

# Classificatori Bayesiani

Prof. Matteo Golfarelli  
Alma Mater Studiorum - Università di Bologna

## Classificatori Bayesiani

- Rappresentano un approccio probabilistico per risolvere problemi di classificazione
  - ✓ In molte applicazioni la relazione tra i valori degli attributi e quello della classe non è deterministico
    - Rumore dei dati
    - Presenza di caratteristiche del fenomeno non modellate dagli attributi
    - Difficoltà nel quantificare certi aspetti del fenomeno
  - ✓ Per esempio, predire se una persona è a rischio cardiaco dipende fortemente dalla sua dieta e dalla sua attività fisica ma, persone che hanno un'alimentazione sana e si allenano regolarmente possono avere comunque problemi di cuore
    - Esistono altri fattori quali l'ereditarietà, l'abuso di alcool
    - E' difficile dire quanto una dieta sia "sana" e se l'allenamento sia "adeguato"
  - ✓ Tutto ciò introduce incertezza sull'esito della previsione
- I classificatori Bayesiani modellano relazioni probabilistiche tra gli attributi e l'attributo di classificazione

## Richiami di statistica

- **Probabilità condizionata:** probabilità che si verifichi l'evento  $c$  sapendo che si è verificato l'evento  $a$

$$P(C | A) = \frac{P(A, C)}{P(A)}$$

$$P(A | C) = \frac{P(A, C)}{P(C)}$$

$$P(A, C) = P(A | C)P(C)$$

- **Teorema di Bayes:** 
$$P(C | A) = \frac{P(A | C)P(C)}{P(A)}$$

- **Teorema della probabilità assoluta**

✓  $b_i$  indica uno degli  $n$  valori assumibili da  $B$

$$P(A = a) = \sum_{i=1}^n P(A = a, B = b_i) = \sum_{i=1}^n P(A = a | B = b_i)P(B = b_i)$$

## Teorema di Bayes un esempio

- Si supponga che
  - ✓ Un dottore sa che la meningite causa indolenzimento del collo nel 50% dei casi
  - ✓ La probabilità a priori che un paziente abbia la meningite è  $1/50,000$
  - ✓ La probabilità a priori che un qualsiasi paziente soffra di indolenzimento al collo è pari a  $1/20$
- Se un paziente denuncia indolenzimento al collo, qual è la probabilità che abbia la meningite?

$$P(M | I) = \frac{P(I | M)P(M)}{P(I)} = \frac{0.5 \times 1/50000}{1/20} = 0.0002 = 1/5,000$$

*Si supponga che nella triennale gli studenti fumatori siano il 15% mentre nella magistrale il 23%. Se  $1/5$  del numero totale di studenti è iscritto alla magistrale quale è la probabilità che uno studente che fumi sia iscritto alla magistrale?*



## Classificatori Bayesiani

- Sia dato il vettore  $\mathbf{A}=(A_1, A_2, \dots, A_n)$  che descrive il set di attributi e sia  $C$  la variabile di classe
- Se  $C$  è legato in modo non deterministico ai valori assunti da  $\mathbf{A}$  possiamo trattare le due variabili come variabili casuali e catturare le loro relazioni probabilistiche utilizzando  $P(C|\mathbf{A})$
- Durante la fase di training si imparano i legami probabilistici  $P(C|\mathbf{A})$  per ogni combinazione di valori assunti da  $\mathbf{A}$  e  $C$
- Conoscendo queste probabilità un test record  $\mathbf{a}$  può essere classificato trovando la label di classe  $c$  che massimizza la probabilità a posteriori  $P(c|\mathbf{a})$

## Classificatori Bayesiani

| Tid | Refund | Marital Status | Taxable Income | Evade |
|-----|--------|----------------|----------------|-------|
| 1   | Yes    | Single         | 125K           | No    |
| 2   | No     | Married        | 100K           | No    |
| 3   | No     | Single         | 70K            | No    |
| 4   | Yes    | Married        | 120K           | No    |
| 5   | No     | Divorced       | 95K            | Yes   |
| 6   | No     | Married        | 60K            | No    |
| 7   | Yes    | Divorced       | 220K           | No    |
| 8   | No     | Single         | 85K            | Yes   |
| 9   | No     | Married        | 75K            | No    |
| 10  | No     | Single         | 90K            | Yes   |

$X=(\text{Refund}=\text{Yes}, \text{Marital Status}=\text{Married}, \text{Taxable Income}=60\text{K})$

- Risolvere il problema di classificazione significa calcolare  $P(\text{Evade}=\text{No}|X)$  e  $P(\text{Evade}=\text{Yes}|X)$ 
  - ✓ Se  $P(\text{Evade}=\text{No}|X) > P(\text{Evade}=\text{Yes}|X) \rightarrow \text{No}$
  - ✓ Se  $P(\text{Evade}=\text{Yes}|X) > P(\text{Evade}=\text{No}|X) \rightarrow \text{Yes}$

## Classificatori Bayesiani

- Calcolare  $P(C|A)$  per ogni possibile valore di  $C$  e  $A$  richiede un training set molto grande anche per un numero ridotto di attributi
- Il teorema di Bayes è utile in questo caso poiché permette di esprimere la probabilità a posteriori  $P(C|A)$  in termini di  $P(A|C)$ ,  $P(C)$  e  $P(A)$

$$P(C | A_1 A_2 \dots A_n) = \frac{P(A_1 A_2 \dots A_n | C) P(C)}{P(A_1 A_2 \dots A_n)}$$

- ✓ Visto che  $P(A)$  è costante in questa formula il problema di massimizzare la probabilità a posteriori equivale a scegliere il valore di  $C$  che massimizza

$$P(A_1, A_2, \dots, A_n | C) P(C)$$

- **Come stimare  $P(A_1, A_2, \dots, A_n | C)$ ?**
  - ✓ **Naïve Bayes**
  - ✓ **Reti Bayesiane (Bayesian Belief Network)**

## Naïve Bayes

- Assumono l'indipendenza tra gli attributi  $A_i$  quando la classe è nota (indipendenza condizionale):

- ✓ Date tre variabili aleatorie  $X$ ,  $Y$  e  $Z$ ,  $X$  si dice indipendente da  $Y$  dato  $Z$

$$\text{se } P(X|Y,Z) = P(X|Z)$$

- ✓ In altre parole, se si conosce il valore assunto da  $Z$ , conoscere il valore di  $Y$  non influenza il valore assunto da  $X$

- Grazie all'indipendenza stocastica possiamo scrivere:

$$\begin{aligned} P(X, Y | Z) &= \frac{P(X, Y, Z)}{P(Z)} \\ &= \frac{P(X, Y, Z)}{P(Y, Z)} \times \frac{P(Y, Z)}{P(Z)} \\ &= P(X | Y, Z) \times P(Y | Z) \\ &= P(X | Z) \times P(Y | Z) \end{aligned}$$

## Naïve Bayes

- L'assunzione di indipendenza tra gli attributi  $A_i$  quando è nota la classe  $C$  permette di riscrivere:

$$P(A_1, A_2, \dots, A_n | C) = P(A_1 | C) P(A_2 | C) \dots P(A_n | C)$$

- ✓ Invece di dover calcolare la probabilità condizionata per ogni combinazione di valori di  $\mathbf{A}$  è sufficiente calcolare  $P(A_k | C)$  per ogni  $A_k$  e  $C=c_j$ .

- ✓ Il nuovo punto è classificato come  $C=c_{j^*}$  se:

$$j^* = \arg \max_j P(C = c_j) \prod_{k=1}^n P(A_k = a_k | C = c_j)$$

## Naïve Bayes: un esempio

| Tid | Refund | Marital Status | Taxable Income | Evade |
|-----|--------|----------------|----------------|-------|
| 1   | Yes    | Single         | 125K           | No    |
| 2   | No     | Married        | 100K           | No    |
| 3   | No     | Single         | 70K            | No    |
| 4   | Yes    | Married        | 120K           | No    |
| 5   | No     | Divorced       | 95K            | Yes   |
| 6   | No     | Married        | 60K            | No    |
| 7   | Yes    | Divorced       | 220K           | No    |
| 8   | No     | Single         | 85K            | Yes   |
| 9   | No     | Married        | 75K            | No    |
| 10  | No     | Single         | 90K            | Yes   |

- Classe:  $P(C=c_j) = N_j / N$

- ✓ Es.  $P(\text{No}) = 7/10$ ,  
 $P(\text{Yes}) = 3/10$

- Per attributi discreti:

$$P(A=a_i | C=c_j) = N_{ij} / N_j$$

- ✓ dove  $N_{ij}$  è il numero di istanze che assumono il valore  $a_i$  e che appartengono alla classe  $c_j$

- ✓ Esempi:

$$P(\text{Status}=\text{Married} | \text{Evade}=\text{No}) = 4/7$$
$$P(\text{Refund}=\text{Yes} | \text{Evade}=\text{Yes})=0$$

## Stima delle probabilità per attributi continui

- Nel caso l'attributo A sia continuo non è possibile stimare la probabilità per ogni suo valore
  - ✓ **Discretizzare l'attributo in intervalli creando un attributo ordinale**
    - se si usano troppi intervalli il limitato numero di eventi del training set per intervallo rende inaffidabile la previsione
    - se si usano pochi intervalli uno di essi può aggregare valori associabili a classi diverse e determinare quindi un decision boundary sbagliato
  - ✓ **Associare all'attributo una funzione di densità e stimare i parametri della funzione dal training set:**
    - Tipicamente si assume che la probabilità condizionata di un attributo (a valori continui) rispetto all'attributo classe segua una distribuzione normale
    - Quando la densità di probabilità è nota, può essere usata per stimare  $P(A|C)$

## Naïve Bayes: un esempio

| Tid | Refund | Marital Status | Taxable Income | Evade |
|-----|--------|----------------|----------------|-------|
| 1   | Yes    | Single         | 125K           | No    |
| 2   | No     | Married        | 100K           | No    |
| 3   | No     | Single         | 70K            | No    |
| 4   | Yes    | Married        | 120K           | No    |
| 5   | No     | Divorced       | 95K            | Yes   |
| 6   | No     | Married        | 60K            | No    |
| 7   | Yes    | Divorced       | 220K           | No    |
| 8   | No     | Single         | 85K            | Yes   |
| 9   | No     | Married        | 75K            | No    |
| 10  | No     | Single         | 90K            | Yes   |

- Distribuzione normale:
 
$$P(A = a_i | C = c_j) = \frac{1}{\sqrt{2\pi\sigma_j^2}} e^{-\frac{(a_i - \mu_j)^2}{2\sigma_j^2}}$$
- ✓ Una distribuzione per ogni attributo A e per ogni valore  $c_j$  della classe C
- Per (Income, Class=No):
  - ✓ Se Class=No
    - Media  $\mu = 110$
    - Varianza  $\sigma^2 = 2975$

$$P(\text{Income} = 120 | \text{Evade} = \text{No}) = \frac{1}{\sqrt{2\pi(54.54)}} e^{-\frac{(120-110)^2}{2(2975)}} = 0.0072$$

## Naïve Bayes: un esempio

- Dato il test record:

$X=(\text{Refund}=\text{No}, \text{Marital Status}=\text{Married}, \text{Income}=\text{120K})$

$P(\text{Refund}=\text{Yes}|\text{No}) = 3/7$   
 $P(\text{Refund}=\text{No}|\text{No}) = 4/7$   
 $P(\text{Refund}=\text{Yes}|\text{Yes}) = 0$   
 $P(\text{Refund}=\text{No}|\text{Yes}) = 1$   
 $P(\text{Marital Status}=\text{Single}|\text{No}) = 2/7$   
 $P(\text{Marital Status}=\text{Divorced}|\text{No}) = 1/7$   
 $P(\text{Marital Status}=\text{Married}|\text{No}) = 4/7$   
 $P(\text{Marital Status}=\text{Single}|\text{Yes}) = 2/3$   
 $P(\text{Marital Status}=\text{Divorced}|\text{Yes}) = 1/3$   
 $P(\text{Marital Status}=\text{Married}|\text{Yes}) = 0$

For taxable income:  
 If class = No sample mean = 110  
 sample variance=2975

If class = Yes sample mean = 90  
 sample variance=25

- $P(X|\text{Class}=\text{No}) = P(\text{Refund}=\text{No}|\text{Class}=\text{No}) \times P(\text{Married}|\text{Class}=\text{No}) \times P(\text{Income}=\text{120K}|\text{Class}=\text{No}) = 4/7 \times 4/7 \times 0.0072 = 0.0024$
- $P(X|\text{Class}=\text{Yes}) = P(\text{Refund}=\text{No}|\text{Class}=\text{Yes}) \times P(\text{Married}|\text{Class}=\text{Yes}) \times P(\text{Income}=\text{120K}|\text{Class}=\text{Yes}) = 1 \times 0 \times 1.2 \times 10^{-9} = 0$

$P(X|\text{No})P(\text{No}) > P(X|\text{Yes})P(\text{Yes})$

$0.0024 \times 7/10 > 0 \times 3/10$

Quindi  $P(\text{No}|X) > P(\text{Yes}|X) \Rightarrow \text{Class} = \text{No}$

## Correttori

- L'esempio precedente mostra un problema che si riscontra con questi classificatori
  - ✓ Se una delle probabilità condizionate è 0 l'intera espressione assumerà valore 0
- Per evitare questo problema si possono adottare dei correttivi che evitano l'azzeramento

Originale :  $P(A = a_i | C = c_j) = \frac{N_{ij}}{N_j}$

Laplace :  $P(A = a_i | C = c_j) = \frac{N_{ij} + 1}{N_j + c}$

m - estimate :  $P(A = a_i | C = c_j) = \frac{N_{ij} + mp}{N_j + m}$

$c$ : numero delle label di classe

$p$ : probabilità a priori  $P(A=a_i)$

$m$ : parametro (equivalent sample size) che determina l'importanza della probabilità a priori rispetto alla probabilità osservata ( $N_{ij}/N_j$ )

## Proprietà

- Il vantaggio principale del ragionamento probabilistico rispetto a quello logico sta nella possibilità di giungere a descrizioni razionali anche quando non vi è abbastanza informazione di tipo deterministico sul funzionamento del sistema.
- Robusti a punti di rumore isolato
  - ✓ Il rumore è cancellato dall'operazione di media durante il calcolo di  $P(A|C)$
- I classificatori gestiscono dati mancanti non considerando l'evento durante i calcoli
- Sono robusti rispetto ad attributi irrilevanti
  - ✓ Se  $A$  è un attributo irrilevante  $P(A|C)$  è uniformemente distribuito rispetto ai valori di  $C$  e quindi il suo contributo è irrilevante (uguale per tutti i valori  $c_j$ )
$$j^* = \arg \max_j P(C = c_j) \prod_{k=1}^n P(A_k = a_k | C = c_j)$$
- Attributi correlati possono ridurre l'efficacia dato che per essi non vale l'assunzione di indipendenza condizionale
  - ✓ Conviene utilizzare tecniche più sofisticate quali le Bayesian Belief Networks (BBN)

## Proprietà

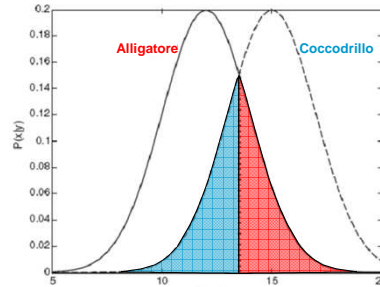
- Forniscono **risultati ottimi** se:
  - ✓ E' rispettata la condizione di indipendenza condizionale
  - ✓ Sono note le distribuzioni di probabilità di  $P(X|Y)$ 
    - ATTENZIONE Quelle vere non quelle inferite dal training set
- Un esempio: si supponga di dover distinguere (classificare) alligatori e coccodrilli in base alla loro lunghezza
  - ✓ Lunghezza media coccodrillo 15 piedi
  - ✓ Lunghezza media alligatore 12 piedi
- Assumendo che la distribuzione delle lunghezze segua una distribuzione gaussiana con  $\sigma=2$  possiamo scrivere

$$P(X | \text{Coccodrillo}) = \frac{1}{\sqrt{2\pi} \cdot 2} \exp\left(-\frac{1}{2} \left(\frac{X-15}{2}\right)^2\right)$$
$$P(X | \text{Alligatore}) = \frac{1}{\sqrt{2\pi} \cdot 2} \exp\left(-\frac{1}{2} \left(\frac{X-12}{2}\right)^2\right)$$



## Classificatori Bayesiani: proprietà

- Il decision boundary sarà posizionato in  $x^*=13.5$ . Questo punto determina il **minimo error rate ottenibile da qualsiasi classificatore**



- L'error rate del classificatore (**errore bayesiano**) è dato dall'area al di sotto la curva di probabilità a posteriori
  - ✓ Per i coccodrilli da 0 a  $x^*$
  - ✓ Per gli alligatori da  $x^*$  a  $\infty$

$$Errorebayesiano = \int_0^{x^*} P(\text{Coccodrillo} | X) dX + \int_{x^*}^{\infty} P(\text{Alligatore} | X) dX$$

## Esercizio

- Si consideri il seguente data set

| ID | A | B | C | Class |
|----|---|---|---|-------|
| 1  | 0 | 0 | 1 | -     |
| 2  | 1 | 0 | 1 | +     |
| 3  | 0 | 1 | 0 | -     |
| 4  | 1 | 0 | 0 | -     |
| 5  | 1 | 0 | 1 | +     |
| 6  | 0 | 0 | 1 | +     |
| 7  | 1 | 1 | 0 | -     |
| 8  | 0 | 0 | 0 | -     |
| 9  | 0 | 1 | 0 | +     |
| 10 | 1 | 1 | 1 | +     |

- Calcolare la probabilità condizionata  $P(A=1|+)$ ,  $P(B=1|+)$ ,  $P(C=1|+)$ ,  $P(A=1|-)$ ,  $P(B=1|-)$ ,  $P(C=1|-)$
- Calcolare il valore dell'attributo di classe per il record  $(A=1, B=1, C=1)$  utilizzando l'approccio Naïve Bayes

