ĐẠI HỌC QUỐC GIA HÀ NỘI
TRƯỜNG ĐẠI HỌC CÔNG NGHỆ

KHAI PHÁ DỮ LIỆU

Phương pháp phân lớp Naïve Bayes

Phan Xuân Hiếu

Bộ môn CHTTT & KTLab,
Khoa Công nghệ thông tin,
Trường Đại học Công nghệ, ĐHQG HN
Email: hieupx@vnu.edu.vn
URL: http://uet.vnu.edu.vn/~hieupx

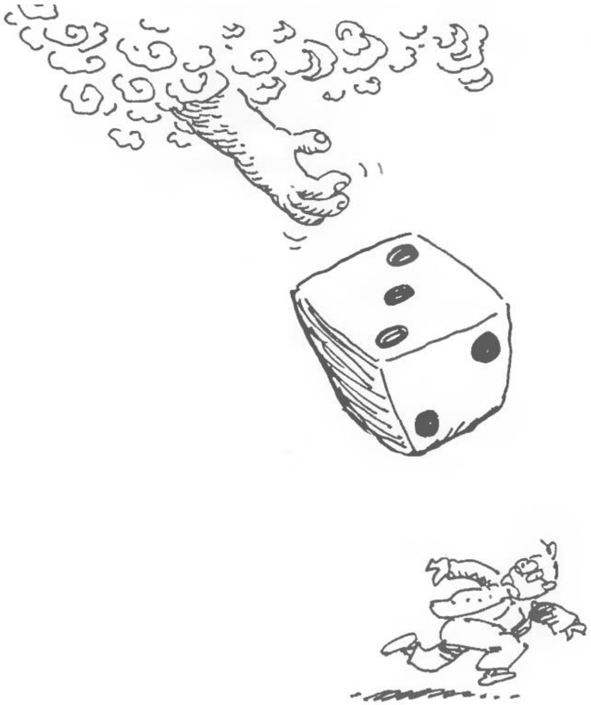
Nội dung bài giảng

- Kiến thức nền tảng
  - Các khái niệm cơ bản trong xác suất
  - Công thức Bayes
  - Phụ thuộc và độc lập
- Phương pháp phân lớp Naïve Bayes
  - Các khái niệm: Prior, Evidence, Likelihood, Posterior
  - Ví dụ minh họa
  - Giả định độc lập (naïve assumption)
  - Mô hình
  - Huấn luyện mô hình
  - Ưu và nhược điểm của phương pháp
  - Ứng dụng cụ thể
- Kết luận

Nội dung bài giảng

- **Kiến thức nền tảng**
  - Các khái niệm cơ bản trong xác suất
  - Công thức Bayes
  - Phụ thuộc và độc lập
- Phương pháp phân lớp Naïve Bayes
  - Các khái niệm: Prior, Evidence, Likelihood, Posterior
  - Ví dụ minh họa
  - Giả định độc lập (naïve assumption)
  - Mô hình
  - Huấn luyện mô hình
  - Ưu và nhược điểm của phương pháp
  - Ứng dụng cụ thể
- Kết luận

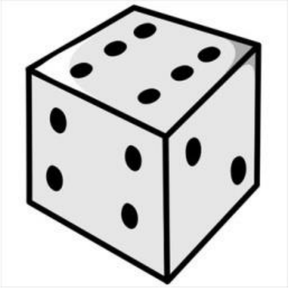
Kiến thức cơ bản về xác suất



Thế giới thực chứa đựng nhiều yếu tố ngẫu nhiên

Các khái niệm cơ bản trong xác suất

- Phép thử: là một thí nghiệm hay quan sát nào đó
- Biến cố sơ cấp: tất cả các kết quả có thể của một phép thử
- Không gian mẫu: Ω - là tập các biến cố sơ cấp
  - Tung con xúc xắc 6 mặt: Ω = {1, 2, 3, 4, 5, 6}
  - Gieo một đồng xu: Ω = {*sấp*, *ngữa*}
  - Gieo 2 đồng xu: Ω = {*sấp-sấp*, *sấp-ngữa*, *ngữa-sấp*, *ngữa-ngữa*}



- Biến cố: là một tập con của không gian mẫu, $A \subseteq \Omega$
  - "tung xúc xắc được số chẵn": A = {2, 4, 6}
- Số biến cố: 2^n, với $n = |\Omega|$
- Ω là biến cố chắc chắn, Ø là biến cố không

Các khái niệm cơ bản trong xác suất (2)

- Biến cố hợp của A và B:

$$A \cup B = \{w : w \in A \,|\, w \in B\}$$

- Biến cố A giao B (còn được viết là biến cố AB):

$$A \cap B = \{w : w \in A \,\&\, w \in B\}$$

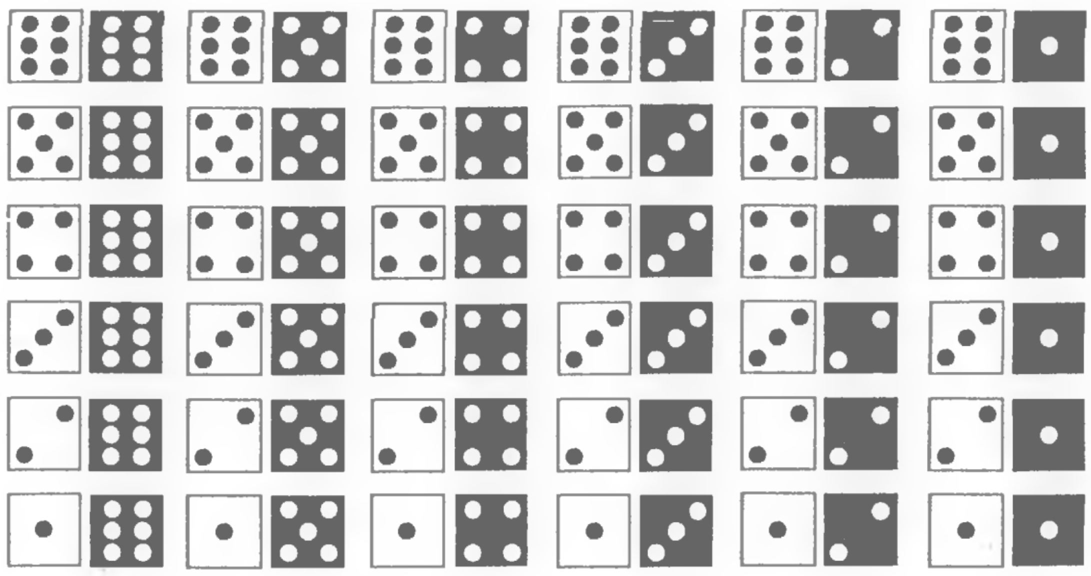
- Biến cố hiệu của A và B:

$$A \setminus B = \{w : w \in A \,\&\, w \notin B\}$$

- Biến cố đối của A:

$$\overline{A} = \{w : w \notin A\}$$

Ví dụ: gieo hai con xúc xắc đen và trắng



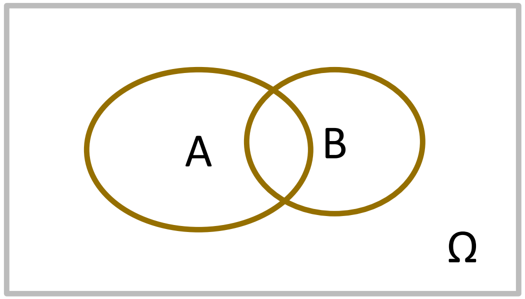
- Biến cố A – tổng số chấm nhỏ hơn hoặc bằng 3: {(1, 2), (1, 1), (2, 1)}
- Biến cố B – tổng số chấm bằng 6: {(1, 5), (2, 4), (3, 3), (4, 2), (5, 1)}
- Biến cố C – xúc xắc trắng là 1: {(1, 1), (1, 2), (1, 3), (1, 4), (1, 5), (1, 6)}
- Biến cố D – xúc xắc đen là 1: {(1, 1), (2, 1), (3, 1), (4, 1), (5, 1), (6, 1)}
- Biến cố C ∪ D? Biến cố C ∩ D?

Một số tính chất cơ bản

- Xác suất của biến cố rỗng, biến cỗ chắc chắn: $P(\varnothing)=0, P(\Omega)=1$
- A là một biến cố:

$$0 \le P(A) \le 1$$

$$P(\overline{A}) = 1 - P(A)$$



- A, B là hai biến cố:

$$P(A\cup B) = P(A) + P(B) - P(AB)$$

  - A và B là biến cố xung khắc: $P(A\cup B) = P(A) + P(B)$

- Xác suất có điều kiện: *xét ví dụ ở slide trước*
  - P(A) = P("tổng số chấm nhỏ hơn hoặc bằng 3") = 3/36 = 1/12
  - P(C) = P("xúc xắc trắng là 1") = 6/36 = 1/6
  - Xác suất của A khi C đã xảy ra gọi là xác suất A điều kiện C, ký hiệu là P(A|C). Khi đó P(A|C) = 2/6 = 1/3

Công thức Bayes

- Với P(B) > 0:

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

- Suy ra:

$$P(AB)=P(A|B)P(B)=P(B|A)P(A)$$

- Công thức Bayes:

$$P(B|A)=\frac{P(AB)}{P(A)}=\frac{P(A|B)P(B)}{P(A)}=\frac{P(A|B)P(B)}{P(AB)+P(A\overline{B})}$$

$$=\frac{P(A|B)P(B)}{P(AB)+P(A\overline{B})}=\frac{P(A|B)P(B)}{P(A|B)P(B)+P(A|\overline{B})P(\overline{B})}$$

Bayes, **Thomas** (**1763**), An eassy towards solving a problem in the doctrine of chances, *Philosophical Transactions of the Royal Society of London*, 53:370-418

Công thức Bayes tổng quát

- Với P(A) > 0 và $\{B_1, B_2, \ldots, B_n\}$ là một hệ đầy đủ các biến cố:
  - Tổng xác suất của hệ bằng 1:

$$\sum_{k=1}^{n} P(B_k) = 1$$

  - Từng đôi một xung khắc:

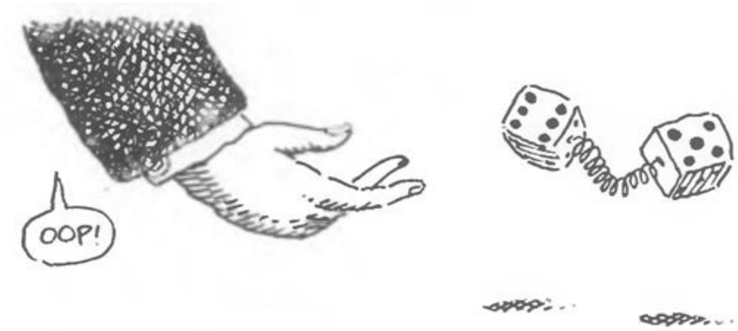
$$P(B_i \cap B_j) = 0$$

- Khi đó ta có:

$$P(B_k \mid A) = \frac{P(A \mid B_k) P(B_k)}{P(A)}$$

$$= \frac{P(A \mid B_k) P(B_k)}{\sum_{i=1}^{n} P(A \mid B_i) P(B_i)}$$

Phụ thuộc và độc lập



Minh họa: tung hai con xúc xắc phụ thuộc lẫn nhau

- Hai biến cố A và B được gọi là độc lập nếu với nhau nếu sự xuất hiện của biến cố này không ảnh hưởng đến xác suất của biến cố kia
- Ví dụ, nếu gieo hai con xúc xắc thì con này không ảnh hưởng đến con còn lại trừ khi chúng được nối với nhau như hình minh họa trên
- Nếu A và B độc lập:

$$P(AB) = P(A)P(B)$$

Nếu P(B) > 0, P(A) > 0 thì dễ thấy A và B độc lập khi và chỉ khi:

$$P(A|B) = P(A) \quad \text{và} \quad P(B|A) = P(B)$$

Luật chuỗi (tổng quát)

- Với ba biến A, B, C:

$$P(ABC) = P(A \mid BC)P(BC) = P(A \mid BC)P(B \mid C)P(C)$$

- Với n biến A_n, A_{n-1}, ..., A_1:

$$P(A_n, A_{n-1}, \ldots, A_1) = P(A_n \mid A_{n-1}, \ldots, A_1)P(A_{n-1}, \ldots, A_1)$$

$$= \prod_{k=1}^{n} P(A_k \mid \bigcap_{j=1}^{k-1} A_j)$$

- Giả sử ta có n biến A_n, A_{n-1}, ..., A_1 và biến B. Và giả sử A_i và A_j từng đôi một độc lập với điều kiện B. Khi đó:

$$P(A_n, A_{n-1}, \ldots, A_1 \mid B) = P(A_n \mid B)P(A_{n-1} \mid B) \ldots P(A_1 \mid B)$$

$$= \prod_{k=1}^{n} P(A_k \mid B)$$

Nội dung bài giảng

- Kiến thức nền tảng
  - Các khái niệm cơ bản trong xác suất
  - Công thức Bayes
  - Phụ thuộc và độc lập
- **Phương pháp phân lớp Naïve Bayes**
  - Các khái niệm: Prior, Evidence, Likelihood, Posterior
  - Ví dụ minh họa
  - Giả định độc lập (naïve assumption)
  - Mô hình
  - Huấn luyện mô hình
  - Ưu và nhược điểm của phương pháp
  - Ứng dụng cụ thể
- Kết luận

Prior, evidence, likelihood, posterior

Likelihood
(conditional)

Prior
(tiền nghiệm)

$$P(B \mid A) = \frac{P(AB)}{P(A)} = \frac{P(A \mid B)P(B)}{P(A)} = \frac{P(A \mid B)P(B)}{P(AB) + P(A\overline{B})}$$

Posterior
(hậu nghiệm)

Evidence

Posterior = Likelihood * Prior / Evidence

Ví dụ minh họa

- Ký hiệu:
  - Bệnh cảm cúm (C), không bị cảm ký hiệu là "koC"
  - Các triệu chứng: ho (H), đau đầu (D), đau mỏi (M)

    Các ký hiệu đối: "koH" (không bị ho), "koD" (không đau đầu), "koM" (không đau mỏi)
- Với các dữ liệu như sau:

Priors

P(C) = 1/40	P(koC) = 39/40
P(D|C) = 1/2	P(D|koC) = 7/78
P(H|C) = 2/3	P(H|koC) = 1/6
P(M|C) = 3/4	P(M| koC) = 1/3

Likelihood (Conditionals)

- Một người bị đau đầu (D), bị đau mỏi (M), và không bị ho (koH), hỏi xác suất bị cảm là bao nhiêu? Nói cách khác:

Posterior → **P(C| D và M và koH) =** bao nhiêu?

Trong phần tính toán sau sẽ dùng ký hiệu: $P(C \mid D \wedge M \wedge \overline{H})$

Giả định về sự độc lập

- Giả sử các triệu chứng độc lập với nhau khi biết trạng thái bệnh, như thế:

  P(<triệu chứng A> | <bị cảm cúm> và <các triệu chứng khác>)

  = P(<triệu chứng A> | <bị cảm cúm>)

  tương tự

  P(<triệu chứng A> | <ko bị cảm cúm> và <các triệu chứng khác>)

  = P(<triệu chứng A> | <ko bị cảm cúm>)

- Đây là một giả định giúp đơn giản hóa quá trình tính toán cho ví dụ này
- Người ta gọi đây là giả định độc lập "đơn giản" (naïve assumption)

Ví dụ minh họa - biến đổi và tính toán

$$P(C \mid D \wedge M \wedge \overline{H}) = \frac{P(D \wedge M \wedge \overline{H} \wedge C)}{P(D \wedge M \wedge \overline{H})}$$

Công thức Bayes

$$= \frac{P(D \wedge M \wedge \overline{H} \mid C)P(C)}{P(D \wedge M \wedge \overline{H})}$$

$$= \frac{P(D \wedge M \wedge \overline{H} \mid C)P(C)}{P(D \wedge M \wedge \overline{H} \wedge C) + P(D \wedge M \wedge \overline{H} \wedge \overline{C})}$$

$P(X) = P(XY) + P(X\overline{Y})$

$$= \frac{P(D \wedge M \wedge \overline{H} \mid C)P(C)}{P(D \wedge M \wedge \overline{H} \mid C)P(C) + P(D \wedge M \wedge \overline{H} \mid \overline{C})P(\overline{C})}$$

$P(XY) = P(X \mid Y)P(Y)$

Ví dụ minh họa - biến đổi và tính toán (2)

$$P(D \wedge M \wedge \overline{H} \mid C)P(C) = P(D \mid M \wedge \overline{H} \wedge C)P(M \wedge \overline{H} \mid C)P(C)$$

$$= P(D \mid M \wedge \overline{H} \wedge C)P(M \mid \overline{H} \wedge C)P(\overline{H} \mid C)P(C)$$

Luật chuỗi

$$= P(D \mid C)P(M \mid C)P(\overline{H} \mid C)P(C)$$

Giả định độc lập

$$= \frac{1}{2}\frac{3}{4}(1-\frac{2}{3})\frac{1}{40} = \frac{1}{320}$$

$$P(D \wedge M \wedge \overline{H} \mid \overline{C})P(\overline{C}) = P(D \mid M \wedge \overline{H} \wedge \overline{C})P(M \wedge \overline{H} \mid \overline{C})P(\overline{C})$$

$$= P(D \mid M \wedge \overline{H} \wedge \overline{C})P(M \mid \overline{H} \wedge \overline{C})P(\overline{H} \mid \overline{C})P(\overline{C})$$

$$= P(D \mid \overline{C})P(M \mid \overline{C})P(\overline{H} \mid \overline{C})P(\overline{C})$$

$$= \frac{7}{78}\frac{1}{3}(1-\frac{1}{6})\frac{39}{40} = \frac{7}{288}$$

Ví dụ minh họa - biến đổi và tính toán (3)

$$\frac{1}{320}$$

$$P(C \mid D \wedge M \wedge \overline{H}) = \frac{P(D \wedge M \wedge \overline{H} \mid C)P(C)}{P(D \wedge M \wedge \overline{H} \mid C)P(C) + P(D \wedge M \wedge \overline{H} \mid \overline{C})P(\overline{C})}$$

$$\frac{1}{320} \qquad \frac{7}{288}$$

$$= 0.1139$$

<u>Kết luận</u>

11.4% có khả năng **bị cảm** nếu cảm thấy **đau đầu, đau mỏi**, và **không bị ho**

Nội dung bài giảng

- Kiến thức nền tảng
  - Các khái niệm cơ bản trong xác suất
  - Công thức Bayes
  - Phụ thuộc và độc lập
- **Phương pháp phân lớp Naïve Bayes**
  - Các khái niệm: Prior, Evidence, Likelihood, Posterior
  - Ví dụ minh họa
  - Giả định độc lập (naïve assumption)
  - **Mô hình**
  - Huấn luyện mô hình
  - Ưu và nhược điểm của phương pháp
  - Ứng dụng cụ thể
- Kết luận

Phát biểu lại bài toán phân lớp

- Phát biểu lại bài toán phân lớp:
  - **C** = $\{c_1, c_2, \ldots, c_K\}$: tập K lớp
  - **X** = $\{\mathbf{x}_i\}$ (i=1,2,…) là không gian các đối tượng cần phân lớp
  - Xây dựng một ánh xạ *f* : **X** → **C**
  - Ánh xạ *f* được gọi là mô hình phân lớp (classification model, classifier)
- Xây dựng mô hình *f* bằng học giám sát
  - **D** = $\{(\mathbf{x}^1, c^1), (\mathbf{x}^2, c^2), \ldots, (\mathbf{x}^N, c^N)\}$ trong đó $\mathbf{x}^n \in \mathbf{X}$, $c^n \in \mathbf{C}$ là tập dữ liệu huấn luyện (training data)
  - Huấn luyện mô hình *f* dựa trên tập huấn luyện **D** sao cho *f* phân lớp chính xác nhất có thể
- Mô hình *f* có thể xây dựng theo:
  - Phương pháp Naïve Bayes **← trong phần này**
  - Phương pháp cây quyết định (decision tree)
  - Phương pháp cực đại hóa entropy (maximum entropy classification)
  - Phương pháp máy vector hỗ trợ (support vector machines)
  - .v.v.

Phương pháp phân lớp Naïve Bayes

- Giả sử C là biến ngẫu nhiên đại diện cho các lớp cần phân loại
  - C nhận một trong K giá trị (K lớp): $\{c_1, c_2, \ldots, c_K\}$
  - Xác suất để C nhận một lớp cụ thể nào đó được ký hiệu bằng các xác suất: $P(C = c_1), P(C = c_2), \ldots, P(C = c_K)$
- Giả sử các đối tượng cần phân lớp **x** có M thuộc tính (Features)
  - Mỗi thuộc tính i được biểu diễn bởi một biến ngẫu nhiên F_i, như vậy các đối tượng **x** được biểu diễn bởi một bộ M biến ngẫu nhiên $\mathbf{F} = (F_1, F_2, \ldots, F_M)$
  - Mỗi thuộc tính/biến ngẫu nhiên F_i có miền xác định $\mathbf{V}^i = \{v^i_1, v^i_2, \ldots, v^i_{Pi}\}$
  - Ký hiệu $P(F_i = v^i | C = c_k)$ là xác suất để thuộc tính/biến ngẫu nhiên F_i nhận giá trị $v^i \in \mathbf{V}^i$ với điều kiện **x** thuộc lớp c_k.
- Các xác suất điều kiện (conditionals) $P(F_i = v^i_j \mid C = c_k)$ và xác suất tiền nghiệm (priors) $P(C = c_k)$ được liệt kê trong bảng ở slide tiếp theo

Ví dụ: P("Messi"=yes|C="thể thao")

K classes $\{c_1, c_2, \ldots, c_K\}$

Priors

	$P(C = c_1)$	$P(C = c_2)$	…	$P(C = c_K)$
F_1	$P(F_1=v^1{}_1 \mid C=c_1)$	$P(F_1=v^1{}_1 \mid C=c_2)$	…	$P(F_1=v^1{}_1 \mid C=c_K)$
F_1	$P(F_1=v^1{}_2 \mid C=c_1)$	$P(F_1=v^1{}_2 \mid C=c_2)$	…	$P(F_1=v^1{}_2 \mid C=c_K)$
F_1	⋮	⋮		⋮
F_1	$P(F_1=v^1{}_{P1} \mid C=c_1)$	$P(F_1=v^1{}_{P1} \mid C=c_2)$	…	$P(F_1=v^1{}_{P1} \mid C=c_K)$
	⋮	⋮		⋮
F_2	$P(F_2=v^2{}_1 \mid C=c_1)$	$P(F_2=v^2{}_1 \mid C=c_2)$	…	$P(F_2=v^2{}_1 \mid C=c_K)$
F_2	$P(F_2=v^2{}_2 \mid C=c_1)$	$P(F_2=v^2{}_2 \mid C=c_2)$	…	$P(F_2=v^2{}_2 \mid C=c_K)$
F_2	⋮	⋮		⋮
F_2	$P(F_2=v^2{}_{P2} \mid C=c_1)$	$P(F_2=v^2{}_{P2} \mid C=c_2)$	…	$P(F_2=v^2{}_{P2} \mid C=c_K)$
	⋮	⋮		⋮
F_M	$P(F_M=v^M{}_1 \mid C=c_1)$	$P(F_M=v^M{}_1 \mid C=c_2)$	…	$P(F_M=v^M{}_1 \mid C=c_K)$
F_M	$P(F_M=v^M{}_2 \mid C=c_1)$	$P(F_M=v^M{}_2 \mid C=c_2)$	…	$P(F_M=v^M{}_2 \mid C=c_K)$
F_M	⋮	⋮		⋮
F_M	$P(F_M=v^M{}_{PM} \mid C=c_1)$	$P(F_M=v^M{}_{PM} \mid C=c_2)$	…	$P(F_M=v^M{}_{PM} \mid C=c_K)$

Likelihood (conditionals)

Ví dụ: P("iPhone"=yes|C="công nghệ")

Xác suất một đối tượng **x** thuộc lớp c_k nào đó được tính như sau:

$$P(C=c_k \mid X=\mathbf{x}) = \frac{P(X=\mathbf{x} \wedge C=c_k)}{P(X=\mathbf{x})}$$

Công thức Bayes

$$= \frac{P(F_1=v^1 \wedge F_2=v^2 \wedge \ldots \wedge F_M=v^M \wedge C=c_k)}{P(F_1=v^1 \wedge F_2=v^2 \wedge \ldots \wedge F_M=v^M)}$$

$$= \frac{P(F_1=v^1 \wedge F_2=v^2 \wedge \ldots \wedge F_M=v^M \wedge C=c_k)}{\sum_{i=1}^{K} P(F_1=v^1 \wedge F_2=v^2 \wedge \ldots \wedge F_M=v^M \wedge C=c_i)}$$

Nhắc lại:
$\{c_1, c_2, \ldots, c_K\}$ là hệ đầy đủ các biến cố

$$= \frac{P(F_1=v^1 \wedge F_2=v^2 \wedge \ldots \wedge F_M=v^M \mid C=c_k)P(C=c_k)}{\sum_{j=1}^{K} P(F_1=v^1 \wedge F_2=v^2 \wedge \ldots \wedge F_M=v^M \mid C=c_j)P(C=c_j)}$$

Giả định độc lập

$$= \frac{\left[\prod_{i=1}^{M} P(F_i=v^i \mid C=c_k)\right]P(C=c_k)}{\sum_{j=1}^{K}\left[\prod_{i=1}^{M} P(F_i=v^i \mid C=c_j)\right]P(C=c_j)}$$

Mô hình phân lớp Naïve Bayes

$$P(C=c_k \mid X=\mathbf{x}) = \frac{P(X=\mathbf{x} \wedge C=c_k)}{P(X=\mathbf{x})}$$

$$= \frac{\left[\prod_{i=1}^{M} P(F_i = v^i \mid C=c_k)\right] P(C=c_k)}{\sum_{j=1}^{K}\left[\prod_{i=1}^{M} P(F_i = v^i \mid C=c_j)\right] P(C=c_j)}$$

$$\propto \left[\prod_{i=1}^{M} P(F_i = v^i \mid C=c_k)\right] P(C=c_k)$$

P(X = **x**) là hằng số đối với các lớp c_k

$$classify(\mathbf{x}) = \arg\max_{c_k} P(C=c_k \mid X=\mathbf{x})$$

$$= \arg\max_{c_k} \left[\prod_{i=1}^{M} P(F_i = v^i \mid C=c_k)\right] P(C=c_k)$$

$$= \arg\max_{c_k} \left[\log P(C=c_k) + \sum_{i}^{M} \log P(F_i = v^i \mid C=c_k)\right]$$

Nội dung bài giảng

- Kiến thức nền tảng
  - Các khái niệm cơ bản trong xác suất
  - Công thức Bayes
  - Phụ thuộc và độc lập
- **Phương pháp phân lớp Naïve Bayes**
  - Các khái niệm: Prior, Evidence, Likelihood, Posterior
  - Ví dụ minh họa
  - Giả định độc lập (naïve assumption)
  - Mô hình
  - **Huấn luyện mô hình**
  - Ưu và nhược điểm của phương pháp
  - Ứng dụng cụ thể
- Kết luận

Huấn luyện mô hình = ước lượng các tham số

$P(C = c_1)$ = **???**	$P(C = c_2)$ = **???**	...	$P(C = c_K)$ = **???**
$P(F_1=v^1{}_1\|C=c_1)$ = **???**	$P(F_1=v^1{}_1\|C=c_2)$ = **???**	...	$P(F_1=v^1{}_1\|C=c_K)$ = **???**
$P(F_1=v^1{}_2\|C=c_1)$ = **???**	$P(F_1=v^1{}_2\|C=c_2)$ = **???**	...	$P(F_1=v^1{}_2\|C=c_K)$ = **???**
⋮	⋮		⋮
$P(F_1=v^1{}_{P1}\|C=c_1)$ = **???**	$P(F_1=v^1{}_{P1}\|C=c_2)$ = **???**	...	$P(F_1=v^1{}_{P1}\|C=c_K)$ = **???**
⋮	⋮		⋮
$P(F_2=v^2{}_1\|C=c_1)$ = **???**	$P(F_2=v^2{}_1\|C=c_2)$ = **???**	...	$P(F_2=v^2{}_1\|C=c_K)$ = **???**
$P(F_2=v^2{}_2\|C=c_1)$ = **???**	$P(F_2=v^2{}_2\|C=c_2)$ = **???**	...	$P(F_2=v^2{}_2\|C=c_K)$ = **???**
⋮	⋮		⋮
$P(F_2=v^2{}_{P2}\|C=c_1)$ = **???**	$P(F_2=v^2{}_{P2}\|C=c_2)$ = **???**	...	$P(F_2=v^2{}_{P2}\|C=c_K)$ = **???**
⋮	⋮		⋮
$P(F_M=v^M{}_1\|C=c_1)$ = **???**	$P(F_M=v^M{}_1\|C=c_2)$ = **???**	...	$P(F_M=v^M{}_1\|C=c_K)$ = **???**
$P(F_M=v^M{}_2\|C=c_1)$ = **???**	$P(F_M=v^M{}_2\|C=c_2)$ = **???**	...	$P(F_M=v^M{}_2\|C=c_K)$ = **???**
⋮	⋮		⋮
$P(F_M=v^M{}_{PM}\|C=c_1)$ = **???**	$P(F_M=v^M{}_{PM}\|C=c_2)$ = **???**	...	$P(F_M=v^M{}_{PM}\|C=c_K)$ = **???**

Ước lượng từ dữ liệu huấn luyện

D = $\{(\mathbf{x}^1, c^1), (\mathbf{x}^2, c^2), \ldots, (\mathbf{x}^N, c^N)\}$ trong đó $\mathbf{x}^i \in \mathbf{X}$ là các đối tượng dữ liệu cần phân lớp, $c^i \in \mathbf{C} = \{c_1, c_2, \ldots, c_K\}$ là lớp tương ứng của **x**

Ước lượng các tham số từ dữ liệu huấn luyện

- Ước lượng xác suất tiền nghiệm (priors) của các lớp $P(C = c_k)$
  - Trường hợp đơn giản nhất: $P(C = c_k) = 1 / K$ (tổng số lớp), có nghĩa là xác suất xuất hiện các lớp là bằng nhau và bằng 1 / K.
  - $P(C = c_k)$ = (số lượng đối tượng **x** thuộc lớp c_k) / (tổng số đối tượng trong tập dữ liệu huấn luyện) = $|\{(\mathbf{x}^j, c^j) \in \mathbf{D}$ với $c^j = c_k\}| / |\mathbf{D}| = |\{(\mathbf{x}^j, c^j) \in \mathbf{D}$ với $c^j = c_k\}| / N$
- Ước lượng các xác xuất điều kiện $P(F_i = v^i \mid C = c_k)$
  - Nếu F_i là thuộc tính rời rạc (dạng chủng loại – categorical)

    $P(F_i = v^i \mid C = c_k)$ = (số lượng đối tượng **x** thuộc lớp c_k có thuộc tính F_i bằng giá trị v^i) / (tổng số đối tượng **x** thuộc lớp c_k)

    $= |\{(\mathbf{x}^j, c^j) \in \mathbf{D}$ với $F_i(\mathbf{x}^j) = v^i$ và $c^j = c_k\}| / |\{(\mathbf{x}^j, c^j) \in \mathbf{D}$ với $c^j = c_k\}|$
  - Nếu F_i là thuộc tính liên tục (continuous-valued), $P(F_i = v^i \mid C = c_k)$ thường được ước lượng dựa trên phân phối chuẩn (Gaussian distribution)

    (tính toán cụ thể trong slide tiếp theo)

- Phân phối chuẩn với kỳ vọng μ và phương sai σ^2:

$$f(x;\mu,\sigma)=\frac{1}{\sigma\sqrt{2\pi}}\exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

  - Kỳ vọng (trung bình): μ = E(X)
  - Phương sai: var(X) = σ^2 = E[(X- μ)^2]
  - Cả kỳ vọng và phương sai có thể ước lượng từ dữ liệu huấn luyện **D**

- Với mỗi thuộc tính liên tục F_i, ta tính kỳ vọng kỳ vọng μ^i_k và phương sai $\sigma^i_k{}^2$ cho lớp c_k:
  - Chia tập dữ liệu **D** theo từng lớp, ví dụ tập $\mathbf{D}_k \subseteq \mathbf{D}$ là tập các đối tượng thuộc lớp c_k.
  - Khi đó:

$$\mu_k^i=\frac{1}{|\mathbf{D}_k|}\sum_{F_i(\mathbf{x})=v^i,\mathbf{x}\in\mathbf{D}_k} v^i$$

  - Phương sai: $\sigma^i_k{}^2$ = E[(X- μ^i_k)^2]

- Khi đó P(F_i = v^i | C = c_k) được tính như sau:

$$P(F_i=v^i\,|\,C=c_k)=f(x;\mu_k^i,\sigma_k^i)=\frac{1}{\sigma_k^i\sqrt{2\pi}}\exp\left(-\frac{(x-\mu_k^i)^2}{2\sigma_k^{i\,2}}\right)$$

Khắc phục vấn đề xác suất điều kiện bằng zero

- Nếu trong dữ liệu huấn luyện không có đối tượng **x** nào có thuộc lớp c_k có thuộc tính F_i nhận một giá trị cụ thể v^i_j, xác suất điều kiện $P(F_i = v^i_j \mid C = c_k)$ sẽ bằng 0 (zero)
- Khi phân lớp, nếu có một đối tượng nào mang thuộc tính-giá trị này thì xác suất phân vào lớp c_k luôn bằng 0.
- Khắc phục bằng cách ước lượng theo công thức sau:

$$P(F_i = v_j^i \mid C = c_k) = \frac{n_{ijk} + mp}{n_k + m}$$

Trong đó:

- n_{ijk} là số đối tượng **x** thuộc lớp c_k mang thuộc tính-giá trị (F_i, v^i_j) trong **D**.
- n_k là số lượng đối tượng **x** thuộc lớp c_k.
- p là một hằng số (p có thể = 1 hoặc = $1/P_i$ với P_i là số giá trị thuộc tính F_i có thể nhận).
- m là số lượng đối tượng x "ảo", m ≥ 1

Nội dung bài giảng

- Kiến thức nền tảng
  - Các khái niệm cơ bản trong xác suất
  - Công thức Bayes
  - Phụ thuộc và độc lập
- **Phương pháp phân lớp Naïve Bayes**
  - Các khái niệm: Prior, Evidence, Likelihood, Posterior
  - Ví dụ minh họa
  - Giả định độc lập (naïve assumption)
  - Mô hình
  - Huấn luyện mô hình
  - **Ưu và nhược điểm của phương pháp**
  - Ứng dụng cụ thể
- Kết luận

Nhược điểm

- Giả định độc lập (naïve assumption)
  - Giả định này sai trong hầu hết các trường hợp thực tế trong đó các thuộc tính trong các đối tượng thường phụ thuộc lẫn nhau
- Vấn đề zero-probability:
  - Đã nêu cách khắc phục ở trên
- Mô hình không được huấn luyện bằng một phương pháp tối ưu mạnh và chặt chẽ
  - Tham số của mô hình là các ước lượng xác suất điều kiện đơn lẻ
  - Không tính đến sự tương tác giữa các ước lượng này

Ưu điểm

- Giả định độc lập (nhược điểm cũng là ưu điểm): hoạt động tốt cho nhiều bài toán/miền dữ liệu và ứng dụng
  - Đơn giản nhưng đủ tốt để giải quyết nhiều bài toán

  Ví dụ: phân lớp văn bản, lọc spam, .v.v.
- Cho phép kết hợp tri thức tiền nghiệm (prior knowledge) và dữ liệu quan sát được (observed data)
  - Tốt khi có sự chênh lệch số lượng giữa các lớp phân loại
- Huấn luyện mô hình (ước lượng tham số) dễ và nhanh

Nội dung bài giảng

- Kiến thức nền tảng
  - Các khái niệm cơ bản trong xác suất
  - Công thức Bayes
  - Phụ thuộc và độc lập
- **Phương pháp phân lớp Naïve Bayes**
  - Các khái niệm: Prior, Evidence, Likelihood, Posterior
  - Ví dụ minh họa
  - Giả định độc lập (naïve assumption)
  - Mô hình
  - Huấn luyện mô hình
  - Ưu và nhược điểm của phương pháp
  - **Ứng dụng cụ thể**
- Kết luận

Phân lớp dữ liệu dạng bảng biểu

Day	Outlook	Temperature	Humidity	Wind	Play Tennis
Day1	Sunny	Hot	High	Weak	*No*
Day2	Sunny	Hot	High	Strong	*No*
Day3	Overcast	Hot	High	Weak	Yes
Day4	Rain	Mild	High	Weak	Yes
Day5	Rain	Cool	Normal	Weak	Yes
Day6	Rain	Cool	Normal	Strong	*No*
Day7	Overcast	Cool	Normal	Strong	Yes
Day8	Sunny	Mild	High	Weak	*No*
Day9	Sunny	Cool	Normal	Weak	Yes
Day10	Rain	Mild	Normal	Weak	Yes
Day11	Sunny	Mild	Normal	Strong	Yes
Day12	Overcast	Mild	High	Strong	Yes
Day13	Overcast	Hot	Normal	Weak	Yes
Day14	Rain	Mild	High	Strong	*No*

age	income	student	credit_rating	_com
<=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

Phân lớp văn bản (document classification)

Tham khảo cuốn sách sau:

Christopher Manning, Prabhakar Raghavan, and Hinrich Schutze, ***Introduction to Information Retrieval***, Cambridge University Press, 2008. (Free)

Chương 13. Text classification and Naïve Bayes

Tham khảo thêm:

http://en.wikipedia.org/wiki/Document_classification

http://en.wikipedia.org/wiki/Naive_Bayes_classifier

Lọc spam (spam filtering)

Tham khảo:

Bayesian spam filtering

http://en.wikipedia.org/wiki/Bayesian_spam_filtering

http://en.wikipedia.org/wiki/Naive_Bayes_classifier

http://en.wikipedia.org/wiki/Email_filtering

Nội dung bài giảng

- Kiến thức nền tảng
  - Các khái niệm cơ bản trong xác suất
  - Công thức Bayes
  - Phụ thuộc và độc lập
- Phương pháp phân lớp Naïve Bayes
  - Các khái niệm: Prior, Evidence, Likelihood, Posterior
  - Ví dụ minh họa
  - Giả định độc lập (naïve assumption)
  - Mô hình
  - Huấn luyện mô hình
  - Ưu và nhược điểm của phương pháp
  - Ứng dụng cụ thể
- **Kết luận**

Kết luận

- Công thức Bayes
- Giả định độc lập (naïve assumption)
- Tri thức tiền nghiệm (prior knowledge)
- Mô hình phân lớp Naïve Bayes
- Huấn luyện mô hình
- Nhược điểm
- Ưu điểm
- Ứng dụng thực tế