S là tập các mẫu thuộc lớp âm và lớp dương

$P_{\oplus}$ là tỷ lệ các mẫu thuộc lớp dương trong S

$p_{\ominus}$ là tỷ lệ các mẫu thuộc lớp âm trong S

$$Entropy(S) = -p_{\oplus} \log_2 p_{\oplus} - p_{\ominus} \log_2 p_{\ominus}$$

$$Gain(S,A) \equiv Entropy(S) - \sum_{v \in Value(A)} \frac{|S_v|}{|S|} Entropy(S_v)$$

Value(A) là tập các giá trị có thể cho thuộc tính A, *và* S_v là tập con của S mà A nhận giá trị v.

Định lí Naive Bayes

Record	A	B	C	Class
1	0	0	0	+
2	0	0	1	-
3	0	1	1	-
4	0	1	1	-
5	0	0	1	+
6	1	0	1	+
7	1	0	1	-
8	1	0	1	-
9	1	1	1	+
10	1	0	1	+

Ứng dụng định lý Naive Bayes để tiên đoán nhãn phân lớp cho dữ liệu X = (A=0, B=1, C=0)

Xác suất tiền định : $P(+) = \frac{1}{2}$; $P(-) = \frac{1}{2}$

Xác suất có điều kiện :

$\mathrm{P}(\mathrm{A} = 0|+) = \frac{2}{5}$; $\mathrm{P}(\mathrm{B} = 1|+) = \frac{1}{5}$; $\mathrm{P}(\mathrm{C} = 0|+) = \frac{1}{5}$;

$\mathrm{P}(\mathrm{A} = 0|-) = \frac{3}{5}$; $\mathrm{P}(\mathrm{B} = 1|-) = \frac{2}{5}$; $\mathrm{P}(\mathrm{C} = 0|-) = 0$;

Với $\mathrm{X} = (\mathrm{A} = 0, \mathrm{B} = 1, \mathrm{C} = 0)$ ta có :

$P(X|+) = \mathrm{P}(\mathrm{A} = 0|+) * \mathrm{P}(\mathrm{B} = 1|+) * \mathrm{P}(\mathrm{C} = 0|+) = \frac{2}{5} * \frac{1}{5} * \frac{1}{5} = \frac{2}{125}$

$P(X|-) = \mathrm{P}(\mathrm{A} = 0|-) * \mathrm{P}(\mathrm{B} = 1|-) * \mathrm{P}(\mathrm{C} = 0|-) = \frac{3}{5} * \frac{2}{5} * 0 = 0$

Suy ra

$$P(+|X) = P(+) * P(X|+) = \frac{1}{2} * \frac{2}{125} = 0.008$$

$$P(-|X) = P(-) * P(X|-) = \frac{1}{2} * 0 = 0$$

Vậy nếu X = (A=0, B=1, C=0) thì thuộc vào lớp +

Cây quyết định :

Cho tập dữ liệu huấn luyện như sau :

RID	age	income	student	credit_rating	Class: bugs_computer
1	youth	high	no	fair	no
2	youth	high	no	excellent	no
3	middle_aged	high	no	fair	yes
4	senior	medium	no	fair	yes
5	senior	low	yes	fair	yes
6	senior	low	yes	excellent	no
7	middle_aged	low	yes	excellent	yes
8	youth	medium	no	fair	no
9	youth	low	yes	fair	yes
10	senior	medium	yes	fair	yes
11	youth	medium	yes	excellent	yes
12	middle_aged	medium	no	excellent	yes
13	middle_aged	high	yes	fair	yes
14	senior	medium	no	excellent	no

a. Sử dụng công thức tính Entropy (Độ hỗn tạp dữ liệu) và Gain (Độ lợi dữ liệu)

$$\mathbf{Entropy}(\mathbf{S}) = \sum_{\mathbf{i}=\mathbf{1}}^{\mathbf{c}} -\mathbf{p_i}\mathbf{log_2}\mathbf{p_i}$$

$$\mathbf{Entropy}([\boldsymbol{a}+, \boldsymbol{b}-]) = \mathbf{0}\ \boldsymbol{if}\ \boldsymbol{a} = \mathbf{0}\ \boldsymbol{or}\ \boldsymbol{b} = \mathbf{0}$$

$$\mathbf{Entropy}([\boldsymbol{a}+, \boldsymbol{b}-]) = \mathbf{1}\ \boldsymbol{if}\ \boldsymbol{a} = \boldsymbol{b}$$

$$\mathbf{Gain}(\mathbf{S}, \mathbf{A}) = \mathbf{Entropy}(\mathbf{S}) - \sum_{\mathbf{V} \in \mathbf{Value}(\mathbf{A})} \frac{|\mathbf{S_V}|}{|\mathbf{S}|} * \mathbf{Entropy}(\mathbf{S_v})$$

Có 14 mẫu trong bảng dữ liệu với 9 mẫu thuộc lớp dương (yes) và 5 mẫu thuộc lớp âm (no) kí hiệu là [9+, 5-]

Với E(S) = E([9+, 5-]) = 0.940

Age		
youth	middle_aged	senior
[2+,3-] E=0.971	[4+,0-] E=0	[3+,2-] E=0.971
Gain(S,Age) = $0.940 - \frac{5}{14} * 0.971 - \frac{4}{14} * 0 - \frac{5}{14} * 0.971 = 0.246$		

Income		
high	medium	low
$[2+, 2-]$ E=1	$[4+, 2-]$ E=0.918	$[3+, 1-]$ E=0.811
Gain(S, Income) = $0.940 - \frac{4}{14} * 1 - \frac{6}{14} * 0.918 - \frac{4}{14} * 0.811 = 0.029$		

Student	
no	yes
$[3+, 4-]$ E=0.985	$[6+, 1-]$ E=0.592
Gain(S, Student) = $0.940 - \frac{7}{14} * 0.985 - \frac{7}{14} * 0.592 = 0.152$	

Credit_rating	
fair	excellent
$[6+, 2-]$ E=0.811	$[3+, 3-]$ E=1
Gain(S, Credit_rating) = $0.940 - \frac{8}{14} * 0.811 - \frac{6}{14} * 1 = 0.048$	

Ta có Gain(S,Age) lớn nhất nên ta chọn Age làm thuộc tính đầu tiên

Age		
youth	middle_aged	senior
{1,2,8,9,11} [2+,3-] E=0.971	{3,7,12,13} [4+,0-] . E=0 <YES>	{4,5,6,10,14} [3+,2-] E=0.971

Với E(S_{youth}) = $Entropy([2+, 3-]) = 0.971$

Income		
high	medium	low
$[0+, 2-]$ E=0	$[1+, 1-]$ E=1	$[1+, 0-]$ E=0
Gain(S, Income) = $0.971 - \frac{2}{5} * 0 - \frac{2}{5} * 1 - \frac{1}{5} * 0 = 0.571$		

Student	
no	yes
$[0+, 3-]$ E=0	$[2+, 0-]$ E=0
Gain(S, Student) = $0.971 - \frac{3}{5} * 0 - \frac{2}{5} * 0 = 0.971$	

Credit_rating	
fair	excellent
$[1+, 2-]$ E=0.918	$[1+, 1-]$ E=1
Gain(S, Credit_rating) = $0.971 - \frac{3}{5} * 0.918 - \frac{2}{5} * 1 = 0.0202$	

Ta có Gain(S, Student) lớn nhất nên ta chọn Age làm thuộc tính

youth	
{1,2,8,9,11} [2+,3-]	
Student	
no	yes
{1,2,8} [0+,3-] . E=0 <NO>	{9,11} [2+,0-] . E=0 <YES>

Với E(S_{Senior}) = $Entropy([3+, 2-]) = 0.971$

Income		
high	medium	low
$[0+, 0-]$ E=0	$[2+, 1-]$ E=0.918	$[1+, 1-]$ E=1
Gain(S, Income) = $0.971 - \frac{0}{5} * 0 - \frac{3}{5} * 0.918 - \frac{2}{5} * 1 = 0.0202$		

Student	
no	yes
$[1+, 1-]$ E=1	$[2+, 1-]$ E=0.918
Gain(S, Student) = $0.971 - \frac{3}{5} * 0.918 - \frac{2}{5} * 1 = 0.0202$	

Credit_rating	
fair	excellent
$[3+, 0-]$ E=0	$[0+, 2-]$ E=0
Gain(S, Credit_rating) = $0.971 - \frac{3}{5} * 0 - \frac{2}{5} * 0 = 0.971$	

Ta có Gain(S, Credit_rating) lớn nhất nên ta chọn Credit_rating làm thuộc tính

senior	
{4,5,6,10,14} [3+,2-]	
Credit_rating	
fair	excellent
{4,5,10} [3+,0-]. E=0 <YES>	{6,14} [0+,2-] . E=0 <NO>

Tóm lại ta có cây quyết định là

Age		
youth	middle_aged	senior
Student	<YES>	Credit_rating

no	yes		fair	excellent
< NO>	<YES>		<YES>	<NO>

b. Vậy Nếu age=senior, income = high, student = yes, credit_rating = fair sẽ có giá trị dự đoán Class: bugs_computer = **YES**

1. Nếu (cột cuối) của hàng này = hàng kia (vd yes = yes thì ra lamda)

Ngược lại kết quả là các thuộc tính khác nhau của 2 hàng

2. Hàm phân biệt là các giá trị của giao của tất cả các ô và hợp của của mỗi giá trị trong ô

3. Rút gọn hàm phân biệt theo luật hút P && (P || Q) = P

VD1

Ví dụ : về ma trận phân biệt

Xét một hệ quyết định

	a	b	c	d
u1	a0	b1	c1	Y
u2	a1	b1	c0	N
u3	a0	b2	c1	N
u4	a1	b1	c1	Y

d: là thuộc tính quyết định

Tìm ma trận phân biệt

	u1	u2	u3	u4
u1	λ			
u2	a,c			
u3	b	λ		
u4	λ	c	a,b	λ

VD2

Ma trận phân biệt

	u1	u2	u3	u4	u5	u6	u7
u1							
u2	λ						
u3	b,c,	b,					
u4	b	b,c	c				
u5	a,b,c	a,b	λ	a,b,c			
u6	a,b,c	a,b	λ	a,b,c,	λ		
u7	λ	λ	a,b,c	a,b	c,	c,	

Hàm phân biệt: f(A) = (b∨c) ∧b∧c∧(a∨b∨c)∧(a∨b)
Rút gọn hàm phân biệt: f(A) = b∧c
Vậy:
Hệ thống có 1 rút gọn là: {b,c}
Nhân của hệ thống: Core = {b,c}

VD3

TABLE I. A DECISION SYSTEM "PLAY SPORT"

	Wind	Temperature	Humidity	Outlook	Play Sport
x_1	Strong	Hot	Normal	Sunny	Yes
x_2	Strong	Mild	Normal	Rain	No
x_3	Weak	Hot	Normal	Rain	No
x_4	Weak	Cool	High	Rain	Yes

DISCERNIBILITY MATRIX OF DECISION SYSTEM "PLAY SPORT"

	x_1	x_2	x_3	x_4
x_1	∅	∅	∅	∅
x_2	b,d	∅	∅	∅
x_3	a,d	∅	∅	∅
x_4	∅	a,b,c	b,c	∅

Hàm phân biệt: $f = (b \vee d) \wedge (a \vee d) \wedge (a \vee b \vee c) \wedge (b \vee c)$.
Rút gọn hàm phân biệt: $f = d \wedge (b \vee c)$.
Các rút gọn tập thuộc tính:
d∧b và d∧c

VD4

Xét một hệ quyết định

	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

a. Tìm ma trận phân biệt
b. Tìm hàm phân biệt của hệ thống
c. Tìm các rút gọn của tập thuộc tính điều kiện.

Ký hiệu : Q: Quốc tịch, V: Vóc dáng, G: Gia cảnh

Ma trận phân biệt

	O1	O2	O3	O4	O5	O6	O7	O8
O1								
O2	λ							
O3	λ	λ						
O4	Q	V,Q	V,Q					
O5	V,G	Q,G	G	λ				
O6	V,Q	Q	Q	λ	λ			
O7	V,Q,G	Q,G	Q,G	λ	λ	λ		
O8	G	V,Q,G	V,G	λ	λ	λ	λ	

Hàm phân biệt: $f(V,Q,G) = Q \wedge (V \vee G) \wedge (V \vee Q) \wedge (V \vee Q \vee G) \wedge G \wedge (Q \vee G)$

Rút gọn hàm phân biệt:

Sử dụng luật hút: $p \wedge (p \vee q) = p$, ta có:

$Q \wedge (V \vee Q) = Q$; $Q \wedge (V \vee Q \vee G) = Q$

$G \wedge (V \vee G) = G$; $G \wedge (Q \vee G) = G$

Vậy: $f(V,Q,G) = Q \wedge G$

Reduct: {Q,G}

Naive Bayes (Thuật toán phân lớp)

- Ví dụ Xét tập mẫu:

Thuộc tính		Lớp
A_1	A_2	
1	0	C_1
0	0	C_1
2	1	C_2
1	2	C_2
0	1	C_1
1	1	??

Xác suất tiền định $P(C_1)=3/5$, $P(C_2)=2/5$.

Xác xuất có điều kiện

$P(A_1=1|C_1)= 1/3$, $P(A_1=1|C_2)= 1/2$,

$P(A_2=1|C_1)= 1/3$, $P(A_2=1|C_2)= 1/2$

Với $X = (A_1=1, A_2=1)$, ta có:

$P(X|C_1) = P(A_1=1|C_1)x\ P(A_2=1|C_1)$

$= (1/3)x(1/3)= 1/9$

$P(X|C_2) = P(A_1=1|C_2)x\ P(A_2=1|C_2)$

$= (1/2)x(1/2)= 1/4$

$\Rightarrow$ $P(C_1|X)=P(C_1)xP(X|C_1)= (3/5)x(1/9)= 1/15$

$P(C_2|X)=P(C_2)xP(X|C_2)= (2/5)x(1/4)= 1/10$

=> $X = (A_1=1, A_2=1)$ thuộc lớp C_2.

B1: Tìm P(c1 | x) và P(c2 | x) //giả sử cái đầu lớn hơn => X thuộc C1

B2: Tìm P(c1) và P(x | c1) tương tự với c2 //vì P(c | x) = P(c) * P(x | c)

B3: Tìm P(x1 | c1) * P(x2 | c1) tương tự với c2 //vì x = x1 + x2

Giải

x = (A1 = 1, A2 = 2)

Ta có P(c1 | x) = P(c1) * P(x | c1)

P(c1) = 3/5

P(x | c1) = P(A1 = 1 | c1) * P(A2 = 1 | c1) = 1/3 * 1/3 = 1/9

=> P(c1 | x) = 3/5 * 1/9 = 1/15

Ta có P(c2 | x) = P(c2) * P(x | c2)

P(c2) = 2/5

P(x | c2) = P(A1 = 1 | c2) * P(A2 = 1 | c2) = 1/2 * 1/2 = 1/4

=> P(c1 | x) = 2/5 * 1/4 = 1/10

Vì P(c1 | x) = 1/15 < P(c2 | x) = 1/10 nên X thuộc lớp c2

Câu 1: Cho một tập dữ liệu như sau:

Record	A	B	C	Class
1	0	0	0	+
2	0	0	1	-
3	0	1	1	-
4	0	1	1	-
5	0	0	1	+
6	1	0	1	+
7	1	0	1	-
8	1	0	1	-
9	1	1	1	+
10	1	0	1	+

Ứng dụng định lý Naïve Bayes để tiên đoán nhãn phân lớp cho dữ liệu X = (A=0, B= 1, C = 0).

Giải

Ta có P(+|X) = P(+) * P(X|+)

P(+) = 5/10 = 1/2

P(X|+) = P(A=0|+) * P(B=1|+) * P(C=0|+)

= 2/5 * 1/5 * 1/5 = 2/15

=> P(+|X) = 1/2 * 2/15 = 1/125

Ta có P(-|X) = P(-) * P(X|-)

P(-) = 5/10 = 1/2

P(X|-) = P(A=0|-) * P(B=1|-) * P(C=0|-)

= 3/5 * 2/5 * 0/5 = 0

=> P(-|X) = 1/2 * 0/15 = 0

Vì P(+ | x) = 1/125 > P(- | x) = 0 nên X thuộc lớp +

Tính entropy

S là tổng số mẫu trong CSDL

x là số mẫu lớp 1

y là số mẫu lớp 2

Entropy(x+, y-) = -(x/S)*log2(x/S) - (y/S)*log2(y/S)

Ví dụ 14 mẫu, 9 dương, 5 âm:

Entropy(9+, 5-) = -(9/14)*log2(9/14) - (5/14)*log2(5/14) = 0.94

* Nếu e(x+,0) hoặc e(0, x-) thì entropy = 0

* Nếu e(x+, x-) thì entropy = 1

Tính Gain

A = [A1, A2]

Gian(S, A) = Entropy(S) - SUM(len(A[i]) / S * entropy(A[i]))

Ví dụ Wind = {Weak, Strong}

S_weak = (6+ , 2-)

S_strong = (3+, 3-)

Gian(S, Wind) = entropy(S) - (len(S_weak) / S * entropy(S_weak) + len(S_strong) / S * entropy(S_trong)

Ta có entropy(S_week) = -6/8*log2(6/8) - 2/8*log2(2/8) = 0.81

entropy(S_strong) = 1

=> Gian(S, Wind) = 0.94 - (8/14 * 0.81 + 6/14 * 1) = 0.048

Xây dựng cây quyết định bằng ID3

B1: Xác định hệ số x, y của entropy(S) và entropy của từng ô thuộc tính

B2: Tính entropy

B3: Tinh Gain(S, từng thuộc tính) để chọn cái max làm root

B4: Vẽ root và nhánh

B5: Tính entropy(nhánh, các thuộc tính còn lại)

B6: Tinh Gain(nhánh, từng thuộc tính còn lại) để chọn cái max làm node của nhánh

B7: Lặp lại cho đến khi gặp điều kiện dừng

Điều kiện dừng:

1. Tất cả thuộc tính đã được đưa vào
2. Full Yes hoặc full No

Ví dụ sau:

Cây quyết định :
Cho tập dữ liệu huấn luyện như sau :

RID	age	income	student	credit_rating	Class: bugs_computer
1	youth	high	no	fair	no
2	youth	high	no	excellent	no
3	middle_aged	high	no	fair	yes
4	senior	medium	no	fair	yes
5	senior	low	yes	fair	yes
6	senior	low	yes	excellent	no
7	middle_aged	low	yes	excellent	yes
8	youth	medium	no	fair	no
9	youth	low	yes	fair	yes
10	senior	medium	yes	fair	yes
11	youth	medium	yes	excellent	yes
12	middle_aged	medium	no	excellent	yes

13	middle_aged	high	yes	fair	yes
14	senior	medium	no	excellent	no

a. Xây dựng cây quyết đinh bằng thuật toán ID3

b. Vậy Nếu age=senior, income = high, student = yes, credit_rating = fair sẽ có giá trị dự đoán Class: bugs_computer = ?

Giải

Xét tập hợp S

Gọi e(S) là viết tắt của entropy(S)

g(S) là viết tắt của gain(S)

len(S) là chiều dài của tập S

x là số thuộc tính phân lớp yes

y là số thuộc tính phân lớp no

A là tập thuộc tính điều kiện

Ta có công thức sau:

e(S) = e(x+, y-) = -x/S*log2(x/S) - y/S * log2(y/S)

g(S -> A) = e(S) - 1/len(S) * SUM(len(S -> A[i]) * e(S -> A[i]))

<1> Tìm node Root

e(S) = e(9+, 5-) = 0.940

lens(S) = 14

g(S -> age) = e(S) - 1/len(S) * SUM(len(S -> youth)*e(S -> youth)
+ len(S -> middle_aged)* e(S -> middle_aged) + len(S -> senior) * e(S -> senior)

len(S->youth) = 5

len(S->middle_aged) = 4

len(S->senior) = 5

e(S->youth) = e(2+, 3-) = 0,971

e(S->middle_aged) = e(4+,0-) = 0

e(S->senior) = e(3+, 2-) = 0,971

=> g(S -> age) = 0.246

.....

Ví dụ xây dựng cây quyết định từ bảng sau:

Bảng dữ liệu huấn luyện (Training data)

Day	Outlook	Temp	Humidity	Wind	PlayTennis
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
D13	Overcast	Hot	Normal	Weak	Yes
D14	Rain	Mild	High	Strong	No

Giải:

Giả sử S là một tập hợp
len(S) là số mẫu của tập S
e(S) là entropy của tập S
g(S) là information gain của tập S
x là số hàng cho thuộc tính quyết định Yes
y _____ No
A là tập hợp các thuộc tính điều kiện

Ta có công thức tính
e(x+, y-) = -x/S * log2(x/S) - y/S * log2(y/S)
g(S, A) = e(S) - 1/len(S) * (len(A[i]*e(A[i])

S = {D1 .. D14}
entropy(S) = entropy(9+, 5-) = -9/14*log2(9/14) - 5/14*log2(5/14) = 0.94

Feature Outlook:

entropy(Sunny) = entropy(2+, 3-) = 0.97

entropy(Overcast) = entropy(4+, 0-) = 0

entropy(Rain) = entropy(3+, 2-) = 0.97

Feature Temp:

entropy(Hot) = entropy(2+, 2-) = 1

entropy(Mild) = entropy(4+, 2-) = 0.91

entropy(Cool) = entropy(3+, 1-) = 0.81

Feature Humidity:

entropy(High) = entropy(3+, 4-) = 0.985

entropy(Normal) = entropy(6+, 1-) = 0.59

Feature Wind:

entropy(Weak) = entropy(6+, 2-) = 0.81

entropy(Strong) = entropy(3+, 3-) = 1

gain(S, outlook) = entropy(S) - len(Sunny) / S *entropy(Sunny)

- len(Overcast) / S *entropy(Overcast)

- len(Rain) / S *entropy(Rain)

= 0.94 - 5 / 14 * 0.97 - 0 - 5/14 * 0.97 = 0.247

gain(S, windy) = entropy(S) - len(Weak) / S *entropy(Weak)

- len(Strong) / S *entropy(Overcast)

= 0.94 - 8 / 14 * 0.81 - 6/14 * 1 = 0,048

...

Vì gain(S, outlook) lớn nhất nên ta chọn làm root

Outlook

(Sunny) **(Overcast 4+,0-)** **(Rain)**

YES

entropy(Sunny, Humidity)

entropy(Sunny, Temp)

entropy(Sunny, Humidity)

Gain(Sunny, Humidity) = 0.97

Gain(Sunny, Temp) = 0.57

Gain(Sunny, Humidity) = 0.019

Vì Gain(Sunny, Humidity) lớn nhất nên node tiếp theo của cạnh Sunny là Humidity

...

Vì nhánh Overcaste có có 4+ và 0- nên node của nhánh là YES

Outlook

(Sunny) **(Overcast)** **(Rain)**

Humidity **YES**

Rain{ D4, D5, D6, D10, D14}
e(S, Rain) = 0.97
Gain(Rain, Temp) = e(S, Rain) - 1/len(S, Rain)
* (len(Rain, Hot)*e(Rain, Hot)
+ len(Rain, Mild)*e(Rain, Mild)
+ len(Rain, Cool) * e(Rain, Cool))

Gain(Rain, Wind) = e(S, Rain) - 1/len(S, Wind)
* (len(Rain, Week)*e(Rain, Week)
+ len(Mild)*e(Rain, Mild)
+ len(Cool) * e(Rain, Cool))