ARTIFICIAL INTELLIGENCE AND MACHINE LEARNING

Introduction to probability theory, Introduction to uncertainty Bayes Theorem, Naïve Bayes Classification

AIM

To familiarize students with the concept of Probability theory, Bayes Theorem, Naïve Bayes Classification

INSTRUCTIONAL OBJECTIVES

This Session is designed to:

1. Understanding Probability Basics
2. Understanding Uncertainty, Probability and Bayesian Thinking
3. Understanding Classification, Components of Naïve Bayes

LEARNING OUTCOMES

At the end of this session, you should be able to:

1. Deep understanding of fundamental concepts in probability theory, including sample spaces, events, and probability measures.
2. Bayes' Theorem and its significance in probabilistic reasoning and decision-making.
3. Naïve Bayes Classification as a simple and effective probabilistic classifier based on Bayes' Theorem.

Probability Theory

Probability Theory: Probability is defined as the chance of happening or occurrences of an event. Generally, the possibility of analyzing the occurrence of any event concerning previous data is called probability. For example, if a fair coin is tossed, what is the chance that it lands on the head? These types of questions are answered under probability.

Probability measures the likelihood of an event's occurrence. In situations where the outcome of an event is uncertain, we discuss the probability of specific outcomes to understand their chances of happening. The study of events influenced by probability falls under the domain of statistics.

Probability Theory Definition

Probability theory studies random events and tells us about their occurrence. The two main approaches for studying probability theory are.

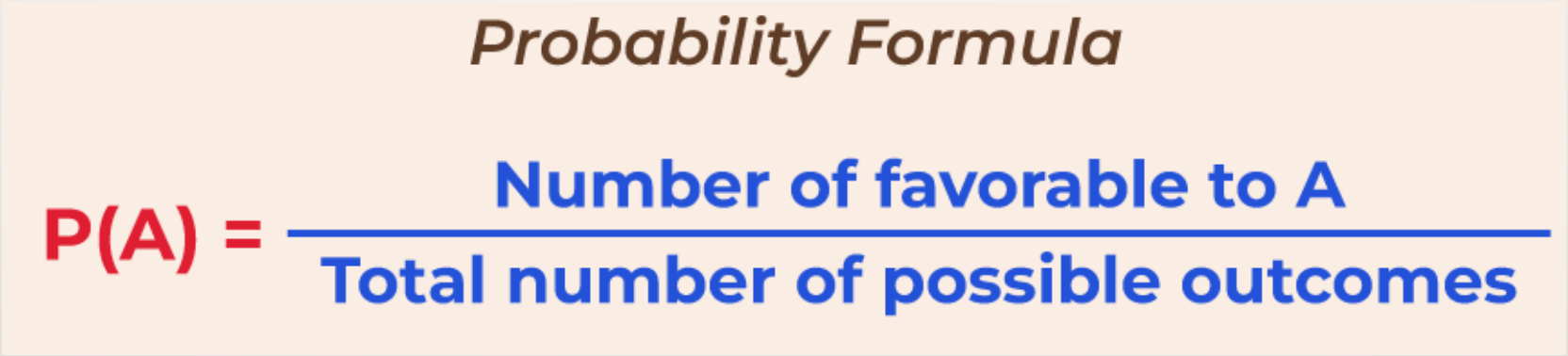
1. Theoretical Probability

2. Experimental Probability

Theoretical Probability

Theoretical Probability deals with assumptions in order to avoid unfeasible or expensive repetition of experiments. The theoretical Probability for an Event A can be calculated as follows:

P(A) = (Number of outcomes favourable to Event A) / (Number of all possible outcomes)



Note: Here we assume the outcomes of an event as equally likely.

Now, as we learn the formula, let's put this formula in our coin-tossing case. In tossing a coin, there are two outcomes: Head or Tail. Hence, The Probability of occurrence of Head on tossing a coin is
P(H) = ½

Similarly, The Probability of the occurrence of a Tail on tossing a coin is
P(T) = ½

The following image shows an unbiased coin that has an equal probability of landing both heads and tails

Experimental Probability

Experimental probability is found by performing a series of experiments and observing their outcomes. These random experiments are also known as trials. The experimental probability for Event A can be calculated as follows:

P(E) = (Number of times event A happened) / (Total number of trials)

The following image shows the Experimental Probability Formula,

Now, as we learn the formula, let's put this formula in our coin-tossing case. If we tossed a coin 10 times and recorded heads for 4 times and a tail 6 times then the Probability of Occurrence of Head on tossing a coin:

P(H) = 4/10

Similarly, the Probability of Occurrence of Tails on tossing a coin:

P(T) = 6/10

Probability Theory Examples

We can study the concept of probability with the help of the example discussed below,

Example: Let's take two random dice and roll them randomly, now the probability of getting a total of 10 is calculated.

Solution:

Total Possible events that can occur (sample space) {(1,1), (1,2),..., (1,6),..., (6,6)}. The total spaces are 36.

Now the required events, {(4,6), (5,5), (6,4)} are all which adds up to 10.
So the probability of getting a total of 10 is = 3/36 = 1/12

Uncertainty In Artificial Intelligence

Uncertainty plays a significant role in artificial intelligence (AI) and machine learning. It refers to the lack of complete information or the presence of randomness and variability in data or in the outcomes of AI models. Dealing with uncertainty is crucial in AI for making informed decisions, handling noisy data, and building robust and reliable systems. Here are some key aspects of uncertainty in artificial intelligence:

1.Types of Uncertainty:

a. **Aleatoric Uncertainty**: This type of uncertainty arises from inherent randomness and variability in data. It is often associated with observations that are subject to random noise. For example, in computer vision, the position of an object in an image may have aleatoric uncertainty due to variations in lighting and camera sensor noise.

b. **Epistemic Uncertainty**: Epistemic uncertainty is related to the lack of knowledge or information. It represents uncertainty that can be reduced with more data or improved models. For example, a machine learning model may have epistemic uncertainty when trying to make predictions in a data-scarce region.

c. **Model Uncertainty**: Model uncertainty encompasses the uncertainty associated with the choice of model architecture and parameters. It reflects the uncertainty in the model's own internal representation of the data. Techniques like Bayesian neural networks and dropout regularization can help quantify model uncertainty.

Bayes' Theorem

Bayes' Theorem is used to determine the conditional probability of an event. It was named after an English statistician, **Thomas Bayes** who discovered this formula in 1763. Bayes Theorem is a very important theorem in mathematics, that laid the foundation of a unique statistical inference approach called the **Bayes' inference. It is used to find the probability of an event, based on prior knowledge of conditions that might be related to that event.**

For example, if we want to find the probability that a white marble drawn at random came from the first bag, given that a white marble has already been drawn, and **there are three bags each containing some white and black marbles, then we can use Bayes' Theorem.**

What is Bayes' Theorem?

Bayes theorem (also known as the Bayes Rule or Bayes Law) is used to determine the conditional probability of event A when event B has already occurred.

The general statement of Bayes' theorem is "**The conditional probability of an event A, given the occurrence of another event B, is equal to the product of the event of B, given A and the probability of A divided by the probability of event B.**" i.e.

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

where,

- ***P(A)*** *and* ***P(B)*** *are the probabilities of events A and B*
- ***P(A|B)*** *is the probability of event A when event B happens*
- ***P(B|A)*** *is the probability of event B when A happens*

Bayes Theorem Statement

Bayes' Theorem for n set of events is defined as,

Let E_1, E_2,…, E_n be a set of events associated with the sample space S, in which all the events E_1, E_2,…, E_n have a non-zero probability of occurrence. All the events E_1, E_2,…, E form a partition of S. Let A be an event from space S for which we have to find probability, then according to Bayes' theorem,

$$P(E_i|A) = P(E_i)P(A|E_i) / \sum P(E_k)P(A|E_k)$$

for k = 1, 2, 3, …., n

Bayes Theorem Formula

For any two events A and B, then the formula for the Bayes theorem is given by: (the image given below gives the Bayes' theorem formula)

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

where,

- **P(A)** and **P(B)** are the probabilities of events A and B also P(B) is never equal to zero.
- **P(A|B)** is the probability of event A when event B happens
- **P(B|A)** is the probability of event B when A happens

Bayes' theorem is also known as the formula for the **probability of "causes". As we know, the E_i's are a partition of the sample space S, and at any given time only one of the events E_i occurs.** Thus we conclude that the Bayes' theorem formula gives the probability of a particular E_i, given the event A has occurred.

Bayes Theorem Derivation

The proof of Bayes' Theorem is given as, according to the conditional probability formula,

$P(E_i|A) = P(E_i \cap A) / P(A)$.....(i)

Then, by using the multiplication rule of probability, we get

$P(E_i \cap A) = P(E_i)P(A|E_i)$.....(ii)

Now, by the total probability theorem,

$P(A) = \sum P(E_k)P(A|E_k)$.....(iii)

Substituting the value of $P(E_i \cap A)$ and P(A) from eq (ii) and eq(iii) in eq(i) we get,

$$P(E_i|A) = P(E_i)P(A|E_i) / \sum P(E_k)P(A|E_k)$$

Naive Bayes Classifiers

A Naive Bayes classifiers, a family of algorithms based on Bayes' Theorem. Despite the "naive" assumption of feature independence, these classifiers are widely utilized for their simplicity and efficiency in machine learning.

What is Naive Bayes Classifiers?

Naive Bayes classifiers are a collection of classification algorithms based on Bayes' Theorem. It is not a single algorithm but a family of algorithms where all of them share a common principle, i.e. every pair of features being classified is independent of each other. To start with, let us consider a dataset.

One of the most simple and effective classification algorithms, the Naïve Bayes classifier aids in the rapid development of machine learning models with rapid prediction capabilities.

Naïve Bayes algorithm is used for classification problems. It is highly used in text classification. In text classification tasks, data contains high dimension (as each word represent one feature in the data). It is used in spam filtering, sentiment detection, rating classification etc. The advantage of using naïve Bayes is its speed. It is fast and making prediction is easy with high dimension of data.

This model predicts the probability of an instance belongs to a class with a given set of feature value. It is a probabilistic classifier. It is because it assumes that one feature in the model is independent of existence of another feature. In other words, each feature contributes to the predictions with no relation between each other. In real world, this condition satisfies rarely. It uses Bayes theorem in the algorithm for training and prediction

Why it is Called Naive Bayes?

The “Naive” part of the name indicates the simplifying assumption made by the Naïve Bayes classifier. The classifier assumes that the features used to describe an observation are conditionally independent, given the class label. The “Bayes” part of the name refers to Reverend Thomas Bayes, an 18th-century statistician and theologian who formulated Bayes’ theorem.
Consider a fictional dataset that describes the weather conditions for playing a game of golf. Given the weather conditions, each tuple classifies the conditions as fit(“Yes”) or unfit(“No”) for playing golf.Here is a tabular representation of our dataset.

	Outlook	Temperature	Humidity	Windy	Play Golf
0	Rainy	Hot	High	False	No
1	Rainy	Hot	High	True	No
2	Overcast	Hot	High	False	Yes
3	Sunny	Mild	High	False	Yes
4	Sunny	Cool	Normal	False	Yes
5	Sunny	Cool	Normal	True	No
6	Overcast	Cool	Normal	True	Yes
7	Rainy	Mild	High	False	No
8	Rainy	Cool	Normal	False	Yes
9	Sunny	Mild	Normal	False	Yes
10	Rainy	Mild	Normal	True	Yes
11	Overcast	Mild	High	True	Yes
12	Overcast	Hot	Normal	False	Yes
13	Sunny	Mild	High	True	No

The dataset is divided into two parts, namely, **feature matrix** and the **response vector**.

- Feature matrix contains all the vectors(rows) of dataset in which each vector consists of the value of **dependent features**. In above dataset, features are 'Outlook', 'Temperature', 'Humidity' and 'Windy'.

- Response vector contains the value of **class variable**(prediction or output) for each row of feature matrix. In above dataset, the class variable name is 'Play golf'.

Assumption of Naive Bayes

The fundamental Naive Bayes assumption is that each feature makes an:

- **Feature independence:** The features of the data are conditionally independent of each other, given the class label.

- **Continuous features are normally distributed:** If a feature is continuous, then it is assumed to be normally distributed within each class.

- **Discrete features have multinomial distributions:** If a feature is discrete, then it is assumed to have a multinomial distribution within each class.

- **Features are equally important:** All features are assumed to contribute equally to the prediction of the class label.

- **No missing data:** The data should not contain any missing values.

With relation to our dataset, this concept can be understood as:

We assume that no pair of features are dependent. For example, the temperature being ‘Hot’ has nothing to do with the humidity or the outlook being ‘Rainy’ has no effect on the winds. Hence, the features are assumed to be **independent**.

Secondly, each feature is given the same weight(or importance). For example, knowing only temperature and humidity alone can’t predict the outcome accurately. None of the attributes is irrelevant and assumed to be contributing **equally** to the outcome.

Bayes' Theorem

Bayes' Theorem finds the probability of an event occurring given the probability of another event that has already occurred. Bayes' theorem is stated mathematically as the following equation:

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

where A and B are events and $P(B) \neq 0$

- Basically, we are trying to find probability of event A, given the event B is true. Event B is also termed as **evidence**.

- P(A) is the **priori** of A (the prior probability, i.e. Probability of event before evidence is seen). The evidence is an attribute value of an unknown instance(here, it is event B).

- P(B) is Marginal Probability: Probability of Evidence.
- P(A|B) is a posteriori probability of B, i.e. probability of event after evidence is seen.

- P(B|A) is Likelihood probability i.e the likelihood that a hypothesis will come true based on the evidence.

Now, with regards to our dataset, we can apply Bayes' theorem in following way:

$$P(y|X) = \frac{P(X|y)P(y)}{P(X)}$$

where, y is class variable and X is a dependent feature vector (of size n) where:

$X=(x1,x2,x3,.....,xn)X=(x1,x2,x3,.....,xn)$Just to clear, an example of a feature vector and corresponding class variable can be: (refer 1st row of dataset)

X = (Rainy, Hot, High, False)
y = No

So basically, $P(y|X)P(y|X)$here means, the probability of "Not playing golf" given that the weather conditions are "Rainy outlook", "Temperature is hot", "high humidity" and "no wind".

With relation to our dataset, this concept can be understood as:

- We assume that no pair of features are dependent. For example, the temperature being 'Hot' has nothing to do with the humidity or the outlook being 'Rainy' has no effect on the winds. Hence, the features are assumed to be **independent**.
- Secondly, each feature is given the same weight(or importance). For example, knowing only temperature and humidity alone can't predict the outcome accurately. None of the attributes is irrelevant and assumed to be contributing **equally** to the outcome.

Now, its time to put a naive assumption to the Bayes' theorem, which is, **independence** among the features. So now, we split **evidence** into the independent parts.
Now, if any two events A and B are independent, then,

P(A,B) = P(A)P(B)

Hence, we reach to the result:

$$P(y|x_1,\ldots,x_n) = \frac{P(x_1|y)P(x_2|y)\ldots P(x_n|y)P(y)}{P(x_1)P(x_2)\ldots P(x_n)}$$

which can be expressed as:

$$P(y|x_1,\ldots,x_n) = \frac{P(y)\prod_{i=1}^{n} P(x_i|y)}{P(x_1)P(x_2)\ldots P(x_n)}$$

Now, as the denominator remains constant for a given input, we can remove that term:

$$P(y|x_1,\ldots,x_n) \propto P(y)\prod_{i=1}^{n} P(x_i|y)$$

Now, we need to create a classifier model. For this, we find the probability of given set of inputs for all possible values of the class variable *y* and pick up the output with maximum probability. This can be expressed mathematically as:

$$y = argmax_y P(y) \prod_{i=1}^{n} P(x_i|y)$$

So, finally, we are left with the task of calculating $P(y)P(y)$and $P(xi|y)P(xi|y)$.
Please note that $P(y)P(y)$ is also called class probability and $P(xi|y)P(xi|y)$ is called conditional probability.

The different naive Bayes classifiers differ mainly by the assumptions they make regarding the distribution of $P(xi|y).P(xi|y)$.

Let us try to apply the above formula manually on our weather dataset. For this, we need to do some precomputations on our dataset.

We need to find$P(xi|yj)P(xi|yj)$for each $xixi$ in X and$yjyj$ in y. All these calculations have been demonstrated in the tables below:

Outlook

	Yes	No	P(yes)	P(no)
Sunny	3	2	2/9	3/5
Overcast	4	0	4/9	0/5
Rainy	3	2	3/9	2/5
Total	9	5	100%	100%

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	100%	100%

Humidity

	Yes	No	P(yes)	P(no)
High	3	4	3/9	4/5
Normal	6	1	6/9	1/5
Total	9	5	100%	100%

Wind

	Yes	No	P(yes)	P(no)
False	6	2	6/9	2/5
True	3	3	3/9	3/5
Total	9	5	100%	100%

Play		**P(Yes)/P(No)**
Yes	9	9/14
No	5	5/14
Total	14	100%

So, in the figure above, we have calculated $P(xi \mid yj)P(xi \mid yj)$ for each $xixi$ in X and $yjyj$ in y manually in the tables 1-4.

For example, probability of playing golf given that the temperature is cool, i.e P(temp. = cool | play golf = Yes) = 3/9.

Also, we need to find class probabilities $P(y)P(y)$ which has been calculated in the table 5. For example, P(play golf = Yes) = 9/14.

So now, we are done with our pre-computations and the classifier is ready!
Let us test it on a new set of features (let us call it today):

today = (Sunny, Hot, Normal, False)

$$P(Yes|today) = \frac{P(SunnyOutlook|Yes)P(HotTemperature|Yes)P(NormalHumidity|Yes)P(NoWind|Yes)P(Yes)}{P(today)}$$

and probability to not play golf is given by:

$$P(No|today) = \frac{P(SunnyOutlook|No)P(HotTemperature|No)P(NormalHumidity|No)P(NoWind|No)P(No)}{P(today)}$$

Since, P(today) is common in both probabilities, we can ignore P(today) and find proportional probabilities as:

$$P(Yes|today) \propto \frac{3}{9}.\frac{2}{9}.\frac{6}{9}.\frac{6}{9}.\frac{9}{14} \approx 0.02116$$

and

$$P(No|today) \propto \frac{3}{5}.\frac{2}{5}.\frac{1}{5}.\frac{2}{5}.\frac{5}{14} \approx 0.0068$$

Now, since

$P(Yes|today)+P(No|today)=1$

These numbers can be converted into a probability by making the sum equal to 1 (normalization):

$$P(Yes|today) = \frac{0.02116}{0.02116+0.0068} \approx 0.0237$$

and

$$P(No|today) = \frac{0.0068}{0.0141+0.0068} \approx 0.33$$

Since

$P(Yes|today) > P(No|today)$ $P(Yes|today) > P(No|today)$ So, prediction that golf would be played is 'Yes'.

The method that we discussed above is applicable for discrete data. In case of continuous data, we need to make some assumptions regarding the distribution of values of each feature. The different naive Bayes classifiers differ mainly by the assumptions they make regarding the distribution of $P(xi|y)$. $P(xi|y)$.

1. What assumption does Naïve Bayes Classification make about the independence of features?

a) Features are completely independent of each other
b) Features are conditionally independent given the class
c) Features are dependent on each other
d) Features are unrelated to the class

Answer: B

2. What does Bayes' Theorem provide a way to calculate?

a) The probability of an event given prior knowledge
b) The likelihood of an event occurring without prior knowledge
c) The frequency of an event in a given population
d) d) The variance of a dataset

Answer: B

THANK YOU

OUR TEAM