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 caratterisitche 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'eridaterietà, 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*

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

$$P(A,C) = P(A \mid C)P(C)$$

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

- Teorema della probabilità assoluta
  - ✓ b_i indica uno degli *n* valori assumibili da B

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

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 paziende denuncia indolenziamento al collo, qual è la probabilità che abbia la meningite?

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

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 **A**=(A_1, A_2,…,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 **A** possiamo trattare le due variabili come variabili casuali e catturare le loro relazioni probabilistiche utilizzando P(C|**A**)
- Durante la fase di training si imparano i legami probabilistici P(C|**A**) per ogni combinazione di valori assunti da **A** e C
- Conoscendo queste probabilità un test record ***a*** può essere classificato trovando la label di classe *c* che massimizza la probabilità a posteriori P(*c*|***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=(Refund=Yes,Marital Status=Married, Taxable Income=60K)

- Risolvere il problema di classificazione significa calcolare P(Evade=No|X) e P(Evade=Yes|X)
  - ✓ Se P(Evade=No|**X**) > P(Evade=Yes|**X**) →No
  - ✓ Se P(Evade=Yes|**X**) > P(Evade=No|**X**) →Yes

Classificatori Bayesiani

ML

- 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**)

$$P(C \mid A_1 A_2 \ldots A_n) = \frac{P(A_1 A_2 \ldots A_n \mid C) P(C)}{P(A_1 A_2 \ldots A_n)}$$

  - ✓ 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

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

- **Come stimare $P(A_1, A_2, \ldots, A_n \mid C)$?**
  - ✓ **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 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:

$$P(X,Y \mid Z) = \frac{P(X,Y,Z)}{P(Z)}$$

$$= \frac{P(X,Y,Z)}{P(Y,Z)} \times \frac{P(Y,Z)}{P(Z)}$$

$$= P(X \mid Y,Z) \times P(Y \mid Z)$$

$$= P(X \mid Z) \times P(Y \mid Z)$$

Naïve Bayes

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

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

  - ✓ Invece di dover calcolare la probabilità condizionata per ogni combinazione di valori di **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:

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

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(No) = 7/10, P(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(Status=Married|Evade=No) = 4/7
    P(Refund=Yes|Evade=Yes)=0

Stima delle probabilità per attributi continui

- Nel caso l'attributo A sia continuo non è possibile stimare la probaibilità 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:

$$P(A = a_i \mid C = c_j) = \frac{1}{\sqrt{2\pi\sigma_j^2}} e^{-\frac{(a_i - \mu_j)^2}{2\sigma_j^2}}$$

  - ✓ Una distribuzione per ogni attributo A e per ogni valore c_j della classe C
- Per (Income, Class=No):
  - ✓ Se Class=No
    - Media μ= 110
    - Varianza σ^2= 2975

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

Naïve Bayes: un esempio

- Dato il test record:

  X=(Refund=No,Marrital Status=Married,Income=120K)

P(Refound=Yes|No) = 3/7
P(Refound=No|No) = 4/7
P(Refound=Yes|Yes) = 0
P(Refound=No|Yes) = 1
P(Marital Status=Single|No) = 2/7
P(Marital Status=Divorced|No) = 1/7
P(Marital Status=Married|No) = 4/7
P(Marital Status=Single|Yes) = 2/3
P(Marital Status=Divorced|Yes) = 1/3
P(Marital Status=Married|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|Class=No) = P(Refund=No|Class=No)
  × P(Married| Class=No)
  × P(Income=120K| Class=No)
  = 4/7 × 4/7 × 0.0072 = 0.0024

- P(X|Class=Yes) = P(Refund=No| Class=Yes)
  × P(Married| Class=Yes)
  × P(Income=120K| Class=Yes)
  = 1 × 0 × 1.2 × 10^{-9} = 0

P(X|No)P(No) > P(X|Yes)P(Yes)

0.0024 × 7/10 > 0 × 3/10

Quindi P(No|X) > P(Yes|X) => Class = 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 eviare questo problema si possono adottare dei correttivi che evitano l'azzeramento

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

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

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

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 piori 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)

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

- 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 alligatore12 piedi

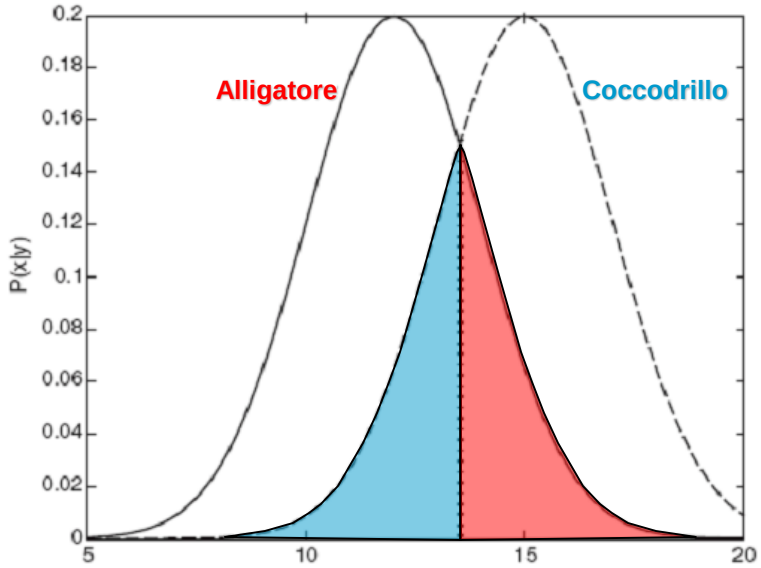
- Assumendo che la distribuzione delle lunghezze segua una distribuzione gaussiana con σ=2 possiamo scrivere

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

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

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 ∞

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

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