


Database and data mining group, Politecnico di Torino 

## **Classificazione**

Elena Baralis  
Politecnico di Torino


Copyright – Tutti i diritti riservati      DATA MINING: CLASSIFICAZIONE - 1      Elena Baralis  
Politecnico di Torino

Database and data mining group, Politecnico di Torino 

## **Classificazione**

- Sono dati
  - insieme di classi
  - oggetti etichettati con il nome della classe di appartenenza (training set)
- L'obiettivo della classificazione è trovare un profilo descrittivo per ogni classe, che permetta di assegnare oggetti di classe ignota alla classe appropriata


Copyright – Tutti i diritti riservati      DATA MINING: CLASSIFICAZIONE - 2      Elena Baralis  
Politecnico di Torino

Database and data mining group, Politecnico di Torino  


## Definizione generale

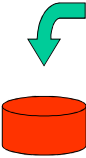
- La classificazione è un processo in due passi
  - Costruzione del modello
    - a partire dal training set, si definisce un modello
    - la qualità del modello è validata mediante una porzione del training set non utilizzata per la costruzione del modello (test set)
  - Uso del modello
    - mediante il modello si assegna l'etichetta di classe a nuovi dati non etichettati
- Il modello può essere definito mediante
  - alberi di decisione
  - reti neurali
  - probabilità
  - regole di associazione
  - support vector machines (SVM)
  - ...

Copyright – Tutti i diritti riservati
DATA MINING: CLASSIFICAZIONE - 3
Elena Baralis  
Politecnico di Torino

Database and data mining group, Politecnico di Torino  


## Esempio

**Creazione del modello**



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

**Classificatore**

```

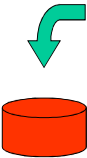
IF Età > 26 and TipoAuto = 'sportiva
THEN Rischio = 'alto'
IF Età ≤ 26 THEN Rischio = 'alto'
IF Età > 26 and TipoAuto <> 'sportiva
THEN Rischio = 'basso'
          
```

Copyright – Tutti i diritti riservati
DATA MINING: CLASSIFICAZIONE - 4
Elena Baralis  
Politecnico di Torino

Database and data mining group, Politecnico di Torino

## Esempio

**Validazione del modello**




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'

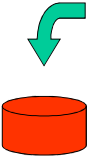


Copyright – Tutti i diritti riservati
DATA MINING: CLASSIFICAZIONE - 5
Elena Baralis  
Politecnico di Torino

Database and data mining group, Politecnico di Torino

## Esempio

**Classificazione**



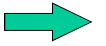
Età	Tipo auto	Classe rischio
35	utilitaria	???

**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'


basso


Copyright – Tutti i diritti riservati
DATA MINING: CLASSIFICAZIONE - 6
Elena Baralis  
Politecnico di Torino

## Preparazione dei dati

- Pulizia dei dati
  - riduzione del rumore
  - gestione dei dati mancanti
- Trasformazione dei dati
  - normalizzazione dell'intervallo di variazione
    - necessaria per alcuni algoritmi (reti neurali, ...)
- Feature selection
  - eliminazione di
    - attributi irrilevanti (esempio: codice fiscale)
    - attributi ridondanti

## Valutazione delle tecniche di classificazione


- Accuratezza
  - qualità della predizione
- Efficienza
  - tempo di costruzione del modello
  - tempo di classificazione
- Scalabilità
  - rispetto alla dimensione del training set
  - rispetto al numero di attributi
- Robustezza
  - gestione di rumore e dati mancanti
- Interpretabilità
  - comprensibilità del modello
  - compattezza del modello

Database and data mining group, Politecnico di Torino  


## Alberi di decisione

- Struttura ad albero in cui
  - i nodi interni denotano un test su un attributo
  - ogni ramo rappresenta un possibile risultato del test
- Punti di forza
  - modello interpretabile
  - buona accuratezza
  - buona efficienza
  - algoritmi recenti con buona scalabilità
- Punti deboli
  - può essere sensibile a dati mancanti


Copyright – Tutti i diritti riservati
DATA MINING: CLASSIFICAZIONE - 9
Elena Baralis  
Politecnico di Torino

Database and data mining group, Politecnico di Torino  


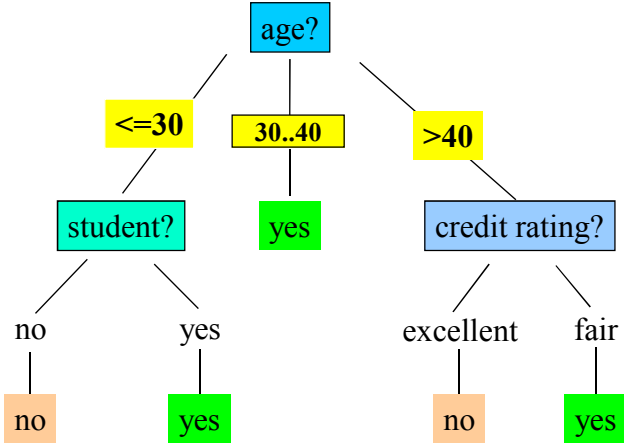
## Alberi di decisione

age	income	student	credit_rating	buys_computer
<=30	high	no	fair	no
<=30	high	no	excellent	no
31...40	high	no	fair	yes
>40	medium	no	fair	yes
>40	low	yes	fair	yes
>40	low	yes	excellent	no
31...40	low	yes	excellent	yes
<=30	medium	no	fair	no
<=30	low	yes	fair	yes
>40	medium	yes	fair	yes
<=30	medium	yes	excellent	yes
31...40	medium	no	excellent	yes
31...40	high	yes	fair	yes
>40	medium	no	excellent	no

Tratto da Han, Kamber, "Data mining: Concepts and Techniques", Morgan Kaufmann 2002  
Copyright – Tutti i diritti riservati
DATA MINING: CLASSIFICAZIONE - 10
Elena Baralis  
Politecnico di Torino


Database and data mining group, Politecnico di Torino  


## Alberi di decisione



Tratto da Han, Kamber, "Data mining; Concepts and Techniques", Morgan Kaufmann 2002

Copyright – Tutti i diritti riservati      DATA MINING: CLASSIFICAZIONE - 11      *Elena Baralis Politecnico di Torino*

Database and data mining group, Politecnico di Torino  


## Alberi di decisione

- Algoritmo base
  - costruzione dell'albero
    - all'inizio tutti i dati di training sono alla radice
    - ripetizione iterativa dei passi
      - selezione dell'attributo "migliore" di partizionamento
      - partizionamento del training set in base all'attributo
    - fino a quando
      - il nodo contiene solo esempi della stessa classe
      - non ci sono più attributi per il partizionamento (classe di maggioranza)
      - non ci sono più dati di training
  - pruning dell'albero
    - riduzione del numero di rami
    - serve per evitare l'overfitting


Copyright – Tutti i diritti riservati      DATA MINING: CLASSIFICAZIONE - 12      *Elena Baralis Politecnico di Torino*

Database and data mining group, Politecnico di Torino  


## Selezione dell'attributo di partizionamento (split)

- Information gain
  - adatto per attributi categorici
  - basato sul concetto di entropia
    - quantità d'informazione necessaria per classificare il training set prima e dopo aver applicato lo split sull'attributo considerato
- Gini index
  - adatto per attributi continui
  - valuta l'insieme di tutti i possibili punti di split per ogni attributo
    - necessarie tecniche per individuare i punti di split "migliori"

Copyright – Tutti i diritti riservati
DATA MINING: CLASSIFICAZIONE - 13
Elena Baralis  
Politecnico di Torino

Database and data mining group, Politecnico di Torino  


## Regole di classificazione

- Ogni percorso nell'albero rappresenta una regola if... then
  - corpo dato dalla congiunzione dei predicati derivati dai nodi interni
  - testa data dall'etichetta di classe del nodo foglia
- Esempio
  - IF *age* = " $\leq 30$ " AND *student* = "no" THEN *buys\_computer* = "no"
  - IF *age* = " $\leq 30$ " AND *student* = "yes" THEN *buys\_computer* = "yes"
  - IF *age* = "31...40" THEN *buys\_computer* = "yes"
  - IF *age* = " $> 40$ " AND *credit\_rating* = "excellent" THEN *buys\_computer* = "yes"
  - IF *age* = " $> 40$ " AND *credit\_rating* = "fair" THEN *buys\_computer* = "no"


Copyright – Tutti i diritti riservati
DATA MINING: CLASSIFICAZIONE - 14
Elena Baralis  
Politecnico di Torino

Database and data mining group, Politecnico di Torino  


## Classificazione associativa

- Modello definito mediante regole di associazione
  - testa della regola è l'etichetta di classe
  - le regole sono selezionate mediante
    - soglie di supporto, confidenza, correlazione
    - database coverage: procedimento di copertura dei dati di training mediante le regole estratte, ordinate secondo opportuni criteri
- Punti di forza
  - modello interpretabile
  - accuratezza maggiore degli alberi di decisione
    - considera contemporaneamente la correlazione di più attributi
  - classificazione efficiente
  - insensibile a dati mancanti
  - buona scalabilità nella dimensione del training set
- Punti di debolezza
  - generazione delle regole può essere lenta
  - scarsa scalabilità nel numero degli attributi

Copyright – Tutti i diritti riservati      DATA MINING: CLASSIFICAZIONE - 15      *Elena Baralis Politecnico di Torino*


Database and data mining group, Politecnico di Torino  


## Classificazione bayesiana

- Basata sul calcolo delle probabilità
- Punti di forza
  - modello di riferimento, se calcolato in modo completo
  - modello facilmente aggiornabile in modo incrementale
  - classificazione efficiente
  - modello con interpretabilità discreta
- Punti di debolezza
  - la generazione del modello di riferimento è computazionalmente intrattabile
    - ipotesi semplificativa: ipotesi naïve
  - ipotesi naïve riduce significativamente l'accuratezza

Copyright – Tutti i diritti riservati      DATA MINING: CLASSIFICAZIONE - 16      *Elena Baralis Politecnico di Torino*




Database and data mining group, Politecnico di Torino  


## Classificazione bayesiana

- Applicazione del teorema di Bayes
  - $P(C|X)$  = probabilità che  $X = \langle x_1, \dots, x_k \rangle$  appartenga alla classe C
  - Assegnazione a X della classe per cui  $P(C|X)$  è massima
  - Teorema di Bayes
 
$$P(C|X) = P(X|C) \cdot P(C) / P(X)$$
    - P(X) costante per tutte le C
    - P(C) probabilità a priori di C
    - P(X|C) **non** può essere computata in modo esaustivo
- Ipotesi naïve
  - indipendenza statistica degli attributi  $x_1, \dots, x_k$
  - non è sempre verificata
    - la qualità del modello può essere affetta
- Reti bayesiane
  - permettono di specificare un sottinsieme di dipendenze tra attributi

Copyright – Tutti i diritti riservati      DATA MINING: CLASSIFICAZIONE - 17      Elena Baralis Politecnico di Torino

Database and data mining group, Politecnico di Torino  


## Classificazione bayesiana

Outlook	Temperature	Humidity	Windy	Class
sunny	hot	high	false	N
sunny	hot	high	true	N
overcast	hot	high	false	P
rain	mild	high	false	P
rain	cool	normal	false	P
rain	cool	normal	true	N
overcast	cool	normal	true	P
sunny	mild	high	false	N
sunny	cool	normal	false	P
rain	mild	normal	false	P
sunny	mild	normal	true	P
overcast	mild	high	true	P
overcast	hot	normal	false	P
rain	mild	high	true	N

Tratto da Han, Kamber, "Data mining: Concepts and Techniques", Morgan Kaufmann 2002  
 Copyright – Tutti i diritti riservati      DATA MINING: CLASSIFICAZIONE - 18      Elena Baralis Politecnico di Torino

Database and data mining group, Politecnico di Torino

## Classificazione bayesiana

<b>outlook</b>	
<b>P(sunny p) = 2/9</b>	<b>P(sunny n) = 3/5</b>
<b>P(overcast p) = 4/9</b>	<b>P(overcast n) = 0</b>
<b>P(rain p) = 3/9</b>	<b>P(rain n) = 2/5</b>
<b>temperature</b>	
<b>P(hot p) = 2/9</b>	<b>P(hot n) = 2/5</b>
<b>P(mild p) = 4/9</b>	<b>P(mild n) = 2/5</b>
<b>P(cool p) = 3/9</b>	<b>P(cool n) = 1/5</b>
<b>humidity</b>	
<b>P(high p) = 3/9</b>	<b>P(high n) = 4/5</b>
<b>P(normal p) = 6/9</b>	<b>P(normal n) = 2/5</b>
<b>windy</b>	
<b>P(true p) = 3/9</b>	<b>P(true n) = 3/5</b>
<b>P(false p) = 6/9</b>	<b>P(false n) = 2/5</b>

**P(p) = 9/14**

**P(n) = 5/14**

Dato da classificare  
 $X = \langle \text{rain, hot, high, false} \rangle$

$P(X|p) \cdot P(p) =$   
 $P(\text{rain}|p) \cdot P(\text{hot}|p) \cdot P(\text{high}|p) \cdot P(\text{false}|p) \cdot P(p)$   
 $= 3/9 \cdot 2/9 \cdot 3/9 \cdot 6/9 \cdot 9/14 = 0.010582$

$P(X|n) \cdot P(n) =$   
 $P(\text{rain}|n) \cdot P(\text{hot}|n) \cdot P(\text{high}|n) \cdot P(\text{false}|n) \cdot P(n)$   
 $= 2/5 \cdot 2/5 \cdot 4/5 \cdot 2/5 \cdot 5/14 = 0.018286$

Tratto da Han, Kamber, "Data mining: Concepts and Techniques", Morgan Kaufmann 2002

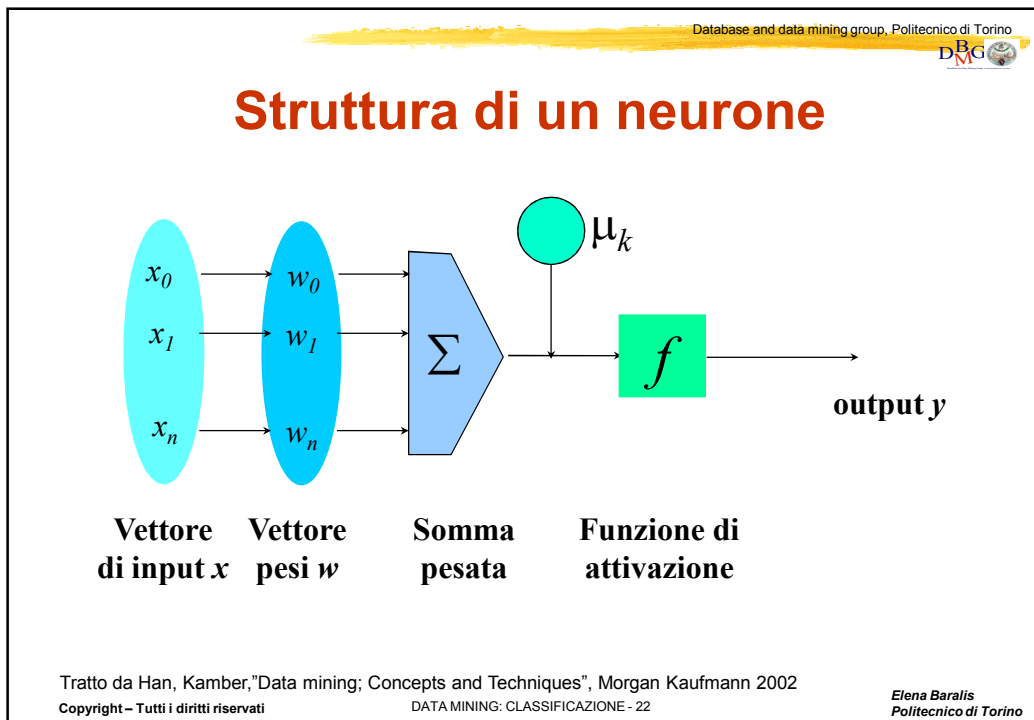
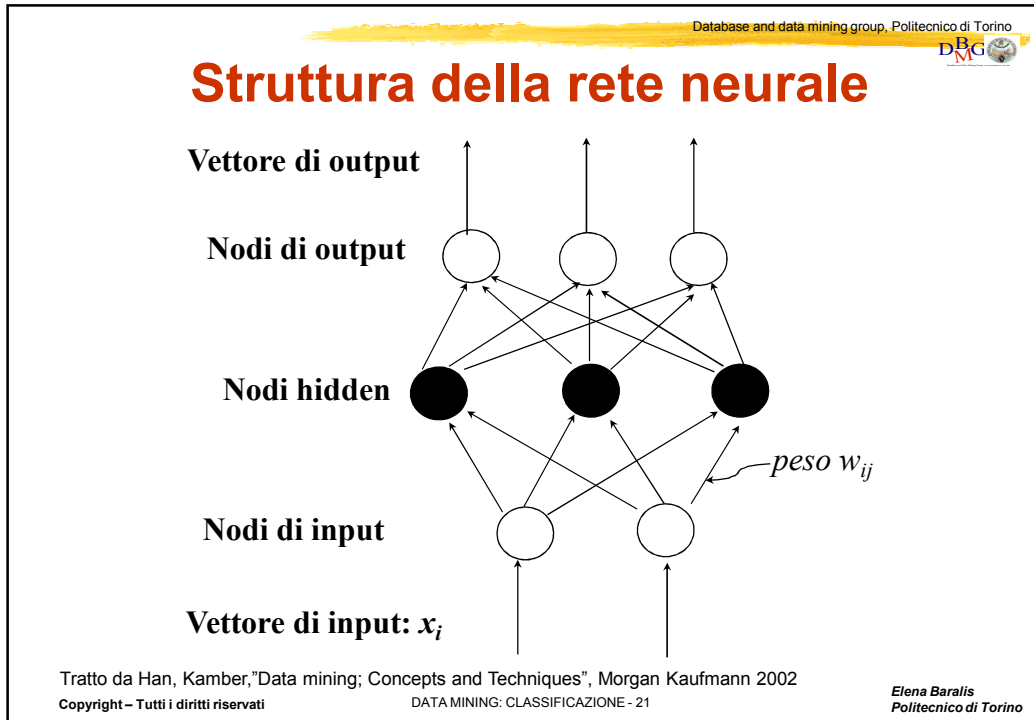
Copyright – Tutti i diritti riservati      DATA MINING: CLASSIFICAZIONE - 19      Elena Baralis Politecnico di Torino


Database and data mining group, Politecnico di Torino

## Reti neurali

- Ispirate alla struttura del cervello umano
  - neuroni come unità di elaborazione
  - sinapsi come rete di collegamento
- Punti di forza
  - accuratezza elevata
  - robusto in presenza di outliers e rumore
  - efficienza di classificazione
  - output discreto o continuo
- Punti di debolezza
  - processo di apprendimento lento
  - modello non interpretabile
  - difficile introdurre conoscenza del dominio applicativo

Copyright – Tutti i diritti riservati      DATA MINING: CLASSIFICAZIONE - 20      Elena Baralis Politecnico di Torino




Database and data mining group, Politecnico di Torino  


## Costruzione della rete neurale

- Obiettivo
  - definire un insieme ottimale di pesi e offset
- Algoritmo di base
  - Assegnazione iniziale di valori casuali a pesi e offset
  - Elaborazione delle istanze del training set una per volta
    - Per ogni neurone, calcolo del risultato dell'applicazione di pesi, offset e funzione di attivazione per l'istanza
    - Propagazione in avanti fino al calcolo dell'output
    - Confronto con l'output atteso e calcolo dell'errore
    - Propagazione all'indietro dell'errore (backpropagation), ricalcolando pesi e offset
  - Il processo termina quando
    - % di accuratezza sopra soglia data
    - % di errore sotto soglia data
    - numero massimo di epoche raggiunto

Copyright – Tutti i diritti riservati      DATA MINING: CLASSIFICAZIONE - 23      Elena Baralis Politecnico di Torino

Database and data mining group, Politecnico di Torino  


## Misure di qualità

- Misura di qualità per il modello di classificazione
 
$$\text{Accuratezza} = \frac{\text{num. dati classificati correttamente}}{\text{num. dati classificati}}$$
  - non adatta per
    - distribuzioni sbilanciate delle etichette di classe
    - importanza diversa delle classi
- Misure di qualità per una singola classe C
 
$$\text{Richiamo} = \frac{\text{num. dati classificati correttamente in } C}{\text{num. dati appartenenti a } C}$$

$$\text{Precisione} = \frac{\text{num. dati classificati correttamente in } C}{\text{num. dati assegnati a } C}$$

Copyright – Tutti i diritti riservati      DATA MINING: CLASSIFICAZIONE - 24      Elena Baralis Politecnico di Torino

## Validazione di un classificatore

- Partizionamento dei dati in
  - dati di training, utilizzati per costruire il modello
  - dati di test, utilizzati per validare il modello generato
- Tecniche di partizionamento
  - partizionamento fisso
    - training set 2/3
    - test set 1/3
    - adatto per dataset molto grandi
  - cross validation (K-fold)
    - divisione del training set in k sottinsiemi, detti folds
    - a rotazione su tutti i K fold, 1 fold è utilizzato per il test e gli altri K-1 per il training
  - leave-one-out
    - cross validation con un solo record di test per volta
    - richiede la generazione di un numero di classificatori pari alla cardinalità del dataset
    - adatto per dataset molto piccoli