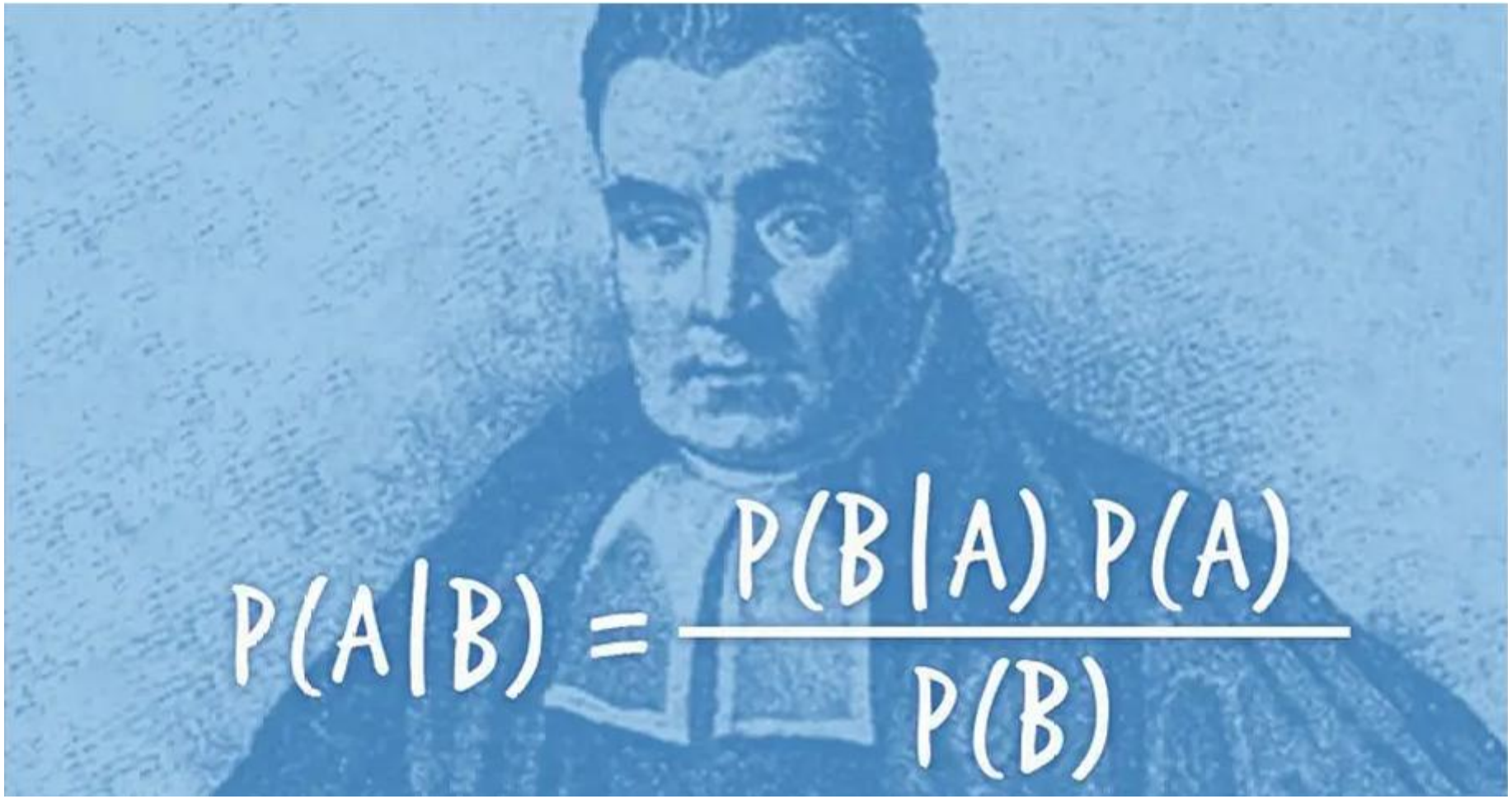
Naïve Bayes Classifier

Giới thiệu



Định lý Bayes

- Tìm ra xác suất xảy ra của một sự kiện khi biết xác suất của một sự kiện khác đã xảy ra.
- Định lý Bayes được phát biểu về mặt toán học như phương trình sau:
  - P(A|B): Xác suất của A xảy ra khi biết B đã xảy ra. *(Kết quả chúng ta cần tìm)*.
  - P(B|A): Xác suất của B xảy ra khi biết A xảy ra. (Khả năng xảy ra)
  - P(A): Xác suất ban đầu của A (Tiên nghiệm)
  - P(B): Tổng xác suất xảy ra của B (Bất kể A có đúng hay không)

$$P(A|B) = \frac{P(B|A) \cdot P(A)}{P(B)}$$

Một số khái niệm

- **Tình huống:** Một email gửi đến có chứa từ "MIỄN PHÍ". Liệu đó có phải là Thư rác?
- **Gán biến:**
  - A: Email là Thư rác.
  - B: Email chứa từ "Miễn phí".
- Dữ liệu:
  - P(A) = 40% (Xác suất rác thông thường)
  - P(B|A) = 80% (Thư rác thường có chữ "Miễn phí")
  - P(B) = 50% (Tổng email có chữ "Miễn phí")

$$P(A|B) = \frac{P(B|A) \cdot P(A)}{P(B)}$$

$$P(\text{Spam}|\text{"Miễn phí"}) = \frac{0.8 \times 0.4}{0.5}$$

NAIVE BAYES CLASSIFIER

- Thuật toán phân loại dựa trên xác suất mạnh mẽ
- Một thuật toán học máy đơn giản nhưng hiệu quả, dựa trên Định lý Bayes và giả định 'ngây thơ' về sự độc lập của các đặc trưng.

ĐỊNH LÝ BAYES 'NGÂY THƠ'

$$P(y \mid x_1, x_2, \ldots, x_n) = \frac{P(y) \prod_{i=1}^{n} P(x_i \mid y)}{P(x_1, x_2, \ldots, x_n)}$$

- $P(y \mid x_1, x_2, \ldots, x_n)$: Xác suất một mẫu thuộc lớp y khi có các đặc trưng $x_1, x_2, \ldots, x_n$. (Đây là điều chúng ta muốn tìm).
- $P(y)$: Xác suất tiền nghiệm của lớp y. (Mức độ phổ biến của lớp y).
- $\prod_{i=1}^{n} P(x_i \mid y)$: Tích của xác suất mỗi đặc trưng x_i xuất hiện khi mẫu thuộc lớp y.
- Giả định "Ngây thơ": Các đặc trưng x_i là độc lập với nhau khi biết lớp Y. (Ví dụ: Sự xuất hiện của từ "mua" độc lập với từ "giảm giá" khi biết email là thư rác).

ĐỊNH LÝ BAYES 'NGÂY THƠ'

$$P(y \mid x_1, x_2, \ldots, x_n) = \frac{P(y) \prod_{i=1}^{n} P(x_i \mid y)}{P(x_1, x_2, \ldots, x_n)}$$

$$P(x_1, x_2, \ldots, x_N \mid y) = P(x_1 \mid y) \cdot P(x_2 \mid y) \cdots P(x_N \mid y)$$

$$P(y \mid x_1, \ldots, x_n) = \frac{P(y) \cdot \prod_{i=1}^{n} P(x_i \mid y)}{P(x_1) \cdot P(x_2) \cdots P(x_n)}$$

$$P(y \mid x_1, \ldots, x_n) \propto P(y) \cdot \prod_{i=1}^{n} P(x_i \mid y)$$

CÁC MÔ HÌNH CỦA NAÏVE BAYES

- Gaussian Naıve Bayes
- Multinomial Naıve Bayes
- Bernouli Naıve Bayes

Gaussian Naive Bayes

- Đây là một thuật toán phân loại thuộc họ **Naïve Bayes**, được dùng khi **các đặc trưng (features) là dữ liệu liên tục**.
- Dữ liệu liên tục => phù hợp với phân phối chuẩn:
- Phân phối chuẩn:
  - Giá trị gần trung bình xuất hiện rất nhiều
  - Giá trị quá lớn hoặc quá nhỏ xuất hiện ít

$$P(x \mid \text{class}) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

Gaussian Naive Bayes

Tính khả năng xảy ra (GaussianLikelihood)

$$P(x \mid \text{class}) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

Tính tiên nghiệm:

- $P(class = 0)$
- $P(class = 1)$

Chuẩn hóa xác suất:

$$P(class \mid x_i) = P(class) \times \prod_{i=1}^{n} P(x_i \mid class)$$

Dự đoán:

- $P(class = 0 \mid x_i) > P(class = 1 \mid x_i)$=> class 0
- $P(class = 1 \mid x_i) > P(class = 1 \mid x_i)$=> class 1

Độ dài cánh hoa	Class
1.3	0
1.4	0
1.5	0
4.5	1
4.6	1
4.7	1
1.0	????
6.4	????
3.3	????

Gaussian Naive Bayes

Độ dài cánh hoa	Class
1.3	0
1.4	0
1.5	0
4.5	1
4.6	1
4.7	1

- Class 0:

$$\mu_0 = \frac{1.4 + 1.3 + 1.5}{3} = 1.4$$

$$\sigma_0^2 = \frac{(1.4 - 1.4)^2 + (1.3 - 1.4)^2 + (1.5 - 1.4)^2}{3} = 0.0067$$

- Class 1:

$$\mu_1 = \frac{4.5 + 4.7 + 4.6}{3} = 4.6$$

$$\sigma_1^2 = \frac{(4.5 - 4.6)^2 + (4.7 - 4.6)^2 + (4.6 - 4.6)^2}{3} = 0.0067$$

Gaussian Naive Bayes

- Tính GaussianLikelihood:
- Class 0:

$$P(1.5|C=0) = \frac{1}{\sqrt{2\pi \times 0.0067}} \times e^{-\frac{(1.5-1.4)^2}{2\times 0.0067}} = 0.247$$

- Class 1:

$$P(1.5|C=1) = \frac{1}{\sqrt{2\pi \times 0.0067}} \times e^{-\frac{(1.5-4.6)^2}{2\times 0.0067}} = 0$$

Gaussian Naive Bayes

- Tính tiên nghiệm:
  - $P(class = 0)$ =0.5
  - $P(class = 1)$ =0.5
- Chuẩn hóa xác suất:
- $P(class = 0 \,|1.5) = P(class = 0) \times P(1.5|class = 0)$
  = 0.5 x 0.247 = 0.1235
- $P(class = 1 \,|1.5) = P(class = 1) \times P(1.5|class = 1)$

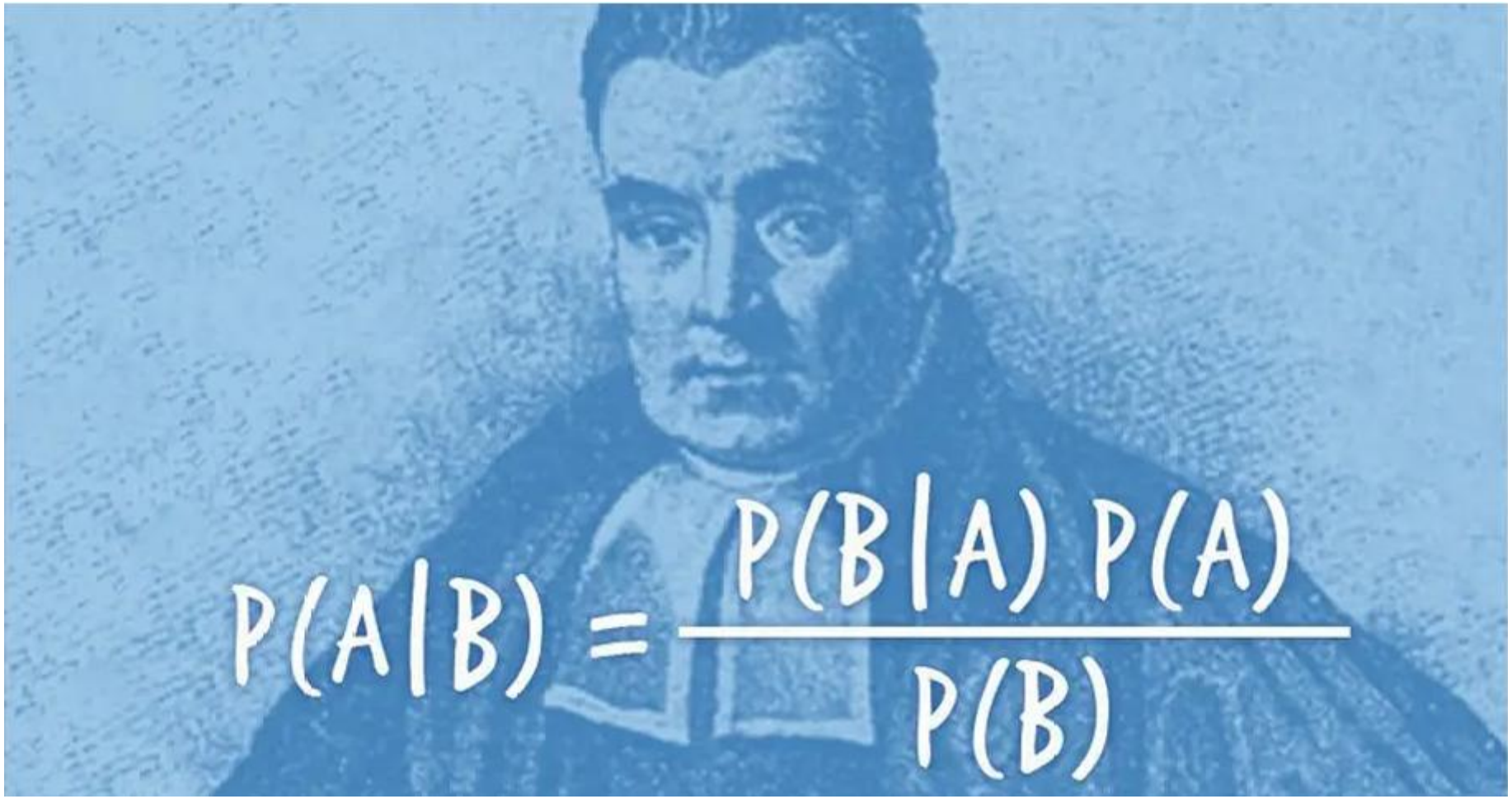
= 0.5 x 0 = 0

- Dự đoán:
  $P(class = 0 \,|1.5) > P(class = 1 \,|1.5)$ => Class 0

TT	Chiều cao	Cân nặng	Cụm
1	160	50	Gầy
2	159	49	Gầy
3	162	52	Bình thường
4	161	51	Bình thường
5	172	72	Bình thường
6	180	85	Mập
7	182	86	Mập
8	170	70	Mập
9	171	71	Mập
10	181	87	Mập
11	**169**	**49**	**?????**

Naıve Bayes Classifier

Giới thiệu



Multinomial Naıve Bayes

- Multinomial Naıve Bayes là một biến thể của thuật toán Bayes Naıve.
- Thường được sử dụng trong các bài toán phân loại văn bản.
- “Multinomial” đề cập đến số lần một từ xuất hiện hoặc tần suất xuất hiện của một danh mục.

Multinomial Naive Bayes

- Ý tưởng chính là giả định rằng mỗi từ trong một thông điệp hoặc đặc trưng là độc lập với nhau.
- Phương pháp này hoạt động bằng cách sử dụng số lượng từ để phân loại văn bản:
- $P(w_i, c) = \frac{count(w_i,c)+1}{N+v}$
- $count(w_i, c)$: số lần từ w_i xuất hiện trong đoạn văn thuộc class c
- N: tổng số từ trong đoạn văn thuộc class c
- v: kích thước của từ vựng (vocab size)

Quy trình xây dựng mô hình Multinomial Naıve Bayes

- Tính xác suất xảy ra của từng đặc trưng:
- Tính tiên nghiệm:
- Tính xác suất
- Dự đoán:

TT	Nội dung	Lớp
1	"buy cheap now"	Spam
2	"limited offer buy"	Spam
3	"meet me now"	Not Spam
4	"let’s catch up"	Not Spam
5	“buy now”	???

Quy trình xây dựng mô hình Multinomial Naıve Bayes

- Tính:
  - Vocal = {buy, cheap, now, limited, offer, meet, me, let's, catch, up} => v = 10
  - $count(buy, c = spam) = 2$
  - $count(cheap, c = spam) = 1$
  - $count(now, c = spam) = 1$
  - $count(limited, c = spam) = 1$
  - $count(offer, c = spam) = 1$
  - => N_spam = 6
  - $count(meet, c = not\ spam) = 1$
  - $count(me, c = not\ spam) = 1$
  - $count(now, c = not\ spam) = 1$
  - $count(let's, c = not\ spam) = 1$
  - $count(catch, c = not\ spam) = 1$
  - $count(up, c = not\ spam) = 1$
  - => N_not spam =6

TT	Nội dung	Lớp
1	"buy cheap now"	Spam
2	"limited offer buy"	Spam
3	"meet me now"	Not Spam
4	"let's catch up"	Not Spam
5	"buy now"	???

Quy trình xây dựng mô hình Multinomial Naıve Bayes

- Tính xác suất xảy ra của từng đặc trưng:
  - $P(x_i|Y) = P(w_i, c) = \frac{count(w_i,c)+1}{N+v}$
  - $P(buy|Y = spam) = \frac{2+1}{6+10} = \frac{3}{16}$
  - $P(now|Y = spam) = \frac{1+1}{6+10} = \frac{2}{16}$
  - $P(buy|Y = not\ spam) = \frac{0+1}{6+10} = \frac{1}{16}$
  - $P(buy|Y = not\ spam) = \frac{1+1}{6+10} = \frac{2}{16}$
- Tính tiên nghiệm
  - $P(Y = spam) = 2/4 = 0.5$
  - $P(Y = not\ spam) = 2/4 = 0.5$
- Tính xác suất
  - $P(Y|x_1, x_2, \dots, x_n) \propto P(Y) \times \prod_i^n P(x_i|Y)$

TT	Nội dung	Lớp
1	"buy cheap now"	Spam
2	"limited offer buy"	Spam
3	"meet me now"	Not Spam
4	"let's catch up"	Not Spam
5	"buy now"	???

- Tính xác suất:
  - $P(Y = spam|buy, now) = 0.5 \times \frac{3}{16} \times \frac{2}{16} = \frac{3}{256}$
  - $P(Y = not\ spam|buy, now) = 0.5 \times \frac{1}{16} \times \frac{2}{16} = \frac{1}{256}$
- Dự đoán:
  - $P(Y = spam|buy, now) > P(Y =$

```
import pandas as pd
from sklearn.feature_extraction.text import CountVectorizer
from sklearn.model_selection import train_test_split
from sklearn.Nave_bayes import MultinomialNB
from sklearn.metrics import accuracy_score

data = {
    "text": [
        "Free money now",
        "Call now to claim your prize",
        "Meet me at the park",
        "Lets catch up later",
        "Win a new car today!",
        "Lunch plans?",
        "Congratulations! You won a lottery",
        "Can you send me the report?",
        "Exclusive offer for you",
        "Are you coming to the meeting?"
    ],
    "label": ["spam", "spam", "not spam", "not spam", "spam", "not spam", "spam", "not
                                                spam", "spam", "not spam"]
}

df = pd.DataFrame(data)

df["label"] = df["label"].map({"spam": 1, "not spam": 0})

X = df["text"]
y = df["label"]
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.3, random_state=
                                                42)

vectorizer = CountVectorizer()
X_train_vectors = vectorizer.fit_transform(X_train)
X_test_vectors = vectorizer.transform(X_test)
```

Phân loại Naive Bayes

- Nếu x=“*A very close game* ”
- Thì y=?

Text	Category
A Great Game	Sports
The Election was over	Not Sports
Very clean match	Sports
A clean but forgettable game	Sports
It was a close election	Not Sports

Naïve Bayes classifier

$$X_{input} = (Rain, Cool, High, Strong)$$

$$Y = ?$$

PlayTennis: training examples

Day	Outlook	Temperature	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

Bernoulli Naıve Bayes

- Bernoulli Naıve Bayes là một biến thể của thuật toán Bayes Naıve.
- Thường được sử dụng khi dữ liệu ở dạng nhị phân

Bernoulli Naive Bayes

- Ý tưởng chính là giả định rằng mỗi từ trong một thông điệp hoặc đặc trưng là độc lập với nhau.
- Phương pháp này hoạt động bằng cách sử dụng sự xuất hiện của từ để phân loại văn bản:
- $P(w_i, c) = \frac{count(w_i,c)+1}{N+2}$
- $count(w_i, c)$: số văn bản có sự xuất hiện của từ w_i thuộc class c
- N: tổng số văn bản thuộc class c
- $+2$: Mỗi đặc trưng (từ) có hai khả năng: xuất hiện (1) hoặc không xuất hiện (0) => cộng thêm 2 để bao quát cả hai khả năng này.

Quy trình xây dựng mô hình Bernoulli Naıve Bayes

- Tính xác suất xảy ra của từng đặc trưng:
- Tính tiên nghiệm:
- Tính xác suất
- Dự đoán:

TT	Nội dung	Lớp
1	"buy cheap now"	Spam
2	"limited offer buy"	Spam
3	"meet me now"	Not Spam
4	"let’s catch up"	Not Spam
5	“buy now”	???

Quy trình xây dựng mô hình Bernoulli Naıve Bayes

- Tính:
  - $count(buy, c = spam) = 2$
  - $count(cheap, c = spam) = 1$
  - $count(now, c = spam) = 1$
  - $count(limited, c = spam) = 1$
  - $count(offer, c = spam) = 1$
  - => N_spam = 2
  - $count(meet, c = not\ spam) = 1$
  - $count(me, c = not\ spam) = 1$
  - $count(now, c = not\ spam) = 1$
  - $count(let's, c = not\ spam) = 1$
  - $count(catch, c = not\ spam) = 1$
  - $count(up, c = not\ spam) = 1$
  - => N_not spam =2

TT	Nội dung	Lớp
1	"buy cheap now"	Spam
2	"limited offer buy"	Spam
3	"meet me now"	Not Spam
4	"let's catch up"	Not Spam
5	"buy now"	???

ID	buy	cheap	now	limited	offer	meet	me	let's	catch	up	Class
M1	1	1	1	0	0	0	0	0	0	0	Spam
M2	1	0	0	1	1	0	0	0	0	0	Spam
M3	0	0	1	0	0	1	1	0	0	0	Not Spam
M4	0	0	0	0	0	0	0	1	1	1	Not Spam

Quy trình xây dựng mô hình Bernoulli Naive Bayes

- Tính xác suất xảy ra của từng đặc trưng:
  - $P(x_i|Y) = P(w_i, c) = \frac{count(w_i,c)+1}{N+2}$
  - $P(buy|Y = spam) = \frac{2+1}{2+2} = \frac{3}{4}$
  - $P(now|Y = spam) = \frac{1+1}{2+2} = \frac{1}{2}$
  - $P(buy|Y = not\ spam) = \frac{0+1}{2+2} = \frac{1}{4}$
  - $P(now|Y = not\ spam) = \frac{1+1}{2+2} = \frac{1}{2}$
- Tính tiên nghiệm
  - $P(Y = spam) = 2/4 = 0.5$
  - $P(Y = not\ spam) = 2/4 = 0.5$
- Tính xác suất
  - $P(Y|x_1, x_2, \dots, x_n) \propto P(Y) \times \prod_i^n P(x_i|Y)$

TT	Nội dung	Lớp
1	"buy cheap now"	Spam
2	"limited offer buy"	Spam
3	"meet me now"	Not Spam
4	"let's catch up"	Not Spam
5	"buy now"	???

- Tính xác suất:
  - $P(Y = spam|buy, now) = 0.5 \times \frac{3}{4} \times \frac{2}{4} = \frac{1.5}{8}$
  - $P(Y = not\ spam|buy, now) = 0.5 \times \frac{1}{4} \times \frac{2}{4} = \frac{0.5}{8}$
- Dự đoán:
  - $P(Y = spam|buy, now)$ > $P(Y =$

Naïve Bayes classifier

$$X_{input} = (Rain, Cool, High, Strong)$$

$$Y = ?$$

PlayTennis: training examples

Day	Outlook	Temperature	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

```
import numpy as np
import pandas as pd
from sklearn.Nave_bayes import BernoulliNB
from sklearn.feature_extraction.text import CountVectorizer

df=pd.read_csv("spam_ham_dataset.csv")
print(df.shape)
print(df.columns)
df= df.drop(["Unnamed: 0"], axis=1)

x = df["text"].values
y = df["label_num"].values

cv = CountVectorizer()

x = cv.fit_transform(x)

from sklearn.model_selection import train_test_split
X_train, X_test, y_train, y_test = train_test_split(x, y, test_size=0.20)
```