Naïve Bayes

Chương II. Phân nhóm dựa trên lý thuyết quyết định Bayes

Ứng dụng Bayes Theorem trong phân lớp (Using Bayes Theorem in Classification)

1. Giới thiệu Bayes Theorem

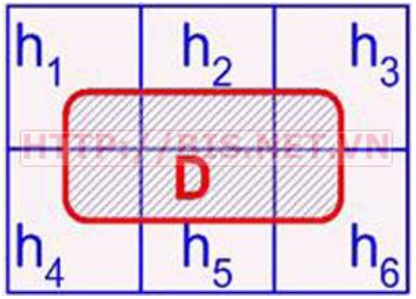
Trong lĩnh vực Data Mining, Bayes Theorem (hay Bayes' Rule) là kỹ thuật phân lớp dựa vào việc tính xác suất có điều kiện. Bayes' Rule được ứng dụng rất rộng rãi bởi tính dễ hiểu và dễ triển khai.

- **Bayes' Rule (CT1):**

$$P(h|D) = P(h) \cdot \frac{P(D|h)}{P(D)}$$

- *Trong đó:*
  - D : Data
  - h : Hypothesis (giả thuyết)
  - P(h) : Xác suất giả thuyết h (tri thức có được về giả thuyết h trước khi có dữ liệu D) và gọi là ***prior*** probability của giả thuyết h.
  - P(D| h): Xác suất có điều kiện D khi biết giả thuyết h (gọi là ***likelihood*** probability).
  - P(D): xác suất của dữ liệu quan sát D không quan tâm đến bất kỳ giả thuyết h nào.(gọi là **prior** probability của dữ liệu D**)**
  - **Tỷ số** $\frac{P(D|h)}{P(D)}$**: Chỉ số liên quan (*irrelevance index) dùng để đo lường sự liên quan giữa 2 biến A và B. Nếu irrelevance index =1*, có nghĩa A và B không liên quan nhau.**
  - **P(h|D) :Xác suất có điều kiện h khi biết D (gọi là *posterior* probability của giả thuyết h)**

1. Giới thiệu Bayes Theorem



- Trong rất nhiều ứng dụng, các giả thuyết hi có thể loại trừ nhau và vì dữ liệu quan sát D là tập con của tập giả thuyết cho nên chúng ta có thể phân rã P(D) như sau (CT2):

$$P(D) = P(D \cap h_1) \cup P(D \cap h_2) \cup \ldots \cup P(D \cap h_k) = \bigcup_j P(D \cap h_j)$$

- Vì $P(D \cap h_j) = P(D \mid h_j) \cdot P(h_j)$ nên (CT1) có thể viết lại như sau (CT3)

$$P(D) = \sum_j P(D \mid h_j) \cdot P(h_j)$$

- Thay P(D) trong (CT2) vào (CT1) ta được (CT4)

$$P(h_i \mid D) = \frac{P(D \mid h_i) \cdot P(h_i)}{\sum_j P(D \mid h_j) \cdot P(h_j)}$$

1. Giới thiệu Bayes Theorem

- **(CT4) gọi là Bayes's Theorem**

$$P(h_i|D) = \frac{P(D|h_i) \cdot P(h_i)}{\sum_j P(D|h_j) \cdot P(h_j)}$$

Ví dụ sau đây mô tả cách tính Bayes's Theorem

- Giả sử ta có dữ liệu quan sát về 250 đối tượng để tìm hiểu mối quan hệ giữa 2 biến thu nhập (income: Low(D1), Medium(D2), High(D3)) và loại xe hơi (Car: Second hand (h1), New (h2)) mà họ đã mua.

Count

Car	Income			Sum
	Low (D1)	Medium (D2)	High (D3)	
Second hand (h1)	25	30	40	95
New (h2)	15	65	75	155
Sum	40	95	115	250

Ví dụ sau đây mô tả cách tính Bayes's Theorem

- Bây giờ giả sử rằng ta chỉ biết phần trăm theo dòng (*Percentage by Row*) và phần trăm theo các biên (*Marginal Percentage hay Percentage by Total*) như sau.

Percentage by Row				
		Income		
Car	Low (D_1)	Medium (D_2)	High (D_3)	Sum
Second hand (h_1)	26%	32%	42%	100%
New (h_2)	10%	42%	48%	100%

		Income		
Car	Low (D1)	Medium (D2)	High (D3)	Sum
Second hand (h1)				38%
New (h2)				62%
Sum	16%	38%	46%	100%

Câu hỏi đặt ra là có thể tính phần trăm theo cột (percentage by column) chỉ dựa vào thông tin từ 2 bảng trên hay không?.

Ví dụ sau đây mô tả cách tính Bayes's Theorem

- ***Bayes Theorem có thể giúp trả lời câu hỏi này như sau:***
  - Trước tiên, ta biểu diễn 2 bảng trên theo ký hiệu trong Bayes' Rule như sau:
  - *Với bảng phần trăm theo dòng (Percentage by Row)*

- *Với bảng phần trăm theo Total (Percentage by Total)*

Percentage by Row				
		Income		
Car	Low (D_1)	Medium (D_2)	High (D_3)	Sum
Second hand (h_1)	$P(D_1 \mid h_1)$	$P(D_2 \mid h_1)$	$P(D_3 \mid h_1)$	100 %
New (h_2)	$P(D_1 \mid h_2)$	$P(D_2 \mid h_2)$	$P(D_3 \mid h_2)$	100 %

Percentage by Total				
		Income		
Car	Low (D1)	Medium (D2)	High (D3)	Sum
Second hand (h1)				$P(h_1)$
New (h2)				$P(h_2)$
Sum	$P(D_1)$	$P(D_2)$	$P(D_3)$	100 %

Ví dụ sau đây mô tả cách tính Bayes's Theorem

- *Bảng phần trăm theo cột (Percentage by Column) được biểu diễn như sau:*

Percentage by Column			
		Income	
Car	Low (D_1)	Medium (D_2)	High (D_3)
Second hand (h_1)	$P(h_1 \mid D_1)$	$P(h_1 \mid D_2)$	$P(h_1 \mid D_3)$
New (h_2)	$P(h_2 \mid D_1)$	$P(h_2 \mid D_2)$	$P(h_2 \mid D_3)$
Sum	100 %	100 %	100 %

- *Sử dụng Bayes' Rule* $P(h|D) = P(h) \cdot \frac{P(D|h)}{P(D)}$ *chúng ta có thể dễ dàng tính các phần trăm theo cột. Chẳng hạn*

$$P(h_1|D_3) = \frac{P(h_1)}{P(D_3)} P(D_3|h_1) = \frac{0.38}{0.46} \cdot 0.42 = 0.35$$

$$P(h_2|D_1) = \frac{P(h_2)}{P(D_1)} P(D_1|h_2) = \frac{0.62}{0.16} \cdot 0.10 = 0.38$$

Ví dụ sau đây mô tả cách tính Bayes's Theorem

- Tương tự như trên, ta tính được tất cả các giá trị trong bảng phần trăm theo cột như sau:

Percentage by Column			
		Income	
Car	Low (D_1)	Medium (D_2)	High (D_3)
Second hand (h_1)	63%	32%	35%
New (h_2)	38%	68%	65%
Sum	100 %	100 %	100 %

2. Ứng dụng Bayes Theorem trong phân lớp dữ liệu (Naïve Bayes Classifier)

- Các ví dụ sau đây minh họa việc sử dụng Bayes Theorem trong việc phân lớp dữ liệu. Bộ phân lớp dữ liệu dựa trên Bayes theorem còn gọi là **Naïve Bayes Classifier**.
- **Ví dụ 1:** Có training data về thời tiết như sau

	Outlook	Temp	Humidity	Windy	Play
1	Sunny	Hot	High	FALSE	No
2	Sunny	Hot	High	TRUE	No
3	Overcast	Hot	High	FALSE	Yes
4	Rainy	Mild	High	FALSE	Yes
5	Rainy	Cool	Normal	FALSE	Yes
6	Rainy	Cool	Normal	TRUE	No
7	Overcast	Cool	Normal	TRUE	Yes
8	Sunny	Mild	High	FALSE	No
9	Sunny	Cool	Normal	FALSE	Yes
10	Rainy	Mild	Normal	FALSE	Yes
11	Sunny	Mild	Normal	TRUE	Yes
12	Overcast	Mild	High	TRUE	Yes
13	Overcast	Hot	Normal	FALSE	Yes
14	Rainy	Mild	High	TRUE	No

2. Ứng dụng Bayes Theorem trong phân lớp dữ liệu (Naïve Bayes Classifier)

- Sử dụng **Naïve Bayes Classifier** để xác định khả năng đến chơi thể thao (Play = “yes” hay “no”) với thời tiết của ngày quan sát được như sau:

Outlook	Temp	Humidity	Windy	Play
Sunny	Cool	High	True	?

- Từ Training data ta có dữ liệu như sau:

Outlook			Temp			Humidity			Windy			Play	
	Yes	*No*		*Yes*	*No*		*Yes*	*No*		*Yes*	*No*	*Yes*	*No*
Sunny	2	3	Hot	2	2	High	3	4	False	6	2	9	5
Overcast	4	0	Mild	4	2	Normal	6	1	True	3	3		
Rainy	3	2	Cool	3	1								
Sunny	2/9	3/5	Hot	2/9	2/5	High	3/9	4/5	False	6/9	2/5	9/14	5/14
Overcast	4/9	0/5	Mild	4/9	2/5	Normal	6/9	1/5	True	3/9	3/5		
Rainy	3/9	2/5	Cool	3/9	1/5								

2. Ứng dụng Bayes Theorem trong phân lớp dữ liệu (Naïve Bayes Classifier)

- Vì thuộc tính phân lớp Play chỉ có 2 giá trị là “yes” (nghĩa là có đến chơi thể thao) và “no”(không đến chơi thể thao) nên ta phải tính Pr(yes|E) và Pr(no|E) như sau. Trong đó E là dữ liệu cần phân lớp (dự đoán)

Outlook	Temp	Humidity	Windy	Play
Sunny	Cool	High	True	?

← ***Evidence E***

$$\Pr[yes \mid E] = \Pr[Outlook = Sunny \mid yes] \times \Pr[Temperature = Cool \mid yes] \times \Pr[Humidity = High \mid yes] \times \Pr[Windy = True \mid yes] \times \frac{\Pr[yes]}{\Pr[E]}$$

Probability of class “yes”

$$= \frac{\frac{2}{9} \times \frac{3}{9} \times \frac{3}{9} \times \frac{3}{9} \times \frac{9}{14}}{\Pr[E]}$$

2. Ứng dụng Bayes Theorem trong phân lớp dữ liệu (Naïve Bayes Classifier)

Outlook	Temp	Humidity	Windy	Play
Sunny	Cool	High	True	?

← *Evidence E*

$$\Pr[No \mid E] = \Pr[Outlook = Sunny \mid No] \times \Pr[Temperature = Cool \mid No] \times \Pr[Humidity = High \mid No] \times \Pr[Windy = True \mid No] \times \frac{\Pr[No]}{\Pr[E]}$$

$$= \frac{\frac{3}{5} \times \frac{1}{5} \times \frac{4}{5} \times \frac{3}{5} \times \frac{5}{14}}{\Pr[E]}$$

Probability of class "No"

Likelihood of the two classes

For "yes" = 2/9 × 3/9 × 3/9 × 3/9 × 9/14 = 0.0053

For "no" = 3/5 × 1/5 × 4/5 × 3/5 × 5/14 = 0.0206

Conversion into a probability by normalization:

P("yes") = 0.0053 / (0.0053 + 0.0206) = 0.205

P("no") = 0.0206 / (0.0053 + 0.0206) = 0.795

2. Ứng dụng Bayes Theorem trong phân lớp dữ liệu (Naïve Bayes Classifier)

- Vì P(“no”) > P(“yes”) nên kết quả dự đoán Play =“no”