Phương pháp học Bayes
Bayesian classification

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.

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

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)

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

Xác suất thống kê

Một vài ví dụ

- Khi tung 1 đồng xu, khả năng nhận mặt ngửa là bao nhiêu?
- Khi tung một hột 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(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?

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

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

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

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".

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} = \frac{likelihood * prior}{normalizing\_constant}$$

Xác suất thống kê

Theo định lí 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 (normalizing 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 xảy ra B khi biết A đã xảy ra (likelihood).

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} = \frac{likelihood * prior}{normalizing\_constant}$$

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

Giải thuật naive Bayes

- Ngây 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

Luật Bayes

Đinh 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 dự báo

Luật Bayes

Đinh lý xác suất Bayes

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

Do giả thiết: " các thuộc tính độc lập nhau"

=>

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

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

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 tương 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

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

Bước 1

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

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								

Ví dụ

Bước 2

Outlook	Temp.	Humidity	Windy	Play
Sunny	Cool	High	True	?

← ***Evidence E***

– Phần tử mới đến,

x' =(Outlook=*Sunny*, Temperature=*Cool*, Humidity=*High*, Wind=True)

xác suất của lớp "yes"

xác suất của lớp "no"

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

Ví dụ

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

Bước 2

Outlook	Temp.	Humidity	Windy	Play
Sunny	Cool	High	True	?

← ***Evidence E***

– Phần tử mới đến,

x' =(Outlook=*Sunny,* Temperature=*Cool,* Humidity=*High,* Wind=True)

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

xác suất của lớp "yes"

Ví dụ

Bước 2

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

xác suất của lớp "yes"

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

Pr[Outlook=*Sunny*|Play=*Yes*] = 2/9

Pr[Temperature=*Cool*|Play=*Yes]* = 3/9

Pr[Humidity=*High*|Play=*Yes*] = 3/9

Pr[Windy=True|Play=*Yes*] = 3/9

Pr[Play=*Yes*] = 9/14

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	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								

- quyết định (play=yes/no)?

P[Yes| E] = (2/9 x 3/9 x 3/9 x 3/9 x 9/14) /P[E]

= 0.0053/P[E]

P[No| E] = 0.0206 /P[E]

=> yes/no?

Outlook	Temp.	Humidity	Windy	Play
Sunny	Cool	High	True	?

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	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								

- quyết định (play=yes/no)?

P[Yes| E] = (2/9 x 3/9 x 3/9 x 3/9 x 9/14) /P[E]

= 0.0053/P[E]

Outlook	Temp.	Humidity	Windy	Play
Sunny	Cool	High	True	?

P[No| E] = 0.0206 /P[E]

Vì P[Yes| E] < P[No |E]=> No

Bài tập- cho tập dữ liệu như bảng

Class:
C1:buys_computer= ‘yes’
C2:buys_computer= ‘no’

Data sample
X =(age=youth, Income=medium, Student=yes Credit_rating= Fair)

RID	age	income	student	credit_rating	Class: buys_compute
1	youth	high	no	fair	no
2	youth	high	no	excellent	no
3	middle_aged	high	no	fair	yes
4	senior	medium	no	fair	yes
5	senior	low	yes	fair	yes
6	senior	low	yes	excellent	no
7	middle_aged	low	yes	excellent	yes
8	youth	medium	no	fair	no
9	youth	low	yes	fair	yes
10	senior	medium	yes	fair	yes
11	youth	medium	yes	excellent	yes
12	middle_aged	medium	no	excellent	yes
13	middle_aged	high	yes	fair	yes
14	senior	medium	no	excellent	no

RID	age	income	student	credit_rating	Class: buys_computer
1	youth	high	no	fair	no
2	youth	high	no	excellent	no
3	middle_aged	high	no	fair	yes
4	senior	medium	no	fair	yes
5	senior	low	yes	fair	yes
6	senior	low	yes	excellent	no
7	middle_aged	low	yes	excellent	yes
8	youth	medium	no	fair	no
9	youth	low	yes	fair	yes
10	senior	medium	yes	fair	yes
11	youth	medium	yes	excellent	yes
12	middle_aged	medium	no	excellent	yes
13	middle_aged	high	yes	fair	yes
14	senior	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 pi là hệ số dương sao cho tổng các $p_i = 1$ (i=1..n)**

Laplace estimator

- ví dụ : thuộc tính *outlook* cho lớp "no"

$$\frac{3+\mu/3}{5+\mu} \qquad \frac{0+\mu/3}{5+\mu} \qquad \frac{2+\mu/3}{5+\mu}$$

Sunny ***Overcast*** ***Rainy***

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								

Laplace estimator

- ví dụ : thuộc tính *outlook* cho lớp "no"

$$\frac{3+\mu/3}{5+\mu} \qquad \frac{0+\mu/3}{5+\mu} \qquad \frac{2+\mu/3}{5+\mu}$$

Sunny ***Overcast*** ***Rainy***

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

$$\frac{2+\mu p_1}{9+\mu} \qquad \frac{4+\mu p_2}{9+\mu} \qquad \frac{3+\mu p_3}{9+\mu}$$

Sunny ***Overcast*** ***Rainy***

Laplace estimator

- ví dụ : thuộc tính *outlook* cho lớp "no"

$$\frac{3+1/3}{5+1} \qquad \frac{0+1/3}{5+1} \qquad \frac{2+1/3}{5+1}$$

Sunny ***Overcast*** ***Rainy***

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

Sunny = 10/18
Overcast = 1/18
Rainy = 7/18

Giá trị thuộc tính nhiễu

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

Outlook	Temp.	Humidity	Windy	Play
?	Cool	High	True	?

Likelihood(yes) = $3/9 \times 3/9 \times 3/9 \times 9/14$ /**P[E]**
= 0.0238
Likelihood(no) = $1/5 \times 4/5 \times 3/5 \times 5/14$ /**P[E]**
= 0.0343

Play tennis dataset

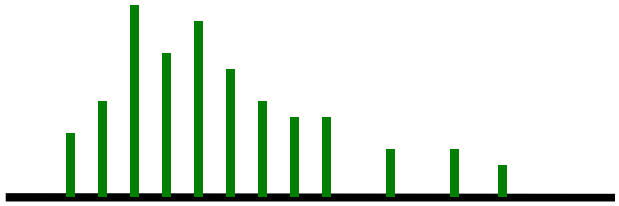
Outlook	Temp	Humidity	Windy	Play
Sunny	Hot	High	False	No
Sunny	Hot	High	True	No
Overcast	Hot	High	False	Yes
Rainy	Mild	High	False	Yes
Rainy	Cool	Normal	False	Yes
Rainy	Cool	Normal	True	No
Overcast	Cool	Normal	True	Yes
Sunny	Mild	High	False	No
Sunny	Cool	Normal	False	Yes
Rainy	Mild	Normal	False	Yes
Sunny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

Play tennis dataset

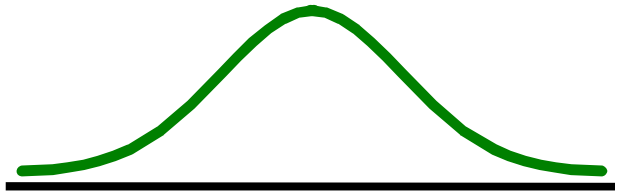
OUTLOOK	TEMPERATURE	HUMIDITY	WINDY	PLAY
sunny	85	85	FALSE	Don't Play
overcast	83	78	FALSE	Play
rain	70	96	FALSE	Play
rain	68	80	FALSE	Play
rain	65	70	TRUE	Don't Play
overcast	64	65	TRUE	Play
sunny	72	95	FALSE	Don't Play
sunny	69	70	FALSE	Play
rain	75	80	FALSE	Play
sunny	75	70	TRUE	Play
overcast	72	90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

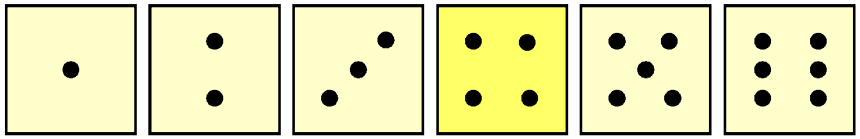
Biến ngẫu nhiên rời rạc

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, 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.

$$P(X \le med(X)) = P(X \ge med(X)) = \frac{1}{2}$$

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

Ví dụ: Tung 2 đồng xu, với X = Số lần xuất hiện mặt hình.

⌦ Bảng phân phối xác suất

X	0	1	2
P	0.25	0.5	0.25

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.

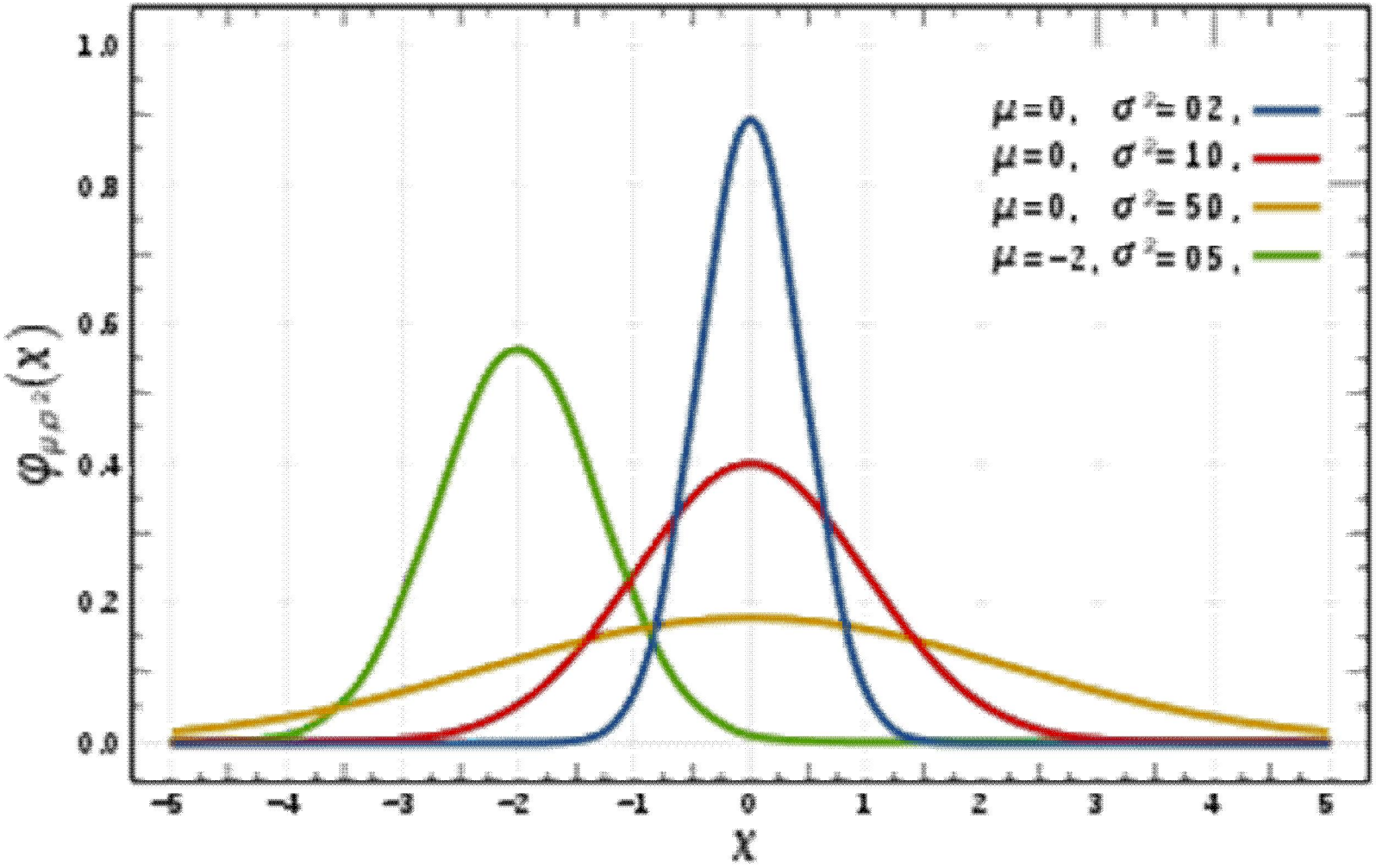
Phương sai thường được ký hiệu là σ^2

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

$$\sigma = \sqrt{\sigma^2} = \sqrt{VarX}$$

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.


1.0
0.8
0.6
0.4
0.2
0.0
$\varphi_{\mu,\sigma^2}(x)$
−6
−5
−4
−3
−2
−1
0
1
2
3
4
5
x
μ=0, σ²=0.2
μ=0, σ²=1.0
μ=0, σ²=5.0
μ=−2, σ²=0.5

Dữ liệu liên tục

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

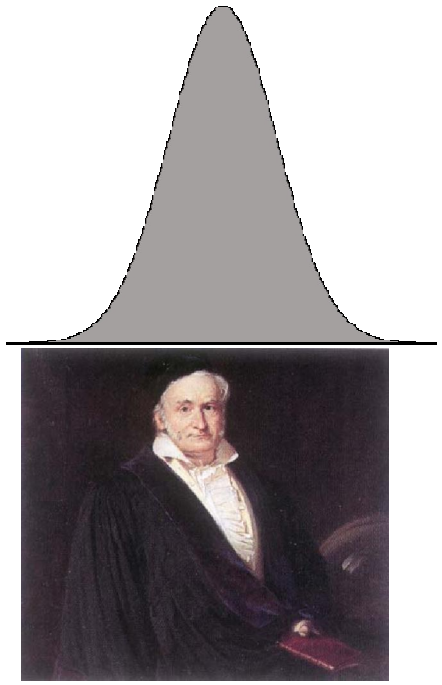
$$\mu = \frac{1}{n}\sum_{i=1}^{n} x_i$$

  - *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

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						

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

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	?

Dữ liệu liên tục

- phân lớp

Outlook	Temp.	Humidity	Windy	Play
Sunny	66	90	true	?

Dữ liệu liên tục

- phân lớp

Outlook	Temp.	Humidity	Windy	Play
Sunny	66	90	true	?

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

Likelihood(yes) = $2/9 \times 0.0340 \times 0.0221 \times 3/9 \times 9/14$ / P[E] = 0.000036 / P[E]

Likelihood(no) = $3/5 \times 0.0291 \times 0.0380 \times 3/5 \times 5/14$ / P[E] = 0.000136 / P[E]

Bernoulli Naive Bayes

Mô hình này được áp dụng cho các loại dữ liệu mà mỗi thành phần là một giá trị binary - bẳng 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:
$p(x_i|c)=p(i|c)^{xi}(1-p(i|c)^{1-xi}$

với $p(i|c)$ 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. Lúc này, 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

Khi đó, $p(x_i|c)$ tỉ lệ với tần suất từ thứ i (hay feature thứ i cho trường hợp tổng quát) xuất hiện trong các văn bản của class c. Giá trị này có thể được tính bằng cách:$\lambda_{ci}=p(x_i|c)=N_{ci}/N_c$

N_{ci} là tổng số lần từ thứ i xuất hiện trong các văn bản của class c

N_c là tổng số từ (kể cả lặp) xuất hiện trong class c.

Nội dung

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

Kết luận

- naïve Bayes
  - cho kết quả tốt trong thực tế mặc dù chịu những giả thiết về tính độc lập thống kê của các thuôc 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 (sử dụng kernel density estimators)

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 !

Bài tập

A	B	C
m	b	t
m	s	t
g	q	t
h	s	t
g	q	t
g	q	f
g	s	f
h	b	f
h	q	f
m	b	f

A=m; B=q; C=?