PHÂN LỚP BAYES

Phân lớp theo mô hình xác xuất

- Xác định xác suất hay dự đoán xác suất là thành viên của lớp.
- Nền tảng: dựa trên định lý Bayes
  - Cho X, Y là các biến bất kỳ (rời rạc, số, cấu trúc)
  - Dự đoán Y từ X
- Lượng giá các tham số của P(X | Y), P(Y) trực tiếp từ dữ liệu huấn luyện.
- Sử dụng định lý Bayes để tính P(Y | X= x)

Định lý Bayes

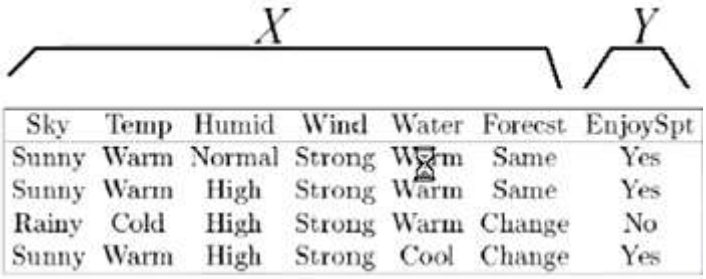
$$P(y \mid x) = \frac{P(x \mid y).P(y)}{P(x)}$$

Cụ thể:

$$P(Y = y_i \mid X = x_j) = \frac{P(X = x_j \mid Y = y_i).P(Y = y_i)}{P(X = x_j)}$$

Phân loại Bayes trên dữ liệu rời rạc

Ví dụ với tập dữ liệu huấn luyện như hình thì



Sky	Temp	Humid	Wind	Water	Forecst	EnjoySpt
Sunny	Warm	Normal	Strong	Warm	Same	Yes
Sunny	Warm	High	Strong	Warm	Same	Yes
Rainy	Cold	High	Strong	Warm	Change	No
Sunny	Warm	High	Strong	Cool	Change	Yes

X: là các giá trị của các thuộc tính

Y: là các thuộc tính phân lớp / các thuộc tính quyết định / các giá trị sẽ tạo ra các lớp

Các bước để phân lớp Bayes:

- Bước 1. Xây dựng mô hình: dựa trên tập dữ liệu huấn luyện đã cho và lượng giá các xác xuất
  P(Y), P(X | Y)
- Bước 2. Phân lớp: dùng định lý Bayes để tính P(Y | X^{new}) , X^{new} được gán vào cho lớp cho giá trị công thức lớn nhất

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

Ví dụ: cho tập dữ liệu huấn luyện

Day	Outloook	Temp	Humidity	Wind	PlayTennis
D1	Sunny	Hot	High	Weak	No
D2	Sunny	Hot	High	Strong	No
D3	Overcast	Hot	High	Weak	Yes
D4	Rain	Mild	High	Weak	Yes
D5	Rain	Cool	Normal	Weak	Yes
D6	Rain	Cool	Normal	Strong	No
D7	Overcast	Cool	Normal	Strong	Yes
D8	Sunny	Mild	High	Weak	No
D9	Sunny	Cool	Normal	Weak	Yes
D10	Rain	Mild	Normal	Weak	Yes
D11	Sunny	Mild	Normal	Strong	Yes
D12	Overcast	Mild	High	Strong	Yes
D13	Overcast	Hot	Normal	Weak	Yes
D14	Rain	Mild	High	Strong	No

Bước 1: xây dựng mô hình

Ước lượng $P(C_i)$ với C_1 = "yes", C_2 = "no" và $P(x_k \mid C_i)$

Ta thu được $P(C_i)$:

$P(C_1) = 9/14 = 0.643$

$P(C_2) = 5/14 = 0.357$

Và

Outlook	
P(sunny | y) = 2/9	**P(sunny | n) = 3/5**
P(overcast | y) = 4/9	**P(overcast | n) = 0**
P(rain | y) = 3/9	**P(rain | n) =2/5**
Temperature	
P(hot | y) = 2/9	**P(hot | n) = 2/5**
P(mild | y) = 3/9	**P(mid | n) = 2/5**
P(cool | y) = 3/9	**P(cool | n) = 1/5**
Humidity	
P(high | y) = 3/9	**P(high | n) = 4/5**
P(normal | y) = 6/9	**P(normal | n) = 1/5**
Windy	
P(strong | y) = 3/9	**P(strong | n) = 3/5**
P(weak | y) = 6/9	**P(weak | n) =2/5**

Bước 2 : Phân lớp

X^{new} = <Outlook = sunny, Temp = cool, Humidity = high, Winndy = strong>

Ta cần tính:

$$P(C_1) * P(X^{new} \mid C_1) = P(C_1)*P(sunny \mid y)*P(cool \mid y)*P(high \mid y)*P(strong \mid y)$$
$$= 9/14*2/9*3/9*3/9*3/9 = 0.005$$

$P(C_2) * P(X^{new} \mid C_2) = P(C_2)*P(sunny \mid n)*P(cool \mid n)*P(high \mid n)*P(strong \mid n)$
$= 5/14*3/5*1/5*4/5*3/5 = 0.021$

$\rightarrow X^{new}$ *thuộc lớp* C_2 *("no")*

Bài tập:

1. Hãy xác định lớp cho mẫu mới sau:

X^{new} = <Outlook = overcast, Temp = cool, Humidity = high, Windy = strong>

Làm trơn Laplace

Để tránh trường hợp giá trị $P(X_k \mid C_i) = 0$ do không có mẫu nào trong DL huấn luyện thỏa mãn tử số, ta làm trơn bằng cách thêm một số mẫu ảo.

Để tránh trường hợp xác suất bằng không như bài tập 1 vừa cho, cho dù thuộc tính high chiếm tới 4/5 thì lớp "no" vẫn luôn bằng 0

Công thức Laplace:

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

Với m – số lớp và r là số giá trị rời rạc của thuộc tính

Ví dụ: với bài tập 1 trên

Ước lượng $P(C_i)$ với C_1 = "yes", C_2 = "no" và $P(x_k \mid C_i)$

Theo công thức làm trơn Laplace :

$P(C_1) = (9+1)/(14+2) = 10/16$

$P(C_2) = (5+1)/(14+2) = 6/16$

Outlook	
P(sunny | y) = 3/12	**P(sunny | n) = 4/8**
P(overcast | y) = 5/12	**P(overcast | n) = 1/8**
P(rain | y) = 4/12	**P(rain | n) =3/8**
Temperature	
P(hot | y) = 4/12	**P(hot | n) = 3/8**
P(mild | y) = 5/12	**P(mid | n) = 3/8**

P(cool | y) = 4/12	**P(cool | n) = 2/8**
Humidity	
P(high | y) = 4/11	**P(high | n) = 5/7**
P(normal | y) = 7/11	**P(normal | n) = 2/7**
Windy	
P(strong | y) = 4/11	**P(strong | n) = 4/7**
P(weak | y) = 7/11	**P(weak | n) = 3/7**

Bước 2 : Phân lớp

X^{new} = <Outlook = sunny, Temp = cool, Humidity = high, Winndy = strong>

Ta tính theo công thức làm trơn Laplace:

$$P(C_1) * P(X^{new} \mid C_1) = P(C_1)*P(sunny \mid y)*P(cool \mid y)*P(high \mid y)*P(strong \mid y) = 0.011$$

$$P(C_2) * P(X^{new} \mid C_2) = P(C_2)*P(sunny \mid n)*P(cool \mid n)*P(high \mid n)*P(strong \mid n) = 5/14*3/5*1/5*4/5*3/5 = 0.005$$

$\rightarrow X^{new}$ *thuộc lớp* C_1 *("yes")*

Bài tập

2. Làm lại bài 1 có áp dụng làm trơn Laplace
3. Một sinh viên trẻ với thu nhập trung bình và mức đánh giá tín dụng bình thường sẽ mua một cái máy tính?

Rec. ID	Age	Income	Student	Credit_Rating	Buy_Computer
1	Young	High	No	Fair	No
2	Young	High	No	Excellent	No
3	Medium	High	No	Fair	Yes
4	Old	Medium	No	Fair	Yes
5	Old	Low	Yes	Fair	Yes
6	Old	Low	Yes	Excellent	No
7	Medium	Low	Yes	Excellent	Yes
8	Young	Medium	No	Fair	No
9	Young	Low	Yes	Fair	Yes
10	Old	Medium	Yes	Fair	Yes
11	Young	Medium	Yes	Excellent	Yes
12	Medium	Medium	No	Excellent	Yes
13	Medium	High	Yes	Fair	Yes
14	Old	Medium	No	Excellent	No

X^{new} = <Age = Young, Income = Medium, Student = Yes, Credit_Rating = Fair>

Y^{new} = <Age=Old, Income = High, Income = High, Student = No, Credit_Rating = Excellent>

4. Cho CSDL huấn luyện

Chills	Runny nose	Headache	Fever	Flu
Y	N	Mild	Y	N
Y	N	No	N	Y
Y	N	Strong	Y	Y
N	Y	Mild	Y	Y
N	N	No	N	N
N	Y	Strong	Y	Y
N	Y	Strong	N	N
Y	Y	Mild	Y	Y

Chills	Runny nose	Headache	Ferver	Flu
Y	N	Mild	Y	?

5. Áp dụng laplace cho bài 3 và 4