Introduzione al Data Mining

Luca Cabibbo, Riccardo Torlone

settembre 2014

Introduzione

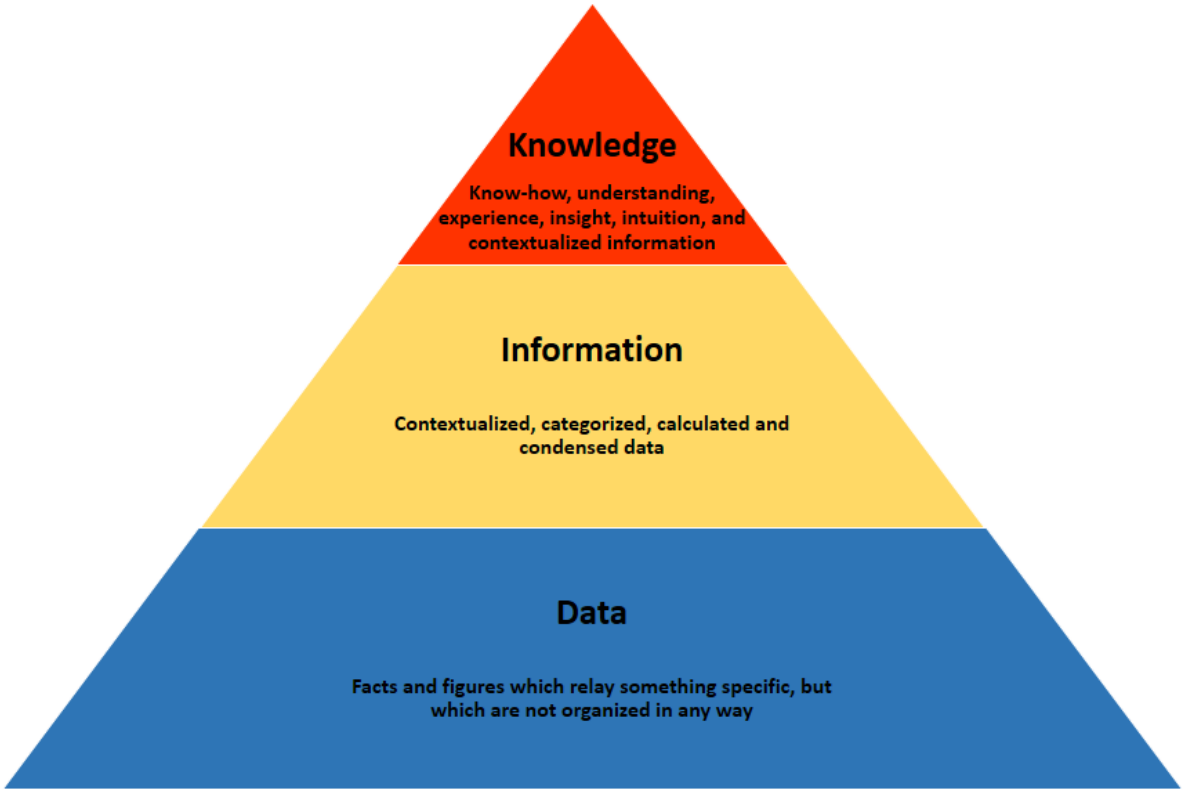
Negli ultimi anni, le organizzazioni hanno effettuato investimenti significativi per migliorare la loro capacità di raccogliere **dati**

- ad esempio, relativi ai propri processi, al comportamento dei clienti, a prestazioni delle campagne di marketing, …
- le organizzazioni possono accedere facilmente anche ad altri dati esterni – ad esempio, tendenze di mercato, notizie nel proprio settore aziendale, mosse della concorrenza, …

Questa vasta disponibilità di **dati** ha aumentato l'interesse in metodi per estrarre dai dati **informazioni** utili e **conoscenza**

- questo è il reame della *data science*
- le applicazioni più significative di questa disciplina riguardano il supporto alle decisioni aziendali – ad esempio, per progettare nuove campagne di marketing

Dati, informazioni e conoscenza



Data science e data mining

Data science

- la data science è un insieme di principi fondamentali che guida l'estrazione di conoscenza da dati

Data mining

- il data mining è l'estrazione di conoscenza da dati, tramite tecnologie che incorporano i principi della data science

Spesso questi due termini vengono usati in modo intercambiabile

- "data science" ha un significato più generale, e fa riferimento ad un insieme di principi
- "data mining" ha un significato più specifico, e fa riferimento ad un insieme di tecniche e metodi

Esempio: l'uragano Frances

Dal NYT, 2004

- Hurricane Frances was on its way, barreling across the Caribbean, threatening a direct hit on Florida's Atlantic coast. Residents made for higher ground, but far away, in Bentonville, Ark., executives at Wal-Mart Stores decided that the situation offered a great opportunity for one of their newest data-driven weapons … predictive technology.
- A week ahead of the storm's landfall, Linda M. Dillman, Wal-Mart's chief information officer, pressed her staff to come up with forecasts based on what had happened when Hurricane Charley struck several weeks earlier. Backed by the trillions of bytes' worth of shopper history that is stored in Wal-Mart's data warehouse, she felt that the company could 'start predicting what's going to happen, instead of waiting for it to happen,' as she put it. (Hays, 2004)

In questo caso può essere utile fare *previsioni* perché

- ci sono dei prodotti che, ovviamente, si venderanno di più, come le bottiglie d'acqua – ma quanto venderanno di più?
- ci sono altri prodotti che, in modo meno ovvio, venderanno di più? ad esempio, torte alle fragole
- ci sono prodotti che, pur essendosi esauriti in concomitanza con l'uragano Charley, non venderanno altrettanto? ad esempio, è il caso di uno specifico film di successo

Esempio: abbandono dei clienti

Supponiamo di essere analisti della MegaTelCo – una grande azienda di telecomunicazioni

- uno dei problemi principali di questa azienda è l'abbandono di clienti (customer churn) del servizio di telefonia cellulare – il 20% degli abbonati abbandona il servizio allo scadere dell'abbonamento – inoltre, acquisire nuovi clienti è difficile
- per questo, l'azienda vuole offrire una promozione mirata "di trattenimento" ad alcuni clienti a cui sta per scadere l'abbonamento, per cercare di ridurre gli abbandoni
- ma a quali specifici clienti va offerta questa promozione?
  - in questo caso è utile fare previsioni circa i clienti che (i) senza offerta, probabilmente abbandoneranno, ma (ii) con l'offerta, probabilmente rimarranno clienti

Data science e decisioni guidate dai dati

La data science riguarda principi, processi e tecniche per comprendere fenomeni tramite l'analisi (automatizzata) di dati

- l'obiettivo fondamentale della data science è migliorare il "decision making", consentendo alle organizzazioni di prendere decisioni guidate dai dati – ovvero, decisioni basate su esperienze effettive e dati storici – e non semplicemente "guidate dall'intuizione"
- di solito, queste decisioni vengono basate sull'estrazione di conoscenza dai dati sotto forma di "modelli" o "pattern"
- l'esperienza ha infatti dimostrato che è possibile estrarre conoscenza utile dai dati per risolvere problemi di business concreti – e, inoltre, che questo può essere fatto in modo sistematico sulla base di un processo di data mining, con fasi ragionevolmente ben definite

Esempio: prestiti

Si consideri, per una certa banca, il problema della concessione di prestiti

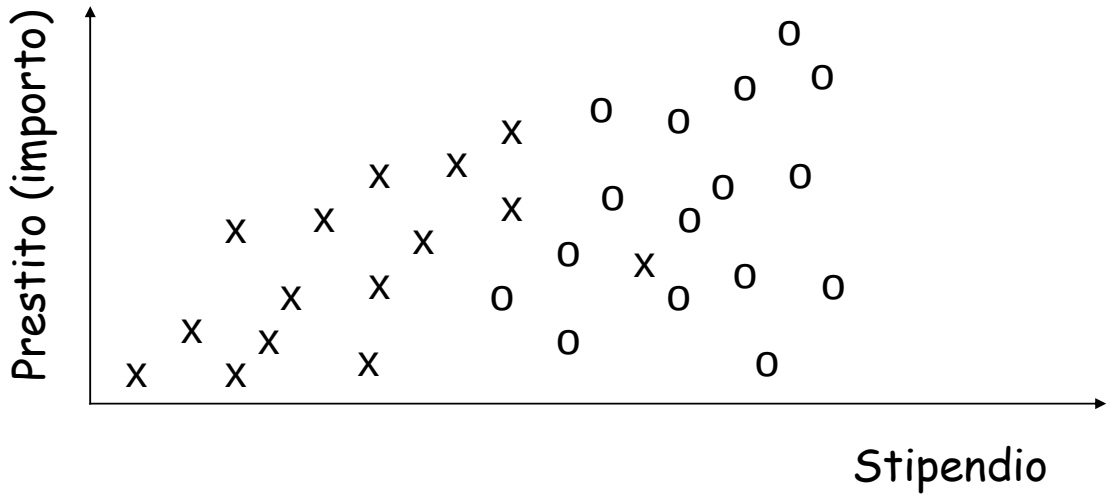
- la banca vuole concedere prestiti solo se si aspetta che il cliente pagherà tutte le rate nei tempi previsti

Questo problema può essere risolto mediante l'estrazione di "pattern" da "dati storici"

- la banca possiede *dati storici* sui prestiti che ha già erogato nel passato – e sa in quali casi le rate sono state pagate regolarmente o meno
- per *pattern* si intende, ad esempio, una regola che descrive in modo succinto delle informazioni estratte dai dati storici per prevedere se un potenziale cliente pagherà le rate di un prestito regolarmente o meno

Esempio: prestiti

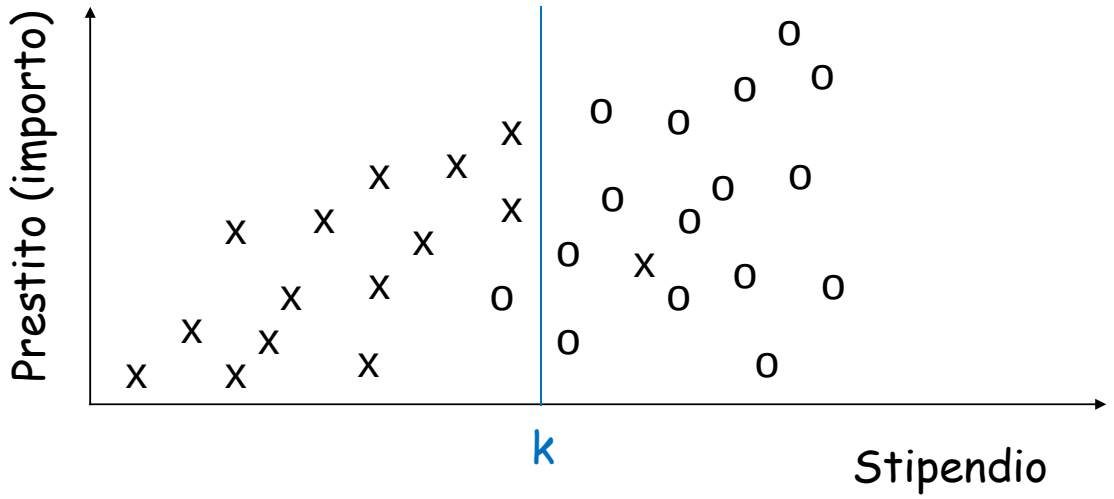
Dati storici sui prestiti



Persone che hanno ricevuto un prestito dalla banca:
x: persone che hanno mancato la restituzione di rate
o: persone che hanno rispettato le scadenze

Esempio: prestiti

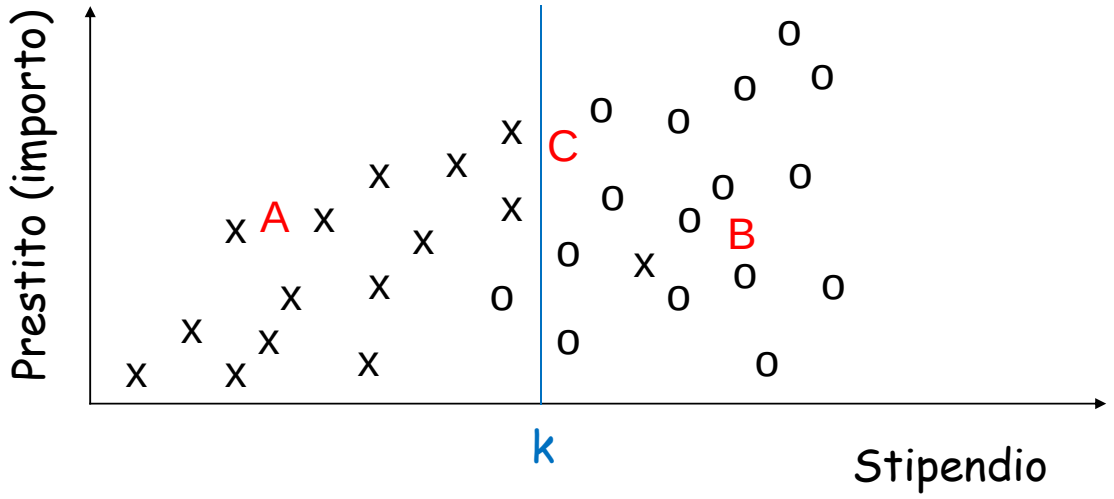
Un possibile pattern estratto da questi dati storici



IF stipendio < k THEN mancato pagamento

Esempio: prestiti

Applicazione del pattern estratto ad altri potenziali clienti



IF stipendio < k THEN mancato pagamento

Dati e pattern

Dati

- insieme di informazioni, tipicamente non strutturate, estratte da una base di dati o da un data warehouse

Pattern

- espressione, in un linguaggio opportuno, che descrive in modo succinto le informazioni estratte dai dati
  - regolarità
  - informazione di alto livello

Caratteristiche dei pattern

Validità (ad esempio, accuratezza)

- una misura del grado di validità del pattern – ad esempio, nel 90% dei casi – si noti che, rispetto all'identificazione dei mancati pagamenti, aumentando il valore k aumentano i "falsi positivi", e diminuendolo aumentano i "falsi negativi"

Novità

- misurata rispetto a variazioni dei dati o della conoscenza estratta

Utilità

- esempio: aumento di profitto atteso dalla banca associato alla regola estratta

Comprensibilità

- misure di tipo sintattico e/o semantico

Problemi fondamentali di data mining

I problemi decisionali di business sono complessi – ciascuno di essi è unico, e va risolto diversamente da altri problemi

- tuttavia, in generale è possibile risolvere un problema decisionale complesso decomponendolo in sottoproblemi o compiti (task), risolvendo questi problemi parziali separatamente, e ricomponendo le soluzioni parziali in una soluzione complessiva del problema iniziale
- inoltre, molti task sono comuni, e le loro tecniche risolutive possono essere usate in problemi diversi – ad es., calcolare da dati storici la probabilità di abbandono di un cliente
- il *data mining* è basato su un insieme di tecniche e algoritmi risolutivi per compiti decisionali comuni
- una capacità critica della *data science* è la capacità di decomporre un problema in parti, in modo tale che ciascuna parte possa essere messa in corrispondenza con un compito di data mining noto

Problemi fondamentali di data mining

Nel corso degli anni sono stati sviluppati un gran numero di algoritmi di data mining

- tuttavia, essi fanno riferimento solo a una manciata di compiti fondamentali
- lo scopo generale è quello di costruire automaticamente, a partire da dati storici, un “modello” – da usare, ad esempio, per fare previsioni su dati correnti e futuri
- ne descriviamo brevemente alcuni

Problemi fondamentali di data mining

Classificazione

- ha lo scopo di predire, per ogni individuo di una popolazione, a quale classe appartiene – tra un certo numero di classi predefinite
- ad esempio, nel caso della MegaTelCo, “cliente propenso ad abbandonare” e “cliente propenso a rimanere”
- può essere di interesse determinare anche la probabilità di appartenenza di un individuo a una classe

Regressione

- predire, per un individuo, il valore numero di una sua variabile
- ad esempio, stimare la quantità che sarà venduta di un prodotto

Problemi fondamentali di data mining

Similarity matching

- identificare individui simili tra loro (con riferimento ad uno scopo specifico)
- ad esempio, per consigliare prodotti – “se ti piace questo prodotto, allora potrebbe piacerti anche quest’altro”

Clustering

- per raggruppare gli individui di una popolazione sulla base della loro similarità (non guidati da uno scopo specifico)
- ad esempio, come attività preliminare ad un compito di classificazione – “in quali classi di suddividono, effettivamente, i miei clienti?”

Problemi fondamentali di data mining

Co-occorrenze (associazioni)

- trovare relazioni comuni tra entità sulla base di transazioni che le riguardano
- ad esempio, analisi di mercato degli acquisti – “quali prodotti vengono comunemente acquistati insieme?”

Profiling (descrizione del comportamento)

- caratterizza il comportamento tipico di un individuo o di una popolazione
- ad esempio, “quel è l’utilizzo tipico del servizio di telefonia mobile per un cliente appartenente a un certo segmento?”

Problemi fondamentali di data mining

Link prediction

- prevede relazioni tra individui, di solito suggerendo che ci debba essere un collegamento quando invece manca
- ad esempio, per suggerire amicizie su un social network

Causal modeling

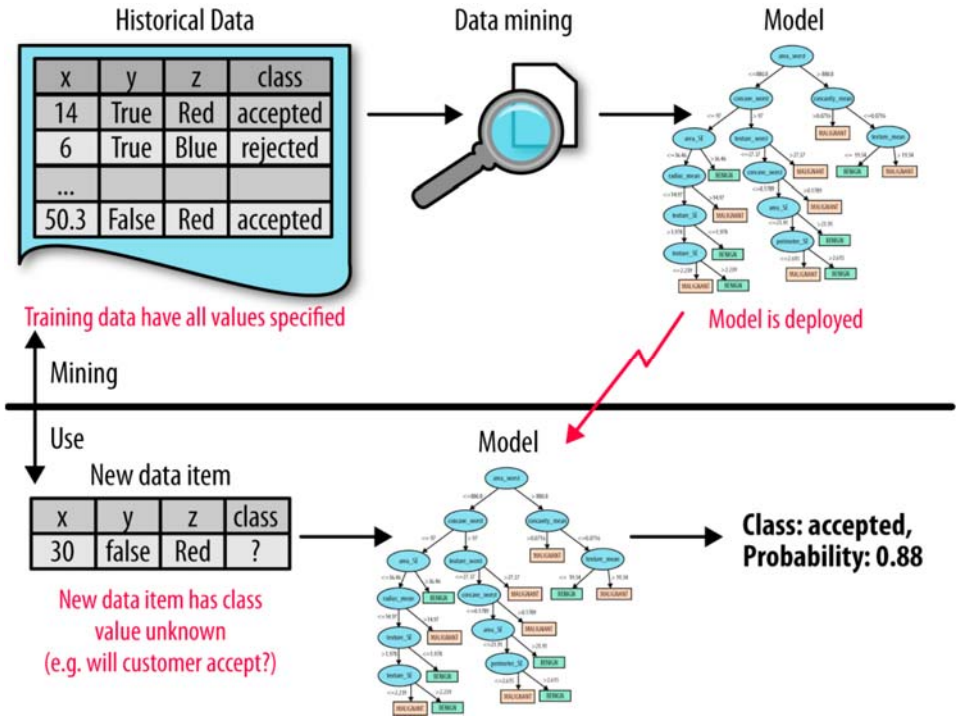
- aiuta a comprendere relazioni di causa-effetto tra eventi o azioni
- ad esempio, si consideri un cliente a cui è stata rivolta una promozione mirata per un prodotto ha poi acquistato quel prodotto – ma l'acquisto è stato effettivamente influenzato dalla promozione, o quel cliente avrebbe acquistato quel prodotto comunque? (si pensi anche all'effetto placebo in medicina)

Metodi supervisionati e non supervisionati

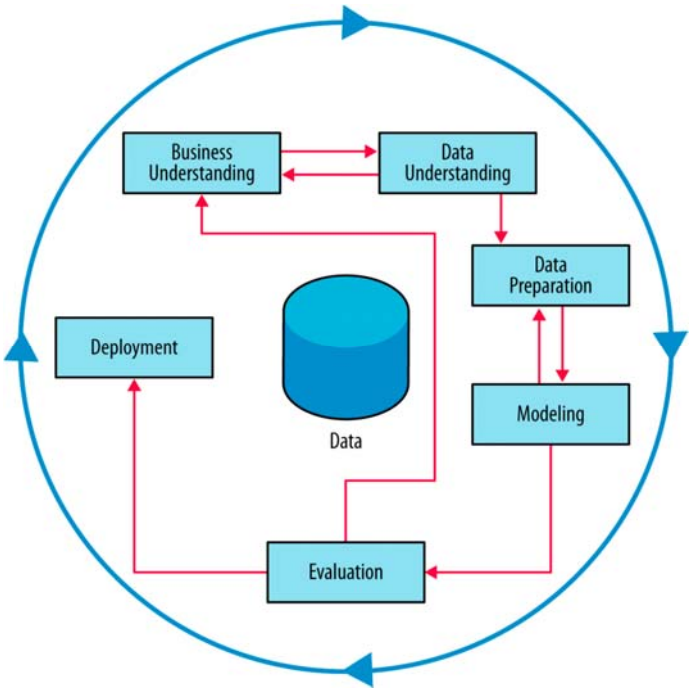
In generale, un compito di data mining ha lo scopo di costruire un modello a partire da dati storici

- un compito (o metodo) si dice *supervisionato* quando i dati di "addestramento" comprendono un insieme di esempi insieme a dati sulla caratteristica indirizzata dal modello
  - nel caso degli abbandoni, i dati storici sui clienti comprendono il fatto se il cliente ha abbandonato o meno
- un compito si dice invece *non supervisionato* quando i dati di addestramento non comprendono dati sulla caratteristica indirizzata dal modello
  - questo succede, ad esempio, nei problemi di clustering e di associazioni

Data mining e uso dei risultati del data mining



Il processo di data mining



Classificazione

La **classificazione** è un compito supervisionato di data mining

- ha lo scopo di predire, per ogni individuo di una popolazione, a quale classe appartiene – tra un certo numero di classi predefinite

Dati del problema

- un insieme di oggetti (*training set*), ciascuno dei quali è caratterizzato da un insieme di attributi
- uno degli attributi è l'attributo obiettivo (*target*) della classificazione – il suo valore appartiene ad un insieme predefinito di valori (*classi*)

Problema

- trovare un modello, basato sugli attributi del training set, per prevedere la classe di appartenenza di altri oggetti di cui si conoscono i diversi attributi (ma non l'attributo target)

Applicazioni

Classificazione tendenze di mercato

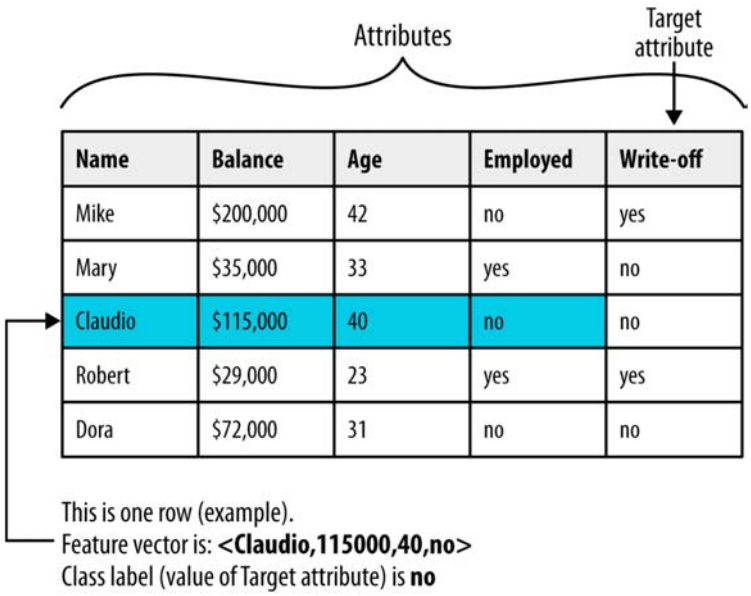
Identificazione del rischio in mutui/assicurazioni

Identificazione automatica di immagini

Efficacia trattamenti medici

Classificazione

Un esempio di training set



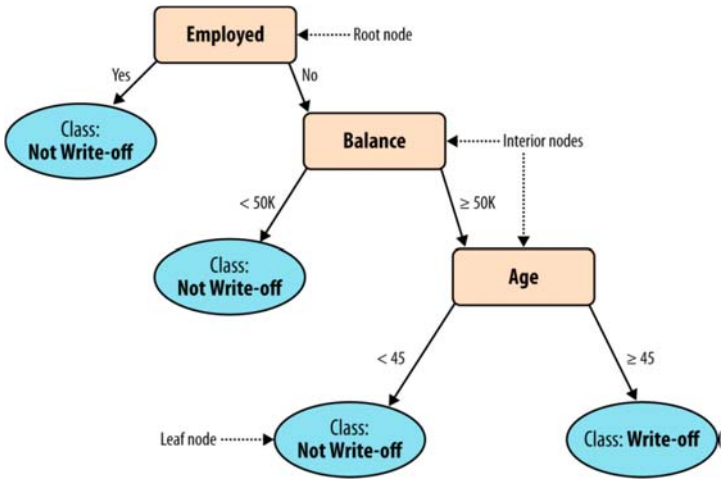
Name	Balance	Age	Employed	Write-off
Mike	$200,000	42	no	yes
Mary	$35,000	33	yes	no
Claudio	$115,000	40	no	no
Robert	$29,000	23	yes	yes
Dora	$72,000	31	no	no

- questo è solo un esempio: di solito i dati di addestramento sono molti di più, ed anche gli attributi sono molti di più!

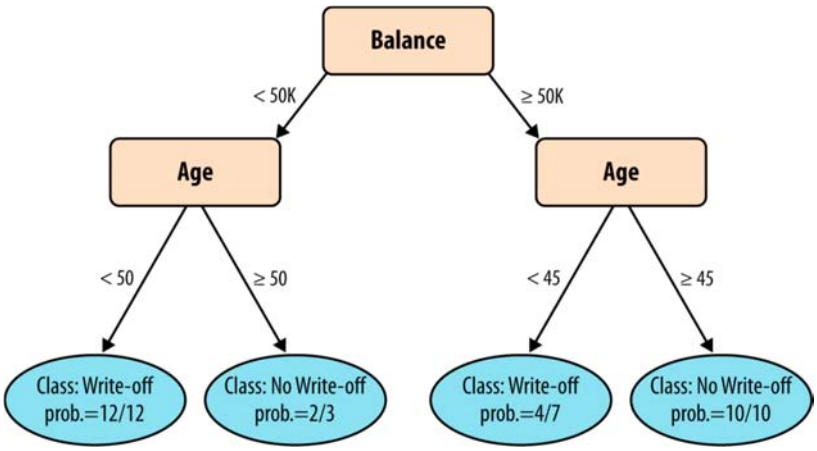
Alberi di classificazione

Il modello usato per la classificazione può essere di tipi diversi

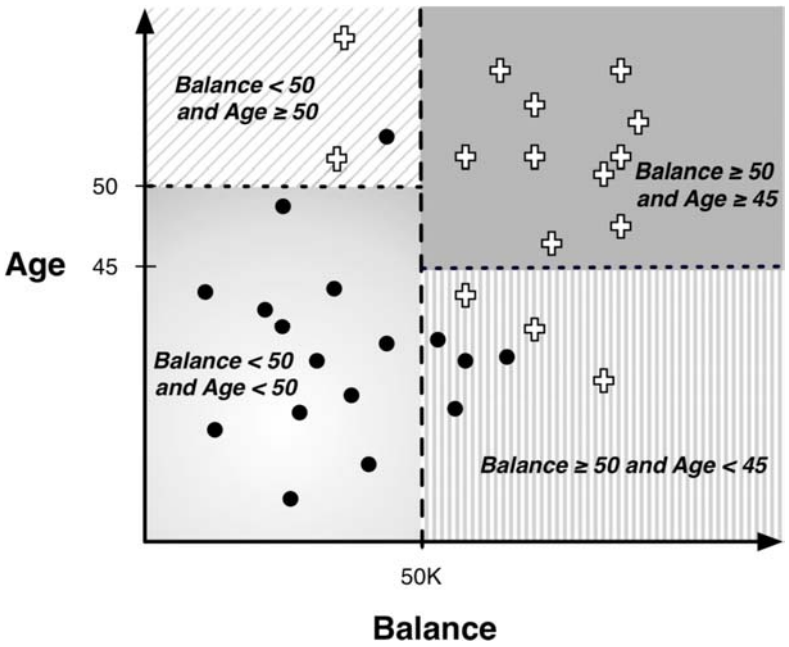
- un caso comune è quello degli *alberi di classificazione* – ovvero, una regola di classificazione che può essere rappresentata mediante una struttura decisionale con la forma ad albero



Un altro albero di classificazione...

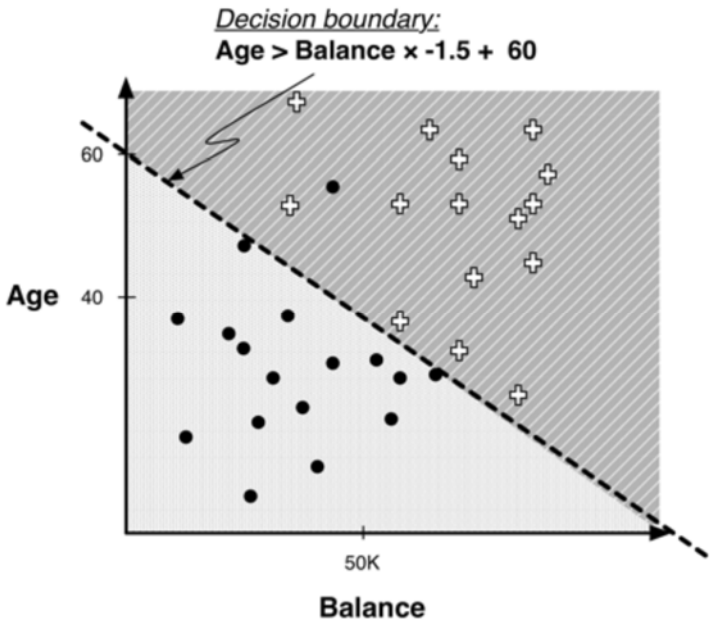


... e una sua interpretazione grafica



Un classificatore lineare

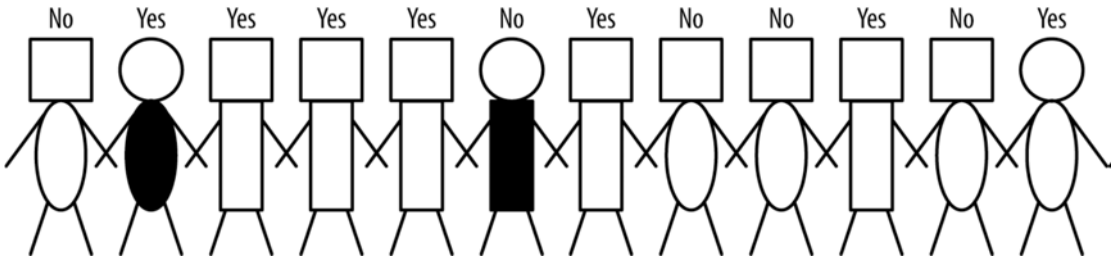
Un *classificatore lineare* è un modello di classificazione basato su una disequazione lineare sugli attributi dell'entità



- esistono anche classificatori non lineari – e di natura ancora diversa

Esercizio

Un problema di classificazione

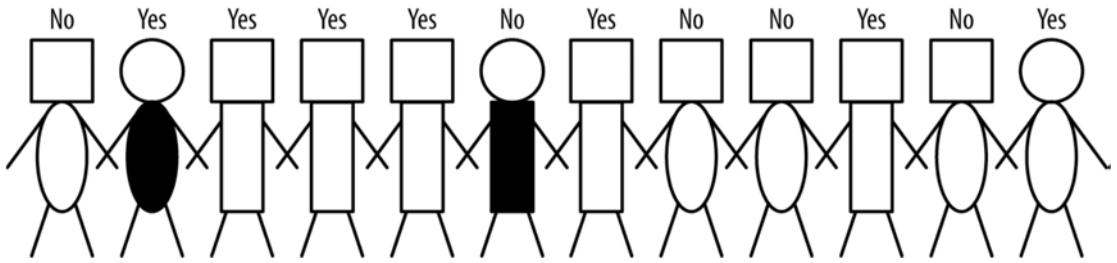


Costruzione degli alberi di classificazione

Ogni albero di classificazione è basato su una successione di condizioni – ciascuna delle quali riguarda un singolo attributo dell'entità

- è bene basare la prima condizione sul singolo attributo più informativo – ovvero quello che, intuitivamente, fornisce più informazioni (degli altri) sull'attributo target
- questa condizione consente di partizionare il training set in insiemi disgiunti – su ciascuno dei quali va applicato, ricorsivamente, questo stesso ragionamento – fino a trovare un albero di classificazione soddisfacente

Esercizio

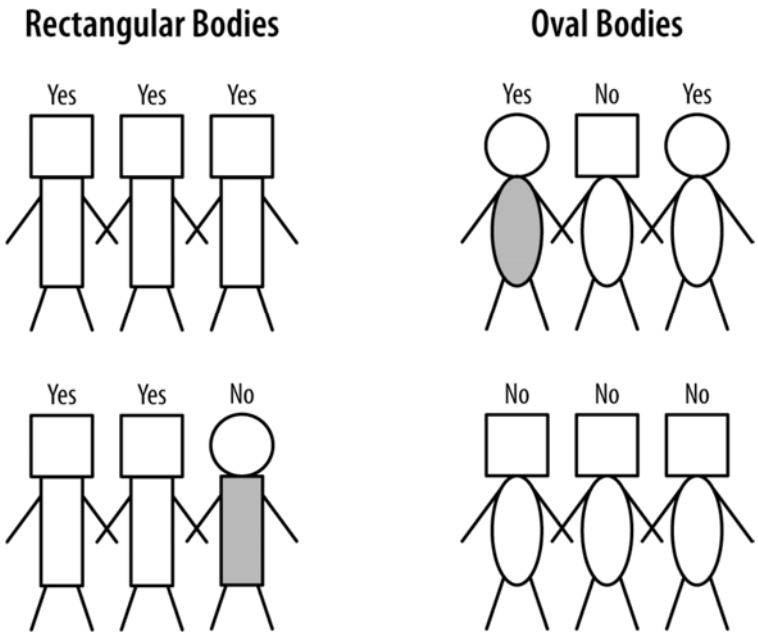


Qual è il singolo attributo più informativo?

- la forma della testa? la forma del corpo? il colore del corpo?

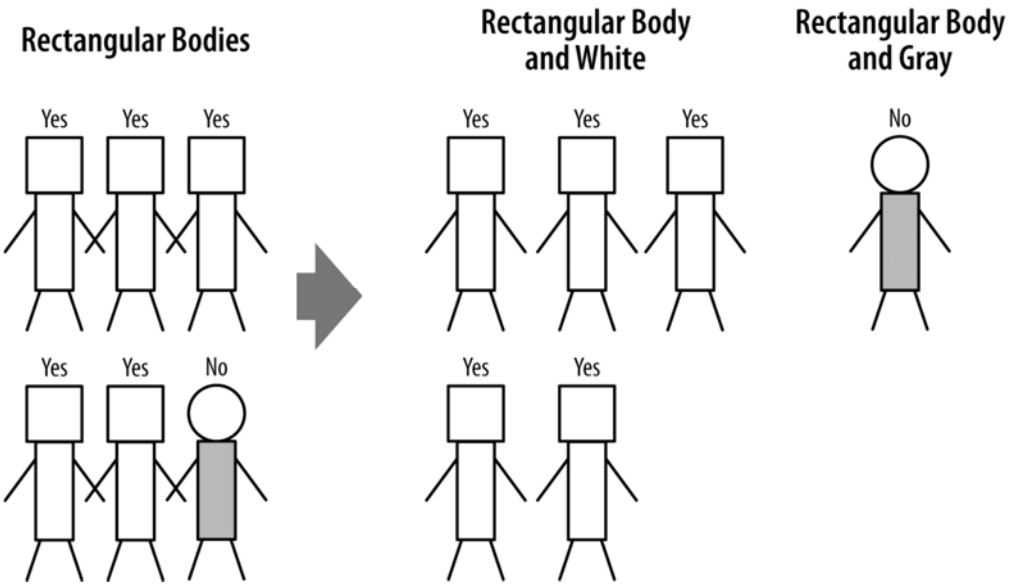
Esercizio

Il primo partizionamento può riguardare la forma del corpo



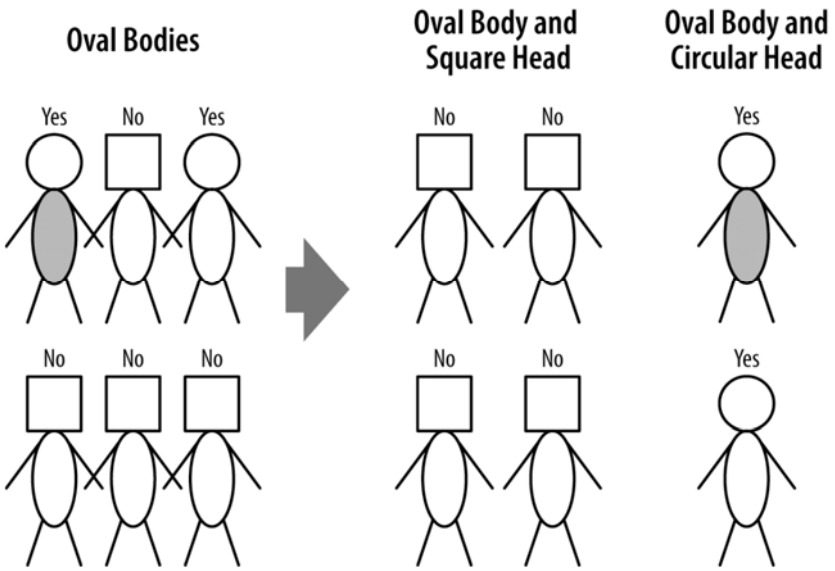
Esercizio

Un successivo partizionamento relativo al corpo rettangolare



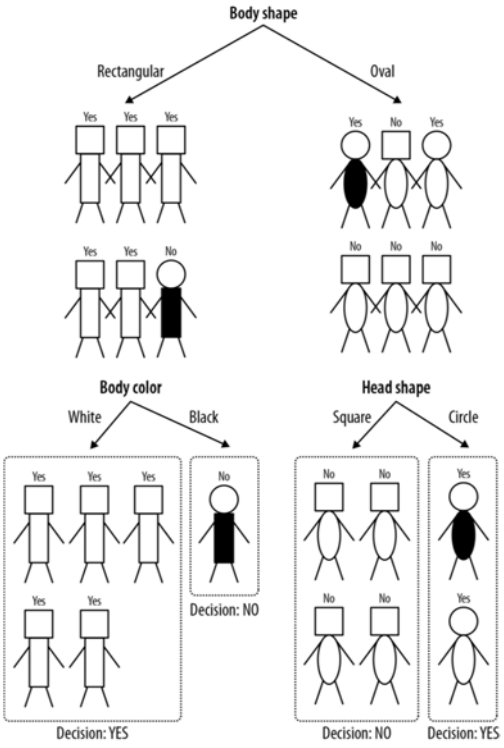
Esercizio

Un successivo partizionamento relativo al corpo ovale



Esercizio

Albero di classificazione complessivo



Associazioni

Le **co-occorrenze** (o **associazioni**) sono un problema di data mining non supervisionato

- lo scopo è scoprire relazioni tra oggetti entità che compaiono nell'ambito di transazioni
- ad esempio, l'analisi di market-basket – quali prodotti vengono comunemente acquistati insieme?

Dati del problema

- un insieme I di *oggetti*
  - ad esempio, prodotti venduti da un supermercato
- un insieme D di *transazioni*
  - ciascuna transazione T è un insieme di oggetti, $T \subseteq I$
  - ad esempio, prodotti acquistati nella stessa transazione di cassa al supermercato

Associazione

Una *regola di associazione* ha la forma

$$X \Rightarrow Y \text{ con } X,Y \subseteq I$$

- si legge: X implica Y
- in questo caso vuol dire: chi compra X compra anche Y
- X e Y possono essere sia singoli oggetti che gruppi di oggetti

Proprietà delle associazioni

Una *regola di associazione* ha la forma

$$X \Rightarrow Y \text{ con } X,Y \subseteq I$$

Proprietà di una regola di associazione $X \Rightarrow Y$

- *supporto* S – misura la rilevanza statistica della regola

$$S = \frac{\text{\# transazioni che contengono } X \cup Y}{\text{\# transazioni in D}}$$

- *confidenza* C – misura la significatività della regola

$$C = \frac{\text{\# transazioni che contengono } X \cup Y}{\text{\# transazioni che contengono X}}$$

Proprietà delle associazioni

In termini di probabilità, le proprietà della regola di associazione $X \Rightarrow Y$ possono essere anche espresse come segue

- *supporto* S – probabilità che una transazione contiene sia X che Y

$$S = \text{prob}(X \cup Y)$$

- *confidenza* C – probabilità che una transazione contiene Y, condizionata al fatto che la transazione contiene X

$$C = \text{prob}(Y|X)$$

Esempio

Latte $\Rightarrow$ Uova

Supporto

- il 2% delle transazioni contiene entrambi gli elementi

Confidenza

- il 30% delle transazioni che contengono latte contiene anche uova

Associazioni

Problema

- determinare tutte le regole di associazione con supporto e confidenza superiori ad una soglia data

Esempio

TRANSACTION ID	OGGETTI ACQUISTATI
2000	A,B,C
1000	A,C
4000	A,D
5000	B,E,F

Assumiamo:

- supporto minimo 50%
- confidenza minima 50%

Esempio

TRANSACTION ID	OGGETTI ACQUISTATI
2000	**A**,B,**C**
1000	**A**,**C**
4000	**A**,D
5000	B,E,F

Regole ottenute:

- A $\Rightarrow$ C supporto 50% confidenza 66.6
- C $\Rightarrow$ A supporto 50% confidenza 100%

Applicazioni

Analisi market basket

- * $\Rightarrow$ Uova
  - cosa si deve promuovere per aumentare le vendite di uova?
- Latte $\Rightarrow$ *
  - quali altri prodotti devono essere venduti da un supermercato che vende latte?
- in generale, gli insiemi X e Y possono comprendere anche più di un oggetto

L'obiettivo di questo problema è trovare regole di associazioni che non sono ovvie – per sfruttarle in modo competitivo

- un esempio classico: Pannolini $\Rightarrow$ * ?

Decomposizione problema

Trovare tutti gli insiemi di item che hanno un supporto minimo (*frequent itemsets*)

Generazione delle regole a partire dai frequent itemsets

Algoritmo fondamentale: APRIORI

[Agrawal, Srikant 1994]

Esempio

Passo 1: estrazione frequent itemsets

TRANSACTION ID	OGGETTI ACQUISTATI
1	A,B,C
2	A,C
3	A,D
4	B,E,F

supporto minimo 50%

FREQUENT ITEMSET	SUPPORTO
{A}	75%
{B}	50%
{C}	50%
{A,C}	50%

Esempio

Passo 2: estrazione regole

- confidenza minima 50%
- Confidenza regola A $\Rightarrow$ C
  - Supporto {A,C} / Supporto {A} = 66.6%
- regole estratte
  - A $\Rightarrow$ C supporto 50%, conf. 66.6%
  - C $\Rightarrow$ A supporto 50%, conf. 100%

Interesse regole estratte

Non sempre tutte le regole con supporto e confidenza superiori ad una certa soglia, anche alta, sono interessanti

Ad esempio, in una scuola con 5000 studenti

- 3000 studenti (60%) giocano a pallacanestro
- 3750 studenti (75%) mangiano cereali a colazione
- 2000 studenti (40%) giocano a pallacanestro e mangiano cereali a colazione

si consideri la regola di associazione

gioca a pallacanestro $\Rightarrow$ mangia cereali

- supporto = 2000/5000 = 40%
- confidenza = 2000/3000 = 66%
- la regola è fuorviante perché il 75% degli studenti mangia cereali!

Interesse regole estratte

In effetti, per determinare l'interesse di una regola di associazione $X \Rightarrow Y$ può essere utile far riferimento anche ad altre proprietà/misure

$$\text{lift} = \frac{\text{prob}(X \cup Y)}{\text{prob}(X)\ \text{prob}(Y)}$$

- affinché una regola sia effettivamente interessante, il lift deve essere maggiore di 1 – nell'esempio è solo 0.88

$$\text{leverage} = \text{prob}(Y|X) - \text{prob}(X)\ \text{prob}(Y)$$

- il leverage serve invece a misurare l'"indipendenza" tra gli oggetti che compaiono nella regola – in questo caso è il 21%

Pattern sequenziali

Una variante dei problemi di associazioni

- dati
  - un insieme di transazioni non anonime – ad esempio, di ciascuna si conosce il cliente
- obiettivo
  - trovare gruppi di oggetti che compaiono in transazioni successive di uno stesso cliente

Esempio

- chi compra test di gravidanza poi compra (con una qualche probabilità) pannolini
- ma ci sono altri acquisti che consentono di predire, in modo tempestivo e con più precisione, una gravidanza?

Pattern sequenziali

Applicazioni

- misura della soddisfazione del cliente
- promozioni mirate
- medicina (sintomi - malattia)

Clustering

Il **clustering** è un altro problema non supervisionato di data mining

- lo scopo è identificare delle classi per raggruppare gli individui di una popolazione sulla base della loro similarità

Dati

- un insieme di *oggetti*

Problema

- trovare una suddivisione degli oggetti in *gruppi* in modo che:
  - gli oggetti in un gruppo siano molto simili tra di loro
  - oggetti in gruppi diversi siano molto diversi
- i gruppi possono essere anche sovrapposti o organizzati gerarchicamente

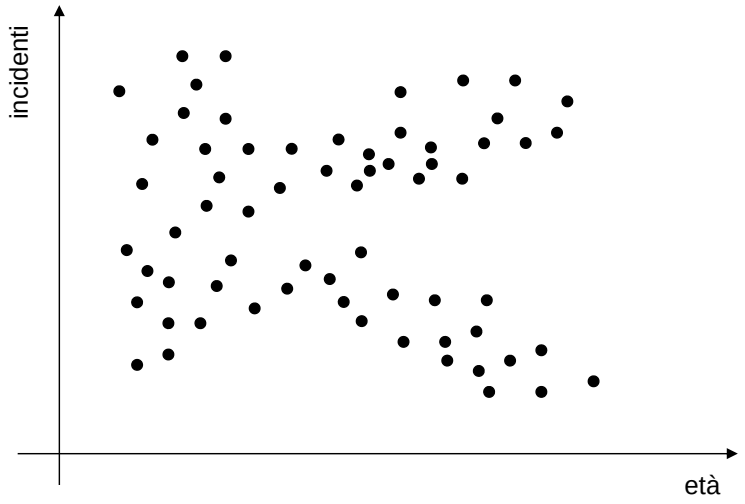
Applicazioni

Identificazione di popolazioni omogenee di clienti in basi di dati di marketing

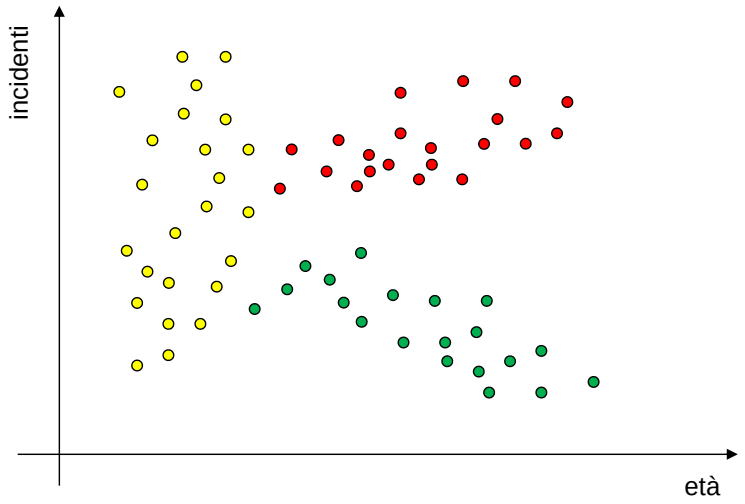
Valutazione dei risultati di esperimenti clinici

Monitoraggio dell'attività di aziende concorrenti

Esempio: dati



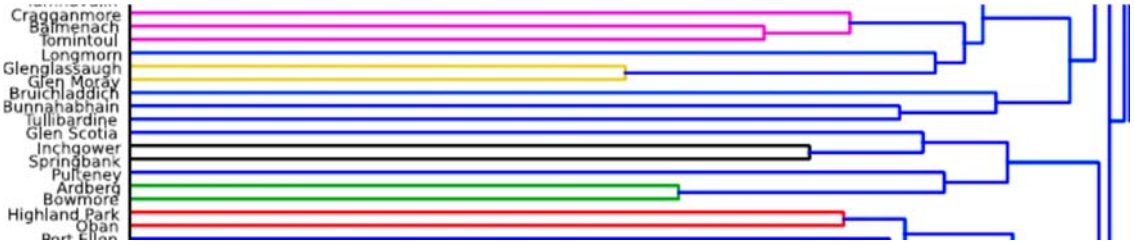
Esempio: clusterizzazione



Esempio

Clustering di whisky scozzesi

- per superare la classificazione basata sulla regione di provenienza – e trovarne una basata sul “gusto”
- ad esempio, un piccolo negozio vuole comunque vendere almeno un prodotto per ciascuno dei “gusti”
- attributi usati per il clustering: color, nose, body, palate, finish



Approcci

Processo

- si determinano i rappresentanti di ogni cluster
- si cercano gli elementi “simili”
- si aggiornano i rappresentanti

Gli algoritmi si differenziano principalmente nella scelta dei rappresentanti

Analisi di sequenze temporali

Un'altra classe di problemi significativi di data mining riguarda l'*analisi di sequenze temporali*

- in questo casi, l'obiettivo può essere
  - trovare sequenze temporali simili ad una sequenza data
  - trovare coppie di sequenze simili
- questo può essere fatto misurando la correlazione tra sequenze
  - può essere utile anche misurare la correlazione di una sequenza con se stessa (ma opportunamente traslata nel tempo)

Applicazioni

Identificazione delle società con comportamento simile di crescita

Determinazione di prodotti con profilo simile di vendita

Identificazione di azioni con andamento simile

Individuazione porzioni onde sismiche non simili per determinare irregolarità geologiche

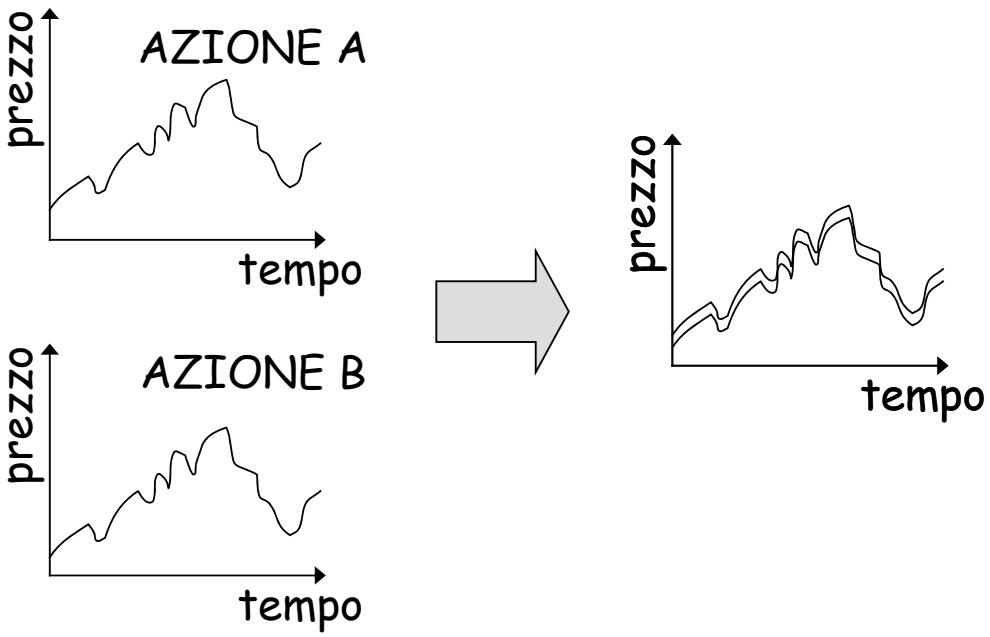
Tecniche

Due tipi di interrogazione

- match completo: la sequenza cercata e le sequenze della base di dati hanno la stessa lunghezza
- match parziale: la sequenza cercata può essere sottosequenza di quelle recuperate dalla base di dati

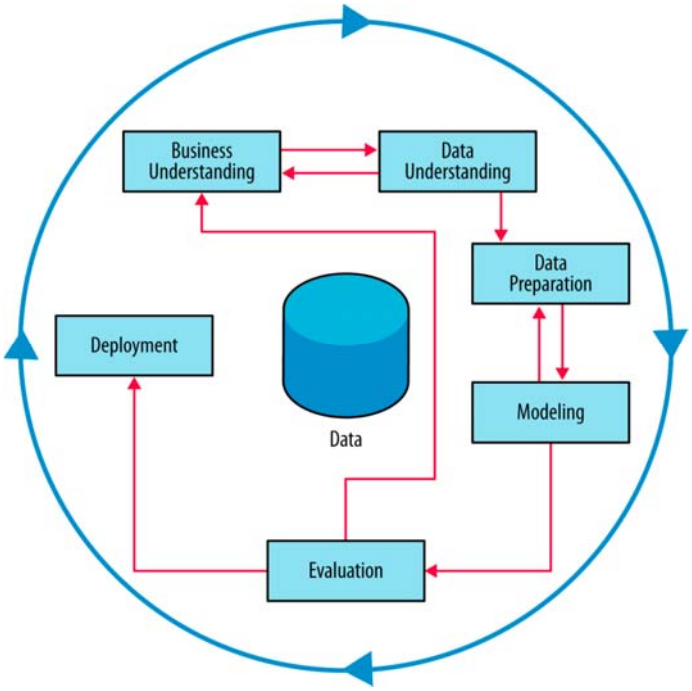
Possibilità di traslazioni, variazioni di scala

Esempio



Il processo di data mining

Il data mining è un processo per risolvere problemi decisionali in modo (ragionevolmente) consistente, ripetibile e oggettivo



Il processo di data mining

Business understanding

- i problemi di business sono complessi, e la loro comprensione è il punto di partenza per la loro soluzione

Data understanding

- è importante anche capire quali sono i dati a disposizione – o quali tra i tanti dati a disposizione sono utili per risolvere il problema in esame
- di solito, la comprensione del problema e quella dei dati sono attività svolte in modo intrecciato, iterativo
- in generale, è iterativo l'intero processo di data mining

Data preparation

- è comune dover pre-elaborare i dati selezionati (ad esempio, tramite conversioni e operazioni di "pulizia" e di normalizzazione dei dati), per poterli effettivamente utilizzare, oppure per ottenere risultati migliori

Il processo di data mining

Modeling

- è l'attività di estrazione di un pattern o di un modello dai dati, che cattura regolarità dei dati che possono essere utilizzate a fini decisionali

Evaluation

- i risultati del data mining vanno poi valutati, ad esempio in termini di validità e affidabilità – per determinare se il modello trovato risolve effettivamente il problema di business iniziale

Deployment

- se il modello trovato è soddisfacente, viene effettivamente messo in uso per risolvere il problema di business di interesse – per realizzare un opportuno ritorno di investimento