Buổi thực hành 4: Naïve Bayes

Bài 1: Cho cơ sở dữ liệu khách hàng đã thu thập được như sau:

ID	Tuổi	Thu nhập	Sinh viên	Đánh giá tín dụng	Mua máy tính
1	youth	high	no	fair	**no**
2	youth	high	no	excellent	**no**
3	middle	high	no	fair	**yes**
4	senior	medium	no	fair	**yes**
5	senior	low	yes	fair	**yes**
6	senior	low	yes	excellent	**no**
7	middle	low	yes	excellent	**yes**
8	youth	medum	no	fair	**yes**
9	youth	low	yes	fair	**yes**
10	senior	medium	yes	fair	**yes**
11	youth	medium	yes	excellent	**yes**
12	middle	medium	no	excellent	**yes**
13	middle	high	yes	fair	**yes**
14	senior	medium	no	excellent	**no**

Dùng giải thuật Naïve Bayes để đưa ra dự đoán việc các khách hàng mới có mua máy tính hay không dựa trên các thuộc tính như sau:

Khách hàng:

S1 : (age = youth, income = medium, student = yes, credit_rating = fair)

S2 : (age = middle, income = high, student = yes, credit_rating = fair)

S3 : (age = youth, income = low, student = no, credit_rating = excellent)

Bài làm

- S1 : (age = youth, income = medium, student = yes, credit_rating = fair)

Ta có: $P(C_{yes}) = 10/14 = 0.714$, $P(C_{no}) = 4/14 = 0.285$

Các xác suất thành phần:

$P(\text{Age} = \text{youth} \mid C_{yes}) = 3/10 = 0.3$

$P(\text{Age} = \text{youth} \mid C_{no}) = 2/4 = 0.5$

$P(\text{income} = \text{medium} \mid C_{yes}) = 5/10 = 0.5$

$P(\text{income} = \text{medium} \mid C_{no}) = 1/4 = 0.25$

P(student = yes | C_{yes}) = 6/10 = 0.6

P(student = yes | C_{no}) = 1/4 = 0.25

P(credit_rating = fair | C_{yes}) = 7/10 = 0.7

P(credit_rating = fair | C_{no}) = 1/4 = 0.25

Vậy ta có:

P(X|C_{yes}) = 0.3 * 0.5 * 0.6 * 0.7 = 0.063

P(X|C_{no}) = 0.5 * 0.25 * 0.25 * 0.25 = 0.0078

P(X|C_{yes}) * P(C_{yes}) = 0.063 * 0.714 = 0.045

P(X|C_{no}) * P(C_{no}) = 0.0078125 * 0.285 = 0.0022

➔ Từ kết quả trên ta thấy P(X|C_{yes}) * P(C_{yes}) có giá trị lớn nhất, do đó thuật toán Naïve Bayes sẽ kết luận rằng khách hàng X sẽ mua máy tính.

- S2 : (age = middle, income = high, student = yes, credit_rating = fair)

Ta có: P(C_{yes}) = 10/14 = 0.714, P(C_{no}) = 4/14 = 0.285

Các xác suất thành phần:

P(Age = middle | C_{yes}) = 4/10 = 0.4

P(Age = middle | C_{no}) = 0/4 = 0

P(income = high | C_{yes}) = 2/10 = 0.2

P(income = high | C_{no}) = 2/4 = 0.5

P(student = yes | C_{yes}) = 6/10 = 0.6

P(student = yes | C_{no}) = 1/4 = 0.25

P(credit_rating = fair | C_{yes}) = 7/10 = 0.7

P(credit_rating = fair | C_{no}) = 1/4 = 0.25

Vậy ta có:

P(X|C_{yes}) = 0.4 * 0.2 * 0.6 * 0.7 = 0.0336

P(X|C_{no}) = 0 * 0.5 * 0.25 * 0.25 = 0

P(X|C_{yes}) * P(C_{yes}) = 0.0336 * 0.714 = 0.024

P(X|C_{no}) * P(C_{no}) = 0 * 0.285 = 0

➔ Từ kết quả trên ta thấy P(X|C_{yes}) * P(C_{yes}) có giá trị lớn nhất, do đó thuật toán Naïve Bayes sẽ kết luận rằng khách hàng X sẽ mua máy tính.

- S3 : (age = youth, income = low, student = no, credit_rating = excellent)

Ta có: $P(C_{yes}) = 10/14 = 0.714$, $P(C_{no}) = 4/14 = 0.285$

Các xác suất thành phần:

$P(\text{Age} = \text{youth} \mid C_{yes}) = 3/10 = 0.3$

$P(\text{Age} = \text{youth} \mid C_{no}) = 2/4 = 0.5$

$P(\text{income} = \text{low} \mid C_{yes}) = 3/10 = 0.3$

$P(\text{income} = \text{low} \mid C_{no}) = 1/4 = 0.25$

$P(\text{student} = \text{no} \mid C_{yes}) = 4/10 = 0.4$

$P(\text{student} = \text{no} \mid C_{no}) = 3/4 = 0.75$

$P(\text{credit\_rating} = \text{excellent} \mid C_{yes}) = 3/10 = 0.3$

$P(\text{credit\_rating} = \text{excellent} \mid C_{no}) = 3/4 = 0.75$

Vậy ta có:

$P(X|C_{yes}) = 0.3 * 0.3 * 0.4 * 0.3 = 0.0108$

$P(X|C_{no}) = 0.5 * 0.25 * 0.75 * 0.75 = 0.0703$

$P(X|C_{yes}) * P(C_{yes}) = 0.0108 * 0.714 = 0.0077$

$P(X|C_{no}) * P(C_{no}) = 0.0703 * 0.285 = 0.020$

➔ Từ kết quả trên ta thấy $P(X|C_{no}) * P(C_{no})$ có giá trị lớn nhất, do đó thuật toán Naïve Bayes sẽ kết luận rằng khách hàng X sẽ không mua máy tính.