Naive Bayes Classifier

(Assignment-Haritha P V)(14-07-24)

- The Naive Bayes algorithm is a supervised learning algorithm, which is based on **Bayes theorem** and used for solving classification problems.
- Naïve Bayes Classifier is one of the simplest and most effective Classification algorithms which helps in building the fast machine learning models that can make quick predictions.
- **It is a probabilistic classifier, which means it predicts on the basis of the probability of an object**.
- Some popular examples of Naïve Bayes Algorithm are **spam filtration, Sentimental analysis, and classifying articles**.
- It is called Naïve because it assumes that the occurrence of a certain feature is independent of the occurrence of other features.
- It is called Bayes because it depends on the principle of Bayes' Theorem.

Bayes' Theorem

- Bayes' theorem is also known as **Bayes' Rule** or **Bayes' law**, which is used to determine the probability of a hypothesis with prior knowledge. It depends on the conditional probability.
- In the context of classification, Bayes' Theorem is used to calculate the probability of a class given a set of features. The "naive" assumption is that all features are independent of each other.

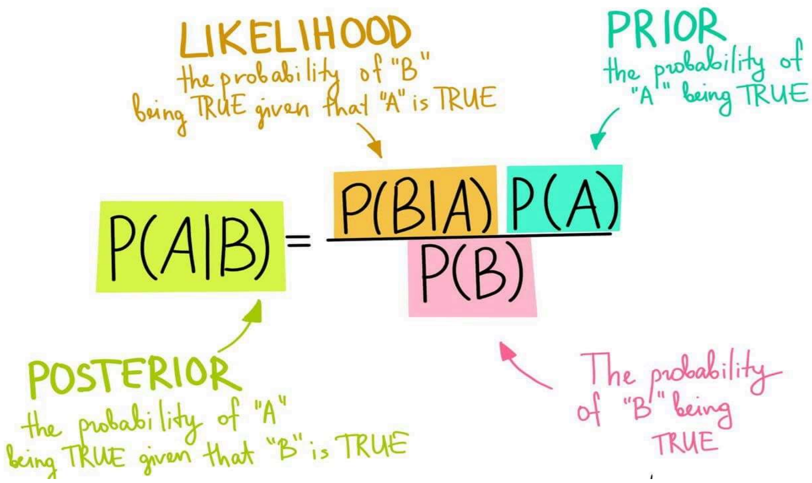
Where,

P(A|B) is Posterior probability: Probability of hypothesis A on the observed event B.

P(B|A) is Likelihood probability: Probability of the evidence given that the probability of a hypothesis is true.

P(A) is Prior Probability: Probability of hypothesis before observing the evidence.

P(B) is Marginal Probability: Probability of Evidence.



Example 1:

From deck of cards:

Probability that a single card drawn is king= **P(king)**=4/52=1/13

If king is a event, single card is a face card

Probability of face given that its a king=**P(Face|King)**=1

Probability of face card=**P(Face**)=12/52=3/13

(since there are 3 faces in each suit of cards(jack,king,queen))

Probability of king given that its a face=**(P(Face|King)* p(king))/p(Face)**

=1*(1/13)/(3/13)=⅓

Probability of getting king in faces is 1/3

Example 2:

Suppose we have a data set with outlook,humidity,and wind which determines whether we should play on that day or not.

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

Probability of playing tennis and not playing:

$$P(PlayTennis = yes) = 9/14 = .64$$
$$P(PlayTennis = no) = 5/14 = .36$$

Frequency table of outlook, wind, humidity

Frequency Table		Play	
		Yes	No
Outlook	Sunny	3	2
	Overcast	4	0
	Rainy	3	2

Frequency Table		Play	
		Yes	No
Wind	Strong	6	2
	Weak	3	3

Frequency Table		Play	
		Yes	No
Humidity	High	3	4
	Normal	6	1

Likelihood Table for Outlook,humidity, wind.

Likelihood Table		Play		
		Yes	No	
Outlook	Sunny	3/10	2/4	5/14
	Overcast	4/10	0/4	4/14
	Rainy	3/10	2/4	5/14
		10/14	4/14	

$P(B|A) = P(Sunny|Yes) = 3/10 = 0.3$

$P(B) = P(Sunny) = 5/14 = 0.36$

$P(A) = P(Yes) = 10/14 = 0.71$

Likelihood of 'Yes' given Sunny is

$$P(A/B) = P(Yes/Sunny) = P(Sunny|Yes)^{*}\ P(Yes) / P(Sunny) = (0.3 \times 0.71) /0.36 = 0.591$$

Similarly Likelihood of 'No' given Sunny is

$$P(A/B) = P(No/Sunny) = P(Sunny|No)^{*}\ P(No) / P(Sunny) = (0.4 \times 0.36) /0.36 = 0.40$$

Likelihood table for Humidity

Likelihood Table		Play		
		Yes	No	
Humidity	High	3/9	4/5	7/14
	Normal	6/9	1/5	7/14
		9/14	5/14	

Likelihood table for Wind

Likelihood Table		Play		
		Yes	No	
Wind	Weak	6/9	2/5	8/14
	Strong	3/9	3/5	6/14
		9/14	5/14	

$$P(Yes|High) = 0.33 \times 0.6 / 0.5 = 0.42$$

$$P(No|High) = 0.8 \times 0.36 / 0.5 = 0.58$$

$$P(Yes|Weak) = 0.67 \times 0.64 / 0.57 = 0.75$$

$$P(No|Weak) = 0.4 \times 0.36 / 0.57 = 0.25$$

What to find? The chance of playing a game in this situation.

Suppose we have a day with the following values

Outlook	=	Rain
Humidity	=	High
Wind	=	Weak
Play	=	?

Likelihood of 'Yes' on that Day = P(Outlook = Rain|Yes)*P(Humidity= High|Yes)* P(Wind= Weak|Yes)*P(Yes)
= 2/9 * 3/9 * 6/9 * 9/14 = 0.0199

Likelihood of 'No' on that Day = P(Outlook = Rain|No)*P(Humidity= High|No)* P(Wind= Weak|No)*P(No)
= 2/5 * 4/5 * 2/5 * 5/14 = 0.0166

P(Yes) = 0.0199 / (0.0199+ 0.0166) = 0.55

P(No) = 0.0166 / (0.0199+ 0.0166) = 0.45

Our model predicts there is a 55% chance there will be a game tomorrow.

Advantages of Naive Bayes Classifier

- Easy to implement and computationally efficient.
- Effective in cases with a large number of features.
- Performs well even with limited training data.
- It performs well in the presence of categorical features.
- For numerical features data is assumed to come from normal distributions

Disadvantages of Naive Bayes Classifier

- Assumes that features are independent, which may not always hold in real-world data.
- Can be influenced by irrelevant attributes.
- May assign zero probability to unseen events, leading to poor generalization.

Applications of Naive Bayes Classifier

- **Spam Email Filtering**: Classifies emails as spam or non-spam based on features.
- **Text Classification**: Used in sentiment analysis, document categorization, and topic classification.
- **Medical Diagnosis:** Helps in predicting the likelihood of a disease based on symptoms.
- **Credit Scoring:** Evaluates creditworthiness of individuals for loan approval.
- **Weather Prediction**: Classifies weather conditions based on various factors.