BÀI GIẢI PHÂN LỚP BAYES

Trần Chí Vỹ

Ngày 11 tháng 11 năm 2025

CÂU 1: PHÂN LỚP THỜI TIẾT (DISCRETE ATTRIBUTES)

Dữ liệu huấn luyện có $N = 14$ mẫu. Số lượng mẫu cho mỗi lớp: $\mathbf{N_{Yes} = 9}$, $\mathbf{N_{No} = 5}$.

Phần (a): Tính Prior và Likelihood

Prior Probability (Xác suất tiên nghiệm) $P(C_i)$

$$P(\text{Play} = \text{Yes}) = \frac{9}{14} \approx 0.643$$
$$P(\text{Play} = \text{No}) = \frac{5}{14} \approx 0.357$$

Likelihood (Khả năng) $P(X|C_i)$

Table 1: Bảng Likelihood cho Câu 1

Thuộc tính	**Giá trị**	**P**(Giá trị|Yes) (**N** = **9**)	**P**(Giá trị|No) (**N** = **5**)
Outlook	Sunny	$3/9 \approx 0.333$	$2/5 = 0.4$
	Overcast	$4/9 \approx 0.444$	$0/5 = 0.0$
	Rainy	$2/9 \approx 0.222$	$3/5 = 0.6$
Temp	Hot	$2/9 \approx 0.222$	$2/5 = 0.4$
	Mild	$4/9 \approx 0.444$	$2/5 = 0.4$
	Cool	$3/9 \approx 0.333$	$1/5 = 0.2$
Humidity	High	$3/9 \approx 0.333$	$4/5 = 0.8$
	Normal	$6/9 \approx 0.667$	$1/5 = 0.2$
Windy	True	$3/9 \approx 0.333$	$2/5 = 0.4$
	False	$6/9 \approx 0.667$	$3/5 = 0.6$

Phần (b): Phân lớp cho mẫu X = (Sunny, Cool, High, True)

$$\begin{aligned}\text{Score}(\text{Yes}) &= P(\text{Sunny}|\text{Yes}) \cdot P(\text{Cool}|\text{Yes}) \cdot P(\text{High}|\text{Yes}) \cdot P(\text{True}|\text{Yes}) \cdot P(\text{Yes}) \\ &= \frac{3}{9} \cdot \frac{3}{9} \cdot \frac{3}{9} \cdot \frac{3}{9} \cdot \frac{9}{14} \\ &\approx 0.00762\end{aligned}$$

$$\begin{aligned}\text{Score}(\text{No}) &= P(\text{Sunny}|\text{No}) \cdot P(\text{Cool}|\text{No}) \cdot P(\text{High}|\text{No}) \cdot P(\text{True}|\text{No}) \cdot P(\text{No}) \\ &= \frac{2}{5} \cdot \frac{1}{5} \cdot \frac{4}{5} \cdot \frac{2}{5} \cdot \frac{5}{14} \\ &\approx 0.00914\end{aligned}$$

Vì $0.00914 > 0.00762$, mẫu được phân lớp là Play = No.

Phần (c): Phân lớp cho mẫu $\mathbf{X}' = (\mathbf{?}, \mathbf{Cool, High, True})$

Sử dụng phương pháp Bỏ qua thuộc tính khuyết (**Outlook**).

$$\begin{aligned}
\text{Score(Yes)} &= P(\text{Cool}|\text{Yes}) \cdot P(\text{High}|\text{Yes}) \cdot P(\text{True}|\text{Yes}) \cdot P(\text{Yes}) \\
&= \frac{3}{9} \cdot \frac{3}{9} \cdot \frac{3}{9} \cdot \frac{9}{14} \\
&\approx 0.0228
\end{aligned}$$

$$\begin{aligned}
\text{Score(No)} &= P(\text{Cool}|\text{No}) \cdot P(\text{High}|\text{No}) \cdot P(\text{True}|\text{No}) \cdot P(\text{No}) \\
&= \frac{1}{5} \cdot \frac{4}{5} \cdot \frac{2}{5} \cdot \frac{5}{14} \\
&\approx 0.0114
\end{aligned}$$

Vì $0.0228 > 0.0114$, mẫu được phân lớp là Play = Yes.

CÂU 2: PHÂN LỚP THỜI TIẾT (CONTINUOUS ATTRIBUTES)

Dữ liệu huấn luyện có $N = 12$ mẫu. Số lượng mẫu cho mỗi lớp: $\mathbf{N_{Yes} = 8}$, $\mathbf{N_{No} = 4}$. Các thuộc tính **Temperature** và **Humidity** là thuộc tính liên tục, cần giả định phân phối chuẩn (Gaussian/Normal Distribution) $P(x|C_i) \sim \mathcal{N}(\mu_i, \sigma_i^2)$.

Phần (a): Tính Prior và Likelihood

Prior Probability $P(C_i)$

$$P(\text{Play} = \text{Yes}) = \frac{8}{12} = \frac{2}{3} \approx 0.667$$
$$P(\text{Play} = \text{No}) = \frac{4}{12} = \frac{1}{3} \approx 0.333$$

Likelihood cho thuộc tính liên tục (Temperature, Humidity)

Ta tính μ (Giá trị trung bình) và σ^2 (Phương sai) cho mỗi lớp:

Lớp Play = Yes (N = 8):

- **Temperature (Yes):** $\{83, 70, 68, 75, 75, 72, 81, 71\}$
- $\mu_{\text{Temp,Yes}} = \frac{83+70+68+75+75+72+81+71}{8} = \frac{595}{8} = 74.375$
- $\sigma^2_{\text{Temp,Yes}} \approx 25.109$
- **Humidity (Yes):** $\{78, 96, 80, 70, 65, 90, 75, 80\}$
- $\mu_{\text{Hum,Yes}} = \frac{78+96+80+70+65+90+75+80}{8} = \frac{634}{8} = 79.25$
- $\sigma^2_{\text{Hum,Yes}} \approx 102.688$

Lớp Play = No (N = 4):

- **Temperature (No):** $\{85, 80, 65, 72\}$
- $\mu_{\text{Temp,No}} = \frac{85+80+65+72}{4} = \frac{302}{4} = 75.5$
- $\sigma^2_{\text{Temp,No}} \approx 62.25$
- **Humidity (No):** $\{85, 90, 70, 95\}$
- $\mu_{\text{Hum,No}} = \frac{85+90+70+95}{4} = \frac{340}{4} = 85.0$
- $\sigma^2_{\text{Hum,No}} \approx 112.5$

Likelihood cho thuộc tính rời rạc (Outlook, Windy)

Table 2: Bảng Likelihood cho thuộc tính rời rạc (Câu 2)

Thuộc tính	**Giá trị**	**P**(Giá trị|Yes) (**N** = **8**)	**P**(Giá trị|No) (**N** = **4**)
Outlook	Sunny	$3/8 = 0.375$	$2/4 = 0.5$
	Overcast	$3/8 = 0.375$	$0/4 = 0.0$
	Rain	$2/8 = 0.25$	$2/4 = 0.5$
Windy	True	$3/8 = 0.375$	$2/4 = 0.5$
	False	$5/8 = 0.625$	$2/4 = 0.5$

Phần (b): Phân lớp cho mẫu $\mathbf{X} = (\mathbf{Sunny}, \mathbf{Temp} = 66, \mathbf{Hum} = 90, \mathbf{Windy} = \mathbf{True})$

Hàm mật độ xác suất cho phân phối chuẩn: $f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$

1. Tính Likelihood cho C = Yes

- $P(\text{Sunny}|\text{Yes}) = 3/8 = 0.375$
- $P(\text{True}|\text{Yes}) = 3/8 = 0.375$
- $P(66|\text{Yes}) = f(66, \mu = 74.375, \sigma^2 = 25.109) \approx 0.019$
- $P(90|\text{Yes}) = f(90, \mu = 79.25, \sigma^2 = 102.688) \approx 0.024$

$$\begin{aligned}\text{Score(Yes)} &= P(\text{Sunny}|\text{Yes}) \cdot P(66|\text{Yes}) \cdot P(90|\text{Yes}) \cdot P(\text{True}|\text{Yes}) \cdot P(\text{Yes}) \\ &\approx 0.375 \cdot 0.019 \cdot 0.024 \cdot 0.375 \cdot \frac{2}{3} \\ &\approx 0.0000427\end{aligned}$$

2. Tính Likelihood cho C = No

- $P(\text{Sunny}|\text{No}) = 2/4 = 0.5$
- $P(\text{True}|\text{No}) = 2/4 = 0.5$
- $P(66|\text{No}) = f(66, \mu = 75.5, \sigma^2 = 62.25) \approx 0.038$
- $P(90|\text{No}) = f(90, \mu = 85.0, \sigma^2 = 112.5) \approx 0.037$

$$\begin{aligned}\text{Score(No)} &= P(\text{Sunny}|\text{No}) \cdot P(66|\text{No}) \cdot P(90|\text{No}) \cdot P(\text{True}|\text{No}) \cdot P(\text{No}) \\ &\approx 0.5 \cdot 0.038 \cdot 0.037 \cdot 0.5 \cdot \frac{1}{3} \\ &\approx 0.000117\end{aligned}$$

3. Kết luận

Vì $0.000117 > 0.0000427$, mẫu $\mathbf{X}$ được phân lớp là Play = No.

CÂU 3: PHÂN LỚP THƯ ĐIỆN TỬ SPAM (MULTINOMIAL NAIVE BAYES)

Bài toán phân lớp email sử dụng mô hình Multinomial Naive Bayes (MNB) vì dữ liệu là số lần xuất hiện (Word Counts). Ta cần áp dụng Kỹ thuật làm mịn Laplace (Laplace Smoothing) để tránh xác suất bằng 0.

Số lượng lớp: $\mathbf{N = 3}$. $N_{\text{S}} = 1$, $N_{\text{N}} = 2$. Vocabulary Size (Kích thước từ vựng): $|\mathbf{V}| = 7$ (từ w_1 đến w_7).

Phần (a): Tính Prior và Likelihood

Prior Probability $P(C_i)$

$$P(\text{S}) = \frac{N_{\text{S}}}{N} = \frac{1}{3} \approx 0.333$$
$$P(\text{N}) = \frac{N_{\text{N}}}{N} = \frac{2}{3} \approx 0.667$$

Likelihood $P(w_j|C_i)$ **(Sử dụng Laplace Smoothing** $\alpha = 1$**)**

Công thức:

$$P(w_j|C_i) = \frac{\text{Count}(w_j, C_i) + \alpha}{\sum_{k=1}^{|V|} \text{Count}(w_k, C_i) + \alpha \cdot |V|}$$

1. Tính tổng số từ trong mỗi lớp:

- $\sum \text{Count}(w, \mathbf{S})$: Từ E3: $1 + 0 + 1 + 1 + 0 + 2 + 0 = 5$
- $\sum \text{Count}(w, \mathbf{N})$: Từ E1+E2: $(1 + 2 + 0 + 1 + 0 + 0) + (0 + 2 + 0 + 0 + 1 + 1 + 1) = 4 + 5 = 9$

2. Tính mẫu số (Denominator):

- Mẫu số cho lớp **S**: $5 + 1 \cdot 7 = 12$
- Mẫu số cho lớp **N**: $9 + 1 \cdot 7 = 16$

3. Tính Likelihood $P(w_j|C_i)$**:**

Table 3: Bảng Likelihood cho Câu 3 (Multinomial Naive Bayes + Laplace Smoothing)

Từ	$\mathbf{P(w_j}$|S) (Mẫu số = 12)	$\mathbf{P(w_j}$|N) (Mẫu số = 16)
w_1	$\frac{1+1}{12} = 2/12 \approx 0.167$	$\frac{(1+0)+1}{16} = 2/16 = 0.125$
w_2	$\frac{0+1}{12} = 1/12 \approx 0.083$	$\frac{(2+2)+1}{16} = 5/16 = 0.3125$
w_3	$\frac{1+1}{12} = 2/12 \approx 0.167$	$\frac{(2+0)+1}{16} = 3/16 = 0.1875$
w_4	$\frac{1+1}{12} = 2/12 \approx 0.167$	$\frac{(0+0)+1}{16} = 1/16 = 0.0625$
w_5	$\frac{0+1}{12} = 1/12 \approx 0.083$	$\frac{(1+0)+1}{16} = 2/16 = 0.125$
w_6	$\frac{2+1}{12} = 3/12 = 0.25$	$\frac{(0+1)+1}{16} = 2/16 = 0.125$
w_7	$\frac{0+1}{12} = 1/12 \approx 0.083$	$\frac{(0+1)+1}{16} = 2/16 = 0.125$

Phần (b): Phân lớp cho dữ liệu E4 $= (w_1 : 1, w_7 : 1)$

Mẫu thử **E4** có tổng số từ là 2: w_1 xuất hiện 1 lần, w_7 xuất hiện 1 lần. Công thức MNB:

$$P(C_i|\mathbf{E4}) \propto P(C_i) \cdot \prod_{j=1}^{|V|} P(w_j|C_i)^{\text{Count}(w_j, \mathbf{E4})}$$

1. Tính khả năng cho C = S

Chỉ xét các từ xuất hiện trong E4 (w_1, w_7) và các từ không xuất hiện (w_2, w_3, w_4, w_5, w_6) có Count = 0.

$$\begin{aligned}\text{Score(S)} &= P(\text{S}) \cdot P(w_1|\text{S})^1 \cdot P(w_7|\text{S})^1 \cdot \prod_{j\in\{2,3,4,5,6\}} P(w_j|\text{S})^0 \\ &= P(\text{S}) \cdot P(w_1|\text{S}) \cdot P(w_7|\text{S}) \\ &= \frac{1}{3} \cdot \frac{2}{12} \cdot \frac{1}{12} \\ &= \frac{2}{432} \approx 0.00463\end{aligned}$$

2. Tính khả năng cho C = N

$$\begin{aligned}\text{Score(N)} &= P(\text{N}) \cdot P(w_1|\text{N})^1 \cdot P(w_7|\text{N})^1 \cdot \prod_{j\in\{2,3,4,5,6\}} P(w_j|\text{N})^0 \\ &= P(\text{N}) \cdot P(w_1|\text{N}) \cdot P(w_7|\text{N}) \\ &= \frac{2}{3} \cdot \frac{2}{16} \cdot \frac{2}{16} \\ &= \frac{8}{768} \approx 0.01042\end{aligned}$$

3. Kết luận

Vì $0.01042 > 0.00463$, mẫu **E4** được phân lớp là Label = N (Not Spam).