DATA SKILLS
UNDERSTANDING THE WORLD

Naïve Bayesian Classification

Di Alessandro rezzani

Sommario

L'algoritmo

La classificazione Bayesiana è una tecnica statistica con la quale si determina la probabilità di un elemento di appartenere a una certa classe. Per esempio, questa tecnica può essere impiegata per stimare la probabilità di un cliente di appartenere alla classe dei compratori di *tablet PC*, dati alcuni attributi del cliente quali: il tipo di lavoro svolto, l'età, il reddito, lo stato civile, ecc.

La tecnica si basa sul teorema di Bayes, matematico e ministro presbiteriano britannico del diciottesimo secolo. Il teorema definisce la probabilità condizionata (o a posteriori) di un evento rispetto ad un altro.

La formula della probabilità condizionata è $P(A|B) = \frac{P(B|A) \times P(A)}{P(B)}$ dove:

P(A|B) è la probabilità condizionata di A rispetto a B
P(B|A) è la probabilità condizionata di A rispetto a B
P(A) è la probabilità "a priori" di A, che non tiene conto di nessuna informazione circa B
P(B) è la probabilità "a priori" di B che non tiene conto di nessuna informazione circa A

Le probabilità "a priori" possono essere stimate attraverso la frequenza campionaria, per quanto riguarda gli attributi discreti, mentre per gli attributi continui si assume che essi siano distribuiti secondo la distribuzione *normale* e si utilizza la funzione di densità per il calcolo delle probabilità.

L'algoritmo *naïve bayesian classifier* assume che l'effetto di un attributo su una data classe è indipendente dai valori degli altri attributi. Tale assunzione, chiamata indipendenza condizionale delle classi, ha lo scopo di semplificare i calcoli e proprio per questo l'algoritmo prende il nome di "naïve". Quando tale assunzione è vera nella realtà, l'accuratezza dell'algoritmo è paragonabile a quella dei *decision tree* e delle *reti neurali*.

L'algoritmo determina la classe di appartenenza in base alle probabilità condizionali per tutte le classi in base agli attributi dei vari elementi. La classificazione corretta si ha quando la probabilità condizionale di una certa classe C rispetto agli attributi A_n ($P\{C|A1, A2, \dots, An\}$) è massima.
La probabilità condizionale è data da:

$$P\{C_j|A\} = P\{C_j|A_1, A_2, \dots, A_n\} = \frac{P\{A_1, A_2, \dots, A_n|C_j\} \times P\{C_j\}}{P\{A_1, A_2, \dots, A_n\}}$$

e visto che assumiamo l'indipendenza degli attributi abbiamo, per ogni classe C_j che:

$$P\{A1, A2, \dots, An|C_j\} = \prod_n P\{An|C_j\}$$

Inoltre massimizzare $P\{C_j|A\}$ equivale a massimizzare il numeratore $P\{A_1, A_2, \dots, A_n|C_j\} \times P\{C_j\}$ che possiamo esprimere anche come $\prod_n P\{An|C_j\} \times P\{C_j\}$

Facciamo ora un esempio concreto di come avviene la naïve bayesian classification, volendo stimare se un soggetto potrà essere acquirente dei nostri prodotti, oppure no. In questo caso, dunque esistono solo due classi: ACQUIRENTE_SI e ACQUIRENTE_NO. I dati, noti, di partenza sono elencati nella tabella seguente, dalla quale si nota che gli attributi su cui lavoreremo sono: reddito annuo lordo, stato coniugale, sesso.

Tabella Errore. Nel documento non esiste testo dello stile specificato.**-1 Data set di training**

NR. progressivo	Stato Coniugale	Sesso	Reddito annulo lordo	ACQUIRENTE
1	Sposato	M	€ 35.000	SI
2	Single	F	€ 47.000	NO
3	Sposato	F	€ 58.000	SI
4	Single	M	€ 31.000	NO
5	Separato	M	€ 70.000	SI
6	Sposato	F	€ 27.000	NO
7	Single	F	€ 36.000	SI
8	Separato	M	€ 50.000	NO
9	Sposato	F	€ 65.000	SI
10	Single	M	€ 18.000	NO
11	***Sposato***	***M***	***€ 40.000***	***??***

Ciò che dobbiamo fare è stimare la classe di appartenenza dell'ultimo elemento (il nr. 11) del quale conosciamo tutti gli attributi, tranne, ovviamente, la classe.

In base ai dati abbiamo le seguenti stime:
$P(C_1)$ = P(ACQUIRENTE_SI) = Numero casi SI/Numero casi = 5/10 = 0.5
$P(C_2)$ = P(ACQUIRENTE_NO) =Numero casi NO/Numero casi =5/10 = 0.5
$P(A_{1\,1}|C_1)$ = P(Stato Coniugale=Sposato|SI) = 3/5 =0,6
$P(A_{1\,2}|C_1)$ = P(Stato Coniugale=Single|SI) = 1/5 =0,2
$P(A_{1\,3}|C_1)$ = P(Stato Coniugale=Separato|SI) = 1/5=0,2
$P(A_{1\,1}|C_2)$ = P(Stato Coniugale=Sposato|NO) = 1/5=0,2
$P(A_{1\,2}|C_2)$ = P(Stato Coniugale=Single|NO) = 3/5=0,6
$P(A_{1\,3}|C_1)$ = P(Stato Coniugale=Separato|NO) = 1/5=0,2
$P(A_{2\,1}|C_1)$ = P(Sesso=M|SI) = 2/5 = 0,4
$P(A_{2\,2}|C_1)$ = P(Sesso =F|SI) = 3/5=0,6
$P(A_{2\,1}|C_1)$ = P(Sesso=M|NO) = 3/5=0,6
$P(A_{2\,2}|C_1)$ = P(Sesso =F|NO) = 2/5= 0,4

Per l'attributo reddito, che è continuo, supponiamo che la distribuzione dei valori sia gaussiana e utilizziamo quindi la seguente formula:

$$P\{A_i|C_j\} = \frac{1}{\sqrt{2\pi} \times \sigma_{i\,j}} \times e^{-\frac{(A_i-\mu_{i\,j})^2}{2\sigma_{i\,j}^2}}$$

Dove:

σ è la deviazione standard del training set

μ è la media del training set

Per esempio, essendo per la classe C_1 (cioè ACQUIRENTE_SI) σ=16360 e μ= 52800 avremo che, per un reddito di 40000 (quello del possibile acquirente, il valore numero 11 nella tabella)

$P(A_{3\,1}|C_1)$ = P(Reddito annuo=40000|SI) $= \frac{1}{\sqrt{2\pi}\times 16361} \times e^{-\frac{(40000-52800)^2}{2\times 16361^2}}$ = 0,2937

$P(A_{3\,1}|C_1)$ = P(Reddito annuo=40000|NO) $= \frac{1}{\sqrt{2\pi}\times 13575} \times e^{-\frac{(40000-34600)^2}{2\times 13575^2}}$ = 0,3685

A questo punto abbiamo tutti gli elementi per calcolare l'appartenenza a ciascuna delle due classi dell'elemento numero 11 della nostra tabella, che presenta gli attributi: Stato Coniugale = Sposato, Sesso = M e Reddito annuo=40000.

Per la classe C_1 ACQUIRENTE_SI avremo che $\prod_n P\{An|C_j\} \times P\{C_j\}$ è uguale a:

P(Stato Coniugale=Sposato|SI) x P(Sesso=M|SI) x P(Reddito annuo=40000|SI) x P(ACQUIRENTE_SI)

=0,2 x 0,6 x 0,2937 x 0,5

=0,0352

Per la classe C_1 ACQUIRENTE_NO avremo che $\prod_n P\{An|C_j\} \times P\{C_j\}$ è uguale a:

P(Stato Coniugale=Sposato|NO) x P(Sesso=M|NO) x P(Reddito annuo=40000|NO) x P(ACQUIRENTE_NO)

=0,6 x 0,4 x 0,3685x 0,5

=0,0221

Avendo dunque che 0,0352 corrispondente alla classe ACQUIRENTE_SI è maggiore di 0,0221 che corrisponde alla classe ACQUIRENTE_NO, la stima della classe di appartenenza per il nostro soggetto è ACQUIRENTE_SI.

Semplice vero?

L'algoritmo possiede i seguenti punti di forza:

- Lavora bene in caso di "rumore" in una parte dati.
- Tende a non considerare gli attributi irrilevanti.
- Il training del modello è molto più semplice rispetto ad altri algoritmi.

Il rovescio della medaglia è rappresentato dall'assunzione dell'indipendenza degli attributi, che può non essere presente nella realtà. Questo limite è comunque superabile attraverso l'uso di altri algoritmi, basati ancora sul teorema di Bayes, quali le reti Bayesiane, oppure di diversa derivazione, come quelli descritti nei prossimi articoli.

Naive Bayes in R

Verifichiamo ora i calcoli utilizzando R con il package e1071 (si veda http://ugrad.stat.ubc.ca/R/library/e1071/html/predict.naiveBayes.html).

Esempio 1

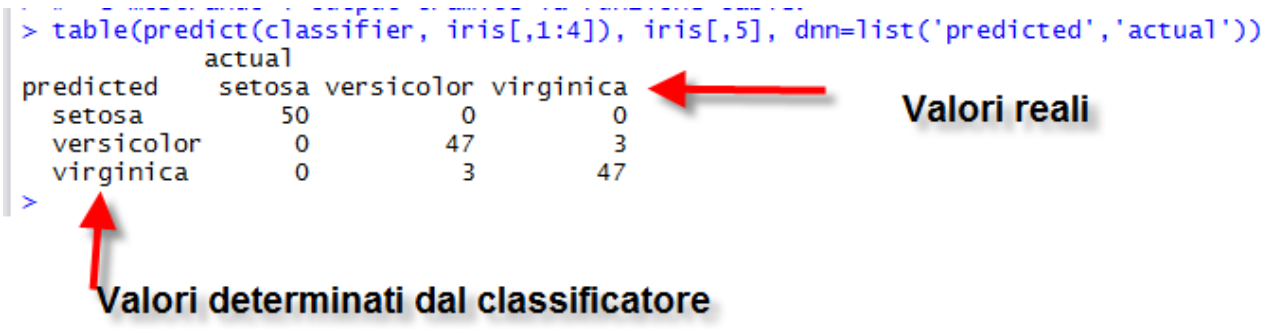
Dapprima vediamo come di utilizza il classificatore, utilizzano un semplice esempio basato sul data set Iris.

```
#installazione el package e1071, contenete l'algoritmo naive bayes
install.packages("e1071")
#caricamento della libraria
library(e1071)

#creazione del classificatore:  naiveBayes  accetta, come primo parametro, una matrice
# o data frame con le variabili numeriche o categoriche.
# Come secondo parametro occorre fornire il vettore delle classi.
# Nell'esempio qui sotto la matrice è costituita dalle prima 4 colonne del dataset IRIS, mentre le classi
# sono presenti nella quinta colonna.
classifier<-naiveBayes(iris[,1:4], iris[,5])

#utilizziamo predict per classificare gli stessi dati di input
#  e mostrando l'output tramite la funzione table.
table(predict(classifier, iris[,1:4]), iris[,5], dnn=list('predicted','actual'))
```

L'output è il seguente:



Come si può notare la maggioranza dei valori è stata correttamente classificata.

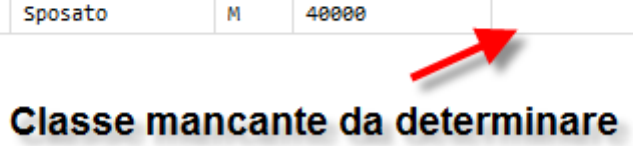
Esempio 2

Utilizziamo ora lo stesso dataset proposto nell'esempio descritto nel paragrafo precedente.

```
#carichiamo il dataset da file csv
clienti <- read.csv(file.path("C:\\R_demos\\", "naive bayes.csv"),sep=";",
stringsAsFactors=TRUE)
```

#visualizziamo il data set
clienti

	ID	Stato.Coniugale	M.F	Reddito_annuo	Acquirente
1	1	Sposato	M	35000	SI
2	2	Single	F	47000	NO
3	3	Sposato	F	58000	SI
4	4	Single	M	31000	NO
5	5	Separato	M	70000	SI
6	6	Sposato	F	27000	NO
7	7	Single	F	36000	SI
8	8	Separato	M	50000	NO
9	9	Sposato	F	65000	SI
10	10	Single	M	18000	NO
11	11	Sposato	M	40000	



#eseguiamo il classificatore
classifier<-naiveBayes(clienti[1:10,2:4], clienti[1:10,5])

#tabella classificazioni
table(predict(classifier, clienti[1:10,2:4]), clienti[1:10,5],
dnn=list('predicted','actual'))

#visualizziamo probabilità condizionali
Classifier

```
Conditional probabilities:
                Stato.Coniugale
clienti[1:10, 5] Separato Single Sposato

              NO       0.2    0.6      0.2
              SI       0.2    0.2      0.6

                M.F
clienti[1:10, 5]   F   M

              NO 0.4 0.6
              SI 0.6 0.4

                Reddito_annuo
clienti[1:10, 5]  [,1]     [,2]
                    NA       NA
              NO 34600 13575.71
              SI 52800 16361.54
```

#predizione sull'elemento senza classe (riga 11)
predict(classifier, clienti[11,2:4])

```
> predict(classifier, clienti[11,2:4])
[1] SI
Levels:  NO SI
>
```

Stessa classificazione ottenuta "a mano"