

Học Máy (IT 4862)

Nguyễn Nhật Quang

quangnn-fit@mail.hut.edu.vn

Trường Đại học Bách Khoa Hà Nội
Viện Công nghệ thông tin và truyền thông
Năm học 2011-2012

Nội dung môn học:

- Giới thiệu chung
- Đánh giá hiệu năng hệ thống học máy
- **Các phương pháp học dựa trên xác suất**
- Các phương pháp học có giám sát
- Các phương pháp học không giám sát
- Lọc cộng tác
- Học tăng cường

Các phương pháp học dựa trên xác suất

- Các phương pháp thống kê cho bài toán phân loại
- Phân loại dựa trên một mô hình xác suất cơ sở
- Việc phân loại dựa trên khả năng xảy ra (probabilities) của các phân lớp
- Các chủ đề chính:
 - Giới thiệu về xác suất
 - Định lý Bayes
 - Xác suất hậu nghiệm cực đại (Maximum a posteriori)
 - Đánh giá khả năng có thể nhất (Maximum likelihood estimation)
 - Phân loại Naïve Bayes
 - Cực đại hóa kỳ vọng (Expectation maximization)

Các khái niệm cơ bản về xác suất

- Giả sử chúng ta có một thí nghiệm (ví dụ: đổ một quân xúc sắc) mà kết quả của nó mang tính ngẫu nhiên (phụ thuộc vào khả năng có thể xảy ra)
- *Không gian các khả năng* S . Tập hợp tất cả các kết quả có thể xảy ra
Ví dụ: $S = \{1, 2, 3, 4, 5, 6\}$ đối với thí nghiệm đổ quân xúc sắc
- *Sự kiện* E . Một tập con của không gian các khả năng
Ví dụ: $E = \{1\}$: kết quả quân súc sắc đổ ra là 1
Ví dụ: $E = \{1, 3, 5\}$: kết quả quân súc sắc đổ ra là một số lẻ
- *Không gian các sự kiện* \mathcal{W} . Không gian (thế giới) mà các kết quả của sự kiện có thể xảy ra
Ví dụ: \mathcal{W} bao gồm tất cả các lần đổ súc sắc
- *Biến ngẫu nhiên* A . Một biến ngẫu nhiên biểu diễn (diễn đạt) một sự kiện, và có một mức độ về khả năng xảy ra sự kiện này

Biểu diễn xác suất

$P(A)$: “Phần của không gian (thế giới) mà trong đó A là đúng”

Không gian sự kiện
của (không gian của
tất cả các giá trị có
thể xảy ra của A)



[<http://www.cs.cmu.edu/~awm/tutorials>]

Các biến ngẫu nhiên 2 giá trị

- Một biến ngẫu nhiên 2 giá trị (nhị phân) có thể nhận một trong 2 giá trị đúng (`true`) hoặc sai (`false`)
- Các tiên đề
 - $0 \leq P(A) \leq 1$
 - $P(\text{true}) = 1$
 - $P(\text{false}) = 0$
 - $P(A \vee B) = P(A) + P(B) - P(A \wedge B)$
- Các hệ quả
 - $P(\text{not } A) \equiv P(\sim A) = 1 - P(A)$
 - $P(A) = P(A \wedge B) + P(A \wedge \sim B)$

Các biến ngẫu nhiên đa trị

Một biến ngẫu nhiên nhiều giá trị có thể nhận một trong số k (>2) giá trị $\{v_1, v_2, \dots, v_k\}$

$$P(A = v_i \wedge A = v_j) = 0 \text{ if } i \neq j$$

$$P(A=v_1 \vee A=v_2 \vee \dots \vee A=v_k) = 1$$

$$P(A = v_1 \vee A = v_2 \vee \dots \vee A = v_i) = \sum_{j=1}^i P(A = v_j)$$

$$\sum_{j=1}^k P(A = v_j) = 1$$

$$P(B \wedge [A = v_1 \vee A = v_2 \vee \dots \vee A = v_i]) = \sum_{j=1}^i P(B \wedge A = v_j)$$

[<http://www.cs.cmu.edu/~awm/tutorials>]

Xác suất có điều kiện (1)

- $P(A | B)$ là phần của không gian (thế giới) mà trong đó A là đúng, với điều kiện (đã biết) là B đúng
- Ví dụ
 - A : Tôi sẽ đi đá bóng vào ngày mai
 - B : Trời sẽ không mưa vào ngày mai
 - $P(A | B)$: Xác suất của việc tôi sẽ đi đá bóng vào ngày mai nếu (đã biết rằng) trời sẽ không mưa (vào ngày mai)

Xác suất có điều kiện (2)

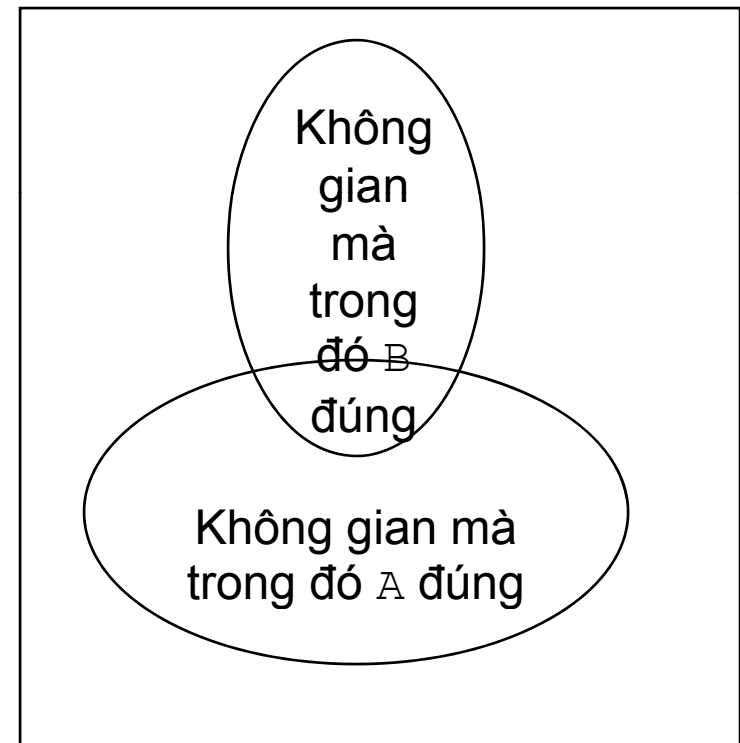
Định nghĩa: $P(A|B) = \frac{P(A, B)}{P(B)}$

Các hệ quả:

$$P(A, B) = P(A|B) \cdot P(B)$$

$$P(A|B) + P(\sim A|B) = 1$$

$$\sum_{i=1}^k P(A = v_i | B) = 1$$



Các biến độc lập về xác suất (1)

- Hai sự kiện A và B được gọi là **độc lập về xác suất** nếu xác suất của sự kiện A là như nhau đối với các trường hợp:
 - Khi sự kiện B xảy ra, hoặc
 - Khi sự kiện B không xảy ra, hoặc
 - Không có thông tin (không biết gì) về việc xảy ra của sự kiện B
- Ví dụ
 - A: Tôi sẽ đi đá bóng vào ngày mai
 - B: Tuấn sẽ tham gia trận đá bóng ngày mai
 - $P(A|B) = P(A)$
 - “Dù Tuấn có tham gia trận đá bóng ngày mai hay không cũng không ảnh hưởng tới quyết định của tôi về việc đi đá bóng ngày mai.”

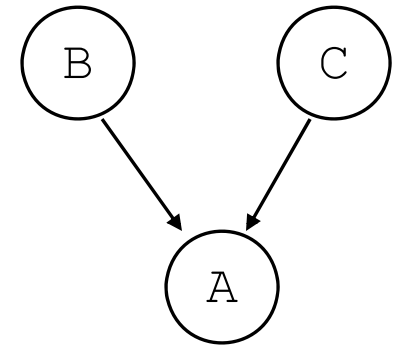
Các biến độc lập về xác suất (2)

Từ định nghĩa của các biến độc lập về xác suất $P(A|B) = P(A)$, chúng ta thu được các luật như sau

- $P(\sim A|B) = P(\sim A)$
- $P(B|A) = P(B)$
- $P(A, B) = P(A) \cdot P(B)$
- $P(\sim A, B) = P(\sim A) \cdot P(B)$
- $P(A, \sim B) = P(A) \cdot P(\sim B)$
- $P(\sim A, \sim B) = P(\sim A) \cdot P(\sim B)$

Xác suất có điều kiện với >2 biến

- $P(A|B, C)$ là xác suất của A đối với (đã biết) B và C
- Ví dụ
 - A: Tôi sẽ đi dạo bờ sông vào sáng mai
 - B: Thời tiết sáng mai rất đẹp
 - C: Tôi sẽ dậy sớm vào sáng mai
 - $P(A|B, C)$: Xác suất của việc tôi sẽ đi dạo dọc bờ sông vào sáng mai, nếu (đã biết rằng) thời tiết sáng mai rất đẹp và tôi sẽ dậy sớm vào sáng mai



$P(A|B, C)$

Độc lập có điều kiện

- Hai biến A và C được gọi là **độc lập có điều kiện** đối với biến B , nếu xác suất của A đối với B bằng xác suất của A đối với B và C
- Công thức định nghĩa: $P(A | B, C) = P(A | B)$
- Ví dụ
 - A : Tôi sẽ đi đá bóng vào ngày mai
 - B : Trận đá bóng ngày mai sẽ diễn ra trong nhà
 - C : Ngày mai trời sẽ không mưa
 - $P(A | B, C) = P(A | B)$
 - Nếu biết rằng trận đấu ngày mai sẽ diễn ra trong nhà, thì xác suất của việc tôi sẽ đi đá bóng ngày mai không phụ thuộc vào thời tiết

Các quy tắc quan trọng của xác suất

■ Quy tắc chuỗi (chain rule)

- $P(A, B) = P(A|B) \cdot P(B) = P(B|A) \cdot P(A)$
- $P(A|B) = P(A, B) / P(B) = P(B|A) \cdot P(A) / P(B)$
- $P(A, B|C) = P(A, B, C) / P(C) = P(A|B, C) \cdot P(B, C) / P(C)$
 $= P(A|B, C) \cdot P(B|C)$

■ Độc lập về xác suất và độc lập có điều kiện

- $P(A|B) = P(A)$; nếu A và B là độc lập về xác suất
- $P(A, B|C) = P(A|C) \cdot P(B|C)$; nếu A và B là độc lập có điều kiện đối với C
- $P(A_1, \dots, A_n|C) = P(A_1|C) \dots P(A_n|C)$; nếu A_1, \dots, A_n là độc lập có điều kiện đối với C

Định lý Bayes

$$P(h | D) = \frac{P(D | h).P(h)}{P(D)}$$

- $P(h)$: Xác suất trước (tiên nghiệm) của giả thiết (phân loại) h
- $P(D)$: Xác suất trước (tiên nghiệm) của việc quan sát được dữ liệu D
- $P(D | h)$: Xác suất (có điều kiện) của việc quan sát được dữ liệu D , nếu biết giả thiết (phân loại) h là đúng
- $P(h | D)$: Xác suất (có điều kiện) của giả thiết (phân loại) h là đúng, nếu quan sát được dữ liệu D
 - **Các phương pháp phân loại dựa trên xác suất sẽ sử dụng xác suất có điều kiện (*posterior probability*) này!**

Định lý Bayes – Ví dụ (1)

Giả sử chúng ta có tập dữ liệu sau (dự đoán 1 người có chơi tennis)?

Ngày	Ngoài trời	Nhiệt độ	Độ ẩm	Gió	Chơi tennis
N1	Nắng	Nóng	Cao	Yếu	Không
N2	Nắng	Nóng	Cao	Mạnh	Không
N3	Âm u	Nóng	Cao	Yếu	Có
N4	Mưa	Bình thường	Cao	Yếu	Có
N5	Mưa	Mát mẻ	Bình thường	Yếu	Có
N6	Mưa	Mát mẻ	Bình thường	Mạnh	Không
N7	Âm u	Mát mẻ	Bình thường	Mạnh	Có
N8	Nắng	Bình thường	Cao	Yếu	Không
N9	Nắng	Mát mẻ	Bình thường	Yếu	Có
N10	Mưa	Bình thường	Bình thường	Yếu	Có
N11	Nắng	Bình thường	Bình thường	Mạnh	Có
N12	Âm u	Bình thường	Cao	Mạnh	Có

[Mitchell, 1997]

Học Máy – IT 4862

Định lý Bayes – Ví dụ (2)

- Dữ liệu D . *Ngoài trời là nắng và Gió là mạnh*
- Giả thiết (phân loại) h . Anh ta chơi tennis
- Xác suất trước $P(h)$. Xác suất rằng anh ta chơi tennis (bất kể *Ngoài trời* như thế nào và *Gió* ra sao)
- Xác suất trước $P(D)$. Xác suất rằng *Ngoài trời là nắng và Gió là mạnh*
- $P(D|h)$. Xác suất *Ngoài trời là nắng và Gió là mạnh*, nếu biết rằng anh ta chơi tennis
- $P(h|D)$. Xác suất anh ta chơi tennis, nếu biết rằng *Ngoài trời là nắng và Gió là mạnh*
 - Chúng ta quan tâm đến giá trị xác suất sau (*posterior probability*) này!

Xác suất hậu nghiệm cực đại (MAP)

- Với một tập các giả thiết (các phân lớp) có thể H , hệ thống học sẽ tìm **giả thiết có thể xảy ra nhất (the most probable hypothesis)** $h \in H$ đối với các dữ liệu quan sát được D
- Giả thiết h này được gọi là giả thiết có xác suất hậu nghiệm cực đại (**Maximum a posteriori – MAP**)

$$h_{MAP} = \arg \max_{h \in H} P(h | D)$$

$$h_{MAP} = \arg \max_{h \in H} \frac{P(D | h).P(h)}{P(D)} \quad (\text{bởi định lý Bayes})$$

$$h_{MAP} = \arg \max_{h \in H} P(D | h).P(h) \quad (P(D) \text{ là như nhau đối với các giả thiết } h)$$

MAP – Ví dụ

- Tập H bao gồm 2 giả thiết (có thể)
 - h_1 : Anh ta chơi tennis
 - h_2 : Anh ta không chơi tennis
- Tính giá trị của 2 xác suất có điều kiện: $P(h_1 | D)$, $P(h_2 | D)$
- Giả thiết có thể nhất $h_{MAP} = h_1$ nếu $P(h_1 | D) \geq P(h_2 | D)$; ngược lại thì $h_{MAP} = h_2$
- Bởi vì $P(D) = P(D, h_1) + P(D, h_2)$ là như nhau đối với cả 2 giả thiết h_1 và h_2 , nên có thể bỏ qua đại lượng $P(D)$
- Vì vậy, cần tính 2 biểu thức: $P(D | h_1) \cdot P(h_1)$ và $P(D | h_2) \cdot P(h_2)$, và đưa ra quyết định tương ứng
 - Nếu $P(D | h_1) \cdot P(h_1) \geq P(D | h_2) \cdot P(h_2)$, thì kết luận là anh ta chơi tennis
 - Ngược lại, thì kết luận là anh ta không chơi tennis

Đánh giá khả năng có thể nhất (MLE)

- Phương pháp MAP: Với một tập các giả thiết có thể H , cần tìm một giả thiết cực đại hóa giá trị: $P(D|h) \cdot P(h)$
- Giả sử (assumption) trong phương pháp **đánh giá khả năng có thể nhất (Maximum likelihood estimation – MLE)**: Tất cả các giả thiết đều có giá trị xác suất trước như nhau: $P(h_i) = P(h_j)$, $\forall h_i, h_j \in H$
- Phương pháp MLE tìm giả thiết cực đại hóa giá trị $P(D|h)$; trong đó $P(D|h)$ được gọi là *khả năng có thể (likelihood)* của dữ liệu D đối với h
- Giả thiết có khả năng nhất (maximum likelihood hypothesis)

$$h_{ML} = \arg \max_{h \in H} P(D|h)$$

MLE – Ví dụ

- Tập H bao gồm 2 giả thiết có thể
 - h_1 : Anh ta chơi tennis
 - h_2 : Anh ta không chơi tennis
- D: Tập dữ liệu (các ngày) mà trong đó thuộc tính *Outlook* có giá trị *Sunny* và thuộc tính *Wind* có giá trị *Strong*
- Tính 2 giá trị khả năng xảy ra (likelihood values) của dữ liệu D đối với 2 giả thiết: $P(D|h_1)$ và $P(D|h_2)$
 - $P(\text{Outlook}=\text{Sunny}, \text{Wind}=\text{Strong}|h_1) = 1/8$
 - $P(\text{Outlook}=\text{Sunny}, \text{Wind}=\text{Strong}|h_2) = 1/4$
- Giả thiết MLE $h_{\text{MLE}}=h_1$ nếu $P(D|h_1) \geq P(D|h_2)$; và ngược lại thì $h_{\text{MLE}}=h_2$
 - Bởi vì $P(\text{Outlook}=\text{Sunny}, \text{Wind}=\text{Strong}|h_1) < P(\text{Outlook}=\text{Sunny}, \text{Wind}=\text{Strong}|h_2)$, hệ thống kết luận rằng:
Anh ta sẽ không chơi tennis!

Phân loại Naïve Bayes (1)

- Biểu diễn bài toán phân loại (classification problem)
 - Một tập học D_{train} , trong đó mỗi ví dụ học x được biểu diễn là một vector n chiều: (x_1, x_2, \dots, x_n)
 - Một tập xác định các nhãn lớp: $C = \{c_1, c_2, \dots, c_m\}$
 - Với một ví dụ (mới) z , thì z sẽ được phân vào lớp nào?
- Mục tiêu: Xác định phân lớp có thể (phù hợp) nhất đối với z

$$c_{MAP} = \arg \max_{c_i \in C} P(c_i | z)$$

$$c_{MAP} = \arg \max_{c_i \in C} P(c_i | z_1, z_2, \dots, z_n)$$

$$c_{MAP} = \arg \max_{c_i \in C} \frac{P(z_1, z_2, \dots, z_n | c_i) \cdot P(c_i)}{P(z_1, z_2, \dots, z_n)} \quad (\text{bởi định lý Bayes})$$

Phân loại Naïve Bayes (2)

- Để tìm được phân lớp có thể nhất đối với $z \dots$

$$c_{MAP} = \arg \max_{c_i \in C} P(z_1, z_2, \dots, z_n | c_i) \cdot P(c_i) \quad (P(z_1, z_2, \dots, z_n) \text{ là như nhau với các lớp})$$

- **Giả sử (assumption) trong phương pháp phân loại Naïve Bayes.** Các thuộc tính là *độc lập có điều kiện* (*conditionally independent*) đối với các lớp

$$P(z_1, z_2, \dots, z_n | c_i) = \prod_{j=1}^n P(z_j | c_i)$$

- Phân loại Naïve Bayes tìm phân lớp có thể nhất đối với z

$$c_{NB} = \arg \max_{c_i \in C} P(c_i) \cdot \prod_{j=1}^n P(z_j | c_i)$$

Phân loại Naïve Bayes – Giải thuật

- Giai đoạn học (training phase), sử dụng một tập học
 - Đối với mỗi phân lớp có thể (mỗi nhãn lớp) $c_i \in C$
 - Tính giá trị xác suất trước: $P(c_i)$
 - Đối với mỗi giá trị thuộc tính x_j , tính giá trị xác suất xảy ra của giá trị thuộc tính đó đối với một phân lớp c_i : $P(x_j | c_i)$
- Giai đoạn phân lớp (classification phase), đối với một ví dụ mới
 - Đối với mỗi phân lớp $c_i \in C$, tính giá trị của biểu thức:

$$P(c_i) \cdot \prod_{j=1}^n P(x_j | c_i)$$

- Xác định phân lớp của z là lớp có thể nhất c^*

$$c^* = \arg \max_{c_i \in C} P(c_i) \cdot \prod_{j=1}^n P(x_j | c_i)$$

Phân lớp Naïve Bayes – Ví dụ (1)

Một sinh viên trẻ với thu nhập trung bình và mức đánh giá tín dụng bình thường sẽ mua một cái máy tính?

Rec. ID	Age	Income	Student	Credit_Rating	Buy_Computer
1	Young	High	No	Fair	No
2	Young	High	No	Excellent	No
3	Medium	High	No	Fair	Yes
4	Old	Medium	No	Fair	Yes
5	Old	Low	Yes	Fair	Yes
6	Old	Low	Yes	Excellent	No
7	Medium	Low	Yes	Excellent	Yes
8	Young	Medium	No	Fair	No
9	Young	Low	Yes	Fair	Yes
10	Old	Medium	Yes	Fair	Yes
11	Young	Medium	Yes	Excellent	Yes
12	Medium	Medium	No	Excellent	Yes
13	Medium	High	Yes	Fair	Yes
14	Old	Medium	No	Excellent	No

Phân lớp Naïve Bayes – Ví dụ (2)

■ Biểu diễn bài toán phân loại

- $z = (\text{Age}=\text{Young}, \text{Income}=\text{Medium}, \text{Student}=\text{Yes}, \text{Credit_Rating}=\text{Fair})$
- Có 2 phân lớp có thể: c_1 (“Mua máy tính”) và c_2 (“Không mua máy tính”)

■ Tính giá trị xác suất trước cho mỗi phân lớp

- $P(c_1) = 9/14$
- $P(c_2) = 5/14$

■ Tính giá trị xác suất của mỗi giá trị thuộc tính đối với mỗi phân lớp

- | | |
|---|--|
| • $P(\text{Age}=\text{Young} c_1) = 2/9;$ | $P(\text{Age}=\text{Young} c_2) = 3/5$ |
| • $P(\text{Income}=\text{Medium} c_1) = 4/9;$ | $P(\text{Income}=\text{Medium} c_2) = 2/5$ |
| • $P(\text{Student}=\text{Yes} c_1) = 6/9;$ | $P(\text{Student}=\text{Yes} c_2) = 1/5$ |
| • $P(\text{Credit_Rating}=\text{Fair} c_1) = 6/9;$ | $P(\text{Credit_Rating}=\text{Fair} c_2) = 2/5$ |

Phân lớp Naïve Bayes – Ví dụ (3)

- Tính toán xác suất có thể xảy ra (likelihood) của ví dụ z đối với mỗi phân lớp

- Đối với phân lớp c_1

$$P(z|c_1) = P(\text{Age}=\text{Young}|c_1).P(\text{Income}=\text{Medium}|c_1).P(\text{Student}=\text{Yes}|c_1).$$

$$P(\text{Credit_Rating}=\text{Fair}|c_1) = (2/9).(4/9).(6/9).(6/9) = 0.044$$

- Đối với phân lớp c_2

$$P(z|c_2) = P(\text{Age}=\text{Young}|c_2).P(\text{Income}=\text{Medium}|c_2).P(\text{Student}=\text{Yes}|c_2).$$

$$P(\text{Credit_Rating}=\text{Fair}|c_2) = (3/5).(2/5).(1/5).(2/5) = 0.019$$

- Xác định phân lớp có thể nhất (the most probable class)

- Đối với phân lớp c_1

$$P(c_1).P(z|c_1) = (9/14).(0.044) = 0.028$$

- Đối với phân lớp c_2

$$P(c_2).P(z|c_2) = (5/14).(0.019) = 0.007$$

→ Kết luận: *Anh ta (z) sẽ mua một máy tính!*

Phân lớp Naïve Bayes – Vấn đề (1)

- Nếu không có ví dụ nào gắn với phân lớp c_i có giá trị thuộc tính $x_j \dots$

$P(x_j | c_i) = 0$, và vì vậy:
$$P(c_i) \cdot \prod_{j=1}^n P(x_j | c_i) = 0$$

- Giải pháp: Sử dụng phương pháp Bayes để ước lượng $P(x_j | c_i)$

$$P(x_j | c_i) = \frac{n(c_i, x_j) + mp}{n(c_i) + m}$$

- $n(c_i)$: số lượng các ví dụ học gắn với phân lớp c_i
- $n(c_i, x_j)$: số lượng các ví dụ học gắn với phân lớp c_i có giá trị thuộc tính x_j
- p : ước lượng đối với giá trị xác suất $P(x_j | c_i)$
 - Các ước lượng đồng mức: $p = 1/k$, với thuộc tính x_j có k giá trị có thể
- m : một hệ số (trọng số)
 - Để bổ sung cho $n(c_i)$ các ví dụ thực sự được quan sát với thêm m mẫu ví dụ với ước lượng p

Phân lớp Naïve Bayes – Vấn đề (2)

■ Giới hạn về độ chính xác trong tính toán của máy tính

- $P(x_j | c_i) < 1$, đối với mọi giá trị thuộc tính x_j và phân lớp c_i
- Vì vậy, khi số lượng các giá trị thuộc tính là rất lớn, thì:

$$\lim_{n \rightarrow \infty} \left(\prod_{j=1}^n P(x_j | c_i) \right) = 0$$

■ Giải pháp: Sử dụng hàm lôgarit cho các giá trị xác suất

$$c_{NB} = \arg \max_{c_i \in C} \left(\log \left[P(c_i) \cdot \prod_{j=1}^n P(x_j | c_i) \right] \right)$$

$$c_{NB} = \arg \max_{c_i \in C} \left(\log P(c_i) + \sum_{j=1}^n \log P(x_j | c_i) \right)$$

Phân loại văn bản bằng NB (1)

■ Biểu diễn bài toán phân loại văn bản

- Tập học D_{train} , trong đó mỗi ví dụ học là một biểu diễn văn bản gắn với một nhãn lớp: $D = \{(d_k, c_i)\}$
- Một tập các nhãn lớp xác định: $C = \{c_i\}$

■ Giai đoạn học

- Từ tập các văn bản trong D_{train} , trích ra tập các từ khóa (keywords/terms): $T = \{t_j\}$
- Gọi $D_{c_i} (\subseteq D_{\text{train}})$ là tập các văn bản trong D_{train} có nhãn lớp c_i
- Đối với mỗi phân lớp c_i
 - Tính giá trị xác suất trước của phân lớp c_i : $P(c_i) = \frac{|D_{c_i}|}{|D|}$
 - Đối với mỗi từ khóa t_j , tính xác suất từ khóa t_j xuất hiện đối với lớp c_i

$$P(t_j | c_i) = \frac{\left(\sum_{d_k \in D_{c_i}} n(d_k, t_j)\right) + 1}{\left(\sum_{d_k \in D_{c_i}} \sum_{t_m \in T} n(d_k, t_m)\right) + |T|}$$

$(n(d_k, t_j))$: số lần xuất hiện của từ khóa t_j trong văn bản d_k

Phân loại văn bản bằng NB (2)

- Giai đoạn phân lớp đối với một văn bản mới d
 - Từ văn bản d , trích ra tập T_d gồm các từ khóa (keywords) t_j đã được định nghĩa trong tập T ($T_d \subseteq T$)
 - **Giả sử (assumption).** Xác suất từ khóa t_j xuất hiện đối với lớp c_i là độc lập đối với vị trí của từ khóa đó trong văn bản

$$P(t_j \text{ ở vị trí } k | c_i) = P(t_j \text{ ở vị trí } m | c_i), \quad \forall k, m$$

- Đối với mỗi phân lớp c_i , tính giá trị likelihood của văn bản d đối với c_i

$$P(c_i) \cdot \prod_{t_j \in T_d} P(t_j | c_i)$$

- Phân lớp văn bản d thuộc vào lớp c^*

$$c^* = \arg \max_{c_i \in C} P(c_i) \cdot \prod_{t_j \in T_d} P(t_j | c_i)$$

Phân loại Naïve Bayes – Tổng kết

- Một trong các phương pháp học máy được áp dụng phổ biến nhất trong thực tế
- Dựa trên định lý Bayes
- Việc phân loại dựa trên các giá trị xác suất của các khả năng xảy ra của các giả thiết (phân loại)
- Mặc dù đặt giả sử về sự độc lập có điều kiện của các thuộc tính đối với các phân lớp, nhưng phương pháp phân loại Naïve Bayes vẫn thu được các kết quả phân loại tốt trong nhiều lĩnh vực ứng dụng thực tế
- Khi nào nên sử dụng?
 - Có một tập huấn luyện có kích thước lớn hoặc vừa
 - Các ví dụ được biểu diễn bởi một số lượng lớn các thuộc tính
 - **Các thuộc tính độc lập có điều kiện đối với các phân lớp**

Cực đại hóa kỳ vọng – EM (1)

- Trong một số bài toán học máy, dữ liệu chỉ có thể quan sát được một phần
 - Đối với một thuộc tính, giá trị của nó không quan sát được (unobservable) – nhưng chúng ta biết dạng tổng quát của hàm phân bố xác suất đối với thuộc tính đó
 - Đối với một thuộc tính, giá trị của nó có lúc quan sát được có lúc không quan sát được (vd: thuộc tính thiếu giá trị)
 - Giá trị của thuộc tính phân loại (nhãn lớp) là không biết

Cực đại hóa kỳ vọng – EM (2)

- Phương pháp học EM thường được sử dụng trong những bài toán tồn tại các biến (thuộc tính) không quan sát được (ẩn)
 - EM tìm các đánh giá có thể nhất (maximum likelihood estimates) của các tham số trong một mô hình xác suất phụ thuộc vào các biến không quan sát được (unobserved variables)
 - Không sử dụng tốc độ học (vd: như phương pháp học mạng nơ-ron nhân tạo)
 - Đảm bảo tìm được một giá trị tối ưu cục bộ (a local optimum) của xác suất likelihood, cùng với các giá trị ước lượng được của các biến không quan sát được

EM – Phát biểu bài toán (1)

- Xét một tập dữ liệu gồm m ví dụ $X = \{x_1, \dots, x_m\}$, trong đó mỗi ví dụ có một phần (một tập các thuộc tính) quan sát được và một phần không quan sát được
 - Phần quan sát được: $Y = \{y_1, \dots, y_m\}$
 - Phần không quan sát được: $Z = \{z_1, \dots, z_m\}$
 - Đối với mỗi ví dụ: $x_i = y_i \cup z_i$
- Tập dữ liệu X được sinh ra (generated) bởi một hàm phân bố xác suất cơ sở – $P(X | \theta)$, và hàm phân bố này phụ thuộc vào (được biểu diễn bằng) một tập các tham số θ (chưa biết, và sẽ được ước lượng!)
- Phần dữ liệu không quan sát được Z được xem như một biến ngẫu nhiên mà hàm phân bố xác suất của nó phụ thuộc vào
 - Các tham số chưa biết giá trị θ , và
 - Phần dữ liệu quan sát được Y

EM – Phát biểu bài toán (2)

- Phương pháp EM lặp lại 2 bước sau đây
 - *Tính toán giá trị kỳ vọng (Expectation step)*. Với các giá trị được ước lượng hiện tại của các tham số θ , tính toán các giá trị kỳ vọng của các biến không quan sát được
 - *Cực đại hóa (Maximization step)*. Với các giá trị kỳ vọng được gán cho các biến không quan sát được (tính ở bước trên – E-step), tính toán lại các đánh giá có thể nhất (maximum likelihood estimates) của các tham số θ
- Ký hiệu $E [P (X | \theta)]$ là giá trị kỳ vọng của khả năng có thể (likelihood) của tập dữ liệu X , đối với các giá trị ước lượng hiện tại của các tham số θ
 - Giá trị (trung bình) kỳ vọng được tính toán dựa trên các giá trị có thể của phần dữ liệu không quan sát được Z ,
 - Gán trọng số (weighting) mỗi giá trị có thể với xác suất xảy ra của giá trị đó
- Giải thuật EM tìm các đánh giá có thể nhất (maximum likelihood estimates) θ^* giúp cực đại hóa (cực bộ) giá trị $E [P (X | \theta)]$

Cực đại hóa kỳ vọng – Giải thuật

- Ký hiệu $\theta^{(i)}$ là các giá trị ước lượng của các tham số θ tại bước lặp thứ i
- Ký hiệu $Q(\theta | \theta^{(i)})$ là một hàm của θ , đối với các đánh giá (ước lượng) hiện thời $\theta^{(i)}$ và phần dữ liệu quan sát được Y (của tập dữ liệu X):

$$Q(\theta | \theta^{(i)}) = E[\ln P(X | \theta) | \theta^{(i)}, Y]$$

- Giải thuật EM tổng quát:

Initialize $\theta^{(0)}$, Thres_Val, $i=0$

Do

$i=i+1$

Compute $Q(\theta | \theta^{(i)})$

// Bước tính kỳ vọng (E-step)

$\theta^{(i+1)} = \operatorname{argmax}_{\theta} Q(\theta | \theta^{(i)})$

// Bước cực đại hóa (M-step)

Until $Q(\theta^{(i+1)} | \theta^{(i)}) - Q(\theta^{(i)} | \theta^{(i-1)}) \leq \text{Thres_Val}$

// Không có cải thiện

Return $\theta^{(i+1)}$

Phân loại văn bản bằng EM (1)

- Trong thực tế, chúng ta thường có một tập nhỏ các văn bản có nhãn lớp và một tập lớn hơn các văn bản không có nhãn lớp

Tập các văn bản $D = D^L \cup D^U$

- D^L : Tập các văn bản có nhãn lớp
 - D^U : Tập các văn bản không có nhãn lớp
 - $|D^L| \ll |D^U|$
- Mục tiêu: Gán nhãn lớp chính xác cho các văn bản trong tập D^U
 - Tại sao chúng ta không xây dựng (học) một bộ phân loại dựa trên tập các văn bản có nhãn trong D^L , và sau đó sử dụng bộ phân loại đó để phân loại các văn bản không có nhãn?
 - Cần khai thác các thông tin ẩn trong tập D^U để nâng cao hiệu năng của bộ phân loại học được
 - Khai thác các phân bố kết hợp của các từ khóa đối với các phân lớp
 - Khai thác sự tương đồng giữa các văn bản (vd: một văn bản có nhãn và một văn bản không có nhãn cùng chứa một số các từ khóa)

Phân loại văn bản bằng EM (2)

- Trong bài toán phân loại văn bản
 - Các tham số chưa biết giá trị là: $P(c_i)$ và $P(t_j | c_i)$
 - Biến không quan sát được là: nhãn lớp của các văn bản trong D^U
- Giải pháp: Kết hợp EM với phân loại Naïve Bayes
- Sử dụng tập dữ liệu có nhãn D^L và phương pháp phân loại Naïve Bayes để khởi tạo các giá trị của các tham số chưa biết giá trị:
 $P^{(0)}(c_i)$ và $P^{(0)}(t_j | c_i)$
- Ở bước lặp thứ k
 - *Bước tính kỳ vọng (E-step)*. Đối với mỗi văn bản không có nhãn d^U và mỗi lớp c_i , sử dụng các giá trị ước lượng hiện tại $P^{(k-1)}(c_i)$ và $P^{(k-1)}(t_j | c_i)$ và phương pháp phân loại Naïve Bayes để đánh giá $P^{(k)}(c_i | d^U)$:

$$P^{(k)}(c_i | d^U) = \frac{P^{(k-1)}(d^U | c_i) \cdot P^{(k-1)}(c_i)}{P^{(k-1)}(d^U)} = \frac{P^{(k-1)}(c_i) \cdot \prod_{t_j} P^{(k-1)}(t_j | c_i)}{\sum_{c_q \in C} P^{(k-1)}(d^U | c_q) \cdot P^{(k-1)}(c_q)}$$

Phân loại văn bản bằng EM (3)

■ Ở bước lặp thứ k (...tiếp tục)

- *Bước cực đại hóa (M-step).* Đối với mỗi từ khóa t_j và mỗi phân lớp c_i , đánh giá lại các tham số của phân bố xác suất (của toàn bộ tập văn bản D)

$$P^{(k)}(t_j | c_i) = \frac{1 + \sum_{d \in D} P^{(k)}(c_i | d) \cdot n(d, t_j)}{|T| + \sum_{d \in D} \sum_{t_q \in T} P^{(k)}(c_i | d) \cdot n(d, t_q)}$$

$$P^{(k)}(c_i) = \frac{1}{|D|} \sum_{d \in D} P^{(k)}(c_i | d)$$

–Đóng góp của giá trị của tần suất xuất hiện từ khóa $n(d, t_j)$ trong văn bản d đối với phân lớp c_i được gán trọng số (weighted) bởi $P(c_i | d)$ – giá trị xác suất của văn bản d thuộc vào lớp c_i

–Đối với một văn bản có nhãn d ($\in D^L$) và nhãn lớp gán với nó c_d ,

$$P(c_i | d) = 1 \text{ if } c_i \equiv c_d, \text{ and } P(c_i | d) = 0 \text{ if } c_i \neq c_d$$

■ Lặp lại 2 bước (*E-step*, *M-step*) cho đến khi thỏa mãn đk hội tụ

Ví dụ: % của thay đổi về nhãn lớp được gán cho các văn bản không có nhãn (trong D^U) giữa 2 bước lặp kế tiếp nhỏ hơn một giá trị ngưỡng.

Các phương pháp học dựa trên xác suất

- Dựa trên lý thuyết xác suất
- Nhằm học (xấp xỉ) một mô hình xác suất sinh ra dữ liệu
- Định lý Bayes đóng vai trò trung tâm
- Kết hợp tri thức tiên nghiệm (các xác suất tiên nghiệm) với dữ liệu quan sát được
- Tính toán trực tiếp các xác suất xảy ra đối với các giả thiết (phân lớp)
- Cung cấp các giải thuật học máy có tính ứng dụng thực tế, như phân loại Naïve Bayes hoặc EM
- Cung cấp cơ sở lý thuyết (khái niệm) để hiểu và đánh giá các giải thuật học máy khác