Khoa Công Nghệ Thông Tin
Trường Đại Học Cần Thơ

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.

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 %

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?

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

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

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

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** (/ 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 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 (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.

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

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

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 sự báo

Luật Bayes

Đinh lý xác suất Bayes

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

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

=>

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

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 (learning Phase)- xây dựng mô hình sẳn dùng (tính sẳn 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)

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

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"

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

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	?

Bài tập- cho tập dữ liệu như bảng

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

Data sample
X =(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
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

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 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