NAIVE BAYES:-

→ Supervised ML Algorithm (classification)

→ Text classification.

ex. Mail < spam / not spam.

Probability → chances of Outcome.

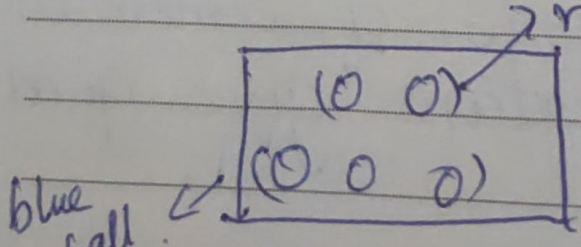
Independent Event:-

Toss a coin : H / T.

$P(H) = 0.5$

$P(T) = 0.5$

Dependent Event:-



1st event (A)

Date .....

Prob of getting a blue

$$P(A) = 2/5$$

Event (B):- Prob of getting red when event A is already taken place

$$\underbrace{P(B/A)}_{\text{Conditional Probability}} = 1/4$$

NB→ Conditional Probability.

↳ Bayes Theorem:-

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

$P(A) = 2/5$, $P(B/A) = 1/4$.

Prob. of both the events [P(A) & P(B/A)] to happen.

$$P(B \cap A) \rightarrow P(A) \times P(B/A)$$

$$= 2/5 \times 1/4 = 1/10.$$

Date .....

$$P(B/A) = \frac{P(B \cap A)}{P(A)} = \frac{(1/10)}{(2/5)} = 1/4$$

Conditional probability.

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

↓

Prob. of event A when event B is always happened.

$$P(B \cap A) = P(A \cap B)$$

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

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

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

} Bayes Theorem

↓

Used by NB classifier

Why called Naive Bayes?

Bayes → Bayes Theorem.

Naive → It assume that each i/p

Date

variables are independent.

let (x_1, x_2, x_3, x_4, y)

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

$$\div \; P(x_1)\,P(x_2)\,P(x_3)\,P(x_4)$$

↳ constant (denominator)

$$= \prod_{i=1}^{n} P(x_i/y) \times P(y)$$

probability of all the features

o/p will be $y = \arg\max\left(\prod_{i=1}^{n} P(x_i/y) \times P(y)\right)$

Ex:

Outlook i/p x_1.

	Yes	No	P(Y)	P(NO)
Sunny	2	3	2/9	3/5
Overcast	4	0	4/9	0/5
Rainy	3	2	3/9	2/5
	9	5		

Date .....

Temperature x_2 i/p

	Yes	No	P(y)	P(N)
Hot	2	2	2/9	2/5
mild	4	2	4/9	2/5
cold	3	1	3/9	1/5
	9	5		

Play y o/p

		P(y)
Yes	9	9/14
No	5	5/14

Pblm Stmt:-

Today is sunny & Hot, whether will play or not?

Today (sunny, hot)

P(y) } → P(y) for sunny = 2/9
P(N) } P(N) for ~~sunny~~ hot = 2/9

$P(yes^{y} / today^{x}) = P(sunny^{x_1} / yes) \times P(hot / yes) \times P(yes)$

P(Today) → Constant.

Date .....

$= 2/9 \times 2/9 \times 9/14 = 0.031$

P(NO):

P(NO) - Sunny = 3/5, P(NO) - Hot = 2/5

P(NO) = 5/14

$P(NO/today) = 3/5 \times 2/5 \times 5/14$

$= 0.08571$

P(y) when sunny & hot → 0.031

P(N) → 0.08571

Normalize it $P(y) = \frac{0.031}{0.031 + 0.08571} \approx 0.27$

$P(N) = 1 - P(y) = 1 - 0.27 = 0.73$

O/P ⇒ No. (max value is P(N)).

Player will not play when its hot & Sunny.

Date .....

Text classification:-

Preprocessing for text data:-

→ Remove punctuation.

→ Remove stop words. (is, was, were are..)

→ Change everything into Lower case.

→ Convert Text data to numerical vectors.

Text to numerical vectors:-

i) Bag of words: (BOW)

→ Way of extracting features from text.

Ex: Sentence 1: He is a good boy.
S2: She is a good girl
S3: Boy & girl are good.

Step 1: Remove stop words:-

Date .....

stop words are words which does not add much meaning to the sentence.

S1 : good boy
S2 : good girl
S3 : good boy girl

Step 2: Convert text to numerical vectors (Feature vector creation).

	good	boy	girl
S1	1	1	0
S2	1	0	1
S3	1	1	1

Disadvantage of BOW.

→ It cannot identify the semantics of a word in a sentence.

→ Feature Dimension.

→ doest not capture the relationship between features.

Date .....

(2) TF - IDF (Term Frequency x Inverse Document frequency)

→ Quantifies the importance or relevence of string representation.

TF: measures how frequently a word appears in a document.

$$= \frac{\text{no. of. repetition of words in sentence}}{\text{Total. no. of. words in sentence.}}$$

Step1: Remove stop words.

S_1 : good boy
S_2 : good girl.
S_3 : good boy girl.

Step2: Calculate TF

	S_1	S_2	S_3
good	1/2	1/2	1/3
boy	1/2	0	1/3
girl	0	1/2	1/3

Date .....

IDF : measure how important a word is

$$= \log\left(\frac{\text{no. of. sentences}}{\text{No. of. Sentences containing word}}\right)$$

Word	IDF
good	$\log(3/3) = 0$
boy	$\log(3/2) =$
girl	$\log(3/2) =$

Step 3: Final Vector creation.

$\Rightarrow$ TF x IDF

	good	boy	girl
S_1	0	$1/2 \times \log(3/2)$	0
S_2	0	0	$1/2 \times \log(3/2)$
S_3	0	$1/3 \times \log(3/2)$	$1/3 \times \log(3/2)$

Date .....

Types:

① Gaussian NB:-

⇒ Used for Continuous features.

→ It is used in classification and it assumes that I/P features follow a normal distribution.

② Multinominal NB:-

→ It is used for discrete counts.

→ Used when the features represents the frequency.

→ Used for text classification.

③ Bernoulli NB:-

→ the input features are binary (0's and 1's)

→ Used for text classifications.