# A white background with black text Description automatically generated

# A black and yellow text Description automatically generated

# 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 :

Xác suất có điều kiện :

; ; ;

; ; ;

Với ta có :

\* \* =

\* \* =

Suy ra

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 |

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

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= | [4+,0-]  E=0 | [3+,2-]  E= |
| Gain(S,Age) = | | |

|  |  |  |
| --- | --- | --- |
| Income | | |
|  |  |  |
| E=1 | E= | E= |
| Gain(S, Income) = | | |

|  |  |
| --- | --- |
| Student | |
| no | yes |
| E= | E= |
| Gain(S, Student) = | |

|  |  |
| --- | --- |
| Credit\_rating | |
|  |  |
| E= | E=1 |
| Gain(S, Credit\_rating) = | |

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() =**

|  |  |  |
| --- | --- | --- |
| Income | | |
|  |  |  |
| E=0 | E= | E= |
| Gain(S, Income) = | | |

|  |  |
| --- | --- |
| Student | |
| no | yes |
| E= | E= |
| Gain(S, Student) = | |

|  |  |
| --- | --- |
| Credit\_rating | |
|  |  |
| E= | E=1 |
| Gain(S, Credit\_rating) = | |

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() =**

|  |  |  |
| --- | --- | --- |
| Income | | |
|  |  |  |
| E=0 | E= | E= |
| Gain(S, Income) = | | |

|  |  |
| --- | --- |
| Student | |
| no | yes |
| E=1 | E= |
| Gain(S, Student) = | |

|  |  |
| --- | --- |
| Credit\_rating | |
|  |  |
| E= | E=0 |
| Gain(S, Credit\_rating) = | |

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 |  |
| {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**

[A screenshot of a computer

Description automatically generated](https://www.blogger.com/blog/post/edit/4242632031927375983/6544185123914510527)

**VD2**

[A screenshot of a computer

Description automatically generated](https://www.blogger.com/blog/post/edit/4242632031927375983/6544185123914510527)

**VD3**[A screenshot of a test

Description automatically generated](https://www.blogger.com/blog/post/edit/4242632031927375983/6544185123914510527)

**VD4**

[A screenshot of a computer

Description automatically generated](https://www.blogger.com/blog/post/edit/4242632031927375983/6544185123914510527)

[A screenshot of a computer

Description automatically generated](https://www.blogger.com/blog/post/edit/4242632031927375983/6544185123914510527)

### [Naive Bayes (Thuật toán phân lớp)](https://laptrinhthidau.blogspot.com/2024/04/thuat-toan-phan-lop-naive-bayes.html)

[A math equations and numbers

Description automatically generated with medium confidence](https://blogger.googleusercontent.com/img/a/AVvXsEjcNko2sCO4oLcPyNHexGnopKzYStDHrC8dMRrtFGCoN1fJvFnu8H-fKq5JMXsb1DVBYErI7LIXYCVr9uGggIYsCMussPNxk6OA5KmfHZj2iXIukvgZgV3oR3_DHcMaFmK_VOY3HH33on7aA-nN2VwUbEOlVIWBKAUPmIHSFAY6wQXEyogYMQyyHgzg5wE)

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

[A white sheet of paper with black text

Description automatically generated](https://blogger.googleusercontent.com/img/b/R29vZ2xl/AVvXsEgEeHsnspbA7aUsvAMUb67izknIC9LQ6NfryESjF5GJTWCTiDxGC1mUo0muY-NJdqzUSesQ-cy_Aix-cDWL8BiZwat4qFWtPI7hkv9qQq2oeizp-2fe_efj82CB_ysrEWIzikJXKJ29vX_4PQKUd6M6_eDksBytHWizqZIE3EW78tmvVXAq1umvVs6DrLI/s572/naivebayes.png)

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:

[A screenshot of a computer screen

Description automatically generated](https://www.blogger.com/blog/post/edit/4242632031927375983/1815375613888226323)

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) )