



Phương pháp học Bayes Bayesian classification

Đỗ Thanh Nghị - dtnghi@cit.ctu.edu.vn
Trần Nguyễn Minh Thư - tnmthu@cit.ctu.edu.vn

Cần Thơ - 2015

Nội dung

- Giới thiệu về Bayesian classification
- Kiến thức về xác suất thống kê
- Giải thuật học của naive Bayes
- Kết luận và hướng phát triển

Bayesian classification

■ Phương pháp học Bayes – bayesian classification

- Phân loại này được đặt theo tên của **Thomas Bayes** (1702-1761), người đề xuất các định lý Bayes
- Giải thuật học có giám sát (supervised learning) - xây dựng mô hình phân loại dựa trên dữ liệu tập học đã có nhãn (lớp)
- Mạng Bayes (Bayesian network), **Bayes ngây thơ (naive Bayes)**
- Giải quyết các vấn đề về phân loại, gom nhóm, etc.

3

Bayesian classification

■ Phương pháp học Bayes ứng dụng thành công

● Phân loại thư rác

Cho một email, dự đoán xem đó là thư rác hay không

● Chẩn đoán y tế

Cho một danh sách các triệu chứng, dự đoán xem bệnh nhân có bệnh X hay không

● Thời tiết

Dựa vào nhiệt độ, độ ẩm, vv ... dự đoán nếu nó sẽ mưa vào ngày mai

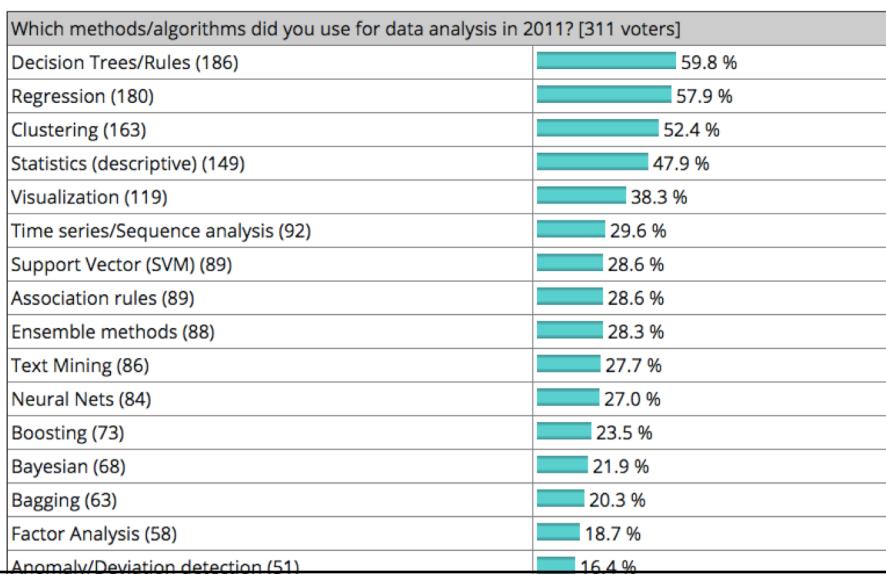
4

Bayesian classification

- ❑ Phương pháp Bayesian là hệ thống **ham học**
- ❑ Dựa vào **các đặc trưng** đưa ra kết luận **nhãn** của đối tượng mới đến
- ❑ Khi đưa ra một tập huấn luyện, hệ thống **ngay lập tức** phân tích dữ liệu và **xây dựng một mô hình**. Khi cần phân loại một đối tượng mới đến, hệ thống sử dụng mô hình đã xây dựng để xác định đối tượng mới.
- ❑ Phương pháp Bayesian (ham học) có xu hướng phân loại các trường hợp nhanh hơn KNN (lười học)

Kỹ thuật DM thành công (2011)

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



Nội dung

- Giới thiệu về Bayesian classification
- Kiến thức về xác suất thống kê
- Giải thuật học của naive Bayes
- Kết luận và hướng phát triển

7

Xác suất thống kê



Một vài ví dụ

- Khi tung 1 đồng xu, khả năng nhận mặt ngửa là bao nhiêu?
- Khi tung một hộp xúc xắc, khả năng xuất hiện mặt “6 nút” là bao nhiêu?

P (h) : ký hiệu xác suất của giả thuyết h

Xác suất thống kê



Xác suất xuất hiện mặt ngửa:

$$P(\text{ngửa}) = 0.5$$

Xác suất xuất hiện mặt có 6 nút:

$$P(6) = 1/6$$

Xác suất thống kê

name	laptop	phone
Kate	PC	Android
Tom	PC	Android
Harry	PC	Android
Annika	Mac	iPhone
Naomi	Mac	Android
Joe	Mac	iPhone
Chakotay	Mac	iPhone
Neelix	Mac	Android
Kes	PC	iPhone
B'Elanna	Mac	iPhone

Xác suất mà một người được lựa chọn ngẫu nhiên **sử dụng iPhone** là bao nhiêu?

Xác suất mà một người được lựa chọn ngẫu nhiên **sử dụng iPhone** khi người này có **sử dụng một máy tính xách tay Mac** là bao nhiêu?

Xác suất thống kê

name	laptop	phone
Kate	PC	Android
Tom	PC	Android
Harry	PC	Android
Annika	Mac	iPhone
Naomi	Mac	Android
Joe	Mac	iPhone
Chakotay	Mac	iPhone
Neelix	Mac	Android
Kes	PC	iPhone
B'Elanna	Mac	iPhone

Xác suất mà một người được lựa chọn ngẫu nhiên **sử dụng iPhone** là bao nhiêu?

Xác suất mà một người được lựa chọn ngẫu nhiên **sử dụng iPhone khi người này có sử dụng một máy tính xách tay Mac** là bao nhiêu?

Xác suất của A với điều kiện B xảy ra được định nghĩa như sau :

$$P(A/B) = \frac{P(AB)}{P(B)}$$

Xác suất thống kê

Xác suất của A với điều kiện B xảy ra được định nghĩa như sau :

$$P(A/B) = \frac{P(AB)}{P(B)}$$

name	laptop	phone
Kate	PC	Android
Tom	PC	Android
Harry	PC	Android
Annika	Mac	iPhone
Naomi	Mac	Android
Joe	Mac	iPhone
Chakotay	Mac	iPhone
Neelix	Mac	Android
Kes	PC	iPhone
B'Elanna	Mac	iPhone

Xác suất mà một người được lựa chọn ngẫu nhiên sử dụng iPhone?

$$P(\text{iPhone}) = 5/10 = 0.5$$

Xác suất mà một người được lựa chọn ngẫu nhiên sử dụng iPhone khi người này sử dụng một máy tính xách tay Mac?

$$P(\text{iPhone} | \text{mac}) = \frac{P(\text{mac} \cap \text{iPhone})}{P(\text{mac})}$$

$$P(\text{mac} \cap \text{iPhone}) = \frac{4}{10} = 0.4 \quad P(\text{mac}) = \frac{6}{10} = 0.6$$

$$P(\text{iPhone} | \text{mac}) = \frac{0.4}{0.6} = 0.667$$

Định lý Bayes

Xác suất của A với điều kiện B xảy ra được định nghĩa như sau :

$$P(A / B) = \frac{P(AB)}{P(B)}$$

Định lý Bayes bắt nguồn từ xác suất có điều kiện. Định lý Bayes được đặt theo tên **Rev. Thomas Bayes** (/ bɛɪz /; 1702-1761), người đầu tiên đã cho thấy làm thế nào để sử dụng thông tin mới để cập nhật những thông tin trước đó.

Xác suất thống kê

Định lý Bayes cho phép tính xác suất xảy ra của một sự kiện ngẫu nhiên A khi biết sự kiện liên quan B đã xảy ra. Xác suất này được ký hiệu là $P(A|B)$, và đọc là "xác suất của A nếu có B ".

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} = \frac{\text{likelihood} * \text{prior}}{\text{normalizing_constant}}$$

Xác suất thống kê

Theo định lý Bayes, xác suất xảy ra A khi biết B sẽ phụ thuộc vào 3 yếu tố:

- Xác suất xảy ra A của riêng nó, không quan tâm đến bất kỳ thông tin nào của B. Kí hiệu là $P(A)$. Đại lượng này còn gọi là tiên nghiệm (**prior**)
- Xác suất xảy ra B của riêng nó, không quan tâm đến A. Kí hiệu là $P(B)$. Đại lượng này còn gọi là hằng số chuẩn hóa (**normalising constant**)
- Xác suất xảy ra B khi biết A xảy ra. Kí hiệu là $P(B|A)$ và đọc là "xác suất của B nếu có A". Đại lượng này gọi là khả năng (**likelihood**) xảy ra B khi biết A đã xảy ra.

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} = \frac{\text{likelihood} * \text{prior}}{\text{normalizing_constant}}$$

Xác suất thống kê

Định lý Bayes

$$P(H|E) = \frac{P(E|H).P(H)}{P(E)}$$

Evidence E = [E1, E2, ..., En] thuộc tính của dữ liệu cần dự báo
Event H: giá trị lớp/ nhãn của dữ liệu E cần sự báo

H	The probability of a hypothesis
E	Conditional on a new piece of evidence
P(H E)	The probability of a hypothesis conditional on a new evidence
P(E H)	The probability of the evidence given the hypothesis
P(H)	The prior probability of the hypothesis
P(E)	The prior probability of the evidence

Nội dung

- Kiến thức về xác suất thống kê
- Giới thiệu về Bayesian classification
- Giải thuật học của naive Bayes
- Kết luận và hướng phát triển

17

Giải thuật naive Bayes

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

■ Ngày thơ

- các thuộc tính (biến) có độ quan trọng như nhau
- các thuộc tính (biến) độc lập thống kê

■ Nhận xét

- Giả thiết các thuộc tính độc lập không bao giờ đúng
- nhưng trong thực tế, naive Bayes cho kết quả khá tốt ☺

18

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

Luật Bayes

Định lý xác suất Bayes

$$P(H|E) = \frac{P(E|H).P(H)}{P(E)}$$

Evidence E = [E1,E2,...,En] có n giá trị thuộc tính của dữ liệu cần dự báo

Event H: giá trị lớp/ nhãn của dữ liệu E cần sự bão

19

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

Luật Bayes

Định lý xác suất Bayes

$$P[H|E] = \frac{P[E|H]P[H]}{P[E]}$$

Do giả thiết: “các thuộc tính độc lập nhau”

=>

$$P(H|E) = \frac{P(E_1|H).P(E_2|H)...P(En|H).P(H)}{P(E)}$$

Evidence E = [E1,E2,...,En] có n thuộc tính của dữ liệu cần dự báo
Event H: giá trị lớp/ nhãn của dữ liệu E cần dự báo

20

Bayes thơ ngây

Bước 1

Học (learning Phase)- xây dựng mô hình sẵn dùng
(tính sẵn xác suất xuất hiện của tất cả các trường hợp)

Bước 2

Khi có đối tượng/sự kiện mới xuất hiện cần phân loại : xác định nhãn của đối tượng mới đến thông qua giá trị xác suất lớn nhất tính được.

Ví dụ:

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

Ví dụ: Dữ liệu weather, dựa trên các thuộc tính (Outlook, Temp, Humidity, Windy), quyết định (play/no)

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

23

Dữ liệu weather, dựa trên các thuộc tính (Outlook, Temp, Humidity, Windy), quyết định (play/no)

Bước 1

$$P(H|E) = \frac{P(E_1|H).P(E_2|H)...P(E_n|H).P(H)}{P(E)}$$

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

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

Ví dụ

Bước 2

Outlook	Temp.	Humidity	Windy	Play
Sunny	Cool	High	True	?

← **Evidence E**

- Phản tử mới đến,

$x' = (\text{Outlook}=\text{Sunny}, \text{Temperature}=\text{Cool}, \text{Humidity}=\text{High}, \text{Wind}=\text{True})$

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

xác suất
của lớp
“yes”

25

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

Ví dụ

Bước 2

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

xác suất
của lớp
“yes”

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

$$\begin{aligned} P(\text{Outlook}=\text{Sunny} | \text{Play}=\text{Yes}) &= 2/9 \\ P(\text{Temperature}=\text{Cool} | \text{Play}=\text{Yes}) &= 3/9 \\ P(\text{Humidity}=\text{High} | \text{Play}=\text{Yes}) &= 3/9 \\ P(\text{Wind}=\text{True} | \text{Play}=\text{Yes}) &= 3/9 \\ P(\text{Play}=\text{Yes}) &= 9/14 \end{aligned}$$

26

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

Dữ liệu weather, dựa trên các thuộc tính (Outlook, Temp, Humidity, Windy), quyết định (play/no)

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

■ quyết định (play=yes/no)?

$$\begin{aligned} P[\text{Yes} | E] &= (2/9 \times 3/9 \times 3/9 \times 3/9 \times 9/14) / P[E] \\ &= 0.0053 / P[E] \end{aligned}$$

$$P[\text{No} | E] = 0.0206 / P[E]$$

Outlook	Temp.	Humidity	Windy	Play
Sunny	Cool	High	True	?

=> yes/no?

27

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

Dữ liệu weather, dựa trên các thuộc tính (Outlook, Temp, Humidity, Windy), quyết định (play/no)

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

■ quyết định (play=yes/no)?

$$\text{Likelihood(yes)} = 2/9 \times 3/9 \times 3/9 \times 3/9 \times 9/14 = 0.0053$$

$$\text{Likelihood(no)} = 3/5 \times 1/5 \times 4/5 \times 3/5 \times 5/14 = 0.0206$$

$$\text{Likelihood(yes)} = 0.0053 / (0.0053 + 0.0206) = 0.205$$

$$\text{Likelihood(no)} = 0.0206 / (0.0053 + 0.0206) = 0.795$$

=> yes/no?

Outlook	Temp.	Humidity	Windy	Play
Sunny	Cool	High	True	?

Bài tập- cho tập dữ liệu như bảng

	age	income	student	credit rating	buys computer
Class: C1:buys_computer= 'yes'	<=30	high	no	fair	no
C2:buys_computer= 'no'	<=30	high	no	excellent	no
	30...40	high	no	fair	yes
	>40	medium	no	fair	yes
	>40	low	yes	fair	yes
	>40	low	yes	excellent	no
Data sample X =(age<=30, Income=medium, Student=yes Credit_rating= Fair)	31...40	low	yes	excellent	yes
	<=30	medium	no	fair	no
	<=30	low	yes	fair	yes
	>40	medium	yes	fair	yes
	<=30	medium	yes	excellent	yes
	31...40	medium	no	excellent	yes
	31...40	high	yes	fair	yes
	>40	medium	no	excellent	no

age	income	student	credit_rating	buys_computer
<=30	high	no	fair	no
<=30	high	no	excellent	no
30...40	high	no	fair	yes
>40	medium	no	fair	yes
>40	low	yes	fair	yes
>40	low	yes	excellent	no
31...40	low	yes	excellent	yes
<=30	medium	no	fair	no
<=30	low	yes	fair	yes
>40	medium	yes	fair	yes
<=30	medium	yes	excellent	yes
31...40	medium	no	excellent	yes
31...40	high	yes	fair	yes
>40	medium	no	excellent	no

Xác suất = 0

- giá trị của thuộc tính không xuất hiện trong tất cả các lớp sử dụng *Laplace estimator*
- xác suất không bao giờ có giá trị 0
- Cộng thêm cho tử một giá trị là $p_i\mu$ và mẫu số giá trị μ để tính xác suất. μ **hằng số dương** và p_i là **hệ số dương** sao cho tổng các $p_i = 1$ ($i=1..n$)

31

Laplace estimator

- ví dụ : thuộc tính *outlook* cho lớp “no”

$$\begin{array}{c}
 \frac{3+\mu/3}{5+\mu} \qquad \qquad \frac{0+\mu/3}{5+\mu} \qquad \qquad \frac{2+\mu/3}{5+\mu} \\
 \text{\textit{Sunny}} \qquad \qquad \text{\textit{Overcast}} \qquad \qquad \text{\textit{Rainy}}
 \end{array}$$

Outlook		Temperature		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								

32

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

Laplace estimator

- ví dụ : thuộc tính *outlook* cho lớp “no”

$$\begin{array}{ccc} \frac{3+\mu/3}{5+\mu} & \frac{0+\mu/3}{5+\mu} & \frac{2+\mu/3}{5+\mu} \\[10pt] \textit{Sunny} & \textit{Overcast} & \textit{Rainy} \end{array}$$

- trọng số có thể không bằng nhau, nhưng tổng phải là 1
- thuộc tính *outlook* cho lớp “Yes”

$$\begin{array}{ccc} \frac{2+\mu p_1}{9+\mu} & \frac{4+\mu p_2}{9+\mu} & \frac{3+\mu p_3}{9+\mu} \\[10pt] \textit{Sunny} & \textit{Overcast} & \textit{Rainy} \end{array}$$

33

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

Laplace estimator

- ví dụ : thuộc tính *outlook* cho lớp “no”

$$\begin{array}{ccc} \frac{3+1/3}{5+1} & \frac{0+1/3}{5+1} & \frac{2+1/3}{5+1} \\[10pt] \textit{Sunny} & \textit{Overcast} & \textit{Rainy} \end{array}$$

Outlook		
	Yes	No
Sunny	2	3
Overcast	4	0
Rainy	3	2
Sunny	2/9	3/5
Overcast	4/9	0/5
Rainy	3/9	2/5

$$\begin{aligned} \text{Sunny} &= 10/18 \\ \text{Overcast} &= 1/18 \\ \text{Rainy} &= 7/18 \end{aligned}$$

34

- Giới thiệu về Bayesian classification
- [Giải thuật học của naive Bayes](#)
- kết luận và hướng phát triển

Giá trị thuộc tính nhiều

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

Outlook	Temp.	Humidity	Windy	Play
?	Cool	High	True	?

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

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

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

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

35

Play tennis dataset

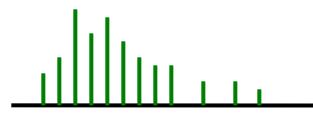
Outlook	Temp	Humidity	Windy	Play
Sunny	Hot	High	False	No
Sunny	Hot	High	True	No
Overcast	Hot	High	False	Yes
Rainy	Mild	High	False	Yes
Rainy	Cool	Normal	False	Yes
Rainy	Cool	Normal	True	No
Overcast	Cool	Normal	True	Yes
Sunny	Mild	High	False	No
Sunny	Cool	Normal	False	Yes
Rainy	Mild	Normal	False	Yes
Sunny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

Play tennis dataset

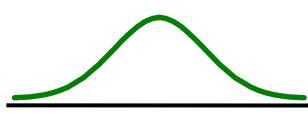
OUTLOOK	TEMPERATURE	HUMIDITY	WINDY	PLAY
sunny	85	85	FALSE	Don't Play
sunny	80	90	TRUE	Don't Play
overcast	83	78	FALSE	Play
rain	70	96	FALSE	Play
rain	68	80	FALSE	Play
rain	65	70	TRUE	Don't Play
overcast	64	65	TRUE	Play
sunny	72	95	FALSE	Don't Play
sunny	69	70	FALSE	Play
rain	75	80	FALSE	Play
sunny	75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



Biên ngẫu nhiên rời rạc



Biên ngẫu nhiên liên tục

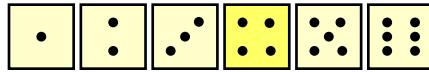
Biên ngẫu nhiên rời rạc

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 $Y = 0, 1, 2, 3, 4, hoặc 5$



Biên 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: Toss 2 đồng xu, với $X =$ Số lần xuất hiện mặt hình.

⇒ Bảng phân phối xác suất

X	0	1	2
P	0.25	0.5	0.25

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

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.

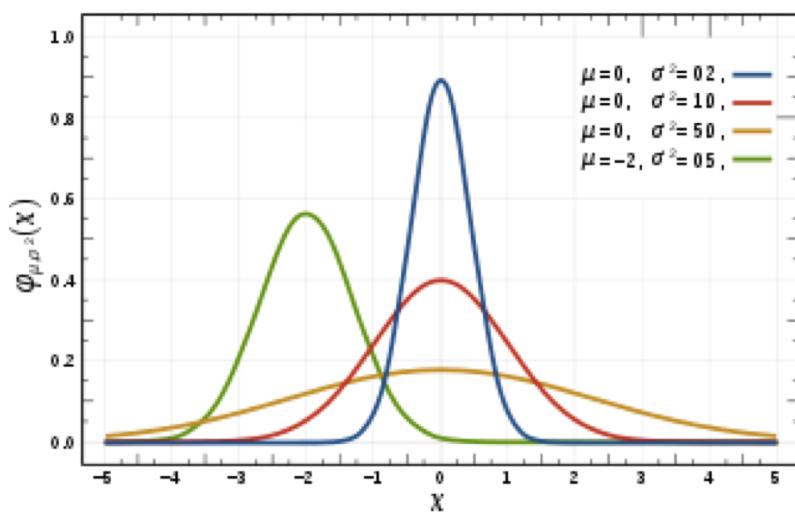
Phương sai thường được ký hiệu là σ^2

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.



Dữ liệu liên tục

- giả sử các thuộc tính có phân phối *Gaussian*
- hàm mật độ xác suất được tính như sau

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

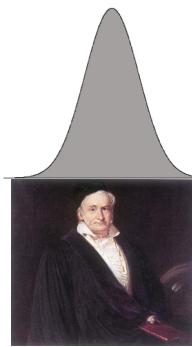
- standard deviation σ

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

- hàm mật độ xác suất $f(x)$

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

Karl Gauss, 1777-1855
great German mathematician



45

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
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
overcast	4/9	0/5	std dev	6.2	7.9	std dev	10.2	9.7	true
rainy	3/9	2/5							

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						

■ 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
$$\sigma^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \mu)^2$$

Phuong 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 $f(x)$

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

48

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

49

- 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	?

50

- 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	?

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	?

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

$\text{Likelihood(yes)} = 2/9 \times 0.0340 \times 0.0221 \times 3/9 \times 9/14 = 0.000036$
$\text{Likelihood(no)} = 3/5 \times 0.0291 \times 0.0380 \times 3/5 \times 5/14 = 0.000136$
$\text{Likelihood(yes)} = 0.000036 / (0.000036 + 0.000136) = 20.9\%$
$\text{Likelihood(no)} = 0.000136 / (0.000036 + 0.000136) = 79.1\%$

52