



SNS COLLEGE OF ENGINEERING

Kurumbapalayam (Po), Coimbatore – 641 107

An Autonomous Institution

Accredited by NBA – AICTE and Accredited by NAAC – UGC with 'A' Grade
Approved by AICTE, New Delhi & Affiliated to Anna University, Chennai

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

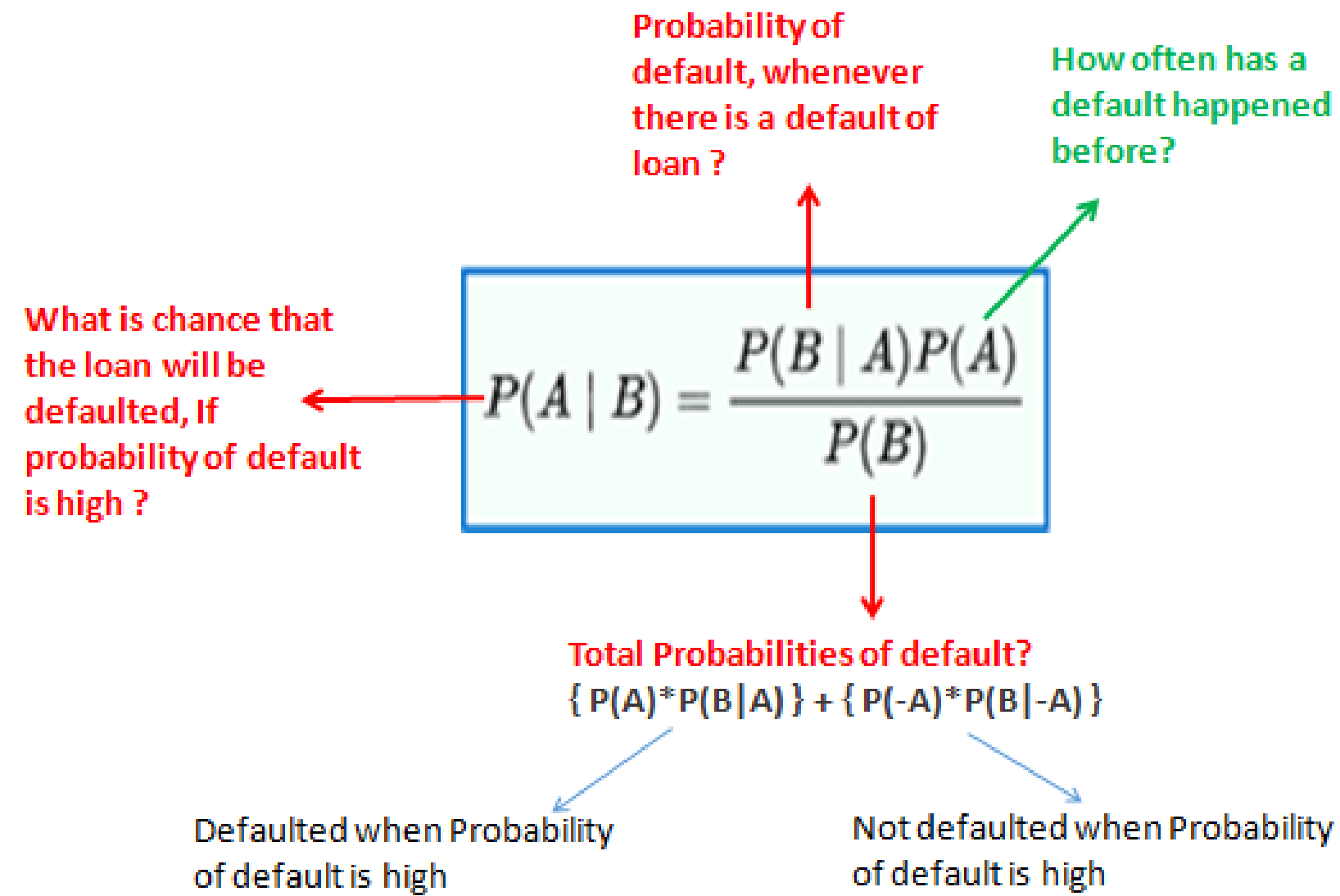
COURSE NAME : 19CS501 Introduction to Machine Learning

III YEAR /V SEMESTER

Unit 2- SUPERVISED LEARNING

Topic : Bayesian Learning, Naïve Bayes







Bayesian Learning for Machine Learning-

The Famous Coin Flip Experiment



- ❑ When we flip a coin, there are two possible outcomes — heads or tails. Of course, there is a third rare possibility where the coin balances on its edge without falling onto either side, which we assume is not a possible outcome of the coin flip for our discussion.
- ❑ We conduct a series of coin flips and record our observations i.e. the number of the heads (or tails) observed for a certain number of coin flips.
- ❑ In this experiment, we are trying to determine the fairness of the coin, using the number of heads (or tails) that we observe



Frequentist Statistics

- ❑ Let's think about how we can determine the fairness of the coin using our observations in the above-mentioned experiment. Once we have conducted a sufficient number of coin flip trials, we can determine the frequency or the probability of observing the heads (or tails).
- ❑ If we observed heads and tails with equal frequencies or the probability of observing heads (or tails) is 0.5, then it can be established that the coin is a fair coin. Failing that, it is a biased coin.
- ❑ Let's denote p as the probability of observing the heads. Consequently, as the quantity that p deviates from 0.5 indicates how biased the coin is, p can be considered as the degree-of-fairness of the coin.
- ❑ Testing whether a hypothesis is true or false by calculating the probability of an event in a prolonged experiment is known as frequentist statistics.
- ❑ As such, determining the fairness of a coin by using the probability of observing the heads is an example of frequentist statistics (a.k.a. frequentist approach).



Frequentist Statistics

- ✓ How confident are we of p being 0.6?
- ✓ How confident are of p being 0.55?
- ✓ Which of these values is the accurate estimation of p ?
- ✓ Will p continue to change when we further increase the number of coin flip trails?

Number of coin flips	Number of heads	Probability of observing heads (p)
10	6	0.6
50	29	0.58
100	55	0.55
200	94	0.47
500	245	0.49



Frequentist Statistics-Cases

- ✓ An experiment with an infinite number of trials guarantees p with absolute accuracy (100% confidence). Yet, it is not practical to conduct an experiment with an infinite number of trials and we should stop the experiment after a sufficiently large number of trials. However, deciding the value of this sufficient number of trials is a challenge when using frequentist statistics.
- ✓ If we can determine the confidence of the estimated p value or the inferred conclusion in a situation where the number of trials are limited, this will allow us to decide whether to accept the conclusion or to extend the experiment with more trials until it achieves sufficient confidence.



Some Terms to Understand

- ✓ **Random variable** (Stochastic variable) — In statistics, the random variable is a variable whose possible values are a result of a random event. Therefore, each possible value of a random variable has some probability attached to it to represent the likelihood of those values.
- ✓ **Probability distribution** — The function that defines the probability of different outcomes/values of a random variable. The continuous probability distributions are described using probability density functions whereas discrete probability distributions can be represented using probability mass functions.
- ✓ **Conditional probability** — This is a measure of probability $P(A|B)$ of an event A given that another event B has occurred.
- ✓ **Joint probability distribution**



Introduction to Bayesian Learning

- ✓ Imagine a situation where your friend gives you a new coin and asks you the fairness of the coin (or the probability of observing heads) without even flipping the coin once. In fact, you are also aware that your friend has not made the coin biased. In general, you have seen that coins are fair, thus you expect the probability of observing heads is 0.5. In the absence of any such observations, you assert the fairness of the coin only using your past experiences or observations with coins.
- ✓ Suppose that you are allowed to flip the coin 10 times in order to determine the fairness of the coin. Your observations from the experiment will fall under one of the following cases:
 - Case 1: observing 5 heads and 5 tails.
 - Case 2: observing h heads and $10-h$ tails, where $h \neq 10-h$.



Introduction to Bayesian Learning

- ✓ If case 1 is observed, you are now more certain that the coin is a fair coin, and you will decide that the probability of observing heads is 0.5 with more confidence. If case 2 is observed, you can either:
 1. Neglect your prior beliefs since now you have new data and decide the probability of observing heads is $h/10$ by solely depending on recent observations.
 2. Adjust your belief accordingly to the value of h that you have just observed, and decide the probability of observing heads using your recent observations.

The first method suggests that we use the frequentist method, where we omit our beliefs when making decisions. However, the second method seems to be more convenient because 10 coins are insufficient to determine the fairness of a coin. Therefore, we can make better decisions by combining our recent observations and beliefs that we have gained through our past experiences



Bayes' Theorem

- ❑ Bayes' theorem describes how the conditional probability of an event or a hypothesis can be computed using evidence and prior knowledge.
- ❑ It is similar to concluding that our code has no bugs given the evidence that it has passed all the test cases, including our prior belief that we have rarely observed any bugs in our code.
- ❑ However, this intuition goes beyond that simple hypothesis test where there are multiple events or hypotheses involved (let us not worry about this for the moment).

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



Bayes' Theorem

- ❑ **$P(\theta)$** — Prior Probability is the probability of the hypothesis θ being true before applying the Bayes' theorem. Prior represents the beliefs that we have gained through past experience, which refers to either common sense or an outcome of Bayes' theorem for some past observations
- ❑ **$P(X|\theta)$** — Likelihood is the conditional probability of the evidence given a hypothesis. The likelihood is mainly related to our observations or the data we have. If it is given that our code is bug-free, then the probability of our code passing all test cases is given by the likelihood.
- ❑ **$P(X)$** — Evidence term denotes the probability of evidence or data. This can be expressed as a summation (or integral) of the probabilities of all possible hypotheses weighted by the likelihood of the same.



Bayes' Theorem

we can write $P(X)$ as: $P(X) = \sum_{\theta \in \Theta} P(X|\theta)P(\theta)$ Θ is the set of all the hypotheses.

For the continuous θ , we write $P(X)$ as an integration:

$$P(X) = \int_{\theta} P(X|\theta)P(\theta)d\theta$$

- 1) observing no bugs in our code or
- 2) 2) observing a bug in our code. Therefore we can denotes evidence as follows:

$$P(X) = P(X|\theta)P(\theta) + P(X|\neg\theta)P(\neg\theta)$$

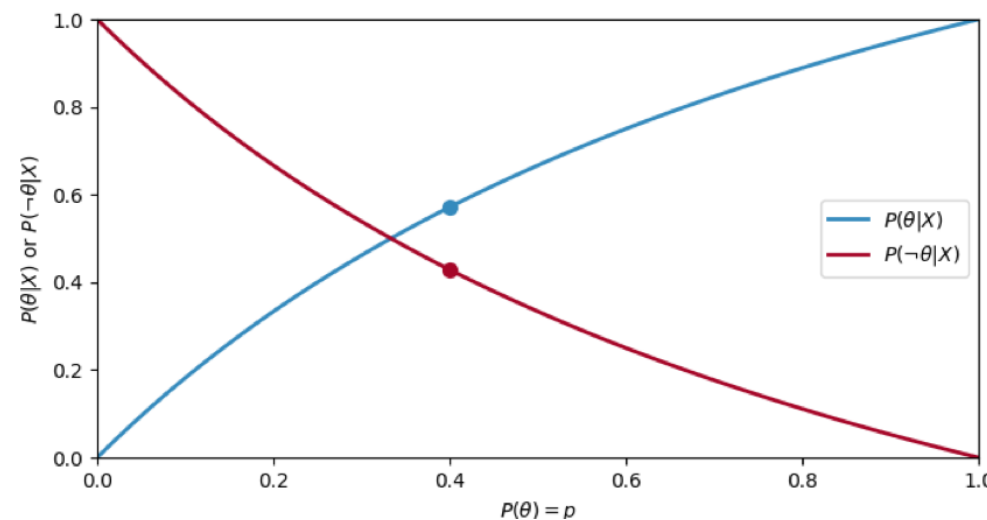
$P(\theta|X)$ — Posteriori probability denotes the conditional probability of the hypothesis θ after observing the evidence X . This is the probability of observing no bugs in our code given that it passes all the test cases.

$$P(\theta|X) = \frac{1 \times p}{0.5(1 + p)}$$

Maximum a Posteriori (MAP)

We can use MAP to determine the valid hypothesis from a set of hypotheses. According to MAP, the hypothesis that has the maximum posterior probability is considered as the valid hypothesis. Therefore, we can express the hypothesis θ_{MAP} that is concluded using MAP as follows:

$$\begin{aligned} \theta_{MAP} &= \operatorname{argmax}_{\theta} P(\theta_i|X) \\ &= \operatorname{argmax}_{\theta} \left(\frac{P(X|\theta_i)P(\theta_i)}{P(X)} \right) \end{aligned}$$

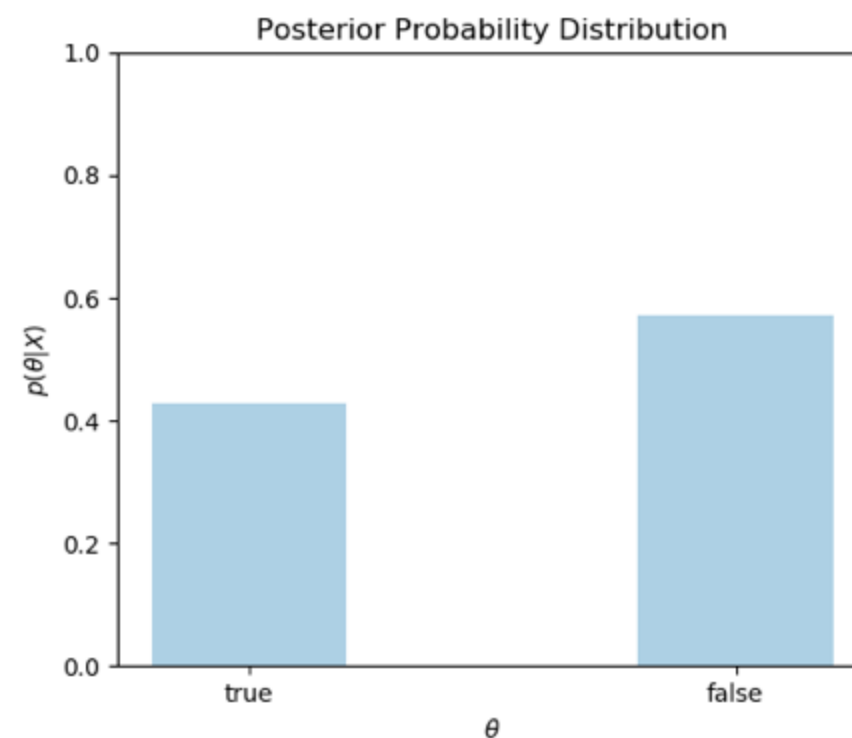
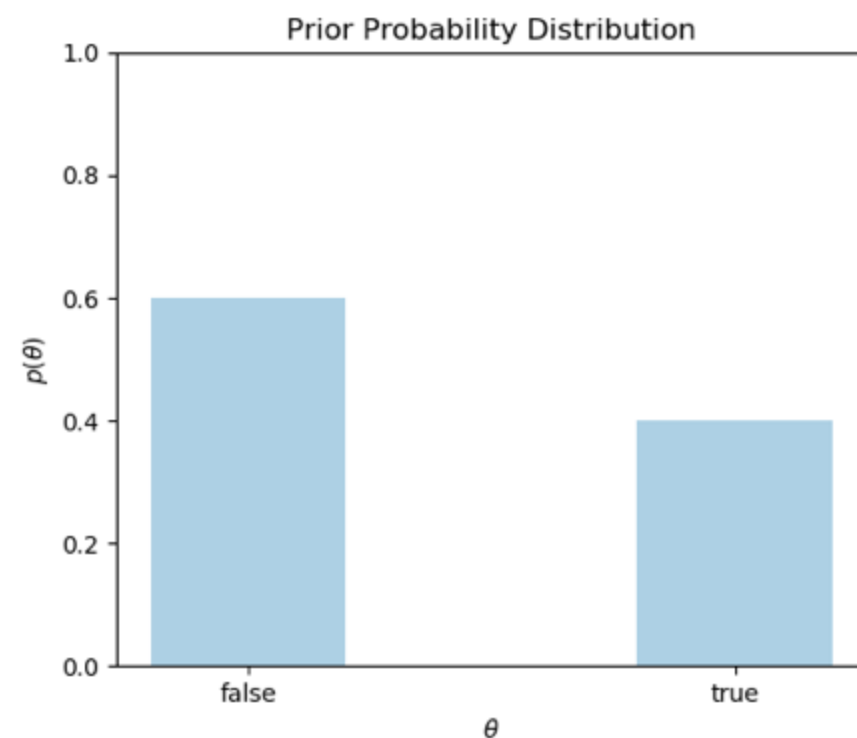


$$\begin{aligned} MAP &= \operatorname{argmax}_{\theta} \left\{ \theta : P(\theta|X) = \frac{0.4}{0.5(1+0.4)}, \neg\theta : P(-\theta|X) = \frac{0.5(1-0.4)}{0.5(1+0.4)} \right\} \\ &= \operatorname{argmax}_{\theta} \left\{ \theta : P(\theta|X) = 0.57, \neg\theta : P(-\theta|X) = 0.43 \right\} \\ &= \theta \implies \text{No bugs present in our code} \end{aligned}$$



Bayesian Learning

- We defined that the event of not observing bug is θ and the probability of producing a bug-free code $P(\theta)$ was taken as p . However, the event θ can actually take two values — either true or false — corresponding to not observing a bug or observing a bug respectively.
- Therefore, observing a bug or not observing a bug are not two separate events, they are two possible outcomes for the same event θ





Binomial Likelihood



- The likelihood for the coin flip experiment is given by the probability of observing heads out of all the coin flips given the fairness of the coin. As we have defined the fairness of the coins (θ) using the probability of observing heads for each coin flip, we can define the probability of observing heads or tails given the fairness of the coin $P(y|\theta)$ where $y = 1$ for observing heads and $y = 0$ for observing tails.

Accordingly:

$$P(y = 1|\theta) = \theta$$

$$P(y = 0|\theta) = (1 - \theta)$$

$$P(Y = y|\theta) = \begin{cases} \theta, & \text{if } y = 1 \\ 1 - \theta, & \text{otherwise} \end{cases}$$

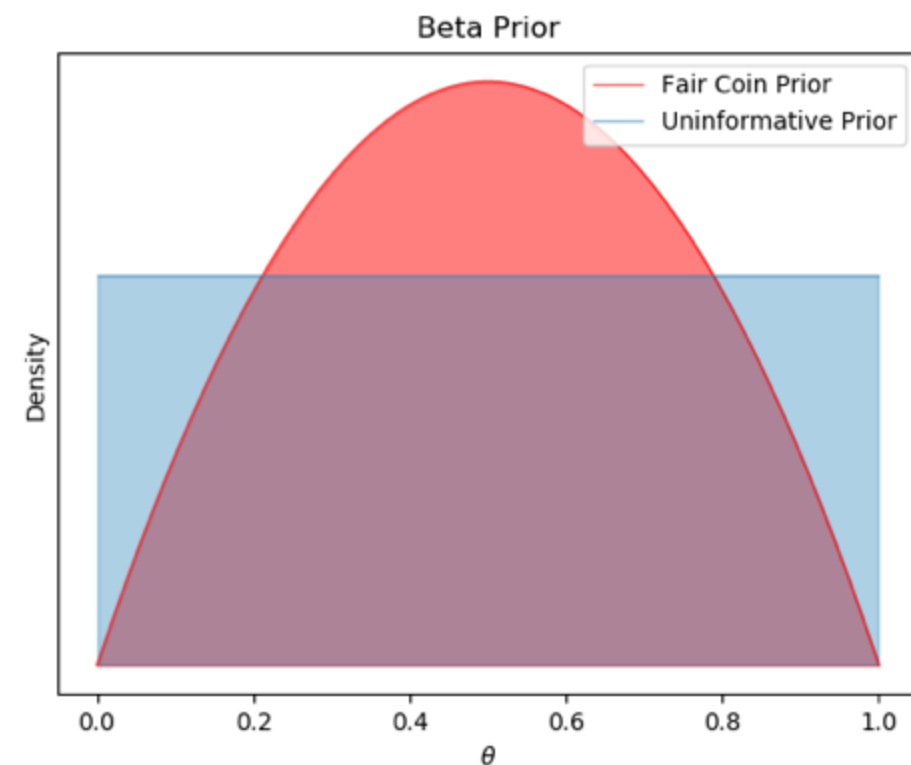
$$P(Y = y|\theta) = \theta^y \times (1 - \theta)^{1-y}$$

$$P(\theta) = \frac{\theta^{\alpha-1}(1 - \theta)^{\beta-1}}{B(\alpha, \beta)}$$

Beta Prior Distribution

- The prior distribution is used to represent our belief about the hypothesis based on our past experiences. We can choose any distribution for the prior if it represents our belief regarding the fairness of the coin. For this example, we use Beta distribution to represent the prior probability distribution as follows

$$P(\theta) = \frac{\theta^{\alpha-1}(1-\theta)^{\beta-1}}{B(\alpha, \beta)}$$





Posterior Distribution

- I previously mentioned that Beta is a conjugate prior and therefore the posterior distribution should also be a Beta distribution. Let us now try to derive the posterior distribution analytically using the Binomial likelihood and the Beta prior.
- First of all, consider the product of Binomial likelihood and Beta prior

$$\begin{aligned} P(X|\theta) \times P(\theta) &= P(N, k|\theta) \times P(\theta) \\ &= \binom{N}{k} \theta^k (1 - \theta)^{N-k} \times \frac{\theta^{\alpha-1} (1 - \theta)^{\beta-1}}{B(\alpha, \beta)} \\ &= \frac{\binom{N}{k}}{B(\alpha, \beta)} \times \theta^{(k+\alpha)-1} (1 - \theta)^{(N+\beta-k)-1} \end{aligned}$$

assignmentsolutionguru.com

posterior Probability


Likelihood

Prior Probability

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

Evidence

**NAIVE BAYES
CLASSIFIER**





Naïve Bayes Classifier Algorithm



- Naïve Bayes algorithm is a supervised learning algorithm, which is based on Bayes theorem and used for solving classification problems.
- It is mainly used in text classification that includes a high-dimensional training dataset.
- Naïve Bayes Classifier is one of the simple 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.



Why is it called Naïve Bayes?

- **Naïve:** It is called Naïve because it assumes that the occurrence of a certain feature is independent of the occurrence of other features. Such as if the fruit is identified on the bases of **color, shape, and taste, then red, spherical, and sweet fruit** is recognized as an apple. Hence each feature individually contributes to identify that it is an apple without depending on each other.
- **Bayes:** 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.
- The formula for Bayes' theorem is given as:

Where,

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

- $P(A|B)$ 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



Working of Naïve Bayes' Classifier



- Suppose we have a dataset of weather conditions and corresponding target variable "Play". So using this dataset we need to decide that whether we should play or not on a particular day according to the weather conditions. So to solve this problem, we need to follow the below steps:
 - I. Convert the given dataset into frequency tables.
 - II. Generate Likelihood table by finding the probabilities of given features.
 - III. Now, use Bayes theorem to calculate the posterior probability.
- **Problem: If the weather is sunny, then the Player should play or not?**



Applications of Naïve Bayes Classifier:

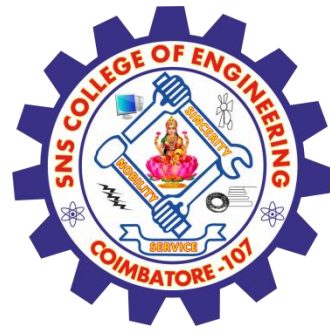


- It is used for Credit Scoring.
- It is used in medical data classification.
- It can be used in real-time predictions because Naïve Bayes Classifier is an eager learner.
- It is used in Text classification such as Spam filtering and Sentiment analysis.



Types of Naïve Bayes Model

- **Gaussian:** The Gaussian model assumes that features follow a normal distribution. This means if predictors take continuous values instead of discrete, then the model assumes that these values are sampled from the Gaussian distribution.
- **Multinomial:** The Multinomial Naïve Bayes classifier is used when the data is multinomial distributed. It is primarily used for document classification problems, it means a particular document belongs to which category such as Sports, Politics, education, etc.
The classifier uses the frequency of words for the predictors.
- **Bernoulli:** The Bernoulli classifier works similar to the Multinomial classifier, but the predictor variables are the independent Booleans variables. Such as if a particular word is present or not in a document. This model is also famous for document classification tasks.



Merits & Demerits

➤ Advantages of Naïve Bayes Classifier:

- ✓ Naïve Bayes is one of the fast and easy ML algorithms to predict a class of datasets.
- ✓ It can be used for Binary as well as Multi-class Classifications.
- ✓ It performs well in Multi-class predictions as compared to the other Algorithms.
- ✓ It is the most popular choice for text classification problems.

➤ Disadvantages of Naïve Bayes Classifier:

- ✓ Naive Bayes assumes that all features are independent or unrelated, so it cannot learn the relationship between features.



Assessment



- Python Implementation of the Naïve Bayes algorithm



REFERENCES



1. Tom M. Mitchell, “Machine Learning”, McGraw-Hill Education (India) Private Limited, 2713.
2. Trevor Hastie, Robert Tibshirani, Jerome Friedman, “The Elements of Statistical Learning: Data Mining, Inference, and Prediction”, Springer; Second Edition, 2709.

THANK YOU