Machine Learning

Pietro Garofalo

April 25, 2022

Contents

Chapter 1

Introduzione alla teoria Bayesiana nel machine learning

Abbiamo un classico problema di classificazione, si hanno C_n classi ciascuna etichettata con un label y_n.
Supponiamo di avere un determinato numero di features $\mathbf{X} = (x_1, x_2, ..., x_n)$ che descrivono le mie classi, il problema che ci poniamo è di trovare la :

$$P(y|\mathbf{X})$$

Ossia la probabilità di avere la classe targata y date features $\mathbf{X}$ infatti data questa la scelta che faremo fra tutte le classi è quella t.c la probabilità di cui sopra è massima !
Trovare tale probabilità però è difficile, entra quindi in gioco il teorema di Bayes :

ATTENZIONE !

Teorema di Bayes :

$$P(y|\mathbf{X}) = \frac{P(\mathbf{X}|y)P(y)}{P(\mathbf{X})}$$

Tale teorema ci aiuta nel compito di determinare la probabilità che cerchiamo, facciamo un esempio per chiarire.

x_1	x_2	y
0	0	0
0	1	1
1	2	1
0	0	0
2	2	0
1	1	0
0	2	1
2	0	0
2	1	0
1	0	0

Se ti dovessi dare $\mathbf{X} = (0, 2)$ mi sapresti dire a quale classe appartiene ?
Applichiamo il th. di Bayes e vediamo, calcoliamoci tutte le quantità che ci servono :

$$
\begin{aligned}
P(y=0) &= \frac{n(y=0)}{n(y=1)+n(y=0)} = \frac{6}{10} \\
P(y=1) &= \frac{n(y=1)}{n(y=1)+n(y=0)} = \frac{4}{10} \\
P(\mathbf{X} = (0,2)|y=0) &= 0 \\
P(\mathbf{X} = (0,2)|y=1) &= \frac{1}{4}
\end{aligned}
$$

Troviamo quindi usando formula di Bayes :

$$
\begin{aligned}
P(y=0|\mathbf{X}) &= 0 \\
P(Y=1|\mathbf{X}) &= \frac{1}{10}
\end{aligned}
$$

La risposta alla domanda iniziale sarà y = 1, semplice no?
Assolutamente no ! il problema è che qui avevamo solo 10 features e 10 campioni, pensa se ne hai 10000 e passa, come fai a controllare tutti i campioni per trovare la combinazione giusta ? E se non la trovi dal tuo dataset di campioni essa risulta zero?

1.1 Metodo di Naive-Bayes

ATTENZIONE !

La assunzione fondamentale è che le features $x_1, x_2, ..., x_n$ siano indipendenti così che possiamo scrivere :

$$P(\mathbf{X} = (\mathbf{0}, \mathbf{2})|\mathbf{y} = \mathbf{0}) = P(x_1 = 0|y = 0) * P(x_2 = 2|y = 0)$$

Trasformando la probabilità in una produttoria sarà ora molto più facile trovarla, basta per esempio un semplice istogramma (come vedremo in un esempio) ed il gioco è fatto !