

KHAI PHÁ DỮ LIỆU

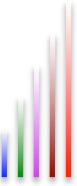
Bài 7

Phân lớp bằng NAÏVE BAYES

Mai Xuân Hùng



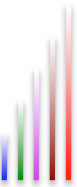
Nội dung



- Đặt vấn đề
- Thuật toán Bayes
- Ví dụ minh họa



Đặt vấn đề

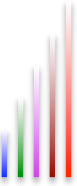


Tên khách	Tuổi	Nghề nghiệp	Mục đích sử dụng	Laptop đã chọn
Tú	Trên 40	Bác sĩ	Đánh văn bản	Acer
Tuấn	18-22	Sinh viên	Học tập	Samsung
Tâm	31-40	Kỹ sư	Thiết kế đồ họa	Dell
Tùng	18-22	Sinh viên		
Trung	31-40	Kỹ sư		
Lâm	Trên 40	Kỹ sư		
Vũ	18-22	Sinh viên		
Minh	31-40	Bác sĩ		
Đạt	18-22	Sinh viên		
Phước	Trên 40	Bác sĩ	Đánh văn bản	
Thiện	18-22	Sinh viên	Học tập	???

Thiện nên mua
máy tính của
hãng nào ???



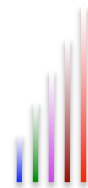
Phân lớp bằng Bayes



- Dự đoán xác suất là thành viên của 1 lớp cho mẫu mới
- Nền tảng: dựa vào định lý Bayes
 - Cho X , Y là các biến bất kì
 - Dự đoán Y từ X
- Lượng giá các tham số của $P(X|Y)$, $P(Y)$ trực tiếp từ tập dữ liệu huấn luyện



Phân lớp Bayes



- Bài toán phân lớp có thể hình thức hóa bằng **xác suất a-posteriori**:

$$P(C|X) = \text{xác suất mẫu}$$

$X = \langle x_1, \dots, x_k \rangle$ thuộc về lớp C

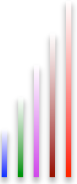
- Ví dụ

$$P(\text{class} = N \mid \text{outlook} = \text{sunny}, \text{windy} = \text{true}, \dots)$$

- **Ý tưởng**: Gán cho mẫu X nhãn phân lớp là C sao cho $P(C|X)$ là lớn nhất



Phân lớp Bayes



- Định lý Bayes

$$P(y | x) = \frac{P(x | y) \cdot P(y)}{P(x)}$$

- Cụ thể

$$P(Y = y_i | X = x_j) = \frac{P(X = x_j | Y = y_i) P(Y = y_i)}{P(X = x_j)}$$

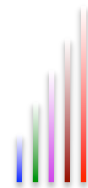
Biến bất kỳ

Giá trị thứ i

2



Ví dụ 1



- Cho tập huấn luyện

Thời tiết	Nhiệt độ	Độ ẩm	Gió	Đi chơi?
Nắng	Nóng	Cao	Yếu	No
Nắng	Nóng	Cao	Mạnh	No
U ám	Nóng	Cao	Mạnh	Yes
Mưa	Mát	Cao	Yếu	Yes
Mưa	Lạnh	Cao	Mạnh	No
Mưa	Lạnh	Bình thường	Mạnh	No
U ám	Lạnh	Bình thường	Yếu	Yes
Nắng	Mát	Cao	Yếu	No
Nắng	Lạnh	Bình thường	Yếu	Yes



Ví dụ 1(tt)

- Ước lượng $P(C_i)$ với $C_1 = \text{“Yes”}$, $C_2 = \text{“No”}$
- Ta thu được $P(C_i)$

$$P(C_1) = 4/9 \quad P(C_2) = 5/9$$

- Với thuộc tính Thời tiết, ta có các giá trị:
Nắng, U ám, Mưa
- Với thuộc tính Nhiệt độ, ta có các giá trị:
Nóng, Mát, Lạnh
- Ta tính $P(\text{Thời tiết}|C_i)$ và $P(\text{Nhiệt độ}|C_i)$ với từng giá trị của thuộc tính

Ví dụ 1(tt)

- $P(\text{Nắng}|C_i)$ là:

Thời tiết	
$P(\text{Nắng} \text{Yes}) = 1/4$	$P(\text{Nắng} \text{No}) = 3/5$

- $P(\text{U ám}|C_i)$ là:

Thời tiết	
$P(\text{Trời u ám} \text{Yes}) = 2/4$	$P(\text{u ám} \text{No}) = 0/5$

- $P(\text{Mưa}|C_i)$ là:

Thời tiết	
$P(\text{Mưa} \text{Yes}) = 1/4$	$P(\text{Mưa} \text{No}) = 2/5$

Ví dụ 1(tt)

- $P(\text{Nóng}|C_i)$ là:

Nhiệt độ	
$P(\text{Nóng} \text{Yes}) = 1/4$	$P(\text{Nóng} \text{No})=2/5$

- $P(\text{Mát}|C_i)$ là:

Nhiệt độ	
$P(\text{Mát} \text{Yes}) = 1/4$	$P(\text{Mát} \text{No})=1/5$

- $P(\text{Lạnh}|C_i)$ là:

Nhiệt độ	
$P(\text{Lạnh} \text{Yes}) = 2/4$	$P(\text{Lạnh} \text{No})=2/5$

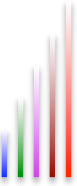
Ví dụ 1(tt)

Hôm nay trời Nắng và

Có nên đi
chơi
không?



Ví dụ 1(tt)



Ta có bảng:

Nắng	Nóng	Đi chơi
1/4	1/4	Yes
3/5	2/5	No

Ta có tỉ lệ sau:

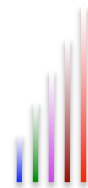
$$P(\text{Yes}|\text{Nắng}, \text{Nóng}) = 1/4 * 1/4 * 4/9 = 0.028$$

$$P(\text{No}|\text{Nắng}, \text{Nóng}) = 3/5 * 2/5 * 5/9 = 0.133$$

→ chọn không đi chơi



Ví dụ 2



❖ Phân lớp X:

- ✓ một mẫu chưa thấy $X = \{mưa, nóng, cao\}$
- ✓ một mẫu chưa thấy $X = \{u ẩm, mát, bình thường, yếu\}$

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



Ví dụ 2(tt)

- Ước lượng $P(x_i|C)$

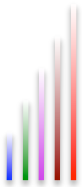
$$P(p) = 9/14$$

$$P(n) = 5/14$$

Thời tiết	
$P(\text{nắng} p) = 2/9$	$P(\text{nắng} n) = 3/5$
$P(\text{u ám} p) = 4/9$	$P(\text{u ám} n) = 0$
$P(\text{mưa} p) = 3/9$	$P(\text{mưa} n) = 2/5$
Nhiệt độ	
$P(\text{nóng} p) = 2/9$	$P(\text{nóng} n) = 2/5$
$P(\text{ấm áp} p) = 4/9$	$P(\text{ấm áp} n) = 2/5$
$P(\text{mát} p) = 3/9$	$P(\text{mát} n) = 1/5$

Độ ẩm	
$P(\text{cao} p) = 3/9$	$P(\text{cao} n) = 4/5$
$P(\text{vừa} p) = 6/9$	$P(\text{vừa} n) = 1/5$
Gió	
$P(\text{có} p) = 3/9$	$P(\text{có} n) = 3/5$
$P(\text{không} p) = 6/9$	$P(\text{không} n) = 2/5$

Ví dụ 2(tt)



- **Phân lớp X :**

O một mẫu chưa thấy $X = \langle \text{mưa}, \text{nóng}, \text{cao}, \text{không} \rangle$

O $P(X|p) \cdot P(p) =$

$$P(\text{mưa}|p) \cdot P(\text{nóng}|p) \cdot P(\text{cao}|p) \cdot P(\text{không}|p) \cdot P(p) = \\ 3/9 \cdot 2/9 \cdot 3/9 \cdot 6/9 \cdot 9/14 = 0.010582$$

O $P(X|n) \cdot P(n) =$

$$P(\text{mưa}|n) \cdot P(\text{nóng}|n) \cdot P(\text{cao}|n) \cdot P(\text{không}|n) \cdot P(n) = \\ 2/5 \cdot 2/5 \cdot 4/5 \cdot 2/5 \cdot 5/14 = \mathbf{0.018286}$$

O Mẫu X được phân vào lớp n (không chơi tennis)



Thuật toán NAÏVE BAYES



Ưu điểm :

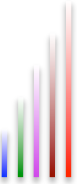
- Dễ dàng cài đặt
- Thời gian thi hành tương tự như cây quyết định
- Đạt kết quả tốt trong phần lớn các trường hợp

Nhược điểm :

- Giả thiết về tính độc lập điều kiện của các thuộc tính làm giảm độ chính xác



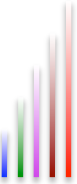
Bài tập



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



Tìm luật phân lớp cho mẫu X



- $X1 = \{\text{thời tiết} = \text{nắng}, \text{độ ẩm} = \text{cao}\}$
- $X2 = \{\text{thời tiết} = \text{nắng}, \text{độ ẩm} = \text{vừa}\}$
- $X3 = \{\text{thời tiết} = \text{U ám}\}$
- $X4 = \{\text{thời tiết} = \text{mưa}, \text{gió} = \text{không}\}$
- $X5 = \{\text{thời tiết} = \text{mưa}, \text{gió} = \text{có}\}$



Kết quả

