

Giá trị thuộc tính nhiều

- học : bỏ qua dữ liệu nhiễu
- phân lớp : bỏ qua các thuộc tính nhiễu
- ví dụ :

Outlook	Temp.	Humidity	Windy	Play
?	Cool	High	True	?

$$\text{Likelihood(yes)} = 3/9 \times 3/9 \times 3/9 \times 9/14 = 0.0238$$

$$\text{Likelihood(no)} = 1/5 \times 4/5 \times 3/5 \times 5/14 = 0.0343$$

$$\text{Likelihood(yes)} = 0.0238 / (0.0238 + 0.0343) = 0.41$$

$$\text{Likelihood(no)} = 0.0343 / (0.0238 + 0.0343) = 0.59$$

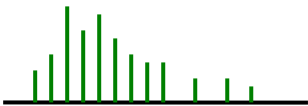
Play tennis dataset

Outlook	Temp	Humidity	Windy	Play
Sunny	Hot	High	False	No
Sunny	Hot	High	True	No
Overcast	Hot	High	False	Yes
Rainy	Mild	High	False	Yes
Rainy	Cool	Normal	False	Yes
Rainy	Cool	Normal	True	No
Overcast	Cool	Normal	True	Yes
Sunny	Mild	High	False	No
Sunny	Cool	Normal	False	Yes
Rainy	Mild	Normal	False	Yes
Sunny	Mild	Normal	True	Yes
Overcast	Mild	High	True	Yes
Overcast	Hot	Normal	False	Yes
Rainy	Mild	High	True	No

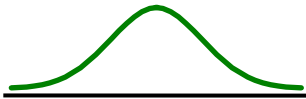
Play tennis dataset

OUTLOOK	TEMPERATURE	HUMIDITY	WINDY	PLAY
sunny	85	85	FALSE	Don't Play
sunny	80	90	TRUE	Don't Play
overcast	83	78	FALSE	Play
rain	70	96	FALSE	Play
rain	68	80	FALSE	Play
rain	65	70	TRUE	Don't Play
overcast	64	65	TRUE	Play
sunny	72	95	FALSE	Don't Play
sunny	69	70	FALSE	Play
rain	75	80	FALSE	Play
sunny	75	70	TRUE	Play
overcast	72	90	TRUE	Play
overcast	81	75	FALSE	Play
rain	71	80	TRUE	Don't Play

The numeric weather data with summary statistics											
outlook			temperature		humidity		windy		play		
	yes	no	yes	no	yes	no	yes	no	yes	no	
sunny	2	3	83	85	86	85	false	6	2	9	5
overcast	4	0	70	80	96	90	true	3	3		
rainy	3	2	68	65	80	70					
			64	72	65	95					
			69	71	70	91					
			75		80						
			75		70						
			72		90						
			81		75						



Biến ngẫu nhiên rời rạc



Biến ngẫu nhiên liên tục

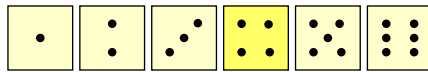
Biến ngẫu nhiên rời rạc

Có miền giá trị là tập hữu hạn hoặc vô hạn đếm được

Ví dụ

Tung một con xúc sắc 2 lần

Đặt X là số lần mặt 4 điểm xuất hiện. X có thể nhận các giá trị 0, 1, hoặc 2.



Tung đồng xu 5 lần

Đặt Y là số lần xuất hiện mặt hình.

Thì $Y = 0, 1, 2, 3, 4, \text{ hoặc } 5$



Biến ngẫu nhiên liên tục

Có miền giá trị là R hoặc một tập con của R .

Ví dụ

- Chiều cao, cân nặng.
- Thời gian để hoàn thành 1 công việc.

Biến ngẫu nhiên liên tục

Số trung vị: Là giá trị của BNN chia phân phối xác suất thành 2 phần có xác suất bằng nhau.

$$P(X \leq \text{med}(X)) = P(X \geq \text{med}(X)) = \frac{1}{2}$$

Số mode: Là giá trị của BNN có xác suất lớn nhất.

Ví dụ: Tung 2 đồng xu, với $X = \text{Số lần xuất hiện mặt hình}$.

☒ Bảng phân phối xác suất

X	0	1	2
P	0.25	0.5	0.25

$\text{Mod}(X) = 1$ Vì $P(X = 1) = 0.5$

Biến ngẫu nhiên liên tục

Phương sai: Biểu thị độ phân tán của các giá trị của biến ngẫu nhiên xung quanh giá trị trung bình của nó. Nếu phương sai bé thì các giá trị của X tập trung gần trung bình.

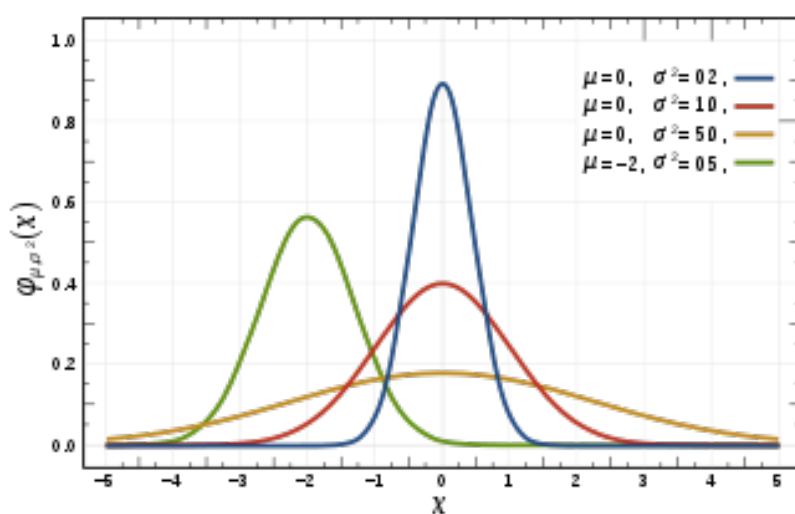
Phương sai thường được ký hiệu là σ^2

Độ lệch chuẩn: Là căn bậc hai của phương sai.

$$\sigma = \sqrt{\sigma^2} = \sqrt{\text{Var}X}$$

Phân phối chuẩn, còn gọi là **phân phối Gauss**, là một phân phối xác suất cực kì quan trọng trong nhiều lĩnh vực. Nó là họ phân phối có dạng tổng quát giống nhau, chỉ khác tham số vị trí (giá trị trung bình μ) và tỉ lệ (phương sai σ^2).


Phân phối chuẩn tắc (*standard normal distribution*) là phân phối chuẩn với giá trị trung bình bằng 0 và phương sai bằng 1 (đường cong màu đỏ trong hình bên phải). Phân phối chuẩn còn được gọi là **đường cong chuông** (*bell curve*) vì đồ thị của mật độ xác suất có dạng chuông.



- *mean* μ

- *standard deviation* σ

- hàm mật độ xác suất $f(x)$



46

[illegible]

The numeric weather data with summary statistics											
outlook			temperature		humidity		windy		play		
	yes	no	yes	no	yes	no	yes	no	yes	no	
sunny	2	3	83	85	86	85	false	6	2	9	5
overcast	4	0	70	80	96	90	true	3	3		
rainy	3	2	68	65	80	70					
			64	72	65	95					
			69	71	70	91					
			75		80						
			75		70						
			72		90						
			81		75						
Outlook			Temp.		Humidity		Windy		Play		
Sunny			66		90		true		?		

- Giới thiệu về Bayesian classification
- Giải thuật học của naive Bayes
- kết luận và hướng phát triển

Dữ liệu liên tục

- *mean* μ $\mu = \frac{1}{n} \sum_{i=1}^n x_i$

$$\mu = (83 + 70 + 68 + 64 + 69 + 75 + 75 + 72 + 81) / 9 = 73$$

- *standard deviation* σ $\sigma^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \mu)^2$

Phương sai:

$$\sigma^2 = 1/8 * [(83-73)^2 + (70-73)^2 + (68-73)^2 + (64-73)^2 + (69-73)^2 + (75-73)^2 + (75-73)^2 + (72-73)^2 + (81-73)^2] = 38.44$$

- hàm mật độ xác suất $f(x)$

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

- Giới thiệu về Bayesian classification
- Giải thuật học của naive Bayes
- kết luận và hướng phát triển

Dữ liệu liên tục

Outlook			Temperature			Humidity			Windy			Play	
	yes	no		yes	no		yes	no		yes	no	yes	no
sunny	2	3		83	85		86	85	false	6	2	9	5
overcast	4	0		70	80		96	90	true	3	3		
rainy	3	2		68	65		80	70					
				64	72		65	95					
				69	71		70	91					
				75			80						
				75			70						
				72			90						
				81			75						
sunny	2/9	3/5	mean	73	74.6	mean	79.1	86.2	false	6/9	2/5	9/14	5/14
overcast	4/9	0/5	std. dev.	6.2	7.9	std. dev.	10.2	9.7	true	3/9	3/5		
rainy	3/9	2/5											

■ ví dụ : $f(\text{temperature} = 66 | \text{yes}) = \frac{1}{\sqrt{2\pi}6.2} e^{-\frac{(66-73)^2}{2*6.2^2}} = 0.0340$

50

- Giới thiệu về Bayesian classification
- Giải thuật học của naive Bayes
- kết luận và hướng phát triển

Dữ liệu liên tục

Outlook			Temperature			Humidity			Windy			Play	
	yes	no		yes	no		yes	no		yes	no	yes	no
sunny	2	3		83	85		86	85	false	6	2	9	5
overcast	4	0		70	80		96	90	true	3	3		
rainy	3	2		68	65		80	70					
				64	72		65	95					
				69	71		70	91					
				75			80						
				75			70						
				72			90						
				81			75						
sunny	2/9	3/5	mean	73	74.6	mean	79.1	86.2	false	6/9	2/5	9/14	5/14
overcast	4/9	0/5	std. dev.	6.2	7.9	std. dev.	10.2	9.7	true	3/9	3/5		
rainy	3/9	2/5											

■ ví dụ :

Outlook	Temp.	Humidity	Windy	Play
Sunny	66	90	true	?

51

- Giới thiệu về Bayesian classification
- Giải thuật học của naive Bayes
- kết luận và hướng phát triển

Dữ liệu liên tục

■ phân lớp

Outlook	Temp.	Humidity	Windy	Play
Sunny	66	90	true	?

52

- Giới thiệu về Bayesian classification
- Giải thuật học của naive Bayes
- kết luận và hướng phát triển

Dữ liệu liên tục

■ phân lớp

Outlook	Temp.	Humidity	Windy	Play
Sunny	66	90	true	?

$$f(\text{temperature} = 66 | \text{Yes}) = \frac{1}{\sqrt{2\pi}(6.2)} e^{-\frac{(66-73)^2}{2(6.2)^2}} = 0.0340$$

$$\text{Likelihood}(\text{yes}) = 2/9 \times 0.0340 \times 0.0221 \times 3/9 \times 9/14 = 0.000036$$

$$\text{Likelihood}(\text{no}) = 3/5 \times 0.0291 \times 0.0380 \times 3/5 \times 5/14 = 0.000136$$

$$\text{Likelihood}(\text{yes}) = 0.000036 / (0.000036 + 0.000136) = 20.9\%$$

$$\text{Likelihood}(\text{no}) = 0.000136 / (0.000036 + 0.000136) = 79.1\%$$

53

Bernoulli Naive Bayes

Mô hình này được áp dụng cho các loại dữ liệu mà mỗi thành phần là một giá trị binary - bằng **0 hoặc 1**. Ví dụ: cũng với loại văn bản nhưng thay vì đếm tổng số lần xuất hiện của 1 từ trong văn bản, ta chỉ cần quan tâm từ đó có xuất hiện hay không

Khi đó, $p(x_i|c)$ được tính bằng:

$$p(x_i|c) = p(i|c)^{x_i} (1 - p(i|c))^{1-x_i}$$

$p(i|c)$ có thể được hiểu là xác suất từ thứ i xuất hiện trong các văn bản của class c .

Multinomial Naive Bayes

Mô hình này chủ yếu được sử dụng trong phân loại văn bản mà feature vectors được tính bằng [Bags of Words](#).

Mỗi văn bản được biểu diễn bởi một vector có độ dài d chính là số từ trong từ điển.

Giá trị của thành phần thứ i trong mỗi vector chính là số lần từ thứ i xuất hiện trong văn bản đó

Nội dung

- Giới thiệu về Bayesian classification
- Giải thuật học của naive Bayes
- Kết luận và hướng phát triển

56

Kết luận

- Giới thiệu về Bayesian classification
- Giải thuật học của naive Bayes
- kết luận và hướng phát triển

- naïve Bayes
 - cho kết quả tốt trong thực tế mặc dù chịu những giả thiết về tính độc lập thống kê của các thuộc tính
 - phân lớp không yêu cầu phải ước lượng một cách chính xác xác suất
 - dễ cài đặt, học nhanh, kết quả dễ hiểu
 - sử dụng trong phân loại text, spam, etc
 - tuy nhiên khi dữ liệu có nhiều thuộc tính dư thừa thì naïve Bayes không còn hiệu quả
 - dữ liệu liên tục có thể không tuân theo phân phối chuẩn (=> kernel density estimators)

57

Hướng phát triển

■ naïve Bayes

- chọn thuộc tính con từ các thuộc tính ban đầu
- chỉ sử dụng các thuộc tính con để học phân lớp
- mạng Bayes : mối liên quan giữa các thuộc tính
- tìm kiếm thông tin (ranking)



Cám ơn !