Naïve Bayes Classifier

Sample Dataset (Allectronic dataset)

RID	age	income	student	credit_rating	buys_computer(Class)
1	youth	high	no	fair	no
2	youth	high	no	excellent	no
3	middle_aged	high	no	fair	yes
4	senior	medium	no	fair	yes
5	senior	low	yes	fair	yes
6	senior	low	yes	excellent	no
7	middle_aged	low	yes	excellent	yes
8	youth	medium	no	fair	no
9	youth	low	yes	fair	yes
10	senior	medium	yes	fair	yes
11	youth	medium	yes	excellent	yes
12	middle_aged	medium	no	excellent	yes
13	middle_aged	high	yes	fair	yes
14	senior	medium	no	excellent	no

Encoded Dataset

1. age
    - youth = 0
    - middle_aged = 1
    - senior = 2
2. income
    - low = 0
    - medium = 1
    - high = 2
3. student
    - no = 0
    - yes = 1
4. credit_rating
    - fair = 0
    - excellent = 1
5. buys_computer(Class)
    - no = 0
    - yes = 1

age	income	student	credit_rating	buys_computer(Class)
0	2	0	0	0
0	2	0	1	0
1	2	0	0	1
2	1	0	0	1
2	0	1	0	1
2	0	1	1	0
1	0	1	1	1
0	1	0	0	0
0	0	1	0	1
2	1	1	0	1
0	1	1	1	1
1	1	0	1	1
1	2	1	0	1
2	1	0	1	0

*Exclude RID column because it is not an important attribute to classify whether a person buys a computer.

Bayes' Theorem

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

In this case,

- There are two classes – ***no*** and ***yes*** which is C_1 and C_2 respectively.
- There are four attributes and the result vector looks like X = (x_1, x_2, x_3, x_4) which is known as X = (age, income, student, credit_rating).

Step by Step Calculation

Step (1) Calculating prior probabilities – P(C_i) for i = 1 and 2

P(C_1) = P(buys_computer = no) = $\frac{5}{14}$ = 0.357

P(C_2) = P(buys_computer = yes) = $\frac{9}{14}$ = 0.643

Step (2) Calculating conditional probabilities P($x_j|C_i$)

Here, +1 is known as Laplace correction used to avoid probability value of zero.

1. P($x_1|C_1$) = P(age | $\text{buys\_computer} = no$)
   a. P($\text{age} = \text{youth}$ | $\text{buys\_computer} = no$) = $\frac{3}{5}$
   b. P($\text{age} = \text{middle\_aged}$ | $\text{buys\_computer} = no$) = $\frac{0+1}{5}$
   c. P($\text{age} = \text{senior}$ | $\text{buys\_computer} = no$) = $\frac{2}{5}$
2. P($x_1|C_2$) = P(age | $\text{buys\_computer} = yes$)
   a. P($\text{age} = \text{youth}$ | $\text{buys\_computers} = yes$) = $\frac{2}{9}$
   b. P($\text{age} = \text{middle\_aged}$ | $\text{buys\_computer} = yes$) = $\frac{4}{9}$
   c. P($\text{age} = \text{senior}$ | $\text{buys\_computer} = yes$) = $\frac{3}{9}$
3. P($x_2|C_1$) = P(income | $\text{buys\_computer} = no$)
   a. P($\text{income} = \text{low}$| $\text{buys\_computers} = no$) = $\frac{1}{5}$
   b. P($\text{income} = \text{medium}$ | $\text{buys\_computer} = no$) = $\frac{2}{5}$
   c. P($\text{income} = \text{high}$ | $\text{buys\_computer} = no$) = $\frac{2}{5}$
4. P($x_2|C_2$) = P(income | $\text{buys\_computer} = yes$)
   a. P($\text{income} = \text{low}$| $\text{buys\_computers} = yes$) = $\frac{3}{9}$

b. P(income $=$ medium | buys_computer $= yes$) = $\frac{4}{9}$

c. P(income $=$ high | buys_computer $= yes$) = $\frac{2}{9}$

5. P($x_3|C_1$) = P(student | buys_computer $= no$)

a. P(student $=$ no | buys_computer $= no$) = $\frac{4}{5}$

b. P(student $=$ yes | buys_computer $= no$) = $\frac{1}{5}$

6. P($x_3|C_2$) = P(student | buys_computer $= yes$)

a. P(student $=$ no | buys_computer $= yes$) = $\frac{3}{9}$

b. P(student $=$ yes | buys_computer $= yes$) = $\frac{6}{9}$

7. P($x_4|C_1$) = P(credit_rating | buys_computer $= no$)

a. P(credit_rating $=$ fair | buys_computer $= no$) = $\frac{2}{5}$

b. P(credit_rating $=$ excellent | buys_computer $= no$) = $\frac{3}{5}$

8. P($x_4|C_2$) = P(credit_rating | buys_computer $= yes$)

a. P(credit_rating $=$ fair | buys_computer $= yes$) = $\frac{6}{9}$

b. P(credit_rating $=$ excellent | buys_computer $= yes$) = $\frac{3}{9}$

Step (3) Classification a test vector

X_test = (**age**=youth, **income**=medium, **student**=yes, **credit_rating**=fair)

1. P(X_test | buys_computer = no)
   = P(age $=$ youth | buys_computer $= no$) x
   P(income $=$ medium | buys_computer $= no$) x
   P(student $=$ yes | buys_computer $= no$) x
   P(credit_rating $=$ fair | buys_computer $= no$)
   = $\frac{3}{5}$ x $\frac{2}{5}$ x $\frac{1}{5}$ x $\frac{2}{5}$
   = 0.019
2. P(X_test | buys_computer = yes)
   = P(age $=$ youth | buys_computer $= yes$) x
   P(income $=$ medium | buys_computer $= yes$) x
   P(student $=$ yes | buys_computer $= yes$) x
   P(credit_rating $=$ fair | buys_computer $= yes$)
   = $\frac{2}{9}$ x $\frac{4}{9}$ x $\frac{6}{9}$ x $\frac{6}{9}$
   = 0.0439

Step(4) According to Bayes Theorem,

1. P(X_test | buys_computer = no) x P(buys_computer = no)
   = 0.019 x 0.357
   = 0.00678

2. P(X_test | buys_computer = yes) x P(buys_computer = yes)
   = 0.0439 x 0.643
   = 0.0282

Therefore, the prediction class for X_test = (**age**=youth, **income**=medium, **student**=yes, **credit_rating**=fair) is yes according to the naïve Bayesian classifier.