KHOA CNTT & TRUYỀN THÔNG
BM KHOA HỌC MÁY TÍNH

Phương pháp học Bayes
Bayesian classification

PGS. TS. Đỗ Thanh Nghị
TS. Trần Nguyễn Minh Thư
tnmthu@ctu.edu.vn

1

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

2

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

3

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

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)

5

Kỹ thuật DM

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

6

6

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

7

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 :

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

8

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?

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

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

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

9

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)

10

Định lý Bayes

Đinh lý 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

11

11

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

12

12

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

13

13

Định lý 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"

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

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

14

14

Bayes thơ ngây

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

xây dựng mô hình sẳn dùng (tính sẳn 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 tượng mới đến thông qua giá trị xác suất lớn nhất tính được

15

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

16

16

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								

17

17

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"***

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

18

18

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"

19

19

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

20

20

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?

21

21

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	?

22

22

Bài tập- cho tập dữ liệu như bảng

Class:
C1:buys_computer= 'yes'
C2:buys_computer= 'no'

X_1 =(age<=30, Income=medium, Student=yes Credit_rating= Fair)

age	income	student	credit_rating	buys_computer
<=30	high	no	fair	no
<=30	high	no	excellent	no
31…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

X2:= (Age: 31-40; income=high, student=yes;credit =Fair

23

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

X_1 =(age<=30, Income=medium, Student=yes Credit_rating=Fair)

P[Yes| X1] = ?

P[No| X1] = ?

age	income	student	credit_rating	buys_computer
<=30	high	no	fair	no
<=30	high	no	excellent	no
31…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

24

X_1 =(age<=30, Income=medium, Student=yes Credit_rating=Fair)

X2:= (Age: 31-40; income=high, student=yes;credit =Fair

Age	Yes	No	Income	Yes	No	Student	Yes	No	credit	Yes	No		
<=30	2/9	3/5	high	2/9	2/5	No	3/9	4/5	fair	6/9	2/5	9/14	5/14
31..40	4/9	0/5	Medium	4/9	2/5	Yes	6/9	1/5	exc	3/9	3/5		
>40	3/9	2/5	low	3/9	1/5								

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

P[Yes| X1] = ?

P[No| X1] = ?

age	income	student	credit_rating	buys_computer
<=30	high	no	fair	no
<=30	high	no	excellent	no
31…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

25

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)**

26

26

Laplace estimator – Ước lượng Laplace

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

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

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								

27

27

Laplace estimator – Ước lượng Laplace

- 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

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

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

28

28

Laplace estimator – Ước lượng Laplace

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

$\frac{2+\mu p_1}{9+\mu}$	$\frac{4+\mu p_2}{9+\mu}$	$\frac{3+\mu p_3}{9+\mu}$
Sunny	***Overcast***	***Rainy***

Đề xuất giá trị p1, p2, p3 và **μ**

29

29

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

30

30

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

31

31

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

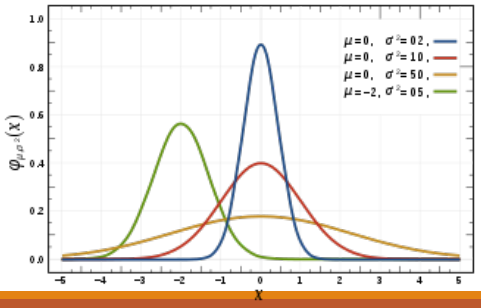
Outlook	Temp	Humidity	Windy	Play
Sunny	85	85	False	**No**
Sunny	80	90	True	**No**
Overcast	83	86	False	**Yes**
Rainy	70	96	False	**Yes**
Rainy	68	80	False	**Yes**
Rainy	65	70	True	**No**
Overcast	64	65	True	**Yes**
Sunny	72	95	False	**No**
Sunny	69	70	False	**Yes**
Rainy	75	80	False	**Yes**
Sunny	75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

32

Dữ liệu liên tục

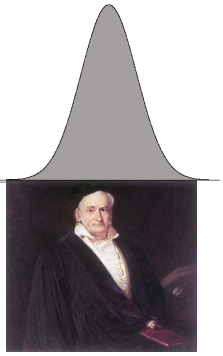
Phân phối chuẩn, còn gọi là **phân phối Gauss**, là một phân phối xác suất cực kì quan trọng trong nhiều lĩnh vực. Nó là họ phân phối có dạng tổng quát giống nhau, chỉ khác tham số *vị trí* (giá trị trung bình μ) và *tỉ lệ* (phương sai σ^2).

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.



33

Dữ liệu liên tục



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

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

Karl Gauss, 1777-1855
great German
mathematician

- *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*

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

https://www.mathsisfun.com/data/standard-deviation.html 34

34