# Module 2 - Week 2 - Exercise

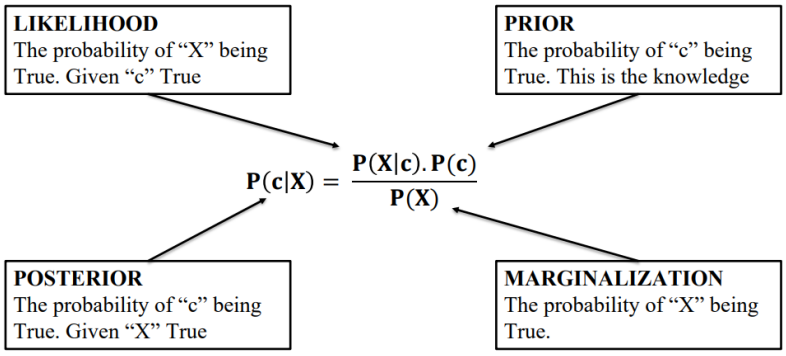
## Probability - Naive Bayes Classifier

TimeSeries Team

Ngày 15 tháng 7 năm 2025

## I. Naive Bayes Classifier

### 1.1. Bayes's Rule




Hình 1: Bayes' Theorem và các thành phần xác suất

Công thức Bayes cho phép chúng ta cập nhật xác suất xảy ra của một giả thuyết $c$ khi có bằng chứng quan sát được $X$. Cụ thể:

$$P(c|X) = \frac{P(X|c) \cdot P(c)}{P(X)}$$

- **Posterior** $P(c|X)$: Là xác suất giả thuyết $c$ đúng sau khi đã quan sát dữ kiện $X$. Đây là giá trị mà ta muốn tính toán.
- **Likelihood** $P(X|c)$: Là xác suất để dữ kiện $X$ xảy ra nếu giả thuyết $c$ đúng. Đây là mức độ phù hợp giữa dữ kiện và giả thuyết.
- **Prior** $P(c)$: Là xác suất tiên nghiệm của giả thuyết $c$ trước khi quan sát dữ kiện. Đây là kiến thức sẵn có hoặc giả định ban đầu.
- **Marginal likelihood** $P(X)$: Là xác suất quan sát thấy dữ kiện $X$ bất kể giả thuyết nào. Đây còn được gọi là giá trị chuẩn hoá (marginalization).

Tóm lại, định lý Bayes là nền tảng trong thống kê suy diễn, học máy (machine learning), và giúp đưa ra quyết định dựa trên bằng chứng quan sát được.

### 1.2. Maximum A Posterior (MAP)

Trong thống kê Bayes, **Maximum A Posteriori (MAP)** là phương pháp ước lượng tham số $\theta$ có xác suất hậu nghiệm cao nhất sau khi đã quan sát dữ liệu.

$$\theta_{\text{MAP}} = \arg\max_{\theta} P(\theta|x_1, x_2, \ldots, x_N)$$

Từ công thức trên ta có thể hiểu rằng, ***ý tưởng của Maximum A Posteriori (MAP)*** là chọn ra giá trị của $\theta$ sao cho xác suất hậu nghiệm $P(\theta|X)$ là lớn nhất trong số các khả năng $x_1, x_2, \ldots, x_N$ có thể xảy ra. Áp dụng định lý Bayes:

$$\theta_{\text{MAP}} = \arg\max_{\theta} \frac{P(x_1, x_2, \ldots, x_N|\theta) \cdot P(\theta)}{P(x_1, x_2, \ldots, x_N)}$$

hay:

$$P(\theta|X) = \frac{P(X|\theta) \cdot P(\theta)}{P(X)}$$

ta thấy rằng mẫu số $P(X)$ không phụ thuộc vào $\theta$, nên không ảnh hưởng đến việc so sánh các giá trị của $P(\theta|X)$ khi tìm giá trị lớn nhất. Vì thế, khi so sánh giữa $P(C_1|X)$ và $P(C_2|X)$, ta có thể bỏ qua mẫu số $P(X)$, vì cả hai đều có cùng mẫu này.

Nói cách khác, do ta chỉ quan tâm đến việc xác suất nào lớn hơn (tức là so sánh tỉ lệ), nên mẫu số chung không ảnh hưởng đến kết quả. Điều này cho phép ta rút gọn biểu thức và tập trung vào tích $P(X|\theta) \cdot P(\theta)$ khi áp dụng MAP:

$$\theta_{\text{MAP}} = \arg\max_{\theta} P(x_1, x_2, \ldots, x_N|\theta) \cdot P(\theta)$$

### 1.3. Maximum Likelihood Estimation (MLE)

Trong phân loại bằng Naive Bayes, ta giả định các đặc trưng đầu vào là **độc lập có điều kiện (conditionally independent).** Khi đó, xác suất một lớp $c$ cho dữ liệu $\mathbf{x} = (x_1, x_2, \ldots, x_N)$ được tính như sau:

$$P(c|\mathbf{x}) \propto P(\mathbf{x}|c) \cdot P(c) = P(x_1|c) \cdot P(x_2|c) \cdots P(x_N|c) \cdot P(c)$$

$$\Rightarrow c_{\text{MAP}} = \arg\max_{c} P(c|\mathbf{x}) = \arg\max_{c} P(c) \cdot \prod_{i=1}^{N} P(x_i|c)$$

Phương pháp này đơn giản, hiệu quả trong thực tế và thường được dùng trong phân loại văn bản (text classification), phát hiện spam, hoặc các bài toán phân loại nhị phân.

***Lưu ý mở rộng:*** Khi số lượng đặc trưng $x$ tăng lên, tích các xác suất $P(x_i|c)$ trong mô hình Naive Bayes có thể trở nên rất nhỏ (gần bằng 0), gây ra hiện tượng *underflow* trong quá trình tính toán. Điều này làm giảm độ chính xác của mô hình. Để khắc phục, ta thường áp dụng logarit lên các xác suất, biến tích thành tổng:

$$\log P(x_1, x_2, \ldots, x_N|c) = \sum_{i=1}^{N} \log P(x_i|c)$$

Điều này không làm thay đổi thứ tự so sánh xác suất giữa các lớp, nhưng giúp việc tính toán trở nên ổn định và chính xác hơn trong thực tế.

### 1.4. Naive Bayes Classifier for Continuous Data

Trong trường hợp dữ liệu đầu vào là liên tục (continuous), thay vì ước lượng xác suất có điều kiện $P(x_i|c)$ bằng tần suất như với dữ liệu rời rạc (discrete), ta giả định rằng các đặc trưng liên tục này tuân theo phân phối xác suất cụ thể, phổ biến nhất là **phân phối chuẩn** (Gaussian).

Với mỗi đặc trưng $x_i$ ứng với lớp $c$, ta ước lượng trung bình $\mu_{ic}$ và phương sai $\sigma_{ic}^2$ từ dữ liệu huấn luyện. Xác suất có điều kiện được tính bằng công thức:

$$P(x_i|c) = \frac{1}{\sqrt{2\pi\sigma_{ic}^2}} \exp\left(-\frac{(x_i - \mu_{ic})^2}{2\sigma_{ic}^2}\right)$$

Sau đó, các xác suất này được kết hợp với xác suất tiên nghiệm $P(c)$ và áp dụng MAP tương tự như mô hình Naive Bayes thông thường:

$$c^* = \arg\max_c \; P(c) \prod_{i=1}^{N} P(x_i|c)$$

# II. Naive Bayes Classifier - Exercises

## 2.1. Exercise 1: PLAY TENNIS

Training Samples

| 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 | Overcast | Mild | High | Weak | No |
| D9 | Sunny | Cool | Normal | Weak | Yes |
| D10 | Rain | Mild | Normal | Weak | Yes |

Trong bài tập này, chúng ta sử dụng mô hình *Naive Bayes* để dự đoán khả năng chơi tennis dựa trên các đặc trưng thời tiết như *Outlook, Temperature, Humidity, Wind*. Tập dữ liệu huấn luyện gồm 10 dòng, và mục tiêu là dự đoán nhãn cho ngày D11. **Bước 1: Tính xác suất tiên nghiệm (prior probability)**

- $P(\text{Yes}) = \frac{6}{10}$ (Có 6 dòng "Yes")
- $P(\text{No}) = \frac{4}{10}$

**Bước 2: Tính các xác suất có điều kiện (likelihood) dựa trên bảng tần suất:**
Ví dụ cho $P(\text{Outlook} = \text{Sunny}|\text{PlayTennis} = \text{Yes}) = \frac{1}{6}$

**Bước 3: Áp dụng công thức Naive Bayes (MAP inference)**
Giả sử đầu vào mới là:

D11: Sunny, Cool, High, Strong

**Xác suất cho lớp "Yes"**:

$$P(\text{Yes}|X) \propto P(\text{Sunny}|\text{Yes}) \cdot P(\text{Cool}|\text{Yes}) \cdot P(\text{High}|\text{Yes}) \cdot P(\text{Strong}|\text{Yes}) \cdot P(\text{Yes})$$

$$= \frac{1}{6} \cdot \frac{3}{6} \cdot \frac{2}{6} \cdot \frac{1}{6} \cdot \frac{6}{10} = 0.0028$$

**Xác suất cho lớp "No"**:

$$P(\text{No}|X) \propto P(\text{Sunny}|\text{No}) \cdot P(\text{Cool}|\text{No}) \cdot P(\text{High}|\text{No}) \cdot P(\text{Strong}|\text{No}) \cdot P(\text{No})$$

$$= \frac{2}{4} \cdot \frac{1}{4} \cdot \frac{3}{4} \cdot \frac{2}{4} \cdot \frac{4}{10} = 0.0188$$

**Kết luận:** Vì $P(\text{No}|X) > P(\text{Yes}|X)$ nên ta dự đoán: **Không chơi tennis (No)** vào ngày D11.

### 2.2. Exercise 2: TRAFFIC DATA (MULTI-LABEL CLASSIFICATION)

Training Samples

| Day | Season | Fog | Rain | Class |
|---|---|---|---|---|
| Weekday | Spring | None | None | On Time |
| Weekday | Winter | None | Slight | On Time |
| Weekday | Winter | None | None | On Time |
| Holiday | Winter | High | Slight | Late |
| Saturday | Summer | Normal | None | On Time |
| Weekday | Autumn | Normal | None | Very Late |
| Holiday | Summer | High | Slight | On Time |
| Sunday | Summer | Normal | None | On Time |
| Weekday | Winter | High | Heavy | Very Late |
| Weekday | Summer | None | Slight | On Time |
| Saturday | Spring | High | Heavy | Cancelled |
| Weekday | Summer | High | Slight | On Time |
| Weekday | Winter | Normal | None | Late |
| Weekday | Summer | High | None | On Time |
| Weekday | Winter | Normal | Heavy | Very Late |
| Saturday | Autumn | High | Slight | On Time |
| Weekday | Autumn | None | Heavy | On Time |
| Holiday | Spring | Normal | Slight | On Time |
| Weekday | Spring | Normal | None | On Time |
| Weekday | Spring | Normal | Heavy | On Time |

Trong bài toán này, chúng ta sử dụng thuật toán *Naive Bayes* để phân loại trạng thái **Class** của phương tiện giao thông (On Time, Late, Very Late, Cancelled) dựa trên các đặc trưng như *Day, Season, Fog, Rain*.

**Bước 1: Tính xác suất tiên nghiệm (prior probabilities):**

- $P(\text{On Time}) = \frac{14}{20}$
- $P(\text{Late}) = \frac{2}{20}$
- $P(\text{Very Late}) = \frac{3}{20}$
- $P(\text{Cancelled}) = \frac{1}{20}$

**Bước 2: Tính xác suất có điều kiện (likelihood) từ bảng tần suất.**
Ví dụ:

$$P(\text{Fog} = \text{High}|\text{Class} = \text{Very Late}) = \frac{1}{3}$$

**Bước 3: Pha kiểm tra với đầu vào mới:**

| Day | Season | Fog | Rain | Class |
|---|---|---|---|---|
| Weekday | Winter | High | Heavy | ? |

**Tính toán xác suất hậu nghiệm cho từng lớp:**

- **On Time**:
$$P(\text{On Time}|X) \propto \frac{9}{14} \cdot \frac{2}{14} \cdot \frac{4}{14} \cdot \frac{2}{14} \cdot \frac{14}{20} = 0.0026$$
- **Late**:
$$P(\text{Late}|X) \propto \frac{1}{2} \cdot \frac{2}{2} \cdot \frac{1}{2} \cdot \frac{0}{2} \cdot \frac{2}{20} = 0.0000$$
- **Very Late**:
$$P(\text{Very Late}|X) \propto \frac{3}{3} \cdot \frac{2}{3} \cdot \frac{1}{3} \cdot \frac{2}{3} \cdot \frac{3}{20} = 0.0222$$
- **Cancelled**:
$$P(\text{Cancelled}|X) \propto \frac{0}{1} \cdot \frac{0}{1} \cdot \frac{1}{1} \cdot \frac{1}{1} \cdot \frac{1}{20} = 0.0000$$

**Kết luận:** Vì xác suất hậu nghiệm cao nhất là với lớp **Very Late** nên hệ thống sẽ dự đoán trạng thái:

Very Late

### 2.3. Exercise 3: IRIS CLASSIFICATION

Trong bài toán này, chúng ta sẽ phân loại hoa Iris dựa trên đặc trưng chiều dài **(Length)** bằng mô hình Naive Bayes với giả định mỗi lớp có phân phối chuẩn (*Gaussian distribution*).

**Dữ liệu huấn luyện:**

| Length | Class |
|---|---|
| 1.4 | 0 |
| 1.0 | 0 |
| 1.3 | 0 |
| 1.9 | 0 |
| 2.0 | 0 |
| 1.8 | 0 |
| 3.0 | 1 |
| 3.8 | 1 |
| 4.1 | 1 |
| 3.9 | 1 |
| 4.2 | 1 |
| 3.4 | 1 |

**Tính toán thông số:**

- **Lớp 0:** $n_0 = 6$, $\mu_0 = 1.56$, $\sigma_0^2 = 0.128$
- **Lớp 1:** $n_1 = 6$, $\mu_1 = 3.73$, $\sigma_1^2 = 0.172$

**Hàm mật độ xác suất chuẩn:**

$$f(x|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

**Pha kiểm tra:** Với đầu vào $x = 3.4$

- Với **lớp 0**:

$$P(x|C=0) = \frac{1}{\sqrt{2\pi \cdot 0.128}} \exp\left(-\frac{(3.4-1.56)^2}{2 \cdot 0.128}\right) \approx 2.18 \times 10^{-6}$$

$$P(C=0|x) \propto P(x|C=0) \cdot P(C=0) = 2.18 \times 10^{-6} \cdot \frac{6}{12} = 1.09 \times 10^{-6}$$

- Với **lớp 1**:

$$P(x|C=1) = \frac{1}{\sqrt{2\pi \cdot 0.172}} \exp\left(-\frac{(3.4-3.73)^2}{2 \cdot 0.172}\right) \approx 0.697$$

$$P(C=1|x) \propto P(x|C=1) \cdot P(C=1) = 0.697 \cdot \frac{6}{12} = 0.3486$$

**Kết luận:** Vì $P(C=1|x)$ lớn hơn $P(C=0|x)$ nên ta phân loại $x = 3.4$ thuộc **Class 1**.

Dự đoán: Class = 1