NAIVE BAYES CLASSIFIER

Trần Thị Nam Phương - Quản Lượng

December 2020

1 Định lý Bayes

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

2 Naive Bayes

2.1 Thuật toán Naive Bayes

Cho tập dữ liệu huấn luyện D với các dữ liệu X (còn gọi là các mẫu X). Có thể hiểu mỗi mẫu X là 1 dòng, và cũng là 1 vector với n thành phần là các giá trị của các thuộc tính (cột).
Giả sử có m lớp (Class): C_i , với $1 \leq i \leq m$

Thuật toán Navie Bayes chỉ đơn giản là tìm giá trị lớn nhất trong các $P(C_i) \prod_{k=1}^{n} P(x_k|C_i)$ tính được. Từ đó có thể dự đoán được mẫu mới thuộc vào Class C_i nào, mẫu mới sẽ thuộc vào Class C_i có $P(C_i) \prod_{k=1}^{n} P(x_k|C_i)$ là lớn nhất.

Kí hiệu:

$$argmax_{C_i} P(C_k) \prod_{k=1}^{n} P(x_k|C_i)$$

Nguồn gốc
Ta sẽ đi tìm hiểu nguồn gốc của Naive Bayes. Naive Bayes xuất phát từ Định lý Bayes. Để cho dễ hình dung ta sẽ cho m = 2. Tức có 2 lớp Class là C_1 và C_2.
Áp dụng Định lý Bayes, ta có:

$$P(C_1|X) = \frac{P(X|C_1) \times P(C_1)}{P(X)}$$

$$P(C_2|X) = \frac{P(X|C_2) \times P(C_2)}{P(X)}$$

Để biết mẫu mới X thuộc vào lớp C_1 hay C_2, chúng ta chỉ cần so sánh xem $P(C_1|X)$ và $P(C_2|X)$ cái nào lớn hơn, thì X sẽ thuộc vào lớp C_i đó.

- Nếu $P(C_1|X) > P(C_2|X)$ ==> X thuộc Class C_1
- Nếu $P(C_2|X) > P(C_1|X)$ ==> X thuộc Class C_2

Nhưng ở đây ta quan sát thấy $P(C_1|X)$ và $P(C_2|X)$ có cùng mẫu số với nhau là $P(X)$. Vì vậy dễ dàng nhận thấy rằng để so sánh $P(C_1|X)$ và $P(C_2|X)$ ta chỉ cần tính toán phần tử số của chúng:

$$P(C_i) \times P(X|C_i) = P(C_i) \prod_{k=1}^{n} P(x_k|C_i)$$

Sau cùng là so sánh chúng với nhau để tìm giá trị lớn nhất, rồi gán nhãn cho mẫu mới. Và đây chính là Naive Bayes Classifier.

Thuật toán chung

- B1: Tính $P(C_i)$, lập bảng tính tất cả các $P(x_k|C_i)$ trong tập dữ liệu huấn luyện.
- B2: Phân lớp cho mẫu mới X_{new}.
  Xem $P(C_i) \prod_{k=1}^{n} P(x_k|C_i)$ nào có giá trị lớn nhất ==> X thuộc class C_i đó.

2.1.1 Multinomial Naive Bayes

(Mẫu có thành phần là các giá trị rời rạc)

Thuật toán

- B1: Tính $P(C_i)$, lập bảng tính tất cả các $P(x_k|C_i)$ trong tập dữ liệu huấn luyện.

  $$P(C_i) = \frac{|C_{i,D}|}{|D|}$$

  Trong đó:
  $|C_{i,D}|$: Lực lượng trong tập huấn luyện D thuộc về Class C_i
  $|D|$: Lực lượng của tập huấn luyện D.

  $$P(x_k|C_i) = \frac{|C_{i,D,x_k}|}{|C_{i,D}|}$$

  Trong đó:
  $|C_{i,D,x_k}|$: Lực lượng trong tập $C_{i,D}$ mà nó nhận giá trị là x_k

- B2: Phân lớp cho mẫu mới X_{new}.
  Xem $P(C_i)\prod_{k=1}^{n} P(x_k|C_i)$ nào có giá trị lớn nhất ==> X thuộc class C_i đó.

  LƯU Ý: Do $X = \{x_1, x_2, \ldots, x_n\}$ *có các thành phần là các giá trị rời rạc. Vì vậy cách tính xác suất có điều kiện:*
  $P(X|C_i) = P(x_1, x_2, \ldots, x_n|C_i) = \prod_{k=1}^{n} P(x_k|C_i)$

Ví dụ 1 Cho tập dữ liệu huấn luyện (tập train) về việc chúng ta có nên đi chơi thể thao hay không? Dựa trên các thông tin: thời tiết ngoài trời (Outlook), nhiệt độ (Temperature) , độ ẩm (Humidity), sức gió (Windy).

Outlook	Temperature	Humidity	Windy	Play?
sunny	hot	high	weak	No
sunny	hot	high	strong	No
overcast	hot	high	weak	Yes
rain	mild	high	weak	Yes
rain	cool	normal	weak	Yes
rain	cool	normal	strong	No
overcast	cool	normal	strong	Yes
sunny	mild	high	weak	No
sunny	cool	normal	weak	Yes
rain	mild	normal	weak	Yes
sunny	mild	normal	strong	Yes
overcast	mild	high	strong	Yes
overcast	hot	normal	weak	Yes
rain	mild	high	strong	No

Với mẫu mới $X_{new} = \{Outlook = sunny, Temperature = cool, Humidity = high, Windy = strong\}$. Hãy dự đoán xem X_{new} thuộc Class nào?

Bài làm:

- **B1: Tính** $P(C_i)$ **, lập bảng tính tất cả các** $P(X_k|C_i)$ **trong tập dữ liệu huấn luyện.**
  - Ta đặt:
    C_1 = "yes", C_2 = "no"

  Ta thu được:
  $P(C_1) = P(C_{yes}) = 9/14$
  $P(C_2) = P(C_{no}) = 5/14$

  - Lập bảng tính $P(X_k|C_i)$:
  Ta nhìn vào ta sẽ thấy được thuộc tính Outlook nó sẽ có 3 giá trị chính là sunny, overcast, rain. $P(sunny|yes)$ chính là ta đi đếm số lượng class yes

có mặt giá trị sunny, sau đó đem chia cho tổng số lượng class yes. Tương tự cho các thuộc tính còn lại. Từ đây ta lập được bảng tính xác suất có điều kiện như sau:

Outlook	
$P(sunny\|yes) = 2/9$	$P(sunny\|no) = 3/5$
$P(overcast\|yes) = 4/9$	$P(overcast\|no) = 0$
$P(rain\|yes) = 3/9$	$P(rain\|no) = 2/5$
Temperature	
$P(hot\|yes) = 2/9$	$P(hot\|no) = 2/5$
$P(mild\|yes) = 4/9$	$P(mild\|no) = 2/5$
$P(cool\|yes) = 3/9$	$P(cool\|no) = 1/5$
Humidity	
$P(high\|yes) = 3/9$	$P(high\|no) = 4/5$
$P(normal\|yes) = 6/9$	$P(normal\|no) = 1/5$
Windy	
$P(strong\|yes) = 3/9$	$P(strong\|no) = 3/5$
$P(normal\|yes) = 6/9$	$P(normal\|no) = 1/5$

B2: Phân lớp cho mẫu mới X_{new} đưa vào để test.
$X_{new} = \{Outlook = sunny, Temperature = cool, Humidity = high, Windy = strong\}$

Vì X có thành phần là các giá trị rời rạc *(xem LƯU Ý ở mục 2.1.1)* nên :

$$P(X|yes) = P(sunny|yes) \times P(cool|yes) \times P(high|yes) \times P(strong|yes)$$
$$= 2/9 \times 3/9 \times 3/9 \times 3/9$$
$$= 2/243$$

$$P(X|no) = P(sunny|no) \times P(cool|no) \times P(high|no) \times P(strong|no)$$
$$= 3/5 \times 1/5 \times 4/5 \times 3/5$$
$$= 36/625$$

Suy ra, áp dụng Định lý Bayes ở mục 1 ta có:
$P(X|yes) \times P(yes) = 2/243 \times 9/14 = 0.0052$
$P(X|no) \times P(no) = 36/625 \times 5/14 = 0.0205$

$==> P(X|no) \times P(no) > P(X|yes) \times P(yes)$

$==> X_{new}$ thuộc Class "no" (đồng nghĩa với việc, ngoài trời nắng (sunny), nhiệt độ mát mẻ (cool), độ ẩm cao (high), gió mạnh (strong). Sẽ đưa ra quyết định là không đi chơi thể thao.

CHÚ Ý:

Để tránh trường hợp giá trị $P(x_k|C_i) = 0$, suy ra lúc này nó sẽ làm cho $P(X|C_i) = 0$ do không có mẫu nào trong tập dữ liệu huấn luyện thỏa mãn tử số. Khi các xác suất thành phần nhân với nhau lớn, chỉ cần tồn tại 1 thành phần có giá trị bằng 0, sẽ đẩy xác suất về 0 . Điều này có thể làm cho việc dự đoán bị sai. Ta làm trơn bằng cách thêm một số mẫu ảo. Khi đó ta làm trơn theo Laplace:

$$P(C_i) = \frac{|C_{i,D}| + 1}{D + m}$$

$$P(x_k|C_i) = \frac{|C_{i,D,x_k}| + 1}{|C_{i,D}| + r}$$

Trong đó: m là số lớp, r là số giá trị rời rạc của thuộc tính.

Ví dụ 2 Cho tập dữ liệu huấn luyện (tập train) như Ví dụ 1.
Với mẫu mới $X_{new} = \{Outlook = overcast, Temperature = cool, Humidity = high, Windy = strong\}$. Hãy dự đoán xem X_{new} thuộc Class nào?

Bài làm:

Class: C1 = yes, C2 = no

- Cách làm thông thường:

$P(C_1) \times P(X|C_1) = P(yes) \times P(overcast|yes) \times P(cool|yes) \times P(high|yes) \times P(strong|yes) = 9/14 \times 4/9 \times 3/9 \times 3/9 \times 3/9 = 0.011$

$P(C_2) \times P(X|C_2) = P(no) \times P(overcast|no) \times P(cool|no) \times P(high|no) \times P(strong|no) = 5/14 \times 0 \times 1/5 \times 4/5 \times 3/5 = 0$

Suy ra: $P(C_1) \times P(X|C_1) > P(C_2) \times P(X|C_2)$
Vì vậy X thuộc vào Class C_1 = yes.

Chúng ta để ý rằng $P(overcast|no) = 0$. Nó đã vi phạm phần CHÚ Ý ta nói ở trên, vì vậy ta sẽ làm trơn theo Laplace để tránh trường hợp này.

- Làm trơn theo Laplace:

$P(C_1) = P(yes) = (9 + 1)/(14 + 2) = 10/16$
$P(C_2) = P(no) = (5 + 1)/(14 + 2) = 6/16$

$P(overcast|yes) = 5/12$, $P(overcast|no) = 1/8$
$P(cool|yes) = 4/12$, $P(cool|no) = 2/8$
$P(high|yes) = 4/11$, $P(high|no) = 5/7$
$P(strong|yes) = 4/11$, $P(strong|no) = 4/7$

Ta có:
$P(C_1) \times P(X|C_1) = P(yes) \times P(overcast|yes) \times P(cool|yes) \times P(high|yes) \times P(strong|yes) = 10/16 \times 5/12 \times 4/12 \times 4/11 \times 4/11 = 0.0114$

$P(C_2) \times P(X|C_2) = P(no) \times P(overcast|no) \times P(cool|no) \times P(high|no) \times P(strong|no) = 6/16 \times 1/8 \times 2/8 \times 5/7 \times 4/7 = 0.0047$

Suy ra: $P(C_1) \times P(X|C_1) > P(C_2) \times P(X|C_2)$
Vì vậy X thuộc vào Class C_1 = yes.

2.1.2 Gaussian Naive Bayes

(Mẫu có thành phần là các giá trị liên tục)

Thuật toán

- B1: Tính $P(C_i)$, lập bảng tính tất cả các $P(x_k|C_i)$ trong tập dữ liệu huấn luyện.
- B2: Phân lớp cho mẫu mới X_{new}.
  Xem $P(C_i) \prod_{k=1}^{n} P(x_i|C_i)$ nào có giá trị lớn nhất ==> X thuộc class C_i đó.

  LƯU Ý: Do $X = \{x_1, x_2, \ldots, x_n\}$ *có các thành phần là các giá trị liên tục:*
  $P(x_k|C_i) = \frac{1}{\sqrt{2\Pi\sigma_{C_i}^2}} \times exp\left(\frac{-(x_k - \mu_{C_i})^2}{2\sigma_{C_i}^2}\right)$

  Trong đó:
  Các tham số dưới đây được tính bằng Maximum Likelihood.
  $$\mu_{x_k} = mean(x_k) = \sum_{j=1}^{N} \frac{x_j}{N}$$
  $$\sigma_{x_k}^2 = var(x_k) = \sum_{j=1}^{N} \frac{(x_j - \mu)^2}{N}$$

Ví dụ 3 Cho tập dữ liệu huấn luyện (tập train) có thông số về chiều dài (Length) và chiều rộng (Width) của các bông hoa. Dựa trên thông tin đó sẽ biết được bông hoa đó thuộc loại hoa nào: Hoa Cúc hoặc Hoa Ly.

NOTE: Dây là dữ liệu không có thật, do nhóm tác giả tự nghĩ ra dùng để minh họa, giúp người đọc dễ hình dung về thuật toán mà thôi.

Length	Width	Class?
1	2	Hoa Cúc
2	4	Hoa Cúc
3	6	Hoa Cúc
4	1	Hoa Ly
5	3	Hoa Ly

Với mẫu mới $X_{new} = \{Length = 3, Width = 5\}$. Hãy dự đoán xem X_{new} thuộc Class Hoa Cúc hay Hoa Ly?

Bài làm:

B0: Tính toán chung:

- Vì có 2 Class là Hoa Cúc và Hoa Ly, vì vậy chúng ta chia tập dữ liệu huấn luyện thành 2 tập dữ liệu con (chia theo Class).

Class C_1 = Hoa Cúc:

Length	Width	Class?
1	2	Hoa Cúc
2	4	Hoa Cúc
3	6	Hoa Cúc

Class C_2 = Hoa Ly:

Length	Width	Class?
4	1	Hoa Ly
5	3	Hoa Ly

- Ta đi tính $variance(x_k)$ và $mean(x_k)$ trong mỗi Class C_i.

$$\mu_{x_k} = mean(x_k) = \frac{\sum_{j=1}^{N} x_j}{N}$$
$$\sigma_{x_k}^2 = var(x_k) = \frac{\sum_{j=1}^{N} (x_j - \mu)^2}{N}$$

Class C_1 = Hoa Cúc:

Length	Width
$\mu_1 = \frac{1+2+3}{3} = 2$	$\mu_2 = \frac{2+4+6}{3} = 4$
$\sigma_1^2 = \frac{(1-2)^2+(2-2)^2+(3-2)^2}{3} = 2/3$	$\sigma_1^2 = \frac{(2-4)^2+(4-4)^2+(6-4)^2}{3} = 8/3$

Class C_2 = Hoa Ly:

Length	Width
$\mu_1 = \frac{4+5}{2} = 4.5$	$\mu_2 = \frac{1+3}{2} = 2$
$\sigma_1^2 = \frac{(4-4.5)^2+(5-4.5)^2}{2} = 0.25$	$\sigma_2^2 = \frac{(1-2)^2+(3-2)^2}{2} = 1$

B1, B2: Tính $P(C_i)$, $P(x_k|C_i)$. Dán nhãn cho mẫu mới
$P(C_1) = 3/5$
$P(C_2) = 2/5$

Với mẫu mới X_{new} thì $x_1 = 3, x_2 = 5$. Ta tính được:
Class C1 = Hoa Cúc
$P(x_1 = 3|C_1) = \frac{1}{\sqrt{2\Pi\sigma_1^2}} \times exp\left(\frac{-(x_1-\mu_1)^2}{2\sigma_1^2}\right)$ Trong đó: $\sigma_1^2 = 2/3, \mu_1 = 2$
$P(x_2 = 5|C_1) = \frac{1}{\sqrt{2\Pi\sigma_2^2}} \times exp\left(\frac{-(x_2-\mu_2)^2}{2\sigma_2^2}\right)$ Trong đó: $\sigma_2^2 = 8/3, \mu_2 = 4$

Ta được:
$P(X|C_1) = P(x_1 = 3|C_1) \times P(x_2 = 5|C_1) = 0.2307 \times 0.2025 = 0.0467$

$P(C_1) \times P(X|C_1) = 3/5 \times 0.0467 = 0.02802$

Class C2 = Hoa Ly
$P(x_1 = 3|C_2) = \frac{1}{\sqrt{2\Pi\sigma_1^2}} \times exp\left(\frac{-(x_1-\mu_1)^2}{2\sigma_1^2}\right)$ Trong đó: $\sigma_1^2 = 0.25, \mu_1 = 4.5$
$P(x_2 = 5|C_2) = \frac{1}{\sqrt{2\Pi\sigma_2^2}} \times exp\left(\frac{-(x_2-\mu_2)^2}{2\sigma_2^2}\right)$ Trong đó: $\sigma_2^2 = 1, \mu_2 = 2$

Ta được:
$P(X|C_2) = P(x_1 = 3|C_1) \times P(x_2 = 5|C_1) = 0.00886 \times 0.00443 = 0.0000392$

$P(C_2) \times P(X|C_2) = 2/5 \times 0.0000392 = 0.00001568$

Suy ra: $P(C_1) \times P(X|C_1) > P(C_2) \times P(X|C_2)$
Vì vậy X thuộc vào Class C_1 = Hoa Cúc.

2.1.3 Complement Naive Bayes

Công thức

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

$$argmax_{C_i} P(C_k) \prod_{k=1}^{n} \frac{1}{P(x_k|not\ C_i)}$$

$$P(x_k|not\ C_i) = \frac{number\ x_k\ in\ feature\ k\ without\ class\ C_i\ +\ 1}{total\ number\ feature\ k\ without\ class\ C_i + number\ feature}$$

Ví dụ 4: Cho tập dữ liệu huấn luyện của Ví dụ 1

Outlook	Temperature	Humidity	Windy	Play?
sunny	hot	high	weak	No
sunny	hot	high	strong	No
overcast	hot	high	weak	Yes
rain	mild	high	weak	Yes
rain	cool	normal	weak	Yes
rain	cool	normal	strong	No
overcast	cool	normal	strong	Yes
sunny	mild	high	weak	No
sunny	cool	normal	weak	Yes
rain	mild	normal	weak	Yes
sunny	mild	normal	strong	Yes
overcast	mild	high	strong	Yes
overcast	hot	normal	weak	Yes
rain	mild	high	strong	No

Cùng câu hỏi Ví dụ 1:
Với mẫu mới $X_{new} = \{Outlook = sunny, Temperature = cool, Humidity = high, Windy = strong\}$. Hãy dự đoán xem X_{new} thuộc Class nào?

Bài làm:

Outlook	
$P(sunny\|not\ yes)$ =3+1/5+4=4/9	$P(sunny\|not\ no)$ = 2+1/9+4=3/13
$P(overcast\|not\ yes)$ = 0+1/5+4=1/9	$P(overcast\|not\ no)$ = 4+1/9+4=5/13
$P(rain\|not\ yes)$ =2+1/5+4=3/9	$P(rain\|not\ no)$ = 3+1/9+4=4/13
Temperature	
$P(hot\|not\ yes)$ =2+1/5+4=3/9	$P(hot\|not\ no)$ = 2+1/9+4=3/13
$P(mild\|not\ yes)$ = 2+1/5+4=3/9	$P(mild\|not\ no)$ = 4+1/9+4=5/13
$P(cool\|not\ yes)$ = 1+1/5+4=2/9	$P(cool\|not\ no)$ = 3+1/9+4=4/13
Humidity	
$P(high\|not\ yes)$ = 4+1/5+4=5/9	$P(high\|not\ no)$ = 3+1/9+4=4/13
$P(normal\|not\ yes)$ = 1+1/5+4=2/9	$P(normal\|not\ no)$ = 6+1/9+4=7/13
Windy	
$P(strong\|not\ yes)$ = 3+1/5+4=4/9	$P(strong\|not\ no)$ = 3+1/9+4=4/13
$P(normal\|not\ yes)$ = 1+1/5+4=2/9	$P(normal\|not\ no)$ = 6+1/9+4=7/13

$X_{new} = \{Outlook = sunny, Temperature = cool, Humidity = high, Windy = strong\}$

Vì X có thành phần là các giá trị rời rạc *(xem LƯU Ý ở mục 2.1.1)* nên :

$$
\begin{aligned}
P(X|notyes) &= P(sunny|notyes)\times P(cool|notyes)\times P(high|notyes)\times P(strong|notyes) \\
&= 4/9 \times 2/9 \times 5/9 \times 4/9 \\
&= 160/6561 \\
P(X|notno) &= P(sunny|notno)\times P(cool|notno)\times P(high|notno)\times P(strong|notno) \\
&= 3/13 \times 4/13 \times 4/13 \times 4/13 \\
&= 192/28561
\end{aligned}
$$

$P(yes)/P(X|not\ yes) = 9/14 \times 6561/160 = 26.3611$
$P(no)/P(X|not\ no) = 5/14 \times 28561/192 = 53.1268$

$==> P(no)/P(X|not\ no) > P(yes)/P(X|not\ yes)$

$==>$ X_{new} thuộc Class "no" (đồng nghĩa với việc, ngoài trời nắng (sunny), nhiệt độ mát mẻ (cool), độ ẩm cao (high), gió mạnh (strong). Sẽ đưa ra quyết định là không đi chơi thể thao.

3 Code minh họa

Nhóm tác giả minh họa code bằng Python đầy đủ 3 ví dụ về Multinomial Naive Bayes, Gaussian Naive Bayes và Complement Naive Bayes (Cả file .py lẫn file .ipynp cho mỗi ví dụ).