazioni di dettaglio	Verifica di ipotesi formulate	Individuazione di pattern e ricorrenze nei dati di dettaglio
Generazione di conoscenza (informazione aggregata)	Validazione dell'informazione aggregata	Generazione di conoscenza (modelli, pattern)
Analisi delle vendite per categoria di prodotto, al variare del mese di vendita	Analisi della varianza del ricavo rispetto alla categoria del prodotto	Caratterizzazione dei prodotti con caratteristiche simili rispetto alla vendita

Origini del Data Mining

- Si ispira a idee e tecniche derivate da **diversi settori**
  - Machine learning
  - Statistical learning
  - Pattern recognition
  - Sistemi di gestione dati

- Le tecniche vengono contestualizzate in un **processo analitico complesso, volto al supporto decisionale**

Processo di Data Mining



@Carlo Vercellis. Business Intelligence – Data Mining and Optimization for Decision Making. Wiley, 2009

Esempi di strumenti di Data Mining

- RapidMiner (rapidminer.com)
- Weka (www.cs.waikato.ac.nz/ml/weka/)
- Orange (https://orangedatamining.com)
- KNIME (knime.org, parzialmente commerciale)
- SAS (sas.com)
- IBM SPSS (www.ibm.com/software/analytics/spss/)
- IBM Cognos (www.ibm.com/software/analytics/cognos/)
- QlikView (qlikview.com)

Terminologia

- Input: **dataset** di **individui o istanze**
- **Variabili**: gli **attributi** delle istanze (anche chiamati **feature** o **elementi**)
- Due tipi di variabili:
  - **Categoriche**
    - assumono un numero finito di valori distinti, rappresentano una proprietà qualitativa
    - booleane, ordinali, nominali
  - **Numeriche**:
    - assumono un insieme finito o infinito di valori numerici
    - discrete, continue

Classificazioni tecniche di Data Mining

- **Supervisionate**
  - Sono guidate da un **attributo target**
  - Spiegano i valori dell'attributo target (**variabile dipendente**) rispetto ai valori di un insieme di variabili predittive (**variabili indipendenti**)
  - Principali approcci
    - **Classificazione**
    - **Regressione**

- **Non supervisionate**
  - Non sono guidate da un **attributo target**
  - Non distinguono quindi tra variabili dipendenti e indipendenti
  - Principali approcci
    - **Clustering**
    - **Pattern discovery**

Classificazione

- **L’attributo target è categorico** e associa ogni individuo a una **classe** (un gruppo)
- Il dataset di input viene chiamato **training set**
- Obiettivo interpretativo
  - Creare un **modello che spieghi le relazioni tra l’attributo target e gli altri attributi** partendo dalle istanze del **training set**
  - Viene utilizzato un **test set** per **validare** il modello, cioè determinarne **l’accuratezza**
- Obiettivo predittivo
  - usando il modello, assegnare agli individui di un insieme precedentemente sconosciuto una classe

Classificazione — applicazioni

- Classificare i clienti di una agenzia assicurativa rispetto al rischio
- Classificare le transazioni con carta di credito come legittime o fraudolente
- Classificare una cellula tumorale come benigna o maligna
- Classificare le notizie riportate in un sito web rispetto al dominio di riferimento

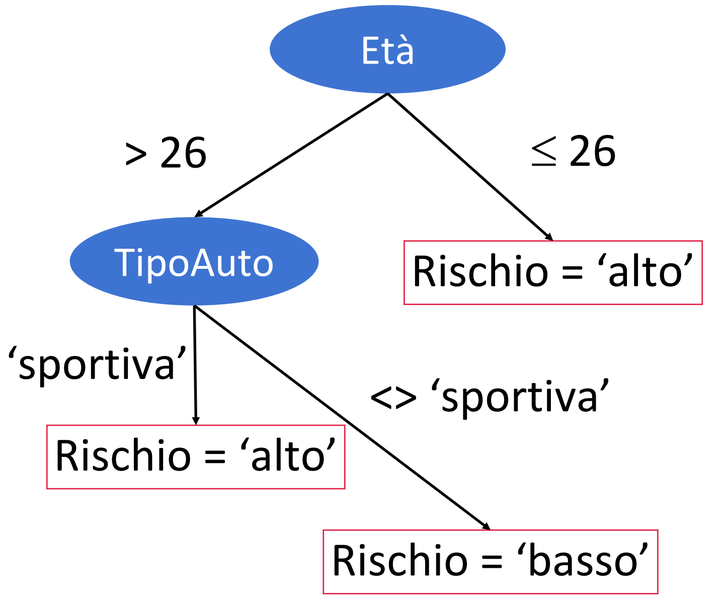
Classificazione — esempio

Training set

attributo target

Età	Tipo auto	Classe rischio
40	familiare	basso
65	sportiva	alto
20	utilitaria	alto
25	sportiva	alto
50	utilitaria	basso

Creazione del modello

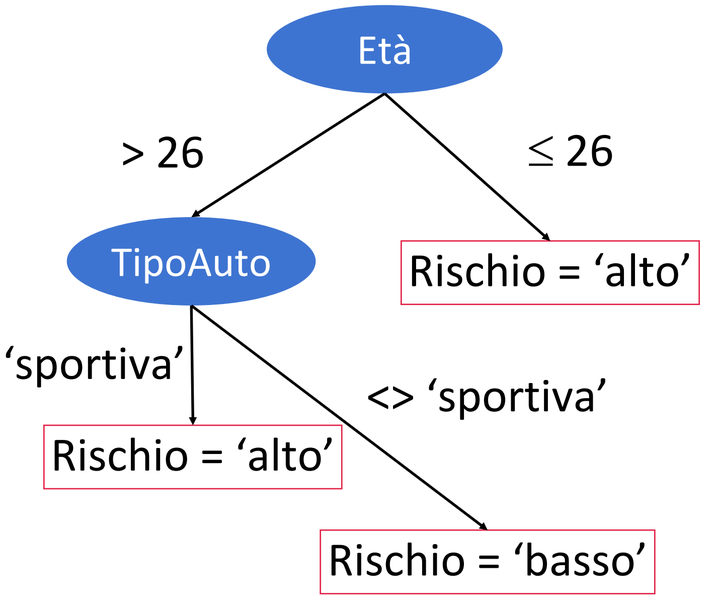


Classificatore

IF Età > 26 and TipoAuto =‘sportiva’
THEN Rischio = ‘alto’
IF Età ≤ 26 THEN Rischio = ‘alto’
IF Età > 26 and TipoAuto <> ‘sportiva
THEN Rischio = ‘basso’

Classificazione — esempio (segue)

Validazione del modello



Test set

attributo target

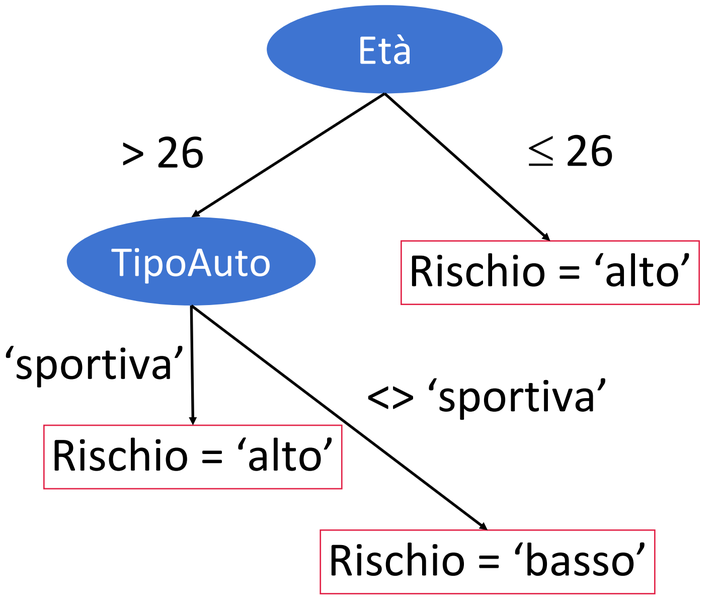
Età	Tipo auto	Classe rischio
45	familiare	basso
50	berlina	basso
20	utilitaria	alto
25	familiare	basso
50	sportiva	alto

Classificatore

IF Età > 26 and TipoAuto =‘sportiva’
THEN Rischio = ‘alto’
IF Età ≤ 26 THEN Rischio = ‘alto’
IF Età > 26 and TipoAuto <> ‘sportiva
THEN Rischio = ‘basso’

Classificazione — esempio (segue)

Previsione



Nuovi dati

attributo target

Età	Tipo auto	Classe rischio
35	utilitaria	???

Classificatore

IF Età > 26 and TipoAuto =‘sportiva’
THEN Rischio = ‘alto’
IF Età ≤ 26 THEN Rischio = ‘alto’
IF Età > 26 and TipoAuto <> ‘sportiva
THEN Rischio = ‘basso’

Rischio = ‘basso’

Classificazione — algoritmi

- Alberi di decisione
- Algoritmi Bayesiani
- Regressione logistica
- Reti neurali
- Support vector machine

Regressione

- **L’attributo target è continuo**
- Obiettivo interpretativo
  - Interpretare la **dipendenza tra la variabile dipendente e le variabili indipendenti,** partendo dalle istanze del training set, **attraverso una funzione**
  - Viene utilizzato un **test set** per **validare** il modello, cioè determinarne **l’accuratezza**
- Obiettivo predittivo
  - predire il valore della variabile target per ogni istanza di un insieme precedentemente sconosciuto, utilizzando la funzione individuata

Regressione — applicazioni

- Predire il totale delle vendite di un nuovo prodotto sulla base della spesa per la relativa campagna pubblicitaria
- Predire la velocità del vento in funzione di temperature, umidità e pressione
- Predire gli anni di vita attesa di una persona rispetto al peso

Regressione — esempio

- Variabile dipendente (attributo target): anni
- Variabile indipendente: peso

- Funzione: anni = 124 – 0.8 x peso
  - Se peso= 50 kg, gli anni di vita attesa sono 84
  - Se peso = 100 kg, gli anni di vita attesa sono 44

Regressione — algoritmi

- Regressione lineare semplice
- Regressione lineare multipla
- Regressione non lineare

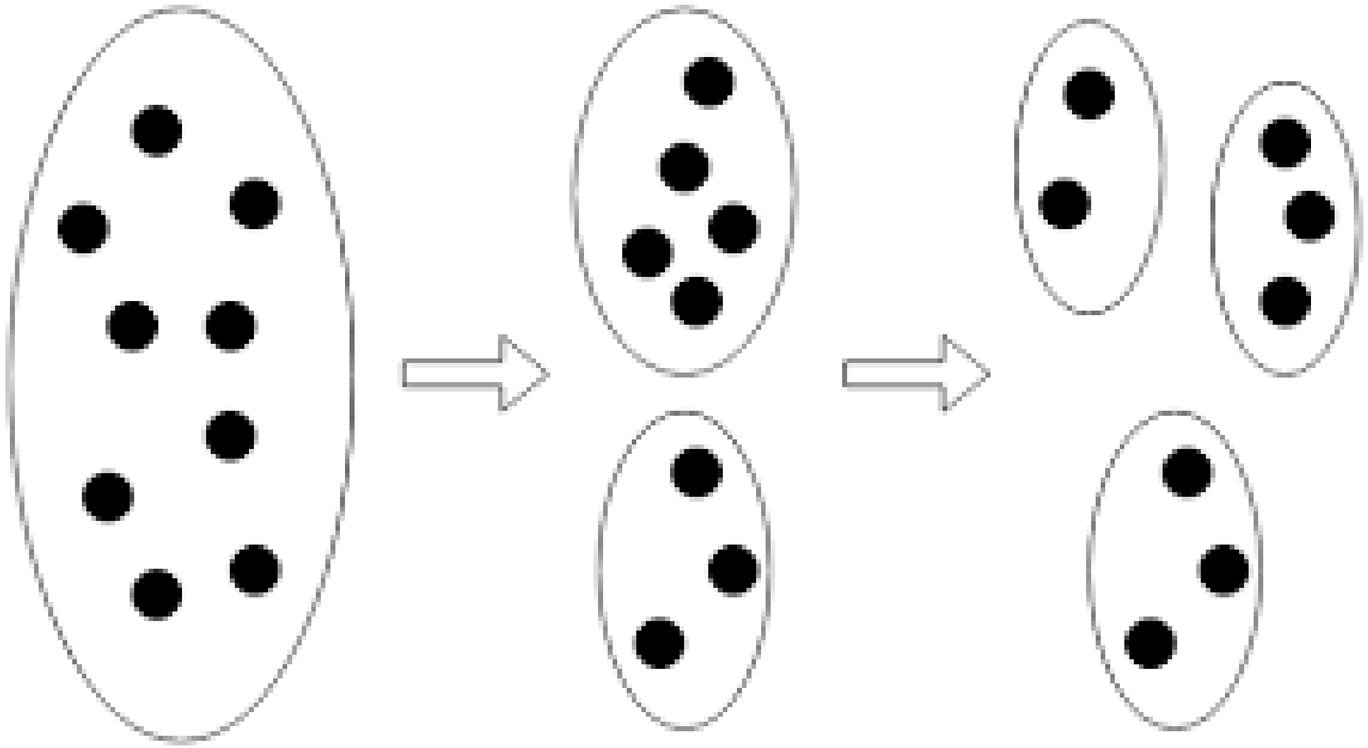
Clustering

- Obiettivo interpretativo
  - Individuare i sotto-gruppi (**cluster**) di una popolazione contenenti individui simili (**omogenei**) rispetto alle variabili che li descrivono
- Viene minimizzata la distanza tra individui appartenenti allo stesso cluster
- Viene massimizzata la distanza tra individui appartenenti a cluster diversi
- Diverse **funzioni di distanza in relazione al tipo di variabili considerate**
  - **Continue:** distanza di Manhattan nello spazio uni-dimensionale (una variabile) distanza Euclidea nello spazio bi-dimensionale (due variabili), distanza di Minkowski (numero arbitrario di variabili)
  - **Categoriche**: coseno, coefficiente di Tanimoto, indice di Jaccard
- Può essere utilizzato come passo iniziale prima di utilizzare altre tecniche di mining su ogni cluster

Clustering — applicazioni

- Clusterizzare i clienti di una azienda rispetto alle informazioni personali e caratteristiche di acquisto in modo da inviare agli individui di ogni cluster informazioni pubblicitarie mirate
- Clusterizzare un insieme di documenti rispetto alle parole chiave che contengono in modo da suddividerli in gruppi di affinità tematica
- Clusterizzare i clienti di una agenzia assicurativa rispetto alle richieste di rimborso presentate
- Clusterizzare gli appartamenti in vendita in una certa zona rispetto alle loro caratteristiche

Clustering — esempio



Clustering — algoritmi

- Algoritmi di partizionamento
  - i cluster rappresentano un partizionamento del dataset iniziale (K-MEANS)
- Algoritmi gerarchici
  - creano una decomposizione gerarchica del dataset
  - i cluster non rappresentano necessariamente un partizionamento del dataset iniziale
- Algoritmi basati sulla densità (DBSCAN)
  - individuazione di aree dense di punti

Pattern discovery

- I pattern più comuni sono chiamati **regole di associazione**
  - IF X THEN Y (o X → Y) dove X e Y rappresentano valori di variabili distinte
  - IF Età > 26 and TipoAuto =‘sportiva’ THEN Rischio = ‘alto’
- Se le istanze contengono un insieme di elementi T, presi da una certa collezione, la regola X → Y significa che se T contiene X allora molto probabilmente contiene anche Y
- Obiettivo informativo
  - Descrivere gli individui nel dataset utilizzando le regole
- Obiettivo predittivo
  - Utilizzare le regole per predire le caratteristiche di nuovi dati

Pattern discovery — applicazioni

- Individuare i prodotti venduti insieme per migliorare le strategie di vendita
- Individuare quali libri vengono comprati da chi compra il libro “La Divina Commedia”
- Individuare se una richiesta di congedo per malattia è sempre collegata a una richiesta di visita a un dottore nello stesso periodo
- Individuare quali pagine web vengono accedute frequentemente nell’ambito della stessa sessione

Pattern discovery — esempio

TID	Prodotti
1	Pane, Coca, Latte
2	Birra, Pane
3	Birra, Coca, Pannolini, Latte
4	Birra, Pane, Pannolini, Latte
5	Coca, Pannolini, Latte

Regole individuate:
{Latte} → {Coca}
{Pannolini, Latte} → {Birra}

Pattern discovery — algoritmi

- Algoritmo APRIORI
- Alberi di decisione

Qualità dei modelli e dei pattern individuati

- Dipende dal tipo di conoscenza acquisita
- Classificazione e regressione
  - Diverse misure basate sulla **matrice di confusione**
- Clustering
  - **Distanza intra-cluster** (deve essere bassa)
  - **Distanza inter-cluster** (deve essere alta)
- Regole di associazione X → Y
  - **Supporto**: frequenza con cui X U Y appare tra tutte le istanze
  - **Confidenza**: frequenza con cui Y appare nelle istanze che includono X

Riepilogo e conclusioni finali

Abbiamo visto:

- **Cosa si intende per Data Mining**
- **Classificazione delle principali tecniche di Data Mining**
- **Esempi di applicazione delle principali tecniche di Data Mining**