KHOA CNTT & TRUYỀN THÔNG
BM KHOA HỌC MÁY TÍNH

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)

Kỹ thuật DM thành công (2011)

Which methods/algorithms did you use for data analysis in 2011? [311 voters]	
Decision Trees/Rules (186)	59.8 %
Regression (180)	57.9 %
Clustering (163)	52.4 %
Statistics (descriptive) (149)	47.9 %
Visualization (119)	38.3 %
Time series/Sequence analysis (92)	29.6 %
Support Vector (SVM) (89)	28.6 %
Association rules (89)	28.6 %
Ensemble methods (88)	28.3 %
Text Mining (86)	27.7 %
Neural Nets (84)	27.0 %
Boosting (73)	23.5 %
Bayesian (68)	21.9 %
Bagging (63)	20.3 %
Factor Analysis (58)	18.7 %
Anomaly/Deviation detection (51)	16.4 %
Social Network Analysis (44)	14.2 %
Survival Analysis (29)	9.32 %
Genetic algorithms (29)	9.32 %
Uplift modeling (15)	4.82 %

Top 10 DM algorithms (2015)

Here are the algorithms:

1. C4.5
2. k-means
3. Support vector machines
4. Apriori
5. EM
6. PageRank
7. AdaBoost
8. kNN
9. Naive Bayes
10. CART

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(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?

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

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

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

Định lý Bayes

Định lý Bayes bắt nguồn từ xác suất có điều kiện. Định lý Bayes được đặt theo tên **Rev. Thomas Bayes** (/ beɪz /; 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 của A với điều kiện B xảy ra được định nghĩa như sau :

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

P(A/B) = P(AB)/P(B)
=> P(AB) = P(A/B)*P(B)

P(B/A)=P(AB)/P(A)
=> P(AB)=(B/A)*P(A)
P(A/B) = (P(B/A)*P(A))/P(B)

Định lý Bayes

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".

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

Định lý Bayes

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 (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.

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

Định lý Bayes

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

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

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

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 sự 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"

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

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

Bayes thơ ngây

Bước 1: học/ huấn luyện mô hình (learning Phase)

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

Bước 2: dự báo/ dự đoán

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

22

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)

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

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

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								

24

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)

Cần xác định: ***xác suất của lớp "yes" và xác suất của lớp "no"***

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

Ví dụ

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

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"

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

P(Outlook=*Sunny*|Play=*Yes*) = 2/9
P(Temperature=*Cool*|Play=*Yes*) = 3/9
P(Huminity=*High*|Play=*Yes*) = 3/9
P(Wind=True|Play=*Yes*) = 3/9
P(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]

Outlook	Temp.	Humidity	Windy	Play
Sunny	Cool	High	True	?

=> yes/no?

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)?

Likelihood(yes) = 2/9 x 3/9 x 3/9 x 3/9 x 9/14 = 0.0053

Likelihood(no) = 3/5 x 1/5 x 4/5 x 3/5 x 5/14 = 0.0206

Likelihood(yes) = 0.0053 / (0.0053 + 0.0206) = 0.205

Likelihood(no) = 0.0206 / (0.0053 + 0.0206) = 0.795

=> yes/no?

Outlook	Temp.	Humidity	Windy	Play
Sunny	Cool	High	True	?

29

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)**

30

Laplace estimator – Ước lượng Laplace

- VD: thuộc tính *outlook* cho lớp "no" => $p_1=p_2=p_3=1/3$; $\mu=1$

$$\frac{3+\mu/3}{5+\mu}$$ ***Sunny***

$$\frac{0+\mu/3}{5+\mu}$$ ***Overcast***

$$\frac{2+\mu/3}{5+\mu}$$ ***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								

31

Laplace estimator – Ước lượng Laplace

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

$$\frac{3+1/3}{5+1}$$ ***Sunny***

$$\frac{0+1/3}{5+1}$$ ***Overcast***

$$\frac{2+1/3}{5+1}$$ ***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

$p_1 = p_2 = p_3 = 1/3$; $\mu=1$

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

32

Laplace estimator – Ước lượng Laplace

- 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***

Đề xuất giá trị p1, p2, p3 và **μ**

Laplace estimator – Ước lượng Laplace

Ước lượng Laplace cho trường hợp sau (**μ**, p_i = ***?)***

	A	B	C
T1	1/7	2/10	5/13
T2	2/7	1/10	3/13
T3	1/7	2/10	0/13
T4	3/7	5/10	5/13

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 × 3/9 × 3/9 × 9/14 = 0.0238
Likelihood(no) = 1/5 × 4/5 × 3/5 × 5/14 = 0.0343
Likelihood(yes) = 0.0238 / (0.0238 + 0.0343) = 0.41
Likelihood(no) = 0.0343 / (0.0238 + 0.0343) = 0.59

Bài tập- cho tập dữ liệu như bảng

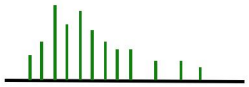
RID	age	income	student	credit_rating	Class: buys_computer
1	youth	high	no	fair	no
2	youth	high	no	excellent	no
3	middle_aged	high	no	fair	yes
4	senior	medium	no	fair	yes
5	senior	low	yes	fair	yes
6	senior	low	yes	excellent	no
7	middle_aged	low	yes	excellent	yes
8	youth	medium	no	fair	no
9	youth	low	yes	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

Class: C1:buys_computer= 'yes' C2:buys_computer= 'no'
Dự đoán nhãn của phần tử X1 =(age=youth, Income=medium, Student=yes, Credit_rating=Fair)
Dự đoán nhãn của phần tử X2 =(age=middle_agged, Student=yes, Credit_rating=Fair)

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

Giá trị rời rạc và liên tục

- **Rời rạc**
  - Màu sắc
  - Giới tính
  - Tôn giáo
- **Liên tục**
  - Chiều cao
  - Cân nặng
  - Thời gian hoàn thành công việc



Hình 1

Hình 2

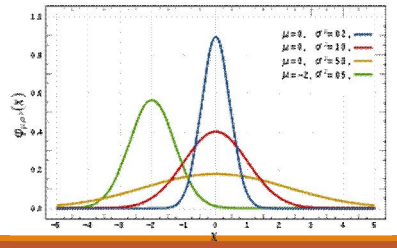
Xác định dữ liệu trong bảng kế tiếp, giá trị của các thuộc tính là giá trị rời rạc hay liên tục?

Outlook	Temp	Humidity	Windy	Play
Sunny	85	85	False	**No**
Sunny	80	90	True	**No**
Overcast	83	86	False	**Yes**
Rainy	70	96	False	**Yes**
Rainy	68	80	False	**Yes**
Rainy	65	70	True	**No**
Overcast	64	65	True	**Yes**
Sunny	72	95	False	**No**
Sunny	69	70	False	**Yes**
Rainy	75	80	False	**Yes**
Sunny	75	70	True	**Yes**
Overcast	72	90	True	**Yes**
Overcast	81	75	False	**Yes**
Rainy	71	91	True	**No**

Dữ liệu liên tục

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).

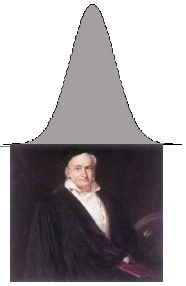
Phân phối chuẩn tắc (*standard normal distribution*) là phân phối chuẩn với giá trị trung bình bằng 0 và phương sai bằng 1 (đường cong màu đỏ trong hình). Phân phối chuẩn còn được gọi là **đường cong chuông** (*bell curve*) vì đồ thị của mật độ xác suất có dạng chuông.



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

Dữ liệu liên tục



- Giả sử các thuộc tính có phân phối *Gaussian*
- hàm mật độ xác suất ***f(x)*** được tính như sau

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

- *Mean* μ $\quad \mu = \frac{1}{n}\sum_{i=1}^{n} x_i$
- *Phương sai (Variance)* σ^2 $\quad \sigma^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \mu)^2$
- *Độ lệch chuẩn -standard deviation: căn bậc 2 của phương sai* $\quad \sigma = \sqrt{\sigma^2}$

Karl Gauss, 1777-1855 great German mathematician

https://www.mathsisfun.com/data/standard-

Outlook	Temp.	Humidity	Windy	Play
Sunny	66	90	true	?

Outlook	Temp	Humidity	Windy	Play
Sunny	85	85	False	**No**
Sunny	80	90	True	**No**
Overcast	83	86	False	**Yes**
Rainy	70	96	False	**Yes**
Rainy	68	80	False	**Yes**
Rainy	65	70	True	**No**
Overcast	64	65	True	**Yes**
Sunny	72	95	False	**No**
Sunny	69	70	False	**Yes**
Rainy	75	80	False	**Yes**
Sunny	75	70	True	**Yes**
Overcast	72	90	True	**Yes**
Overcast	81	75	False	**Yes**
Rainy	71	91	True	**No**

Bước 1: huấn luyện mô hình

Outlook	Temp	Humidity	Windy	Play
Sunny	85	85	False	**No**
Sunny	80	90	True	**No**
Overcast	83	86	False	**Yes**
Rainy	70	96	False	**Yes**
Rainy	68	80	False	**Yes**
Rainy	65	70	True	**No**
Overcast	64	65	True	**Yes**
Sunny	72	95	False	**No**
Sunny	69	70	False	**Yes**
Rainy	75	80	False	**Yes**
Sunny	75	70	True	**Yes**
Overcast	72	90	True	**Yes**
Overcast	81	75	False	**Yes**
Rainy	71	91	True	**No**

Outlook			Temperature		Humidity		Windy			Play	
	Yes	No	Yes	No	Yes	No		Yes	No	Yes	No
Sunny	2	3					False	6	2	9	5
Overcast	4	0					True	3	3		
Rainy	3	2									
Sunny	2/9	3/5	????				False	6/9	2/5	9/14	5/14
Overcast	4/9	0/5					True	3/9	3/5		
Rainy	3/9	2/5									

Temp	Play
85	No
80	No
83	Yes
70	Yes
68	Yes
65	No
64	Yes
72	No
69	Yes
75	Yes
75	Yes
72	Yes
81	Yes
71	No

$$\mu = \frac{1}{n}\sum_{i=1}^{n} x_i \qquad \sigma^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \mu)^2 \qquad f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

Temp	Play
85	No
80	No
83	Yes
70	Yes
68	Yes
65	No
64	Yes
72	No
69	Yes
75	Yes
75	Yes
72	Yes
81	Yes
71	No

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

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

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

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

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

- hàm mật độ xác suất *f(x)* tính khi có phần tử mới xuất hiện $f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$

Temp	Play
85	No
80	No
83	Yes
70	Yes
68	Yes
65	No
64	Yes
72	No
69	Yes
75	Yes
75	Yes
72	Yes
81	Yes
71	No

The numeric weather data with summary statistics													
utlook			temperature			humidity			windy			play	
	yes	no		yes	no		yes	no		yes	no	yes	no
	2	3		83	85		86	85	false	6	2	9	5
	4	0		70	80		96	90	true	3	3		
	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	**σ^2**	**38..44**									

The numeric weather data with summary statistics													
outlook			temperature			humidity			windy			play	
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	**σ²**	**38..44**									

Outlook	Temp.	Humidity	Windy	Play
Sunny	66	90	true	?

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

P[Yes| E] =(P(Outlook=*Sunny*|Play=*Yes*)
x **P(Temp.=*66*|Play=*Yes*)**
x **P(Hum.=*90*|Play=*Yes*)**
x P(Wind=True|Play=*Yes*)
x P(Play=*Yes*)**) /P[E]**

P(Outl=*Sunny*|Play=*Yes*) = 2/9
P(Temp.=*66*|Play=*Yes*) = ??
P(Hum.=*90*|Play=*Yes*) = ??
P(Wind=True|Play=*Yes*) = 3/9
P(Play=*Yes*) = 9/14

The numeric weather data with summary statistics													
outlook			temperature			humidity			windy			play	
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	**σ²**	**38..44**									

Outlook	Temp.	Humidity	Windy	Play
Sunny	66	90	true	?

P(Outl=*Sunny*|Play=*Yes*) = 2/9
P(Temp.=*66*|Play=*Yes*) = ??
P(Hum.=*90*|Play=*Yes*) = ??
P(Wind=True|Play=*Yes*) = 3/9
P(Play=*Yes*) = 9/14

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

The numeric weather data with summary statistics													
outlook			temperature			humidity			windy			play	
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	**σ²**	**38..44**									

P(Outl=*Sunny*|Play=*Yes*) = 2/9
P(Temp.=*66*|Play=*Yes*) = 0.034
P(Hum.=*90*|Play=*Yes*) = ??
P(Wind=True|Play=*Yes*) = 3/9
P(Play=*Yes*) = 9/14

Outlook	Temp.	Humidity	Windy	Play
Sunny	66	90	true	?

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

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

The numeric weather data with summary statistics													
outlook			temperature			humidity			windy			play	
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	**σ²**	**38..44**									

P(Outl=*Sunny*|Play=*Yes*) = 2/9
P(Temp.=*66*|Play=*Yes*) = 0.034
P(Hum.=*90*|Play=*Yes*) = ??
P(Wind=True|Play=*Yes*) = 3/9
P(Play=*Yes*) = 9/14

Outlook	Temp.	Humidity	Windy	Play
Sunny	66	90	true	?

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

f(temp=66/Yes) =?
f(temp=66/No) =?

f(humidity=90/Yes) =?
f(humidity=90/No) =?

The numeric weather data with summary statistics													
outlook			temperature			humidity			windy			play	
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	σ^2	**38..44**									

P(Outl=*Sunny*|Play=*Yes*) = 2/9
P(Temp.=*66*|Play=*Yes*) = 0.034
P(Hum.=*90*|Play=*Yes*) = 0.0221
P(Wind=True|Play=*Yes*) = 3/9
P(Play=*Yes*) = 9/14

Outlook	Temp.	Humidity	Windy	Play
Sunny	66	90	true	?

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

f(humidity=90/Yes) =

The numeric weather data with summary statistics													
outlook			temperature			humidity			windy			play	
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	σ^2	**38..44**									

P(Outl=*Sunny*|Play=*Yes*) = 2/9
P(Temp.=*66*|Play=*Yes*) = 0.034
P(Hum.=*90*|Play=*Yes*) = 0.0221
P(Wind=True|Play=*Yes*) = 3/9
P(Play=*Yes*) = 9/14

Outlook	Temp.	Humidity	Windy	Play
Sunny	66	90	true	?

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

f(temp=66/Yes) = 0.034
f(temp=66/No) = 0.0291
f(humidity=90/Yes) = 0.0221
f(humidity=90/No) = 0.0380

Nhãn????

Outlook	Temp.	Humidity	Windy	Play
Sunny	66	90	true	?

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

f(temp=66/Yes) = 0.034
f(temp=66/No) = 0.0291
f(humidity=90/Yes) = 0.0221
f(humidity=90/No) = 0.0380

The numeric weather data with summary statistics													
outlook			temperature			humidity			windy			play	
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	σ^2	**38..44**									

Dữ liệu liên tục

- **Bước 2- dự đoán**

Outlook	Temp.	Humidity	Windy	Play
Sunny	66	90	true	?

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

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

$$p(x_i|c) = \frac{N_{ci}}{N_c}$$

•NciNci là tổng số lần từ thứ ii xuất hiện trong các văn bản của class cc, nó được tính là tổng của tất cả các thành phần thứ ii của các feature vectors ứng với class cc.
•NcNc là tổng số từ (kể cả lặp) xuất hiện trong class cc. Nói cách khác, nó bằng tổng độ dài của toàn bộ các văn bản thuộc vào class cc.

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 lớp **"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 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)

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