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 s		
Vũ	18-22	Sinh		
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***)** = xác suất mẫu

  $X=<x_1,\ldots,x_k>$ thuộc về lớp *C*

- **Ví dụ**

  P(class=***N*** | **outlook=sunny,windy=true,…**)**

- **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

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

- 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

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)

Hôm nay trời Nắng và Nóng

Ví dụ 1(tt)

- Ước lượng $P(C_i)$ với C_1 = "Yes", C_2 = "No"
- Ta thu được $P(C_i)$

  $P(C_1) = 4/9$ $P(C_2) = 5/9$
- Với thuộc tính Thời tiết, ta có các giá trị:

  Nắng, U ám, Mưa
- Với thuộc tính Nhiệt độ, ta có các giá trí:

  Nóng, Mát, Lạnh

➢ Ta tính P(Thời tiết|C_i) và P(Nhiệt độ|C_i) với từng giá trị của thuộc tinh

Ví dụ 1(tt)

- P(Nắng|Ci) là:

Thời tiết	
P(**Nắng**|Yes) = ¼	P(Nắng|No)=3/5

- P(**U ám**|C_i) là:

Thời tiết	
P(Trời u ám|Yes) = 2/4	P(u ám|No)=0/5

- P(**Mưa**|C_i) là:

Thời tiết	
P(**Mưa**|Yes) = 1/4	P(Mưa|No)=2/5

Ví dụ 1(tt)

- P(**Nóng**|C_i) là:

Nhiệt độ	
P(Nóng|Yes) = 1/4	P(Nóng|No)=2/5

- P(**Mát**|C_i) là:

Nhiệt độ	
P(Mát|Yes) = ¼	P(Mát|No)=1/5

- P(**Lạnh**|C_i) là:

Nhiệt độ	
P(Lạnh|Yes) = 2/4	P(Lạnh|No)=2/5

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(**Yes**|Nắng, Nóng) = 1/ 4* 1/ 4*4/9 = 0.028

P(**No**|Nắng, 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,không}*
  - ✓ một mẫu chưa thấy *X* = *{u ám, mát, bình thường, yếu}*

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

Ví dụ 2(tt)

- Ước lượng **P(x_i|C)**

P(p) = 9/14
P(n) = 5/14

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

Độ ẩm	
P(cao | p) = 3/9	P(cao | n) = 4/5
P(vừa | p) = 6/9	P(vừa | n) = 1/5
Gió	
P(có | p) = 3/9	P(có | n) = 3/5
P(không | p) = 6/9	P(fkhông | n) = 2/5

Ví dụ 2(tt)

- **Phân lớp *X*:**

  O một mẫu chưa thấy *X* = <*mưa, nóng, cao, không*>

  O **P(*X*|*p*)*P(*p*) =**

  P(*mưa*|*p*)*P(*nóng*|*p*)*P(*cao*|*p*)*P(*không*|*p*)*P(*p*) =

  3/9·2/9·3/9·6/9·9/14 **=** 0.010582

  O **P(*X*|*n*)·P(*n*) =**

  P(*mưa*|*n*)*P(*nóng*|*n*)*P(*cao*|*n*)*P(*không*|*n*)*P(*n*) =

  2/5·2/5·4/5·2/5·5/14 **= 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

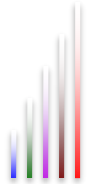
Ví dụ

DAY	Outlook	Temp	Humidity	Windy	Play
D1	Sunny	Hot	High	False	No
D2	Sunny	Hot	High	True	No
D3	Overcast	Hot	High	False	Yes
D4	Rainy	Mild	High	False	Yes
D5	Rainy	Cool	Normal	False	Yes
D6	Rainy	Cool	Normal	True	No
D7	Overcast	Cool	Normal	True	Yes
D8	Sunny	Mild	High	False	No
D9	Sunny	Cool	Normal	False	Yes
D10	Rainy	Mild	Normal	False	Yes
D11	Sunny	Mild	Normal	True	Yes
D12	Overcast	Mild	High	True	Yes
D13	Overcast	Hot	Normal	False	Yes
D14	Rainy	Mild	High	True	No

Bài tập

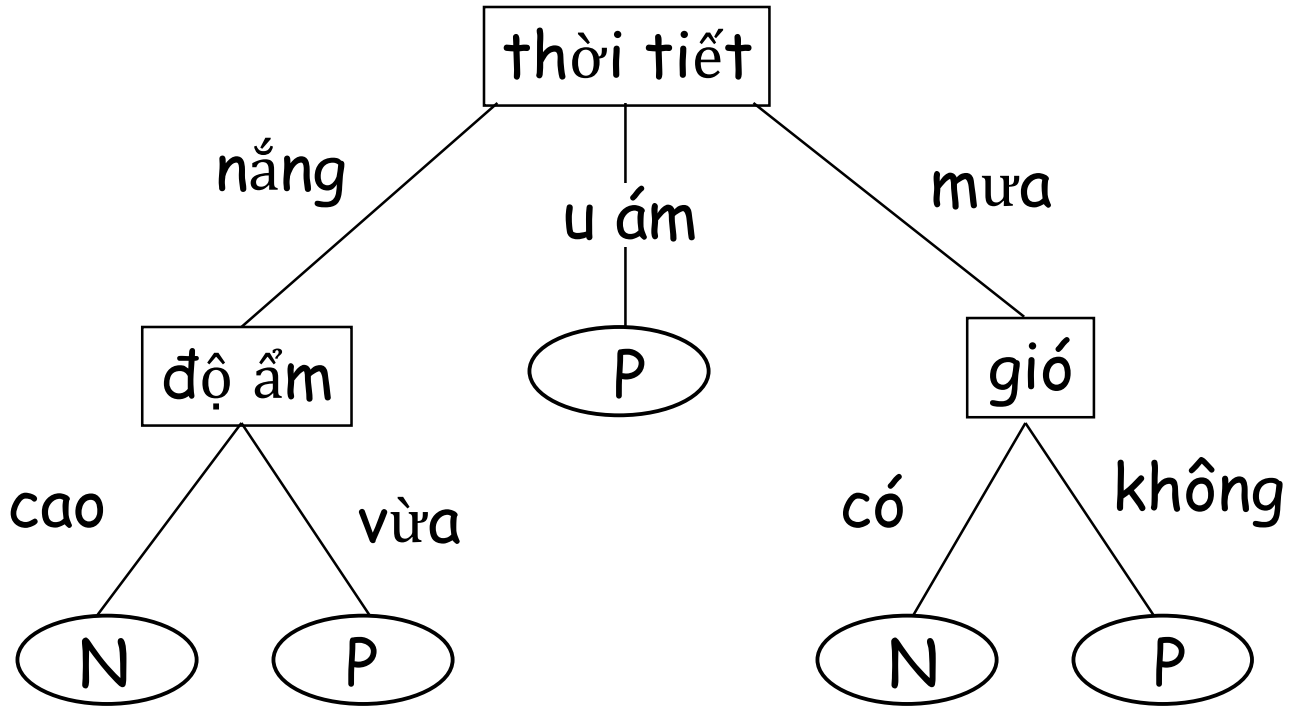
Thời tiết	Nhiệt độ	Độ ẩm	Gió	Lớp
nắng	nóng	cao	không	N
nắng	nóng	cao	không	N
u ám	nóng	cao	không	P
mưa	ấm áp	cao	không	P
mưa	mát	vừa	không	P
mưa	mát	vừa	có	N
u ám	mát	vừa	có	P
nắng	ấm áp	cao	không	N
nắng	mát	vừa	không	P
mưa	ấm áp	vừa	không	P
nắng	ấm áp	vừa	có	P
u ám	ấm áp	cao	có	P
u ám	nóng	vừa	không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

Tìm luật phân lớp cho mẫu X



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

Kết qủa



Thuật toán Bayes – Làm trơn Laplace

- Để tránh trường hợp P(Xk|Ci)=0, áp dụng công thức Laplace
- **$P(Ci)=(|C_{i,D}|+1)/(|D|+m)$**
- **$P(Xk|Ci)=(\#\ C_{i,D}\{xk\}+1)/(|C_{i,D}|+r)$**

Với

- m: số phân lớp
- r: số giá trị rời rạc của thuộc tính

Thuật toán Bayes - Làm trơn Laplace

Outlook	Temperature	Humidity	Wind	Play ball
Sunny	Hot	High	Weak	**No**
Sunny	Hot	High	Strong	**No**
Overcast	Hot	High	Weak	**Yes**
Rainy	Mild	High	Weak	**Yes**
Rainy	Cool	Normal	Weak	**Yes**
Rainy	Cool	Normal	Strong	**No**
Overcast	Cool	Normal	Strong	**Yes**
Sunny	Mild	High	Weak	**No**
Sunny	Cool	Normal	Weak	**Yes**
Rainy	Mild	Normal	Weak	**Yes**
Sunny	Mild	Normal	Strong	**Yes**
Overcast	Mild	High	Strong	**Yes**
Overcast	Hot	Normal	Weak	**Yes**
Rainy	Mild	High	Strong	**No**

Thuật toán Bayes - Làm trơn Laplace

- Áp dụng công thức làm trơn Laplace, phân lớp mẫu X=(Outlook=Overcast, Temp=Cool, Humidity=High, Wind=Strong)

1. P(Play=Yes) = (9+1)/(14+2) = 10/16
2. P(Play=No) = (5+1)/(14+2) = 6/16
3. P(Outlook=Overcast|Play=Yes)=(4+1)/(9+3)=5/12
4. P(Outlook=Overcast|Play=No) = 1/8