Phương pháp Naive Bayes với Laplace Correction

Bước 1: Xác suất phân loại của bộ dữ liệu

Gọi N là tổng số mẫu, C là số lượng lớp và N_c là số mẫu thuộc lớp c. Xác suất tiên nghiệm của lớp c được tính như sau:

$$P(c) = \frac{N_c + 1}{N + C}.$$

Bước 2: Xác suất có điều kiện cho thuộc tính

Cho thuộc tính A và lớp c, xác suất có điều kiện $P(A|c)$ được ước lượng bằng:

$$P(A|c) = \frac{N_{A,c} + 1}{N_c + C_A},$$

trong đó: - $N_{A,c}$ là số mẫu có thuộc tính A và thuộc lớp c. - C_A là số giá trị khả dĩ của thuộc tính A.

Bước 3: Xác suất hậu nghiệm của lớp

Với một tập thuộc tính $\mathbf{A} = \{A_1, A_2, \ldots, A_n\}$, xác suất hậu nghiệm $P(c|\mathbf{A})$ được tính như sau:

$$P(c|\mathbf{A}) \propto P(c) \prod_{i=1}^{n} P(A_i|c).$$

Bước 4: Phân loại

Dự đoán lớp $\hat{c}$ bằng cách chọn lớp có xác suất hậu nghiệm lớn nhất:

$$\hat{c} = \arg\max_{c \in C} P(c|\mathbf{A}).$$

Bài 4 (Bài 1 Đề Cơ sở Trí tuệ nhân tạo năm 2023-2024). Cho bảng quan sát về thời tiết như sau:

# ex.	Weather	Parents	Cash	Exam	Decision
1	sunny	visit	rich	yes	cinema
2	sunny	no-visit	rich	no	tennis
3	windy	visit	rich	no	cinema
4	rainy	visit	poor	yes	cinema
5	rainy	no-visit	rich	no	stay-in
6	rainy	visit	poor	no	cinema
7	windy	no-visit	poor	yes	cinema
8	windy	no-visit	rich	yes	shopping
9	windy	visit	rich	no	cinema
10	sunny	no-visit	rich	no	tennis
11	sunny	no-visit	poor	yes	???

a) Sử dụng độ đo sau để xây dựng cây định danh và tìm bộ luật để phân lớp. Độ đo `Information Gain (IG)`:

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

Trong đó:

- Value(A) là tập tất cả các giá trị có thể có đối với thuộc tính A và S_v là tập con của S mà A có giá trị là v
- Với S bao gồm c lớp, thì Entropy của S được tính bằng công thức sau:

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

Ở đây p_i là tỉ lệ của các mẫu thuộc lớp i trong tập S.

b) Cho biết lớp (Class) của mẫu 11 dựa vào tập luật vừa tìm được?
c) So sánh kết quả ở câu (b) với phương pháp `Naive Bayes`

Công thức Naive Bayes

$$P(C_i \mid X) = \frac{P(X \mid C_i) \cdot P(C_i)}{P(X)}$$

Với:

- $P(C_i)$: Xác suất tiên nghiệm của lớp C_i.
- $P(X \mid C_i)$: Xác suất có điều kiện của các thuộc tính X dựa trên C_i.
- $P(X)$: Xác suất xảy ra của X (là hằng số, không cần tính để so sánh).

Áp dụng Laplace Correction

$$P(\text{Attribute} \mid C_i) = \frac{N_{\text{Attribute},C_i} + k}{N_{C_i} + k}$$

Với:

- $N_{\text{Attribute},C_i}$: Số lượng mẫu trong lớp C_i có giá trị thuộc tính tương ứng.
- N_{C_i}: Tổng số mẫu trong lớp C_i.
- k: Số lượng giá trị khác nhau của thuộc tính.

Bước 1: Tính cho lớp cinema

- Tổng số mẫu thuộc lớp "cinema": $N_{\text{cinema}} = 6$
- Số giá trị khác nhau cho mỗi thuộc tính:
  - Weather: $k = 3$ (sunny, rainy, windy)
  - Parents: $k = 2$ (visit, no-visit)
  - Cash: $k = 2$ (rich, poor)
  - Exam: $k = 2$ (yes, no)

Tính từng thành phần:

$$P(\text{sunny} \mid \text{cinema}) = \frac{N_{\text{sunny,cinema}} + 1}{N_{\text{cinema}} + k} = \frac{1+1}{6+3} = \frac{2}{9}$$
$$P(\text{no-visit} \mid \text{cinema}) = \frac{N_{\text{no-visit,cinema}} + 1}{N_{\text{cinema}} + k} = \frac{1+1}{6+2} = \frac{1}{4}$$
$$P(\text{poor} \mid \text{cinema}) = \frac{N_{\text{poor,cinema}} + 1}{N_{\text{cinema}} + k} = \frac{3+1}{6+2} = \frac{1}{2} = 0.5$$
$$P(\text{yes} \mid \text{cinema}) = \frac{N_{\text{yes,cinema}} + 1}{N_{\text{cinema}} + k} = \frac{3+1}{6+2} = \frac{1}{2} = 0.5$$

Tổng hợp lại:

$$P(X \mid \text{cinema}) = \frac{2}{9} \cdot \frac{1}{4} \cdot \frac{1}{2} \cdot \frac{1}{2} = \frac{1}{72} \approx 0.014$$

Bước 2: Tính cho lớp shopping

- Tổng số mẫu thuộc lớp "shopping": $N_{\text{shopping}} = 1$
- Số giá trị khác nhau cho mỗi thuộc tính:
  - Weather: $k = 3$ (sunny, rainy, windy)
  - Parents: $k = 2$ (visit, no-visit)
  - Cash: $k = 2$ (rich, poor)
  - Exam: $k = 2$ (yes, no)

Tính từng thành phần:

$$P(\text{sunny} \mid \text{shopping}) = \frac{N_{\text{sunny,shopping}} + 1}{N_{\text{shopping}} + k} = \frac{0+1}{1+3} = \frac{1}{4}$$
$$P(\text{no-visit} \mid \text{shopping}) = \frac{N_{\text{no-visit,shopping}} + 1}{N_{\text{shopping}} + k} = \frac{1+1}{1+2} = \frac{2}{3}$$
$$P(\text{poor} \mid \text{shopping}) = \frac{N_{\text{poor,shopping}} + 1}{N_{\text{shopping}} + k} = \frac{0+1}{1+2} = \frac{1}{3}$$
$$P(\text{yes} \mid \text{shopping}) = \frac{N_{\text{yes,shopping}} + 1}{N_{\text{shopping}} + k} = \frac{1+1}{1+2} = \frac{2}{3}$$

Tổng hợp lại:

$$P(X \mid \text{shopping}) = \frac{1}{4} \cdot \frac{2}{3} \cdot \frac{1}{3} \cdot \frac{2}{3} = \frac{4}{108} = \frac{1}{27} \approx 0.037$$

Bước 3: Tính cho lớp tennis

- Tổng số mẫu thuộc lớp "tennis": $N_{\text{tennis}} = 2$
- Số giá trị khác nhau cho mỗi thuộc tính:
  - Weather: $k = 3$ (sunny, rainy, windy)
  - Parents: $k = 2$ (visit, no-visit)
  - Cash: $k = 2$ (rich, poor)
  - Exam: $k = 2$ (yes, no)

Tính từng thành phần:

$$P(\text{sunny} \mid \text{tennis}) = \frac{N_{\text{sunny,tennis}} + 1}{N_{\text{tennis}} + k} = \frac{2+1}{2+3} = \frac{3}{5}$$
$$P(\text{no-visit} \mid \text{tennis}) = \frac{N_{\text{no-visit,tennis}} + 1}{N_{\text{tennis}} + k} = \frac{2+1}{2+2} = \frac{3}{4}$$
$$P(\text{poor} \mid \text{tennis}) = \frac{N_{\text{poor,tennis}} + 1}{N_{\text{tennis}} + k} = \frac{0+1}{2+2} = \frac{1}{4}$$
$$P(\text{yes} \mid \text{tennis}) = \frac{N_{\text{yes,tennis}} + 1}{N_{\text{tennis}} + k} = \frac{0+1}{2+2} = \frac{1}{4}$$

Tổng hợp lại:

$$P(X \mid \text{tennis}) = \frac{3}{5} \cdot \frac{3}{4} \cdot \frac{1}{4} \cdot \frac{1}{4} = \frac{9}{320} \approx 0.028125$$

Bước 4: Tính cho lớp stay-in

Tổng số mẫu thuộc lớp "stay-in": $N_{\text{stay-in}} = 1$. Tính từng thành phần:

$$P(\text{sunny} \mid \text{stay-in}) = \frac{N_{\text{sunny,stay-in}} + 1}{N_{\text{stay-in}} + k} = \frac{0+1}{1+3} = \frac{1}{4}$$
$$P(\text{no-visit} \mid \text{stay-in}) = \frac{N_{\text{no-visit,stay-in}} + 1}{N_{\text{stay-in}} + k} = \frac{1+1}{1+2} = \frac{2}{3}$$
$$P(\text{poor} \mid \text{stay-in}) = \frac{N_{\text{poor,stay-in}} + 1}{N_{\text{stay-in}} + k} = \frac{0+1}{1+2} = \frac{1}{3}$$
$$P(\text{yes} \mid \text{stay-in}) = \frac{N_{\text{yes,stay-in}} + 1}{N_{\text{stay-in}} + k} = \frac{0+1}{1+2} = \frac{1}{3}$$

Tổng hợp lại:

$$P(X \mid \text{stay-in}) = \frac{1}{4} \cdot \frac{2}{3} \cdot \frac{1}{3} \cdot \frac{1}{3} = \frac{2}{108} = \frac{1}{54} \approx 0.0185$$

Bước 5: Tính P(X) với Laplace Correction

Đầu tiên, tính lại xác suất tiên nghiệm P(Ci) với Laplace correction:

$$P(C_i) = \frac{N_{C_i} + 1}{N_{\text{total}} + k}$$

Với:

- $N_{\text{total}} = 10$ (tổng số mẫu)
- $k = 4$ (số lớp khác nhau: cinema, shopping, tennis, stay-in)

Tính xác suất tiên nghiệm cho từng lớp:

$$P(\text{cinema}) = \frac{6+1}{10+4} = \frac{7}{14} = 0.5$$
$$P(\text{shopping}) = \frac{1+1}{10+4} = \frac{2}{14} \approx 0.143$$
$$P(\text{tennis}) = \frac{2+1}{10+4} = \frac{3}{14} \approx 0.214$$
$$P(\text{stay-in}) = \frac{1+1}{10+4} = \frac{2}{14} \approx 0.143$$

Tính lại P(X) với xác suất tiên nghiệm đã điều chỉnh:

$$\begin{aligned} P(X) &= P(X|\text{cinema})P(\text{cinema}) + P(X|\text{shopping})P(\text{shopping}) \\ &\quad + P(X|\text{tennis})P(\text{tennis}) + P(X|\text{stay-in})P(\text{stay-in}) \\ &= 0.014 \cdot 0.5 + 0.037 \cdot 0.143 + 0.028125 \cdot 0.214 + 0.0185 \cdot 0.143 \\ &= 0.007 + 0.005291 + 0.006019 + 0.002646 \\ &= 0.020956 \end{aligned}$$

Bước 6: Tính P($C_i|X$)

Áp dụng công thức Bayes với xác suất tiên nghiệm đã tính

$$P(C_i|X) = \frac{P(X|C_i)P(C_i)}{P(X)}$$

Ta tính được các xác suất hậu nghiệm sau:

$$P(\text{cinema}|X) = \frac{0.014 \cdot 0.5}{0.020956} \approx 0.334$$
$$P(\text{shopping}|X) = \frac{0.037 \cdot 0.143}{0.020956} \approx 0.252$$
$$P(\text{tennis}|X) = \frac{0.028125 \cdot 0.214}{0.020956} \approx 0.287$$
$$P(\text{stay-in}|X) = \frac{0.0185 \cdot 0.143}{0.020956} \approx 0.126$$

Kết luận

Sau khi áp dụng `Laplace correction` cho cả xác suất tiên nghiệm, ta có kết quả:

- P(cinema|X) ≈ 0.334 (33.4%)
- P(shopping|X) ≈ 0.252 (25.2%)
- P(tennis|X) ≈ 0.287 (28.7%)
- P(stay-in|X) ≈ 0.126 (12.6%)

Với `Laplace correction` áp dụng cho cả xác suất tiên nghiệm, `"cinema"` là lựa chọn có xác suất cao nhất (33.4%).