Khoa Công Nghệ Thông Tin
Trường Đại Học Cần Thơ

Phương pháp học Bayes
Bayesian classification

Đỗ Thanh Nghị
dtnghi@cit.ctu.edu.vn

Cần Thơ
12-02-2019

Nội dung

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

Nội dung

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

Bayesian classification

- lớp các giải thuật học
  - dựa trên định lý Bayes
  - mạng Bayes và naive Bayes
  - kết quả sinh ra có thể dịch được
  - giải quyết các vấn đề về phân lớp, gom nhóm, etc.
  - được ứng dụng thành công : phân tích dữ liệu, phân loại text, spam, etc.

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
- **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 có điều kiện khi được cho lớp/nhãn
- nhận xét
  - giả thiết các thuộc tính độc lập không bao giờ đúng
  - nhưng trong thực tế, naive Bayes cho kết quả khá tốt ☺

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			**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								

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			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)

Outlook	Temp.	Humidity	Windy	Play
Sunny	Cool	High	True	?

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

Xác suất :

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

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

Định lý Bayes

- Probability of event H given evidence E :

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

- *A priori* probability of H : $\Pr[H]$
  - Probability of event *before* evidence is seen
- *A posteriori* probability of H : $\Pr[H \mid E]$
  - Probability of event *after* evidence is seen

Luật Bayes

- học phân lớp khi có dữ liệu đến
  - Evidence E = dữ liệu
  - Event H = giá trị lớp của dữ liệu
- naïve :

$$\Pr[H \mid E] = \frac{\Pr[E_1 \mid H]\Pr[E_2 \mid H]\ldots\Pr[E_n \mid H]\Pr[H]}{\Pr[E]}$$

Luật Bayes

Outlook	Temp.	Humidity	Windy	Play
Sunny	Cool	High	True	?

← ***Evidence E***

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

xác suất của lớp "yes"

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

Xác suất = 0

- giá trị của thuộc tính không xuất hiện trong tất cả các lớp ("Humidity = high" của lớp "yes")
  - Probability will be zero! $\Pr[Humidity = High \mid yes] = 0$
  - *A posteriori* probability will also be zero! $\Pr[yes \mid E] = 0$
- sử dụng *Laplace estimator*
- xác suất không bao giờ có giá trị 0

Laplace estimator

- ví dụ : thuộc tính *outlook* cho lớp *yes*

$$\frac{2+\mu/3}{9+\mu} \qquad \frac{4+\mu/3}{9+\mu} \qquad \frac{3+\mu/3}{9+\mu}$$

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

- trọng số có thể không bằng nhau, nhưng tổng phải là 1

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

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

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
P(yes) = 0.0238 / (0.0238 + 0.0343) = 41
P(no) = 0.0343 / (0.0238 + 0.0343) = 59

Dữ liệu liên tục

Outlook	Temperature	Humidity	Windy	Play
sunny	85	85	false	no
sunny	80	90	true	no
overcast	83	78	false	yes
rain	70	96	false	yes
rain	68	80	false	yes
rain	65	70	true	no
overcast	64	65	true	yes
sunny	72	95	false	no
sunny	69	70	false	yes
rain	75	80	false	yes
sunny	75	70	true	yes
overcast	72	90	true	yes
overcast	81	75	false	yes
rain	71	80	true	no

Dữ liệu liên tục

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

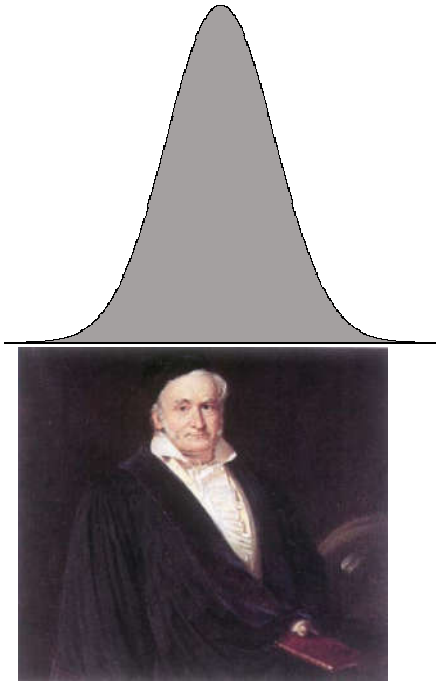
$$\mu = \frac{1}{n}\sum_{i=1}^{n} x_i$$

  - *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

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

- phân lớp

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) = $3/5 \times 0.0291 \times 0.0380 \times 3/5 \times 5/14 = 0.000136$

P(yes) = 0.000036 / (0.000036 + 0. 000136) = 20.9

P(no) = 0.000136 / (0.000036 + 0. 000136) = 79.1

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 có điều kiện (khi được cho nhãn/lớp) 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 (=> 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 !