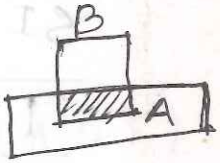
Bayesian and Naive Baysian classifier

Dr. Ohidujjaman Tuhin

Bayesian classification

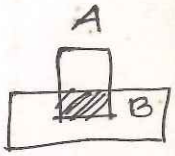
$$P(A/B) = \frac{\text{Common area bet}^n A \text{ and } B}{\text{Total area of } B}$$



$$= \frac{P(A \wedge B)}{P(B)}$$

$$\Rightarrow P(A \wedge B) = P(A/B)\, P(B) \quad \text{---(I)}$$

$$P(B/A) = \frac{\text{Common area bet}^n B \text{ and } A}{\text{Total Area of } A}$$



$$= \frac{P(B \wedge A)}{P(A)}$$

$$= \frac{P(A \wedge B)}{P(A)}$$

$$\Rightarrow P(A \wedge B) = P(B/A)\, P(A) \quad \text{---(II)}$$

from equation (I) and (II) we get —

$$P(A/B) P(B) = P(B/A) P(A)$$

$$\Rightarrow P(A/B) = \frac{P(B/A)\, P(A)}{P(B)}$$

Number of hypothesis is more than one —

$h_1, h_2 \cdots h_n$

$$P(h_1/D) = \frac{P(D/h_1)\, P(h_1)}{P(D)}, \quad P(h_2/D) = \frac{P(D/h_2)\, P(h_2)}{P(D)}$$

$$P(h_n/D) = \frac{P(D/h_n)\, P(h_n)}{P(D)}$$

Maximum Probability $\max(P(h_1/D), P(h_2/D), \cdots, P(h_n/D))$

$$= \max\left\{\frac{P(D/h_1)\, P(h_1)}{P(D)}, \frac{P(D/h_2) P(h_2)}{P(D)}, \frac{P(D/h_n) P(h_n)}{P(D)}\right\}$$

$$\simeq \max\left(P(D/h_1) P(h_1), P(D/h_2) P(h_2) \cdots, P(D/h_n) P(h_n)\right)$$

Output = maximum hypothesis.

Example: 1

SI NO	Age	Income	Student	Credit rating	Buy Computer
1	35	Medium	Yes	Fair	Yes
2	30	High	No	Average	No
3	40	low	Yes	Good	No
4	35	Medium	No	fair	Yes
5	45	low	No	fair	yes
6	35	High	No	Excellent	Yes
7	35	Medium	No	good	No
8	25	Low	No	good	No
9	28	High	No	Average	No
10	35	Medium	Yes	Average	Yes

Decide whether of 35 years and medium income will buy a computer or not using Bayesian classifier.

Ans: Let, h_1 = Customer will buy computer

h_2 = Customer will not buy computer.

D = Customer 35 years and income medium.

$$p(h_1) = \frac{5}{10} = 0.5 \quad , \quad p(h_2) = \frac{5}{10} = 0.5$$

$$p(D/h_1) = \frac{3}{5} = 0.6$$

$$p(D/h_2) = \frac{1}{5} = 0.2$$

$\max\{P(h_1/D), P(h_2/D)\}$

Maximum Probability $= \max(P(D/h_1)P(h_1), P(D/h_2)P(h_2))$

$= \max(0.6 * 0.5, 0.2 * 0.5)$

$= \max(0.30, 0.10)$

$= 0.30$

final output $= h_1 =$ Customer will buy computer

Naive Bayesian classifier:

Conditional independence:

$P(x/y,z)$ where x depends on y and z.

If x and y are conditionally independence

$P(x/y,z) = P(x/z)$

Proof: $P(x,y/z) = \frac{P(x,y,z)}{P(z)}$

$= \frac{P(x,y,z)}{P(y,z)} * \frac{P(y,z)}{P(z)}$

$= P(x/y,z) * P(y/z)$

$= P(x/z) * P(y/z)$

when x and y are conditionally independent for given z.

Number of Attributes

$X = x_1, x_2, x_3 \ldots, x_n$

Number of class / hypothesis

$c_1, c_2, c_3 \ldots c_m$

$$P(X/c_1) = P(x_1, x_2, x_3 \ldots x_n/c_1) = P(x_1/c_1)\, P(x_2/c_1) \ldots P(x_n/c_1)$$

$$P(X/c_2) = P(x_1, x_2, x_3 \ldots x_n/c_2) = P(x_1/c_2)\, P(x_2/c_2) \ldots P(x_n/c_2)$$

$$P(X/c_m) = P(x_1, x_2, x_3 \ldots x_n/c_m) = P(x_1/c_m)\, P(x_2/c_m) \ldots P(x_n/c_m)$$

Maximum probability

$$\max\left(P(c_1/X)\, P(c_2/X) \ldots P(c_m/X)\right)$$

$$\max\left\{\frac{P(X/c_1)\, P(c_1)}{P(X)}, \frac{P(X/c_2)\, P(c_2)}{P(X)} \ldots \frac{P(X/c_m) P(c_m)}{P(X)}\right\}$$

$$\simeq \max\left\{P(X/c_1)\, P(c_1), P(X/c_2)\, P(c_2) \ldots, P(X/c_m) P(c_m)\right\}$$

Output = maximum hypothesis

Example – 1

Let x_1 = 35 years old

x_2 = medium income

c_1 = will buy computer

c_2 = will not buy computer

$$P(c_1) = \frac{5}{10} = 0.5 \qquad P(x_1/c_1) = \frac{4}{5} = 0.8$$

$$P(c_2) = \frac{5}{10} = 0.5 \qquad P(x_1/c_2) = \frac{1}{5} = 0.2$$

$$P(x_2/c_1) = \frac{3}{5} = 0.6 \qquad P(x_2/c_2) = \frac{1}{5} = 0.2$$

Maximum Probability

$$\max\{P(x/c_1)\,P(c_1), P(x/c_2)\,P(c_2)\}$$

$$= \max\{P(x_1/c_1)\,P(x_2/c_1)\,P(c_1), P(x_1/c_2)\,P(x_2/c_2)\,P(c_2)\}$$

$$= \max(0.8 * 0.6 * 0.5,\ 0.2 * 0.2 * 0.5)$$

$$= \max(0.24, 0.020)$$

$$= 0.24$$

output = c_1 = customer will buy computer

Example 2:

Record	Age	Income	student	Credit Rating	Buys computer
r_1	<=30	High	No	fair	No
r_2	<=30	High	No	Excellent	No
r_3	31...40	High	No	fair	Yes
r_4	>40	Medium	No	fair	yes
r_5	>40	Low	yes	fair	yes
r_6	>40	Low	yes	Excellent	No
r_7	31...40	Low	yes	Excellent	yes
r_8	<=30	Medium	No	fair	No
r_9	<=30	Low	yes	fair	yes
r_{10}	>40	Medium	yes	fair	yes
r_{11}	<=30	medium	yes	Excellent	yes
r_{12}	31...40	Medium	No	Excellent	yes
r_{13}	31...40	High	yes	fair	yes
r_{14}	>40	Medium	No	Excellent	No

Whether customer buys computer or not using the tuple

X = (Age = youth <= 30, income = medium, student = yes
credit_rating = fair)

Ans:

Let, e_1 = Buys computer

C_2 = will not buy computer

x_1 = Age = youth <= 30

x_2 = income medium

x_3 = Customer is a student = yes

x_4 = Credit_rating = fair.

$\therefore P(c_1) = \frac{9}{14}$ $P(c_2) = \frac{5}{14}$

$P(x_1/c_1) = \frac{2}{9}$ $P(x_1/c_2) = \frac{3}{5}$

$P(x_2/c_1) = \frac{4}{9}$ $P(x_2/c_2) = \frac{2}{5}$

$P(x_3/c_1) = \frac{6}{9}$ $P(x_3/c_2) = \frac{1}{5}$

$P(x_4/c_1) = \frac{6}{9}$ $P(x_4/c_2) = \frac{2}{5}$

Over calculation:

$P(X/c_1) = P(x_1, x_2, x_3, x_4/c_1)$

$= P(x_1/c_1)\, P(x_2/c_1)\, P(x_3/c_1)\, P(x_4/c_1)$

$= \frac{2}{9} * \frac{4}{5} * \frac{6}{9} * \frac{6}{9}$

$= 0.044$

$P(X/c_2) = P(x_1/c_2)\, P(x_2/c_2)\, P(x_3/c_2)\, P(x_4/c_2)$

$= \frac{3}{5} * \frac{2}{5} * \frac{1}{5} * \frac{2}{5}$

$= 0.019$

$\therefore$ Max $P(c_i/X) \cong \max P(X/c_i)\, P(c_i)$

when $i=1$, $P(c_1/x) \cong P(x/c_1)\, P(c_1)$

$= 0.044 * \frac{9}{14} = 0.028$

when $i=2$, $P(c_2/x) \cong P(x/c_2)\, P(c_2)$

$= 0.019 * \frac{5}{14} = 0.007$

$\therefore \max\{P(X/c_i)\, P(c_i)\} = \max\{P(x_4/c_1)\, P(c_1), P(x/c_2) P(c_2)\}$

$= \max(0.028, 0.007)$

$= 0.028$

$\therefore$ Predition: Buys a computer.

Naive Bayesian

Laplacian - correction.

RID	Age	Income	student	Credit	Ci : buy or not buy
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	medium	NO	fair	NO
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

Decide whether a student with medium income middle aged, and fair credit rating will buy a computer or not.

Ans: Let c_1 = will buy computer

c_2 = will not buy computer

x_1 = student = yes

x_2 = Middle ~~income~~ Aged

x_3 = Midium income

x_4 = fair Credit rating

$\therefore P(c_1) = \frac{9}{14}$ $\quad P(c_2) = \frac{5}{14}$

$P(x_1/c_1) = \frac{6}{9}$ $\quad P(x_1/c_2) = \frac{1}{5}$

$P(x_2/c_1) = \frac{4}{9}$ $\quad P(x_2/c_2) = \frac{0}{5} = 0$

$P(x_3/c_1) = \frac{4}{9}$ $\quad P(x_3/c_2) = \frac{2}{5}$

$P(x_4/c_1) = \frac{6}{9}$ $\quad P(x_4/c_2) = \frac{2}{5}$

$\therefore P(c_1/X) = \frac{P(X/c_1)\ P(c_1)}{P(X)}$

$\propto P(X/c_1)\ P(c_1)$

$= P(x_1\ x_2\ x_3\ x_4/c_1)\ P(c_1)$; [Here $X = x_1, x_2, x_3, x_4$]

$= P(x_1/c_1)\ P(x_2/c_1)\ P(x_3/c_1)\ P(x_4/c_1)\ P(c_1)$

$= \frac{6}{9} * \frac{4}{9} * \frac{4}{9} * \frac{6}{9} * \frac{9}{14}$

$P(c_2/X) = \frac{P(X/c_2)\ P(c_2)}{P(X)}$

$\propto P(X/c_2)\ P(c_2)$

$= P(x_1\ x_2\ x_3\ x_4/c_2)\ P(c_2)$; [Here $X = x_1, x_2, x_3, x_4$]

$= P(x_1/c_2)\ P(x_2/c_2)\ P(x_3/c_2)\ P(x_4/c_2)\ P(c_2)$

$= \frac{1}{5} * \frac{0}{5} * \frac{2}{5} * \frac{2}{5} * \frac{5}{14}$

$= 0$

Therefore, Laplacian correction is needed for "C_2" class.

RID	Age	Income	student	Credit	Ci: buy or not buy
⋮	⋮				⋮
15	middle				No
16	Senior				No
17	youth				No

∴ Now, $P(C_1) = \frac{9}{14+3} = \frac{9}{17}$ $P(C_2) = \frac{5+3}{14+3} = \frac{8}{17}$

$P(x_1/C_1) = \frac{6}{9}$ $P(x_1/C_2) = \frac{1+0}{8}$

$P(x_2/C_1) = \frac{4}{9}$ $P(x_2/C_2) = \frac{1+0}{8}$

$P(x_3/C_1) = \frac{4}{9}$ $P(x_3/C_2) = \frac{2}{8}$

$P(x_4/C_1) = \frac{6}{9}$ $P(x_4/C_2) = \frac{2}{8}$

$$\therefore P(C_1/X) = \frac{P(X/C_1)\ P(C_1)}{P(X)}$$

$$\simeq P(X/C_1)\ P(C_1)$$

$$= P(x_1 x_2 x_3 x_4/C_1)\ P(C_1) \quad ; [\text{Here } X = x_1 x_2 x_3 x_4]$$

$$= P(x_1/C_1)\ P(x_2/C_1)\ P(x_3/C_1)\ P(x_4/C_1)\ P(C_1)$$

$$= \frac{6}{9} * \frac{4}{9} * \frac{4}{9} * \frac{6}{9} * \frac{9}{17}$$

$$= 0.0465$$

$$\therefore P(c_2/X) = \frac{P(X/c_2)\,P(c_2)}{P(X)}$$

$$\cong P(X/c_2)\,P(c_2)$$

$$= P(x_1 x_2 x_3 x_4/c_2)\,P(c_2) \quad ; \left[X = x_1 x_2 x_3 x_4 \text{ Here}\right]$$

$$= P(x_1/c_2)\,P(x_2/c_2)\,P(x_3/c_2)\,P(x_4/c_2)\,P(c_2)$$

$$= \frac{1}{8} * \frac{1}{8} * \frac{2}{8} * \frac{2}{8} * \frac{8}{17}$$

$$= 0.00046$$

$$\therefore \text{Max}\{P(c_i/X)\} \cong \max\{P(X/c_i)\,P(c_i)\}$$

$$= \max\{P(X/c_1)\,P(c_1), P(X/c_2)\,P(c_2)\}$$

$$= \max\{P(c_1/X), P(c_2/X)\}$$

$$= \max(0.0465, 0.00046)$$

$$= 0.0465$$

∴ Prediction: A student with medium income, middle aged and fair credit rating will buy a computer.

Ans:

D Test classification using Naive Bayesian classifier.

Word	sqrt	bit	chip	class
DOC1	42	25	7	Math
DOC2	10	28	45	comp
DOC3	11	25	22	Comp
DOC4	33	40	8	Math
DOC5	28	32	9	Math
DOC6	8	22	30	comp

Classify the above "documentation Text" ~~based~~ " whether computer or math related documentation based on Naive Bayesian classifier.

Ans: Let,

x_1 = sqrt
x_2 = bit
x_3 = chip

c_1 = computer related doc
c_2 = Math related doc

$$p(\text{comp}) = \frac{3}{6} \qquad p(\text{math}) = \frac{3}{6}$$

$$p(c_1) = \frac{3}{6} \qquad p(c_2) = \frac{3}{6}$$

$$p(x_1|c_1) = \frac{10+11+8}{(10+28+45)+(11+25+22)+(8+22+30)}$$

$$= 0.144$$

$$p(x_2|c_1) = \frac{28+25+22}{(10+28+45)+(11+25+22)+(8+22+30)}$$

$$= \frac{75}{201} = 0.373$$

$$P(x_3/c_1) = \frac{45+22+30}{(10+28+45)+(11+25+22)+(8+22+30)}$$

$$= \frac{97}{201} = 0.483$$

$$P(x_1/c_2) = \frac{42+33+28}{(42+25+7)+(33+40+8)+(28+32+9)}$$

$$= \frac{103}{224} = 0.459$$

$$P(x_2/c_2) = \frac{25+40+32}{224}$$

$$= 0.433$$

$$P(x_3/c_2) = \frac{7+8+9}{224}$$

$$= 0.107$$

$\therefore P(X/c_1) = P(x_1 x_2 x_3/c_1)$; [Here $X = x_1 x_2 x_3$]

$= P(x_1/c_1)\, P(x_2/c_1)\, P(x_3/c_1)$

$= 0.144 * .373 * 0.483$

$= 0.0259$

$P(X/c_2) = P(x_1 x_2 x_3/c_2)$

$= P(x_1/c_2)\, P(x_2/c_2)\, P(x_3/c_2)$

$= 0.459 * 0.433 * 0.107$

$= 0.0212$

$$\therefore \text{Max}\{P(X/c_1)\,P(c_1),\ P(X/c_2)\,P(c_2)\}$$

$$= \max\left(0.0259 * \frac{3}{6},\ 0.0212 * \frac{3}{6}\right)$$

$$= \max(0.01295,\ 0.0106)$$

$$= 0.01295$$

Prediction: Computer related document

Ans.