07-DEC-2022

Naive Bayes

Types of event

① Independent : Each event is not affected by other event
② Dependent : Each event affected by another event
③ Mutually exclusive : Event which cannot happen at a single time.

Example

Tossing a coin

① $P(H/T) = \frac{1}{2}$ — 1st time
② $P(H/T) = 1/2$ — 2nd time
③ $P(H/T) = 1/2$ — 3rd time

} Independent Event

Bag of Ball

2 B 3 R ball

$P(B) = \frac{2}{5} \xrightarrow[\text{bag}]{\text{Blue}} P(B) = \frac{1}{4}$

↓ 1st time   ↓ 2nd time.

} Dependent Event

Card Problem

Diamond
Heart
Spade
~~ki~~

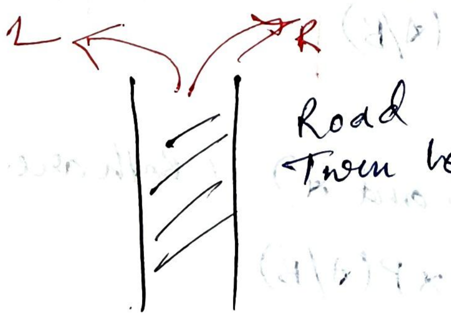
$P(King) = \frac{4}{52}$ (1st time)

↓ one king got removed

$P(King) = \frac{3}{52}$ (2nd time)

} Dependent Event

②



Road
Turn left or Right

} Mutually Exclusive

① King and Queen are mutually exclusive
② Head or tails

Q Difference b/w exclusive and independent events?

Ans

Mutually Exclusive

Two event cannot occur at a $\boxed{\text{same time}}$.
['T' time]

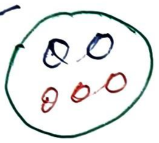
Independent Events

One events remain unaffected by the occurrence of another event. ['T' and 'T-1' time]

Dependent Probability

Q Find out the probability of getting blue ball in consecutive events?

Ans



event 1   event 2

$$\xrightarrow{P(\text{Ans})} \frac{2}{5} \xrightarrow{P(\text{Blue})} \frac{1}{4}$$

$\Downarrow$ $P(A)$ $\Downarrow$ $P(B/A)$ → Dependent Probability

"Event B given A"

"B is happening, A happened already"

$$\boxed{P(A \text{ and } B) = P(A) * P(B/A)}$$ — (Eqn 1)

$$\boxed{P(B/A) = \frac{P(A \text{ and } B)}{P(A)}}$$

↓ Probability of event B given A.

$$\boxed{P(B \text{ and } A) = P(B) * P(A/B)}$$ → (Eqn 2)

> $P(A \text{ and } B) = P(B \text{ and } A)$ (Both are Same)

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

Naive Bayes Formula →

$$\boxed{P(B/A) = \frac{P(B) * P(A/B)}{P(A)}}$$ → Bayes Theorem

> Probability of event "B" given "A"

Q Toss the coin 3 times in a consecutive manner and find out the probability of getting head?

Ans

$P(H) = \frac{1}{2} \times \frac{1}{2} \times \frac{1}{2} = \frac{1}{8}$

Q How to implement this formula in ML?

Ans.

$$\underbrace{x_1 \quad x_2 \quad x_3 \quad x_4 \quad \ldots \quad x_n}_{\text{I/P}} \qquad \underbrace{y}_{\text{O/P}}$$

Independent feature — Dependent feature

$$\left\{ P(B/A) = \frac{P(B) \times P(A/B)}{P(A)} \right\}$$

$$P(y/[x_1, x_2, x_3, \ldots x_n]) = \frac{P(y) \times P([x_1, x_2, x_3, \ldots x_n]/y)}{P([x_1, x_2, x_3 \ldots x_n])}$$

$$P(y/[x_1, x_2, x_3 \ldots x_n]) = \frac{P(y) \times P(x_1/y) \times P(x_2/y) \times P(x_3/y) \ldots \times P(x_n/y)}{P(x_1) \times P(x_2) \times P(x_3) \ldots \times P(x_n)}$$

> Apply this formula on the dataset and this is called probabilistic based approach.

Note: Naive Bayes can only solved classification problem. It cannot be used for Regression Problem.

$$\boxed{P(\text{class}/\text{data}) = \frac{P(\text{data}/\text{class}) \times P(\text{class})}{P(\text{data})}}$$

class = y

data = $[x_1, x_2, x_3, \ldots x_n]$

Example

(Yes/No) (Binary classification)

x_1	x_2	x_3	O/P
			Yes
			No

Yes, $P(Yes/(x_1,x_2,x_3)) = \frac{P(yes) \times P(x_1/yes) \times P(x_2/yes) \times P(x_3/yes)}{\cancel{P(x_1) \times P(x_2) \times P(x_3)}}$

No, $P(No/(x_1,x_2,x_3)) = \frac{P(No) \times P(x_1/No) \times P(x_2/No) \times P(x_3/No)}{\cancel{P(x_1) \times P(x_2) \times P(x_3)}}$

Note: Denominator are same in both the equation. So, we will not consider not denominator in our calculation, because ratio will be same.

Let's assume, | Weightage

$P(Yes/x_i) = 0.13 \rightarrow \frac{0.13}{0.13+0.05} = \frac{0.13}{0.18} = 72\%$

$P(No/x_i) = 0.05 \rightarrow \frac{0.05}{0.13+0.05} = \frac{0.05}{0.18} = 28\%$

Weightage $\rightarrow$ $\boxed{P(Yes/x_i) > P(No/x_i)}$

- So, we will consider final class as yes for that particular row. Similarly, we will calculate for the entire dataset.
- Here, "Yes" weightage is more.

Example with real dataset

Dataset (only categorical features)

Day	Outlook	Temperature	Humidity	Wind	Play Tennis
D1	Sunny	Hot	High	Weak	No
D2	Sunny	Hot	High	Strong	No
D3	Overcast	Hot	High	Weak	Yes
D4	Rain	Mild	High	Weak	Yes
D5	Rain	Cool	Normal	Weak	Yes
D6	Rain	Cool	Normal	Strong	No
D7	Overcast	Cool	Normal	Strong	Yes
D8	Sunny	Mild	High	Weak	No
D9	Sunny	Cool	Normal	Weak	Yes
D10	~~Sunny~~ Rain	Mild	Normal	Weak	Yes
D11	~~Rain~~ Sunny	Mild	Normal	Strong	Yes
D12	~~Sunny~~ Overcast	Mild	High	Strong	Yes
D13	Overcast	Hot	Normal	Weak	Yes
D14	Rain	Mild	High	Strong	No

Target → Play Tennis

$P(Yes) = 9/14$ and $P(No) = 5/14$

Outlook

	Yes	No	P(Yes)	P(No)
Sunny	2	3	2/9	3/5
Overcast	4	0	4/9	0/5
Rain	3	2	3/9	2/5
Total	9	5		

Temperature

	Yes	No	P(Yes)	P(No)
Hot	2	2	2/9	2/5
Mild	4	2	4/9	2/5
Cool	3	1	3/9	1/5
Total	9	5		

> As of now, we are taking 2 features only. We can take all the features also and probability as of yes and no can be calculated in the similar manner.

Humidity

Que. If outlook is [Sunny] and temperature is [Hot], what will be output?

Ans.

$$P(Yes/[Sunny, Hot]) = \frac{P(Yes) \times P(Sunny/Yes) \times P(Hot/Yes)}{P(Sunny) \times P(Hot)}$$

$$P(No/[Sunny, Hot]) = \frac{P(No) \times P(Sunny/No) \times P(Hot/No)}{\cancel{P(Sunny) \times P(Hot)}}$$

> Ignore denominator, as it is same in both the eqn.

$$P(Yes/[Sunny, Hot]) = \frac{(9/14) \times (2/9) \times (2/9)}{} = \frac{2}{63} = 0.031$$

$$P(No/[Sunny, Hot]) = (5/14) \times (3/5) \times (2/5) = \frac{3}{35} = 0.085$$

Weightage of Yes, $= \frac{0.031}{0.031 + 0.085} = 0.27 = 27\%$

Weightage of No, $= \frac{0.085}{0.031 + 0.085} = 0.73 = 73\%$

$$\boxed{\text{Weightage of No} > \text{Weightage of Yes}}$$

Hence, our output will be "No".

Task

① Expand Humidity and Wind calculation like outlook and temperature?
② Solve the particular condition mentioned below:
ⓐ [Sunny, Hot, High, Strong]

Humidity

	Yes	No	P(Yes)	P(No)
High	3	4	3/9	4/5
Normal	6	1	6/9	~~0~~ 1/5
total	9	5		

Wind

	Yes	No	P(Yes)	P(No)
Strong	3	3	3/9	3/5
Weak	6	2	6/9	2/5
total	9	5		

$P(Yes/[Sunny, Hot, High, Strong]) = P(Yes) \times P(Sunny/Yes) \times P(Hot/Yes) \times P(High/Yes) \times P(Strong/Yes)$

$= \left(\frac{9}{14}\right) \times \left(\frac{2}{9}\right) \times \left(\frac{2}{9}\right) \times \left(\frac{3}{9}\right) \times \left(\frac{3}{9}\right)$

$= \left(\frac{9}{14}\right) \times \left(\frac{2}{9}\right) \times \left(\frac{2}{9}\right) \times \left(\frac{1}{3}\right) \times \left(\frac{1}{3}\right)$

$= \frac{36}{30618} = 0.001175$

$P(No/[Sunny, Hot, High, Strong]) = P(No) \times P(Sunny/No) \times P(Hot/No) \times P(High/No) \times P(Strong/\cancel{Yes}\ No)$

$= \left(\frac{5}{14}\right) \times \left(\frac{3}{5}\right) \times \left(\frac{2}{5}\right) \times \left(\frac{4}{5}\right) \times \left(\frac{3}{5}\right)$

$= \frac{90}{8750} = 0.010285$

$$\text{Weight of Yes} = \frac{0.00475}{0.00475 + 0.010285}$$

$$= \frac{0.00475}{0.01146}$$

$$= 0.40 = 40\%$$

$$\text{Weight of No} = \frac{0.010285}{0.00475 + 0.010285}$$

$$= \frac{0.010285}{0.01146}$$

$$= 0.90 = 90\%$$

$$\boxed{\text{weightage of No} > \text{weightage of Yes}}$$

Hence, our output will be "No".

Example with real dataset

categorical + Numerical features

Dataset

City	Gender	Income	Illness
Dallas	Male	40367	No
Dallas	Female	41524	Yes
Dallas	Male	46373	Yes
New York City	Male	98096	No
New York City	Female	102089	No
New York City	Female	100662	No
New York City	Male	47263	Yes
Dallas	Male	56645	No

Illness

P(Yes) = 3/8 , P(No) = 5/8

City

	Yes	No	P(Yes)	P(No)
Dallas	2	2	2/3	2/5
New York City	1	3	1/3	3/5
total	3	5		

Gender

	Yes	No	P(Yes)	P(No)
Male	2	3	2/3	3/5
Female	1	2	1/3	2/5
total	3	5		

Income

Illness		Income	Mean	St. Deviation
Illness	Yes	41524, 46373, 47263	68386.66	42391.5
	No	40367, 98096, 102089, 100662, 56645	79571.8	28972.49

> For numerical feature, like Income we will use mean, standard deviation and normal distribution.

~~Mean~~

Mean, $\mu = \frac{1}{n}\sum_{i=1}^{n} x_i$

Standard deviation, $\sigma = \left\{\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\mu)^2\right\}^{0.5}$

Normal distribution, $f(x) = \frac{1}{\sqrt{2\pi}\,\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$

> We will calculate likelihood of an event using Normal distribution for numerical features.

Q. condition:
city: Dallas, Gender: Female, Income: 100000
find out the output?

Ans

$$f(100000) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

$$P(\text{Income/Yes}) = \frac{1}{42397.5 \times 2.506} \times 2.718^{-\frac{(100000-68386.66)^2}{2\times 42397.5^2}}$$

$$= 0.00000712789$$

$$P(\text{Income/No}) = \frac{1}{28972.49 \times 2.506} \times 2.718^{-\frac{(100000-74571.8)^2}{2\times 28972.49^2}}$$

$$= 0.00000107421$$

Now,

$$P(\text{Yes}/[\text{Dallas, Female, Income}]) = P(\text{Yes}) \times P(\text{Dallas/Yes}) \times P(\text{Female/Yes}) \times P(\text{Income/Yes})$$

$$= \left(\frac{3}{8}\right)\times\left(\frac{2}{3}\right)\times\left(\frac{1}{3}\right)\times 0.00000712789$$

$$= 5.93990833e^{-7}$$

$P(No/[Dallas, Female, Income]) = P(No) \times P(Dallas/No) \times P(Female/No) \times P(Income/No)$

$= \left(\frac{15}{8}\right) \times \left(\frac{2}{5}\right) \times \left(\frac{2}{5}\right) \times (0.00000107421)$

$= 1.07421 e^{-6}$

$= \frac{1.07421}{1000000}$

$= \cancel{0.00000}$

Weightage of Yes $= \frac{5.93 e^{-7}}{1.07 e^{-6} + 5.93 e^{-7}}$

Weightage of No $= \frac{1.07 e^{-6}}{1.07 e^{-6} + 5.93 e^{-7}}$

$\boxed{\text{Weightage of No} > \text{Weightage of Yes}}$

Hence, output will be no and patient will not have a heart disease.