CHƯƠNG 4. PHÂN LỚP DỮ LIỆU

Nội dung

4.1. GIỚI THIỆU
4.2. PHƯƠNG PHÁP CÂY QUYẾT ĐỊNH
4.3. PHƯƠNG PHÁP NAÏVE BAYES

4.1. Giới thiệu

- Bài toán phân lớp (Classification)
- Đối với một tập các bản ghi (instances/records) – gọi là tậphuấn luyện/học (training/learning set)
- Mỗi bản ghi được biểu diễn bằng một tập các thuộc tính, trong đó có một thuộc tính phân lớp (class attribute)
- Tìm/học một hàm cho thuộc tính phân lớp (hàm phân lớp) đối với các giá trị của các thuộc tính khác
- Sử dụng một tập khác các bản ghi với các ví dụ học để kiểm tra độ chính xác của hàm phân lớp học được – gọi là tập kiểm thử (test set)
- Thông thường, tập dữ liệu ban đầu được chia thành 2 tập (không giao nhau): training set (để học hàm phân lớp) và test set (để kiểm thử hàm phân lớp học được)

4.1. Giới thiệu

Phân lớp :

- Cho tập các mẫu đã phân lớp trước, xây dựng mô hình cho từng lớp
- **Mục đích :** Gán các mẫu mới vào các lớp với độ chính xác cao nhất có thể.
- Cho CSDL $D=\{t_1,t_2,\ldots,t_n\}$ và tập các lớp $C=\{C_1,\ldots,C_m\}$, phân lớp là bài toán xác định ánh xạ f : D->C sao cho mỗi t_i được gán vào một lớp.
- **Đầu vào:** một tập các mẫu dữ liệu huấn luyện, với một nhãn phân lớp cho mỗi mẫu dữ liệu
- **Đầu ra:** mô hình (bộ phân lớp) dựa trên tập huấn luyện và những nhãn phân lớp

4.1. Giới thiệu

- **Ví dụ Phân lớp :**
- Phân lớp khách hàng (trong ngân hàng) để cho vay hay không
- Dự đoán tế bào khối u là lành tính hay ác tính
- Phân loại giao dịch thẻ tín dụng là hợp pháp hay gian lận
- Phân loại tin tức thuộc lĩnh vực tài chính, thời tiết, giải trí, thể thao, …
- Dự đoán khi nào sông có lũ

4.1. Giới thiệu

Qui trình phân lớp

Bước 1
Xây dựng mô hình

- **Mỗi bộ dữ liệu** được phân vào một lớp được xác định trước
- Lớp của một bộ dữ liệu được xác định bởi **thuộc tính gán nhãn lớp**
- Tập các bộ dữ liệu huấn luyện - **tập huấn luyện** - được dùng để **xây dựng mô hình**
- Mô hình được biểu diễn bởi **các luật phân lớp**, **các cây quyết định** hoặc **các công thức toán học**

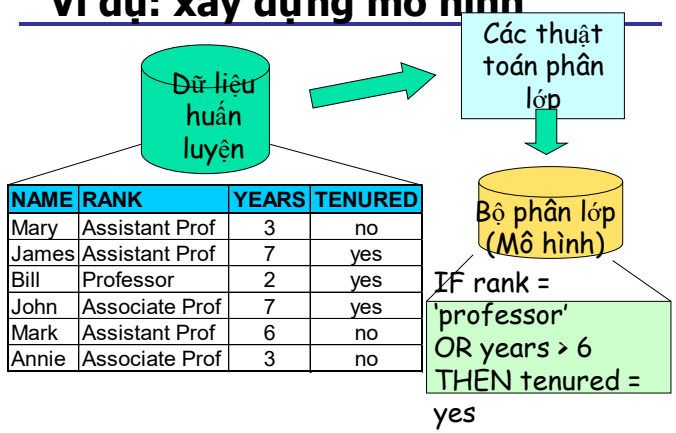
4.1. Giới thiệu

Qui trình phân lớp

Bước 2 sử dụng mô hình

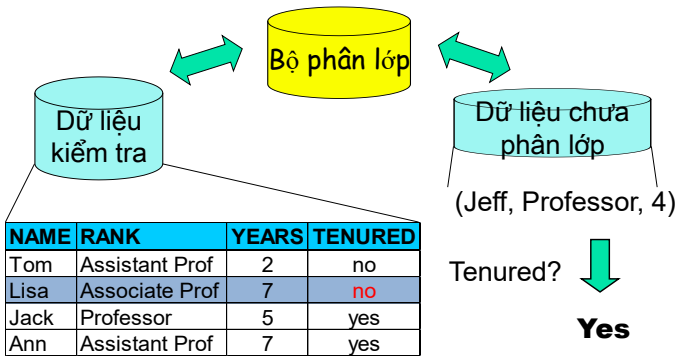
- **Phân lớp cho những đối tượng mới hoặc chưa được phân lớp**
- **Đánh giá độ chính xác của mô hình**
  - Sử dụng tập dữ liệu kiểm tra để xác định độ chính xác của mô hình
  - Tỉ lệ chính xác = phần trăm các bộ dữ liệu được phân lớp đúng bởi mô hình trong số các lần kiểm tra
  - Độ chính xác chấp nhận được -> áp dụng mô hình phân lớp các bộ dữ liệu chưa xác định được nhãn lớp

Ví dụ: xây dựng mô hình



NAME	RANK	YEARS	TENURED
Mary	Assistant Prof	3	no
James	Assistant Prof	7	yes
Bill	Professor	2	yes
John	Associate Prof	7	yes
Mark	Assistant Prof	6	no
Annie	Associate Prof	3	no

Ví dụ: sử dụng mô hình



NAME	RANK	YEARS	TENURED
Tom	Assistant Prof	2	no
Lisa	Associate Prof	7	no
Jack	Professor	5	yes
Ann	Assistant Prof	7	yes

4.1. Giới thiệu

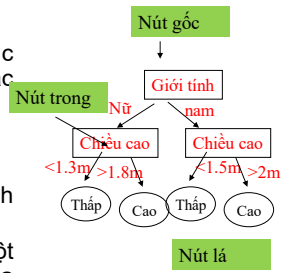
Các kỹ thuật phân lớp :

- Phương pháp dựa trên cây quyết định
- Phương pháp dựa trên luật
- Phương pháp Naïve Bayes
- Phương pháp dựa trên thể hiện
- Mạng Nơron
- SVM (support vector machine)
- Tập thô

4.2. PHƯƠNG PHÁP CÂY QUYẾT ĐỊNH

4.2.1. Định nghĩa

- Cây quyết định là một cấu trúc phân cấp của các nút và các nhánh
- Có 3 loại nút trên cây:
- Nút gốc
- **Nút trong** mang tên thuộc tính của CSDL
- **Nhánh** của cây = đầu ra của một phép kiểm tra, mang giá trị của thuộc tính
- **Nút lá** mang tên lớp c_i



4.2. PHƯƠNG PHÁP CÂY QUYẾT ĐỊNH

4.2.2. Tạo cây quyết định

Hai giai đoạn tạo cây quyết định:

- **Xây dựng cây**
  - Bắt đầu, tất cả các mẫu huấn luyện đều ở gốc
  - Phân chia các bộ dựa trên các thuộc tính được chọn
  - Kiểm tra các thuộc tính được chọn dựa trên một độ đo thống kê hoặc heuristic
- **Thu gọn cây**
  - Xác định và loại bỏ những nhánh nhiễu hoặc tách khỏi nhóm

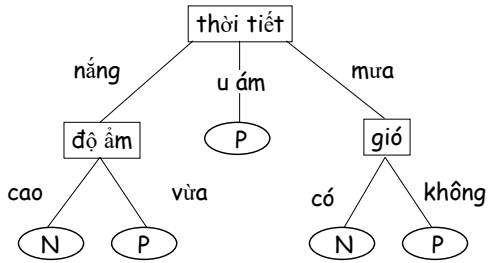
Cây quyết định – Ví dụ tiêu biểu: play tennis?

Tập huấn luyện trích từ Quinlan's ID3

Thời tiết	Nhiệt độ	Độ ẩm	Gió	Lớp
nắng	nóng	cao	không	N
nắng	nóng	cao	không	N
u ám	nóng	cao	không	P
mưa	ấm áp	cao	không	P
mưa	mát	vừa	không	P
mưa	mát	vừa	có	N
u ám	mát	vừa	có	P
nắng	ấm áp	cao	không	N
nắng	mát	vừa	không	P
mưa	ấm áp	vừa	không	P
nắng	ấm áp	vừa	có	P
u ám	ấm áp	cao	có	P
u ám	nóng	vừa	không	P
mưa	ấm áp	cao	có	N

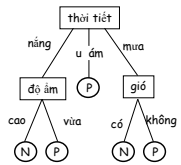
13

Cây quyết định thu được với ID3 (Quinlan 86)



14

Rút luật phân lớp từ cây quyết định



- Mỗi một **đường dẫn** từ gốc đến lá trong cây tạo thành một **luật**
- Nút lá giữ quyết định phân lớp dự đoán
- Các luật tạo được dễ hiểu hơn các cây

IF thời tiết=nắng
AND độ ẩm=vừa
THEN play tennis

15

4.2.3. Một vài tiêu chí xác định điểm chia tốt nhất

- Độ đo này được dựa trên cơ sở lý thuyết thông tin của nhà toán học Claude Shannon. Được sử dụng chủ yếu trong giải thuật ID3
- Độ đo thông tin thu được, được dùng để lựa chọn thuộc tính kiểm định tại mỗi nút trên cây. Độ đo như vậy còn được gọi là *độ đo lựa chọn thuộc tính* hay *độ đo chất lượng phân chia.* Thuộc tính với thông tin thu được cao nhất được chọn là thuộc tính kiểm tra tại nút hiện thời.

16

4.2.3. Một vài tiêu chí xác định điểm chia tốt nhất

- **Chọn thuộc tính có độ lợi thông tin cao nhất**
- **D : tập huấn luyện**
- **$C_{i,D}$: tập các mẫu của D thuộc lớp C_i với i = {1, …, m}**
- **$|C_{i,D}|$, |D| : lực lượng của tập $C_{i,D}$ và D tương ứng**
- **p_i là xác suất để một mẫu bất kỳ của D thuộc về lớp C_i**
- **Thông tin kỳ vọng để phân lớp một mẫu trong D là :**

$$Info(D) = -\sum_{i=1}^{m} p_i \log_2(p_i)$$

$$p_i = \frac{|C_{i,D}|}{|D|}$$

4.2.3. Một vài tiêu chí xác định điểm chia tốt nhất

- Ví dụ
- Trong ví dụ trên có 14 mẫu tin trong đó có 9 mẫu thuộc lớp chơi tennis và 5 mẫu huộc lớp không chơi tenis ta có |D|=14
- Lớp C_1 là lớp có chơi tennis, C_2 là lớp không chơi tennis ta có:
- $|C_{1,D}|$=9,
- $|C_{2,D}|$=5
- thông tin kỳ vọng để phân lớp một mẫu trên D là:

$$Info(9,5) = -\frac{9}{14}\log_2\frac{9}{14} - \frac{5}{14}\log_2\frac{5}{14} = 0.940$$

4.2.3. Một vài tiêu chí xác định điểm chia tốt nhất

- **Thuộc tính *A* có các giá trị :{a_1, a_2, …,a_v}**
- **Dùng thuộc tính A để phân chia tập huấn luyện D thành *v* tập con {D_1, D_2, …, D_v}**
- ***Thông tin cần thiết để phân chia D theo* thuộc tính A :**

$$Info_A(D) = \sum_{j=1}^{v} \frac{|D_j|}{|D|} Info(D_j)$$

- **Độ lợi thông tin (information gain) dựa trên phân chia theo thuộc tính A :**

$$Gain(A) = Info(D) - Info_A(D)$$

4.2.4. Giải thuật ID3

- Thực hiện quá trình tìm kiếm greedy search đối với không gian các cây quyết định có thể
- Xây dựng (học) một cây quyết định theo chiến lược top-down, bắt đầu từ nút gốc
- Ở mỗi nút, thuộc tính kiểm tra (test attribute) là thuộc tính có khả năng phân loại tốt nhất đối với các ví dụ học gắn với nút đó
- Tạo mới một cây con (sub-tree) của nút hiện tại cho mỗi giá trị có thể Của thuộc tính kiểm tra và tập học sẽ được tách ra (thành các tập con) của thuộc tính kiểm tra tương ứng với cây con vừa tạo
- Mỗi thuộc tính chỉ được phép xuất hiện tối đa 1 lần đối với bất kỳ một đường đi nào trong cây
- Quá trình phát triển (học) cây quyết định sẽ tiếp tục cho đến khi…
- Cây quyết định phân loại hoàn toàn (perfectly classifies) các ví dụ học, hoặc tất cả các thuộc tính đã được sử dụng

4.2.4. Giải thuật ID3

Generate_decision_tree (Sinh cây quyết định): Xây dựng cây quyết định từ dữ liệu huấn luyện cho trước.
Đầuvào: Các mẫu huấn luyện *samples*, là các giá trị rời rạc của các thuộc tính;
Tập các thuộc tính attribute-list.
Đầu ra: Cây quyết định.

4.2.3. Giải thuật ID3

Giải thuật
1) **create** một nút N;
2) **if** tất cả các *samples* có cùng lớp C then
3) **return** N là một nút lá với nhãn lớp C;
4) Else **if** attribute-list là rỗng **then**
5) **return** N là một nút lá với nhãn là lớp phổ biến nhất trong *samples*;
6) Else **select** *test-attribute* - là thuộc tính có thông tin thu được cao nhất trong *attribute-list*;
7) Nhãn nút N là *test-attribute*;
8) **for** mỗi giá trị a_i của *test-attribute*
9) Phát triển một nhánh từ nút N với điều kiện *test-attribute= a_i*;
10) Đặt s_i là tập các mẫu trong *samples* có *test-attribute= a_i*;
11) **if** s_i là rỗng **then**
12) gắn một lá với nhãn là lớp phổ biến nhất trong *samples*;
13) **else** gắn một nút được trả lại bởi Generate_decision_tree(s_i, *attribute-list – test-attribute*);

Ví dụ (1)

Thừa nhận:

- Lớp ***P***: plays_tennis = "yes"
- Lớp ***N***: plays_tennis = "no"
- Thông tin cần thiết để phân lớp một mẫu được cho là:

$$Info(p,n) = Info(9,5) = 0.940$$

Ví dụ (2)

Tính cho thuộc tính *thời tiết*:

thời tiết	p_i	n_i	nfo(p_i, n_i)
nắng	2	3	0.971
u ám	4	0	0
mưa	3	2	0.971

Ta có $Info(thoitiet) = \frac{5}{14} Info(2,3) + \frac{4}{14} Info(4,0) + \frac{5}{14} Info(3,2) = 0.694$

Do đó $Gain(thoitiet) = Info(9,5) - Info(thoitiet) = 0.246$

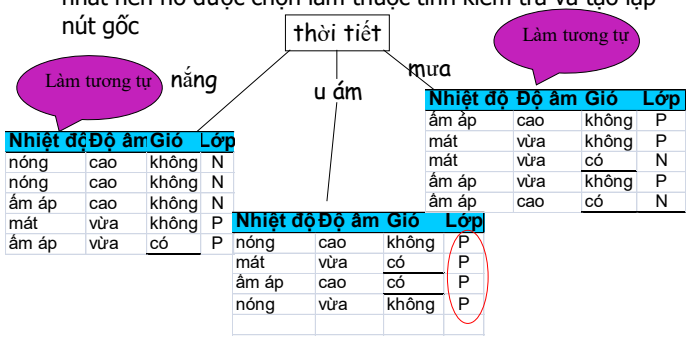
Tương tự $Gain(nhietdo) = 0.029$

$Gain(doam) = 0.151$

$Gain(gio) = 0.048$

Ví dụ (3)

- Như vậy ta thấy thuộc tính thời tiết có độ lợi thông tin lớn nhất nên nó được chọn làm thuộc tính kiểm tra và tạo lập nút gốc

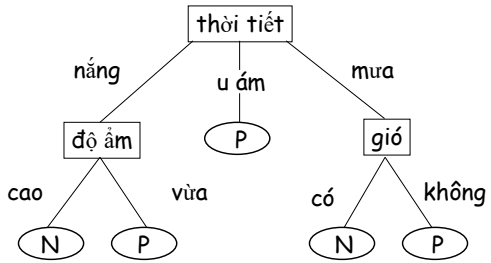


Nhiệt độ	Độ ẩm	Gió	Lớp
nóng	cao	không	N
nóng	cao	không	N
ấm áp	cao	không	N
mát	vừa	không	P
ấm áp	vừa	có	P

Nhiệt độ	Độ ẩm	Gió	Lớp
nóng	cao	không	P
mát	vừa	có	P
ấm áp	cao	có	P
nóng	vừa	không	P

Nhiệt độ	Độ ẩm	Gió	Lớp
ấm ấp	cao	không	P
mát	vừa	không	P
mát	vừa	có	N
ấm áp	vừa	không	P
ấm áp	cao	có	N

Kết quả thu được cây như sau:



Bài tập

Cho tập dữ liệu sau, xây dựng cây quyết định.

age	income	student	credit_rating	buys_computer
<=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

4.2.5.Vấn đề quá phù hợp trong phân lớp bằng cây quyết định

- **Cây tạo được có thể quá phù hợp dữ liệu huấn luyện**
  - Quá nhiều nhánh
  - Độ chính xác kém cho những mẫu chưa biết
- **Lý do quá phù hợp**
  - Dữ liệu nhiễu và tách rời khỏi nhóm
  - Dữ liệu huấn luyện quá ít
  - Các giá trị tối đa cục bộ trong tìm kiếm tham lam (greedy search)

Cách nào để tránh overfitting?

Hai hướng:

- **Rút gọn trước:** ngừng sớm
- **Rút gọn sau:** loại bỏ bớt các nhánh sau khi xây xong toàn bộ cây

Ưu điểm

- **Tính cọ dãn**: phân lớp các tập dữ liệu có hàng triệu mẫu và hàng trăm thuộc tính với tốc độ chấp nhận được
- **Tại sao sử dụng cây quyết định trong khai thác dữ liệu?**
  - Tốc độ phân lớp tương đối nhanh hơn các phương pháp khác
  - Có thể chuyển đổi thành các luật phân lớp đơn giản và dễ hiểu
  - Có thể dùng các truy vấn SQL phục vụ truy cập cơ sở dữ liệu
  - Độ chính xác trong phân lớp có thể so sánh

BÀI TẬP

Dùng thuật toán **ID3** xây dựng cây quyết định và rút ra các luật cho bảng dữ liệu "da rám nắng" sau:

TT	Màu tóc	Chiều cao	Cân nặng	Dùng thuốc?	Kết quả
1	Đen	Tầm thước	Nhẹ	Không	Bị rám
2	Đen	Cao	Vừa phải	Có	Không
3	Râm	Thấp	Vừa phải	Có	Không
4	Đen	Thấp	Vừa phải	Không	Bị rám
5	Bạc	Tầm thước	Nặng	Không	Bị rám
6	Râm	Cao	Nặng	Không	Không
7	Râm	Tầm thước	Nặng	Không	Không
8	Đen	Thấp	Nhẹ	Có	Không

Cho bảng dữ liệu sau:

	Vóc dáng	Quốc tịch	Gia cảnh	Nhóm
O1	Nhỏ	Đức	Độc thân	A
O2	Lớn	Pháp	Độc thân	A
O3	Lớn	Đức	Độc thân	A
O4	Nhỏ	Ý	Độc thân	B
O5	Lớn	Đức	Có gia đình	B
O6	Lớn	Ý	Độc thân	B
O7	Lớn	Ý	Có gia đình	B
O8	Nhỏ	Đức	Có gia đình	B

Sử dụng thuật toán ID3 xây dựng cây quyết định cho bảng dữ liệu trên

•Rút ra các luật từ cây quyết định mới xây dựng.

•Sử dụng các luật đó để phân lớp cho mẫu dữ liệu sau:

Vóc dáng	Quốc tịch	Gia cảnh	Nhóm
Nhỏ	Đức	Có gia đình	
Lớn	Ý	Độc thân	
Nhỏ	pháp	Có gia đình	
Lớn	**Đức**	**Độc thân**	

33

4.3.*Phương pháp Naïve Bayes*

- Naïve Bayes (NB) là phương pháp phân lớp dựa vào xác suất được sử dụng rộng rãi trong lĩnh vực học máy
- **Định lý Bayes**

Cho *X* là mẫu dữ liệu chưa biết nhãn lớp, *H* là giả thuyết mẫu dữ liệu *X* thuộc về lớp *C*. Đối với các bài toán phân lớp, ta cần xác định *P(H|X)* là xác suất xảy ra giả thuyết *H* trên mẫu dữ liệu *X*. *P(H)* là xác suất một mẫu dữ liệu bất kỳ cho trước thuộc về lớp C_i nào đó mà không cần biết trước đặc điểm của mẫu dữ liệu đó.

- Công thức định lý Bayes

$$P(H \mid X) = \frac{P(X \mid H)P(H)}{P(X)}$$

34

III.*Phương pháp Naïve Bayes*

Thời tiết	Nhiệt độ	Độ ẩm	Gió	Chơi tennis?
Nắng	Nóng	Cao	Không	Không
Nắng	Nóng	Cao	Không	Không
U ám	Nóng	Cao	Không	Có
Mưa	ấm áp	Cao	Không	Có
Mưa	Mát	Vừa	Không	Có
Mưa	Mát	Vừa	Có	Không
U ám	Mát	Vừa	Có	Có
Nắng	ấm áp	Cao	Không	Không
Nắng	Mát	Vừa	Không	Có
Mưa	ấm áp	Vừa	Không	Có
Nắng	ấm áp	Vừa	Có	Có
U ám	ấm áp	Cao	Có	Có
U ám	Nóng	Vừa	Không	Có
Mưa	ấm áp	Cao	Có	Không

35

VÍ DỤ

- *Sự kiện H*: Anh ta chơi tennis
- *Sự kiện X*: Thời tiết là *nắng* và có gió.
- *P(H):* Xác suất anh ta chơi tennis bất kể thời tiết như thế nào và có gió hay không.
- *P(X):* Xác suất thời tiết là nắng và có gió.
- *P(X/H):* Xác suất rằng thời tiết là nắng và có gió, nếu biết trước là anh ta chơi tennis.
- *P(H/X):* Xác suất rằng anh ta chơi tennis, nếu biết rằng ngoài trời là nắng và có gió.
- Các xác suất *P(H), P(X), P(X/H)* có thể được tính từ tập dữ liệu cho trước. Xác suất *P(H/X)* được tính theo công thức Bayes dựa vào các xác suất *P(H), P(X), P(X/H).*
- Giá trị xác suất có điều kiện được tính theo công thức Bayes này sẽ được dùng để dự đoán xem anh ta có chơi tennis hay không.

36

III.*Phương pháp Naïve Bayes*

- **Tính độc lập 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

Thuật toán Naïve Bayes

- Giả sử Biểu diễn bài toán phân lớp.
  - Một tập dữ liệu huấn luyện D, trong đó mỗi mẫu X được biểu diễn là một vectơ 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 mẫu dữ liệu chưa biết nhãn lớp X, X sẽ được phân vào lớp nào?
  - Thuật toán Naïve Bayes sẽ tiến hành dự đoán X thuộc vào lớp có xác suất $P(C_i|X)$ cao nhất. Mẫu X được gán vào lớp C_i khi và chỉ khi:
    - $P(C_i|X) > P(C_j|X)$ $(1 \le j \le m, j \ne i)$
- Do vậy cần tìm $P(C_i|X)$ lớn nhất

Thuật toán Naïve Bayes

- Theo định lý Bayes :
- $P(C_i|X) = \frac{P(X|C_i)P(C_i)}{P(X)}$

- Theo tính chất độc lập điều kiện:
- $P(X|C_i) = \prod_{k=1}^{n} P(x_k|C_i)$
- Trong đó:
- $P(x_k|C_i)$: xác suất thuộc tính thứ k mang giá trị x_k khi đã biết X thuộc phân lớp i.

Thuật toán Naïve Bayes

Các bước thực hiện thuật toán Naïve Bayes:

Bước 1: Huấn luyện Naïve Bayes (trên tập dữ liệu huấn luyện), tính $P(C_i)$ và $P(x_k|C_i)$

Bước 2: Phân lớp $X^{new} = (x_1, x_2, \dots, x_n)$, tính xác suất thuộc từng phân lớp khi đã biết trước X^{new}. X^{new} được gán vào lớp có xác suất lớn nhất theo công thức :

$$\max_{C_i \in C}(P(C_i) \prod_{k=1}^{n} P(x_k|C_i))$$

Ví dụ

- . Dùng phương pháp Naïve Bayes dự đoán một mẫu dữ liệu với thông tin sau có thuộc lớp chơi tennis hay không? X^{new} = *(Thời tiết = nắng, Nhiệt độ = mát, Độ ẩm = Cao, Gió = có)*
- **Thực hiện phân lớp theo thuật toán Naïve Bayes:**
- Bước 1: Từ tập dữ liệu huấn luyện tính $P(C_i)$ với i= 1,2. Lớp C_1 = 'có', C_2 = 'không'
- Tính giá trị xác suất cho mỗi phân lớp $P(C_i)$:
- *+ $P(C_1)$ = P(chơi tennis = có) = 9/14 = 0,643*
- *+ $P(C_2)$ = P(chơi tennis = không) = 5/14 = 0,357*

Ví dụ

- Để tính giá trị của $P(X/C_i)$, tiến hành tính các xác suất có điều kiện sau:
- *+ P(Thời tiết = nắng| Chơi tennis = có) = 2/9 ≈ 0,222;*
- *+ P(Thời tiết = nắng | Chơi tennis = không) = 3/5 = 0,600;*
- *+ P(Nhiệt độ = mát | Chơi tennis = có) = 3/9 ≈ 0,333 ;*
- *+ P(Nhiệt độ = mát | Chơi tennis = không) = 1/5 = 0,200;*
- *+ P(Độ ẩm = cao | Chơi tennis = có) = 3/9 ≈ 0,333;*
- *+ P(Độ ẩm = cao | Chơi tennis = không) = 4/5 = 0,800;*
- *+ P(Gió = có | Chơi tennis = có) = 3/9 ≈ 0,333;*
- *+ P(Gió = có | Chơi tennis = không) = 2/5 = 0,4;*

Ví dụ

- Bước 2: Tiến hành phân lớp
- + Sử dụng các xác suất tính được ở trên ta có :
- *P(X|chơi tennis = có) = $P(X/C_1)$ = 0,222*0,333*0,333*0,333 = 0,0082*
- *P(X|chơi tennis = không) = $P(X/C_2)$ = 0,600*0,200*0,800*0,400 = 0,038*
- Mẫu dữ liệu X^{new} sẽ được gán vào lớp có xác suất $P(X/C_i)P(C_i)$ lớn nhất.
- **Đối với phân lớp thứ nhất:**
- *P(C1)*P(X|C1) = 0,643*0,0082 = 0.0053 (1)*
- **Đối với phân lớp thứ hai:**
- *P(C2)*P(X|C2) = 0,357*0,038 = 0.014 (2)*
- Vì (2) > (1) theo công thức trên thì X^{new} thuộc lớp C_2, tức là lớp không chơi tennis.

Bài tập

- **Phân lớp cho mẫu mới sau:**
- Xnew=< thời tiết =nắng, Nhiệt độ = ấm áp, độ ẩm = cao, Gió = có>
- Xnew=< thời tiết = u ám, Nhiệt độ = mát, độ ẩm = vừa, Gió = có>

Bài tập

Bài 1 Cho tập dữ liệu huấn luyện sau:

STT	Tuổi	Thu nhập	Sinh viên	Độ tín nhiệm	Mua máy tính?
1	<30	Cao	Không	Khá tốt	Không
2	<30	Cao	Không	Tốt	Không
3	30-40	Cao	Không	Khá tốt	Có
4	>40	Trung bình	Không	Khá tốt	Có
5	>40	Thấp	Có	Khá tốt	Có
6	>40	Thấp	Có	Tốt	Không
7	30-40	Thấp	Có	Tốt	Có
8	<30	Trung bình	Không	Khá tốt	Không
9	<30	Thấp	Có	Khá tốt	Có
10	>40	Trung bình	Có	Khá tốt	Có
11	<30	Trung bình	Có	Tốt	Có
12	30-40	Trung bình	Không	Tốt	Có
13	30-40	Cao	Có	Khá tốt	Có
14	>40	Trung bình	Không	Tốt	không

Một sinh viên trẻ với mức thu nhập trung bình và mức độ tín nhiệm là khá tốt sẽ mua một máy tính hay không?

Biết X^{new} = *(Tuổi = '< 30', Thu nhập = 'Trung bình', Sinh viên = 'Có', Độ tín nhiệm = 'Khá tốt')* 45