SAPIENZA
UNIVERSITÀ DI ROMA

DIPARTIMENTO DI INGEGNERIA INFORMATICA AUTOMATICA E GESTIONALE
ANTONIO RUBERTI

Sistemi Informativi Aziendali

Umberto Nanni

Cenni al Data Mining

Data Mining

- nasce prima del Data Warehouse
- collezione di tecniche derivanti da Intelligenza Artificiale, Pattern Recognition, e tecniche statistiche di vario tipo (es.: algoritmi genetici, logiche fuzzy, sistemi esperti, reti neurali, etc.)
- obiettivi:
  - descrittivi: individuare schemi di comportamento, rapporti di causa-effetto, classificare individui, etc.
  - predittivi: predire andamenti, classificare individui in base al rischio, etc.

Alcune applicazioni del Data Mining

- Analisi dei dati e Supporto alle Decisioni
- Analisi di mercato e marketing
  - Target Marketing, Customer Relationship Management (CRM), Market Basket Analysis (MBA), segmentazione del mercato
- Analisi e gestione del rischio
  - previsioni di affidabilità, fidelizzazione di utenti, controllo di qualità, …
  - individuazione di frodi e di pattern inusuali (outliers)
- Text Mining
- Web Mining, ClickStream Analysis
- Ingegneria genetica: ricerca sequenze in DNA

Data Mining: regole associative

SE “viene acquistato il prodotto birra” (X),
ALLORA “viene acquistato anche il prodotto pannolino” (Y)

$$X \rightarrow Y$$

Supporto (quale frazione di soggetti verifica la regola):

$$s = \frac{|X \cap Y|}{|all|}$$

$s(X \rightarrow Y) = F(X \wedge Y)$

Confidenza (quale frazione di soggetti soddisfa la regola tra quelli in cui è applicabile):

$$c = \frac{|X \cap Y|}{|X|}$$

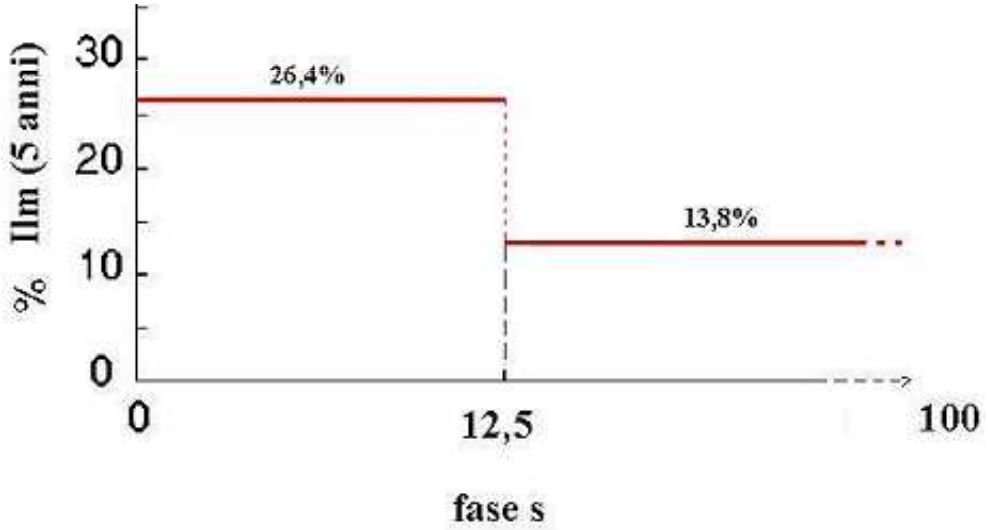
$c(X \rightarrow Y) = F(Y \mid X)$

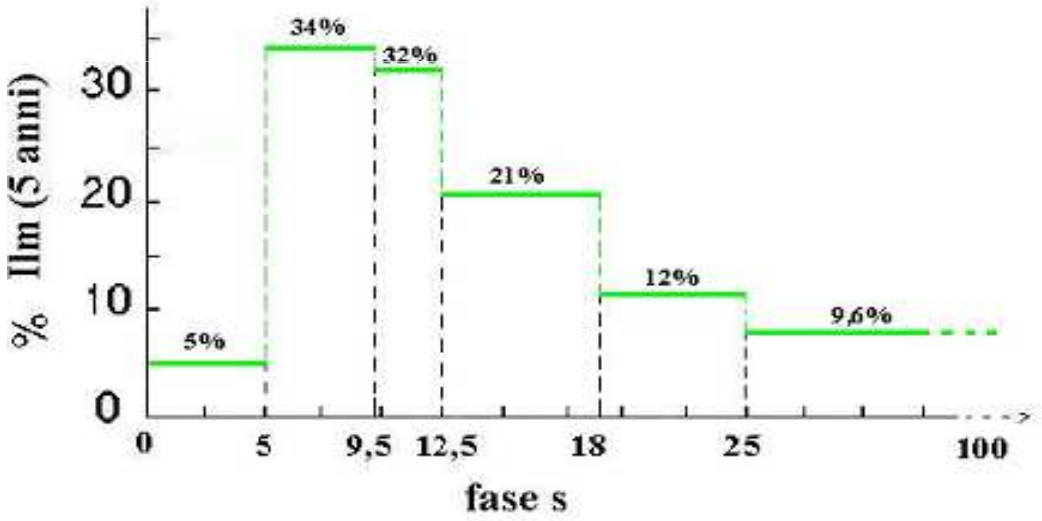
Campi di applicazione: economico (*market basket analysis*), telecomunicazioni, medico, …

Data Mining: clustering

- individuazione di similarità, cogliendo disomogeneità nella distribuzione per definire gruppi omogenei (apprendimento senza supervisione)
- ricerca gruppi (cluster) basata su
  - distribuzione della popolazione
  - una funzione di "distanza"

Esempio: ILM - Intervallo Libero da Malattia (a 5 anni)
(collaborazione con Ist. Regina Elena di Roma)



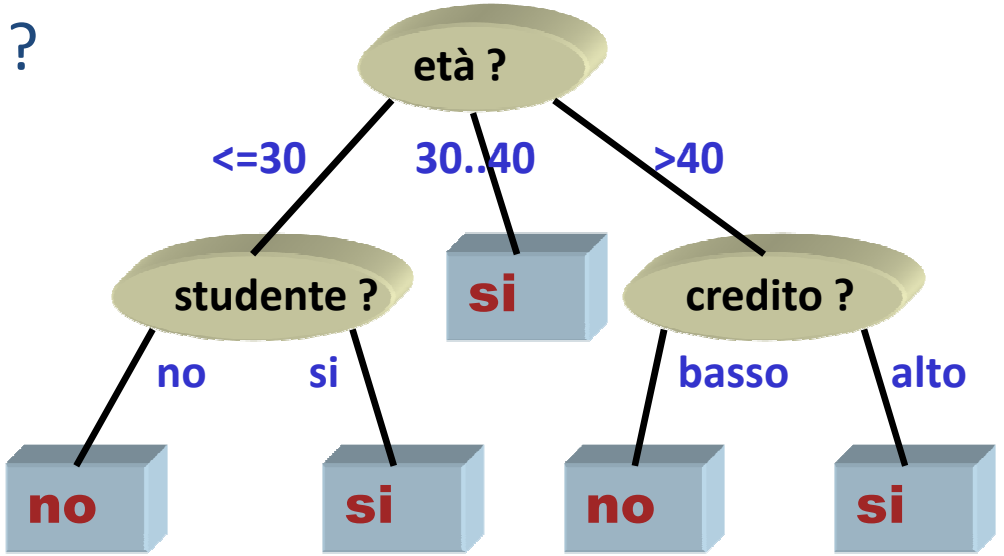


Data Mining: alberi di decisione

determinare le cause di un fenomeno di interesse in ordine di importanza

- nodo interno: test di un attributo
- diramazione: valore (o intervallo) di un attributo
- foglia: assegna una classificazione (decisione finale)

Esempio:
il cliente acquisterà un computer ?



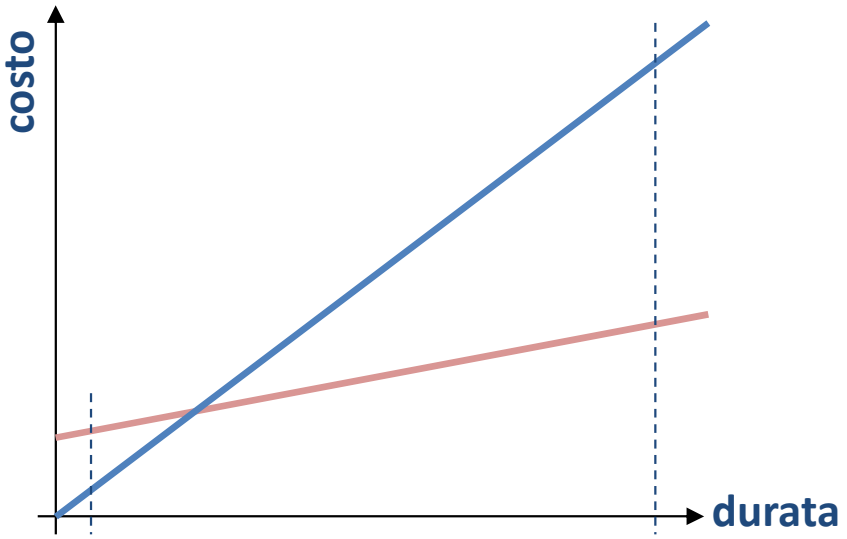
Data Mining: serie temporali

- individuazione pattern ricorrenti / atipici in sequenze temporali
- predizione caratteristiche

Esempio (Least Cost Routing): instradamento traffico telefonico su operatore a costo minimo
(collaborazione con Between – azienda di consulenza)

DOMANDA CHIAVE:
data una telefonata in uscita da un numero interno X diretta verso un numero esterno Y, quanto **durerà** la telefonata ?

Tariffe:
con scatto alla risposta
flat rate



Reti Neurali (Neural Networks)

Alcuni tipici ambiti di applicazione riguardano problemi di riconoscimento:

- OCR (Optical Character Recognition)
- riconoscimento della grafia
- riconoscimento del parlato
- riconoscimento di immagini

Tipico approccio quando è noto INPUT ed OUTPUT, ma poco altro.

Data Mining: risultati “interessanti”

- Semplicità - Ad esempio:
  - lunghezza delle regole (associative)
  - taglia (albero di decisione)
- Certezza - Ad esempio:
  - confidenza (regole associative): $c(X \rightarrow Y) = \#(X \text{ and } Y) / \#(X)$
  - affidabilità della classificazione
- Utilità - Ad esempio:
  - supporto (regole associative): $s(X \rightarrow Y) = \#(X \text{ and } Y) / \#(ALL)$
- Novità - Ad esempio:
  - non nota in precedenza
  - sorprendente
  - sussunzione di altre regole (incluse come casi particolari)