Chương 5 - Học máy

Lê Thanh Hương
Viện CNTT & TT - ĐHBK HN

Nội dung môn học

Chương 1. Tổng quan

Chương 2. Tác tử thông minh

Chương 3. Giải quyết vấn đề

Chương 4. Tri thức và suy diễn

Chương 5. Học máy

- Tổng quan
- Phân lớp Naïve Bayes
- Học cây quyết định
- K láng giềng gần
- Mạng nơron

Phân lớp Naïve Bayes

- Là các phương pháp học phân lớp có giám sát và dựa trên xác suất
- Dựa trên một mô hình (hàm) xác suất
- Việc phân loại dựa trên các giá trị xác suất của các khả năng xảy ra của các giả thiết
- Là một trong các phương pháp học máy thường được sử dụng trong các bài toán thực tế
- Dựa trên định lý Bayes (Bayes theorem)

Các khái niệm cơ bản về xác suất

- Giả sử hành động A_t = Khởi hành từ nhà đến sân bay trước t phút so với giờ khởi hành của chuyến bay
- Hành động A_t cho phép tôi đến sân bay đúng giờ hay không?
- Các vấn đề có thể xảy ra:
  - khả năng quan sát không đầy đủ (ví dụ: về tình hình giao thông trên đường, …)
  - lỗi và nhiễu của các bộ cảm biến (giúp cập nhật thông tin về tình hình giao thông)
  - sự không chắc chắn trong các kết quả của các hành động (ví dụ: lốp bị hết hơi, …)
  - sự phức tạp của việc mô hình hóa và dự đoán tình hình giao thông
- Hành động A_{25} (xuất phát trước 25 phút) sẽ cho phép tôi đến sân bay kịp giờ chuyến bay, nếu:
  - không có tai nạn trên cầu (mà tôi sẽ đi qua),
  - trời không mưa
  - lốp xe tôi vẫn căng,
  - …

Biểu diễn xác suất

P(A): "Phần của không gian 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 Bool

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

Biến ngẫu nhiên nhiều giá trị

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

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

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

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

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

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

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

Xác suất có điều kiện

- P(A|B) là xác suất A đúng, với điều kiện B đúng.
  - 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 trời sẽ không mưa (vào ngày mai)

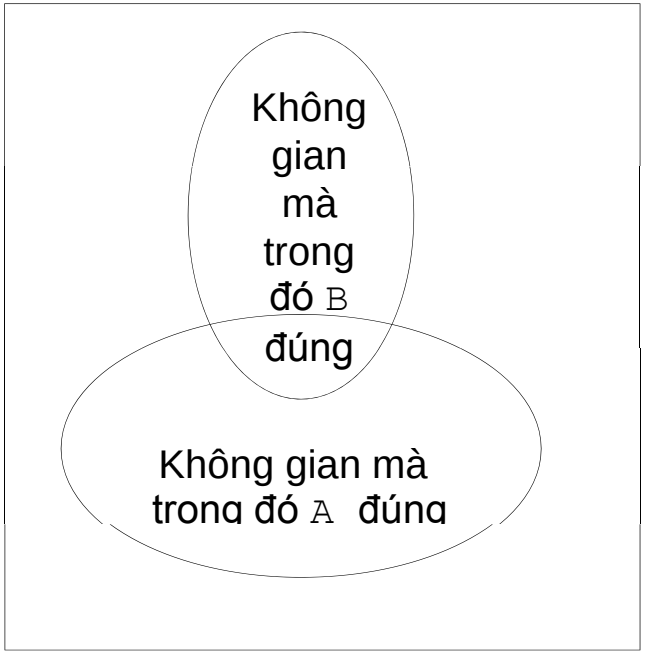
Định nghĩa: $P(A|B) = \frac{P(A,B)}{P(B)}$

Các hệ quả:

$P(A,B) = P(A|B) . P(B)$

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

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



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)

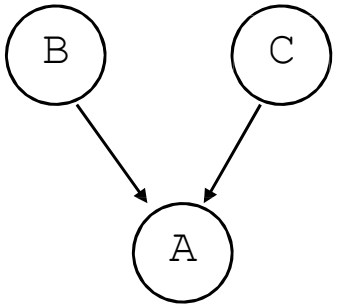
Từ định nghĩa của các biến độc lập về xác suất P(A|B)=P(A), chúng ta thu được các luật như sau

- P(~A|B) = P(~A)
- P(B|A) = P(B)
- P(A,B) = P(A). P(B)
- P(~A,B) = P(~A). P(B)
- P(A,~B) = P(A). P(~B)
- P(~A,~B) = P(~A). P(~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)
  - $P(A,B) = P(A|B).P(B) = P(B|A).P(A)$
  - $P(A|B) = P(A,B)/P(B) = P(B|A).P(A)/P(B)$
  - $P(A,B|C) = P(A,B,C)/P(C) = P(A|B,C).P(B,C)/P(C)$

$= P(A|B,C).P(B|C)$

- 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
  - $P(A,B|C) = P(A|C).P(B|C)$; nếu A và B là độc lập có điều kiện đối với C
  - $P(A_1,\ldots,A_n|C) = P(A_1|C)\ldots P(A_n|C)$; nếu $A_1,\ldots,A_n$ là độc lập có điều kiện đối với C

Định lý Bayes

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

- P(h): Xác suất trước (prior probability) rằng giả thiết (phân lớp) h là đúng
- P(D): Xác suất trước rằng tập dữ liệu D được quan sát
- P(D|h): Xác suất của việc quan sát được tập dữ liệu D, với điều kiện giả thiết h là đúng
- P(h|D): Xác suất của giả thiết h là đúng, với điều kiện tập dữ liệu D được quan sát

Định lý Bayes – Ví dụ (1)

Xét tập dữ liệu sau đây:

Day	Outlook	Temperature	Humidity	Wind	Play Tennis
D1	Sunny	Hot	High	Weak	No
D2	Sunny	Hot	High	Strong	No
D3	Overcast	Hot	High	Weak	Yes
D4	Rain	Mild	High	Weak	Yes
D5	Rain	Cool	Normal	Weak	Yes
D6	Rain	Cool	Normal	Strong	No
D7	Overcast	Cool	Normal	Strong	Yes
D8	Sunny	Mild	High	Weak	No
D9	Sunny	Cool	Normal	Weak	Yes
D10	Rain	Mild	Normal	Weak	Yes
D11	Sunny	Mild	Normal	Strong	Yes
D12	Overcast	Mild	High	Strong	Yes

[Mitchell, 1997]

Định lý Bayes – Ví dụ (2)

- Tập ví dụ `D`. Tập các ngày mà thuộc tính *Outlook* có giá trị *Sunny* và thuộc tính *Wind* có giá trị *Strong*
- Giả thiết (phân lớp) `h`. Anh ta chơi tennis
- Xác suất trước `P(h)`. Xác suất anh ta chơi tennis (không phụ thuộc vào các thuộc tính *Outlook* và *Wind*)
- Xác suất trước `P(D)`. Xác suất của một ngày mà thuộc tính *Outlook* có giá trị *Sunny* và thuộc tính *Wind* có giá trị *Strong*
- `P(D|h)`. Xác suất của một ngày mà thuộc tính *Outlook* có giá trị *Sunny* và Wind có giá trị *Strong*, với điều kiện (nếu biết rằng) anh ta chơi tennis
- `P(h|D)`. Xác suất anh ta chơi tennis, với điều kiện (nếu biết rằng) thuộc tính *Outlook* có giá trị *Sunny* và Wind có giá trị *Strong*

➢ Phương pháp phân lớp Naïve Bayes dựa trên xác suất có điều kiện (*posterior probability*) này!

Cực đại hóa xác suất có điều kiện

- 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(∈H)` đối với các dữ liệu quan sát được `D`
- Giả thiết `h` này được gọi là giả thiết cực đại hóa xác suất có điều kiện (**maximum a posteriori – MAP**)

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

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

(bởi định lý Bayes)

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

(`P(D)` là như nhau đối với các giả thiết `h`)

MAP – Ví dụ

- Tập H bao gồm 2 giả thiết (có thể)
  - h : Anh ta chơi tennis
  - h_2: Anh ta không chơi tennis
- Tính giá trị của 2 xác xuất 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: $P(D|h_1).P(h_1)$ và $P(D|h_2).P(h_2)$, và đưa ra quyết định tương ứng
  - Nếu $P(D|h_1).P(h_1) \geq P(D|h_2).P(h_2)$, thì kết luận là anh ta chơi tennis
  - Ngược lại, thì kết luận là anh ta không chơi tennis

Đánh giá khả năng xảy ra cao nhất

- Phương pháp MAP: Với một tập các giả thiết có thể H, cần tìm một giả thiết cực đại hóa giá trị: $P(D|h).P(h)$
- Giả sử (assumption) trong phương pháp **đánh giá khả năng xảy ra cao 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 xảy ra (likelihood)* của dữ liệu D đối với h
- Giả thiết cực đại hóa khả năng xảy ra (maximum likelihood hypothesis)

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

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 $P(D|h_2)$
  - $P(\text{Outlook=Sunny, Wind=Strong}|h_1) = 1/8$
  - $P(\text{Outlook=Sunny, Wind=Strong}|h_2) = 1/4$
- Giả thiết MLE $h_{MLE}=h_1$ nếu $P(D|h_1) \geq P(D|h_2)$;
- và ngược lại thì $h_{MLE}=h_2$

Ō Bởi vì $P(\text{Outlook=Sunny, Wind=Strong}|h_1) < P(\text{Outlook=Sunny, Wind=Strong}|h_2)$, 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 vectơ `n` chiều: $(x_1, x_2, ..., x_n)$
  - Một tập xác định các nhãn lớp: $C=\{c_1, c_2, ..., c_m\}$
  - Với một ví dụ (mới) `z`, `z` sẽ được phân vào lớp nào?
  - Mục tiêu: Xác định phân lớp có thể (phù hợp) nhất đối với `z`

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

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

$$c_{MAP} = \arg\max_{c_i \in C} \frac{P(z_1, z_2, ..., z_n \mid c_i).P(c_i)}{P(z_1, z_2, ..., z_n)}$$

(bởi định lý Bayes)

Phân loại Naïve Bayes (2)

- Để tìm được phân lớp có thể nhất đối với z…

$$c_{MAP} = \arg\max_{c_i \in C} P(z_1, z_2, \ldots, z_n \mid c_i).P(c_i)$$

($P(z_1, z_2, \ldots, z_n)$ 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

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

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

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

Phân loại Naïve Bayes – Giải thuật

- Giai đoạn học, sử dụng một tập học
- Đối với mỗi phân lớp có thể (mỗi nhãn lớp) 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: $P(x_j|c_i)$
  - Tính giá trị xác suất trước: $P(c_i)$
- Giai đoạn phân lớp, đối với một ví dụ mới
  - Đối với mỗi phân lớp c_i, tính giá trị của biểu thức:

$$P(c_i).\prod_{j=1}^{n} P(x_j \mid c_i)$$

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

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

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?

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 =(Age=Young,Income=Medium,Student=Yes,Credit_Rating=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

  - $P(Age=Young|c_1) = 2/9$;
  - $P(Income=Medium|c_1) = 4/9$;
  - $P(Student=Yes|c_1) = 6/9$;
  - $P(Credit\_Rating=Fair|c_1) = 6/9$;

$P(Age=Young|c_2) = 3/5$

$P(Income=Medium|c_2) = 2/5$

$P(Student=Yes|c_2) = 1/5$

$P(Credit\_Rating=Fair|c_2) = 2/5$

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

$P(z|c_1) = P(\text{Age=Young}|c_1).P(\text{Income=Medium}|c_1).P(\text{Student=Yes}|c_1).P(\text{Credit\_Rating=Fair}|c_1) = (2/9).(4/9).(6/9).(6/9) = 0.044$

  - Đối với phân lớp c_2

$P(z|c_2) = P(\text{Age=Young}|c_2).P(\text{Income=Medium}|c_2).P(\text{Student=Yes}|c_2).P(\text{Credit\_Rating=Fair}|c_2) = (3/5).(2/5).(1/5).(2/5) = 0.019$

- 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

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

Ō 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...

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

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

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

  - $n(c_i)$: 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 $P(x_j \mid c_i)$

    → Các ước lượng đồng mức: $p=1/k$, với thuộc tính f_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
  - $P(x_j \mid c_i) < 1$, 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:

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

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

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

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

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_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

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

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

Phân loại văn bản bằng NB (2)

- Để phân lớp cho một văn bản mới d
- Giai đoạn phân lớp
  - 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

    $P(t_j \text{ ở vị trí } k \mid c_i) = P(t_j \text{ ở vị trí } m \mid c_i), \ \forall k,m$
  - Đối với mỗi phân lớp c_i, tính giá trị likelihood của văn bản d đối với c_i

    $$P(c_i). \prod_{t_j \in T\_d} P(t_j \mid c_i)$$
  - Phân lớp văn bản d thuộc vào lớp c^*

    $$c^* = \arg\max_{c_i \in C} P(c_i). \prod_{t_j \in T\_d} P(t_j \mid c_i)$$