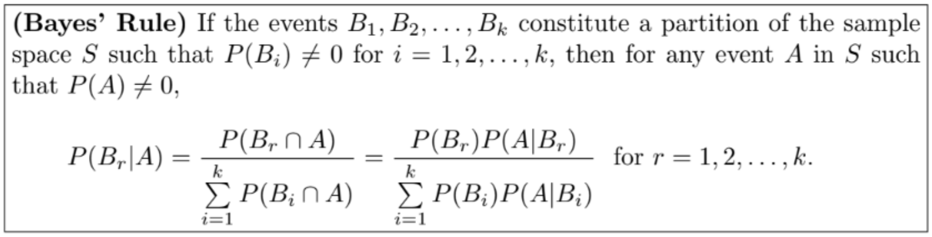
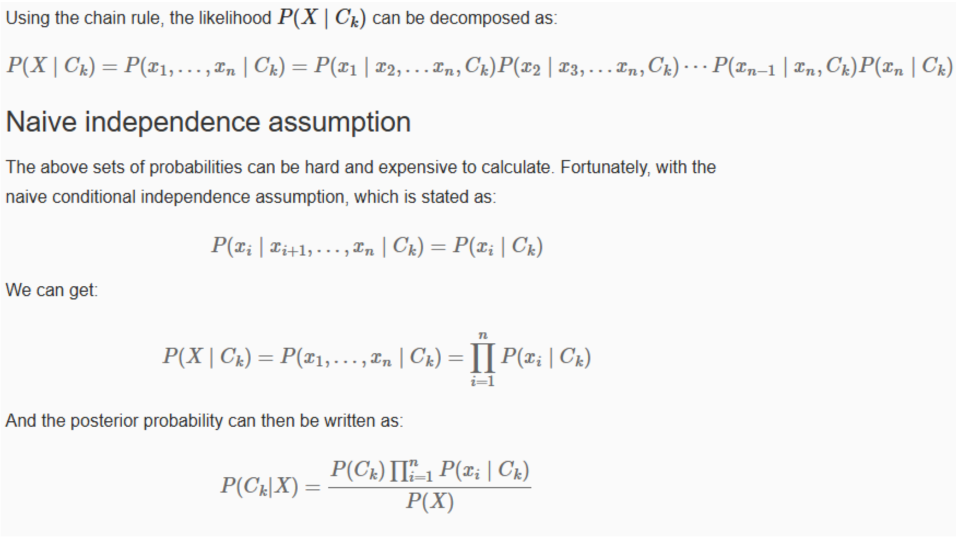
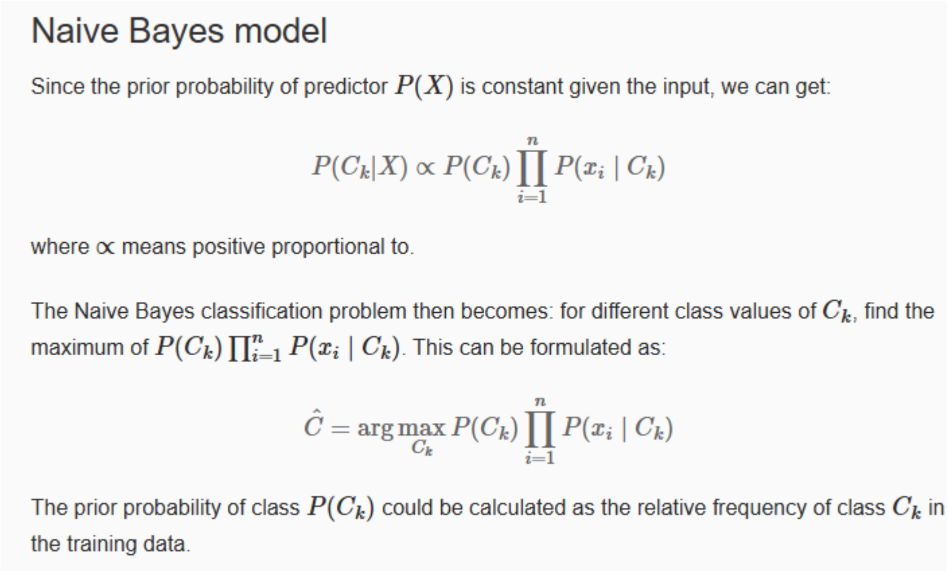
Practical Assignment No. 2	
Title:	Continuous Probability Distribution and Naive Bayes
Problem Statement:	Generate data that follows continuous probability distributions. Implement a Naive Bayes classifier for continuous data using any programming language.
Objective:	To apply mathematical concepts in computer science for solving the problems.
Outcome:	CO505.1: Implement the program to solve the problems using probability.
Software or Hardware Requirements:	Anaconda/Java/GCC
Theory:	**Naive Bayes Classifier**  **(Bayes' Rule)** If the events $B_1, B_2, \ldots, B_k$ constitute a partition of the sample space S such that $P(B_i) \neq 0$ for $i = 1, 2, \ldots, k$, then for any event A in S such that $P(A) \neq 0$, $P(B_r \mid A) = \frac{P(B_r \cap A)}{\sum_{i=1}^{k} P(B_i \cap A)} = \frac{P(B_r)P(A \mid B_r)}{\sum_{i=1}^{k} P(B_i)P(A \mid B_i)}$ for $r = 1, 2, \ldots, k$.  Using the chain rule, the likelihood $P(X \mid C_k)$ can be decomposed as: $P(X \mid C_k) = P(x_1, \ldots, x_n \mid C_k) = P(x_1 \mid x_2, \ldots x_n, C_k) P(x_2 \mid x_3, \ldots x_n, C_k) \cdots P(x_{n-1} \mid x_n, C_k) P(x_n \mid C_k)$ Naive independence assumption The above sets of probabilities can be hard and expensive to calculate. Fortunately, with the naive conditional independence assumption, which is stated as: $P(x_i \mid x_{i+1}, \ldots, x_n \mid C_k) = P(x_i \mid C_k)$ We can get: $P(X \mid C_k) = P(x_1, \ldots, x_n \mid C_k) = \prod_{i=1}^{n} P(x_i \mid C_k)$ And the posterior probability can then be written as: $P(C_k \mid X) = \frac{P(C_k) \prod_{i=1}^{n} P(x_i \mid C_k)}{P(X)}$



Naive Bayes model

Since the prior probability of predictor $P(X)$ is constant given the input, we can get:

$$P(C_k|X) \propto P(C_k) \prod_{i=1}^{n} P(x_i \mid C_k)$$

where ∝ means positive proportional to.

The Naive Bayes classification problem then becomes: for different class values of C_k, find the maximum of $P(C_k) \prod_{i=1}^{n} P(x_i \mid C_k)$. This can be formulated as:

$$\hat{C} = \arg\max_{C_k} P(C_k) \prod_{i=1}^{n} P(x_i \mid C_k)$$

The prior probability of class $P(C_k)$ could be calculated as the relative frequency of class C_k in the training data.

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

- Feature independence
- Continuous features are normally distributed
- Discrete features have multinomial distributions
- Features are equally important
- No missing data

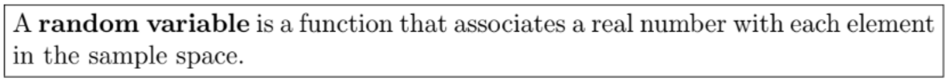
Types of Naive Bayes Model
There are three common types of Naive Bayes Model:

1) Gaussian Naive Bayes classifier
2) Multinomial Naive Bayes
3) Bernoulli Naive Bayes

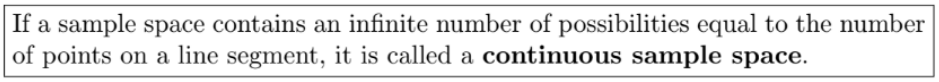
Applications of Naive Bayes Algorithms

- Real-time Prediction
- Multi-class Prediction
- Text classification/ Spam Filtering/ Sentiment Analysis
- Recommendation System

Important Concepts-

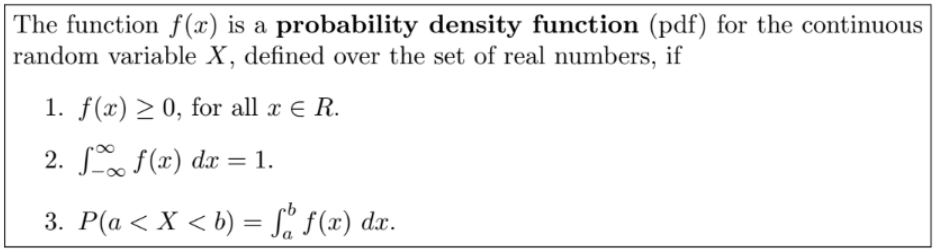


A **random variable** is a function that associates a real number with each element in the sample space.



If a sample space contains an infinite number of possibilities equal to the number of points on a line segment, it is called a **continuous sample space**.

When a random variable can take on values on a continuous scale, it is called a **continuous random variable.**

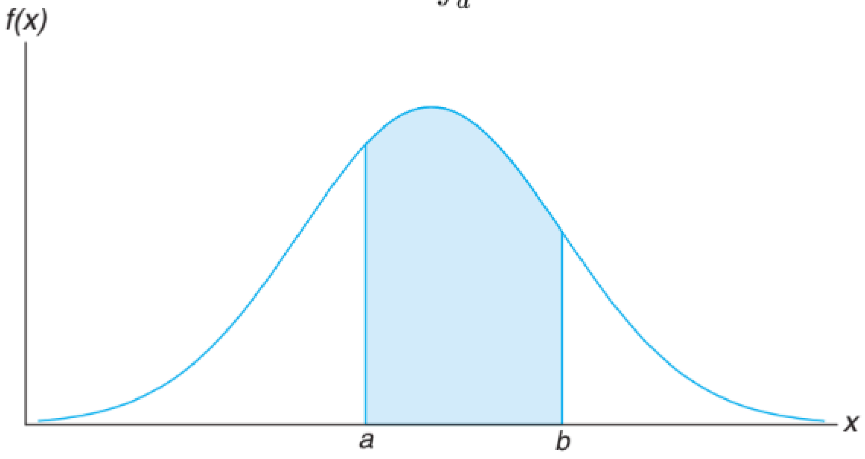


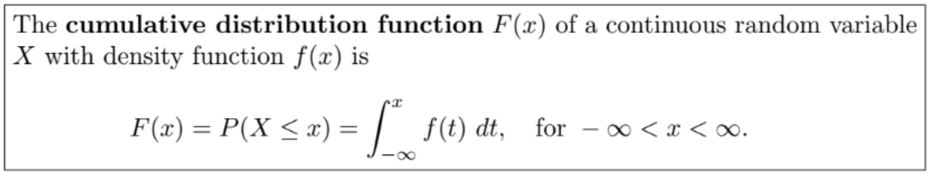
The function $f(x)$ is a **probability density function** (pdf) for the continuous random variable X, defined over the set of real numbers, if

1. $f(x) \geq 0$, for all $x \in R$.
2. $\int_{-\infty}^{\infty} f(x)\ dx = 1$.
3. $P(a < X < b) = \int_a^b f(x)\ dx$.

In Figure below, the probability that X assumes a value between a and b is equal to the shaded area under the density function between the ordinates at x = a and x = b, and from integral calculus is given by

$$P(a < X < b) = \int_a^b f(x)\ dx.$$





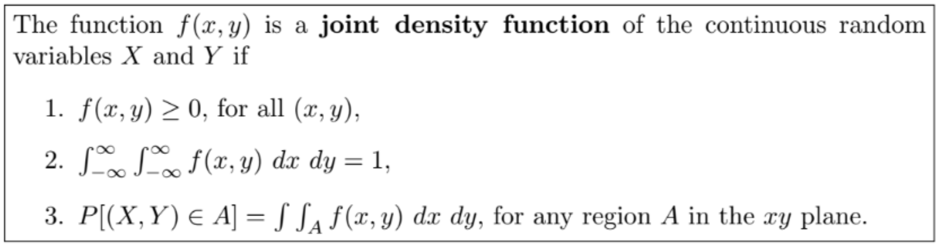
The **cumulative distribution function** $F(x)$ of a continuous random variable X with density function $f(x)$ is

$$F(x) = P(X \leq x) = \int_{-\infty}^{x} f(t)\ dt, \quad \text{for } -\infty < x < \infty.$$

As an immediate consequence of Definition above, one can write the two results

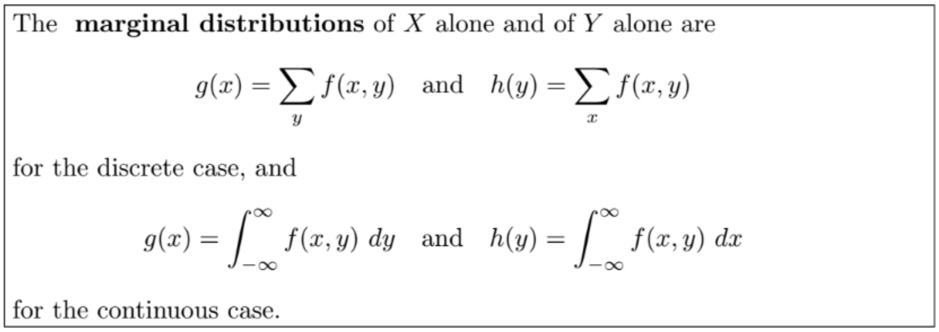
$$P(a < X < b) = F(b) - F(a) \text{ and } f(x) = \frac{dF(x)}{dx},$$

if the derivative exists.



The function $f(x,y)$ is a **joint density function** of the continuous random variables X and Y if

1. $f(x,y) \geq 0$, for all (x,y),
2. $\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x,y)\ dx\ dy = 1$,
3. $P[(X,Y) \in A] = \int\int_A f(x,y)\ dx\ dy$, for any region A in the xy plane.



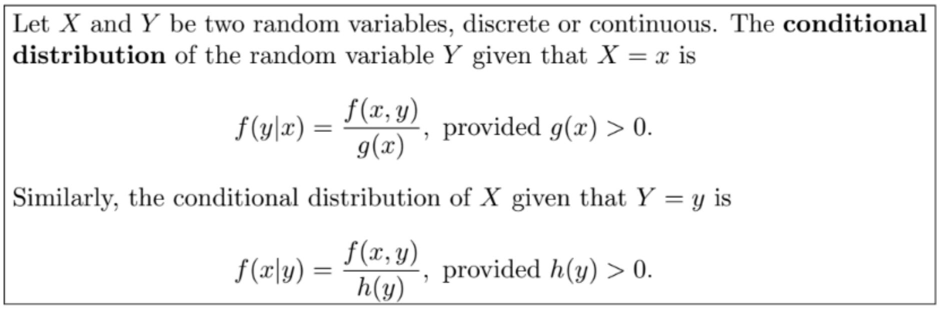
The **marginal distributions** of X alone and of Y alone are

$$g(x) = \sum_y f(x,y) \quad \text{and} \quad h(y) = \sum_x f(x,y)$$

for the discrete case, and

$$g(x) = \int_{-\infty}^{\infty} f(x,y)\ dy \quad \text{and} \quad h(y) = \int_{-\infty}^{\infty} f(x,y)\ dx$$

for the continuous case.

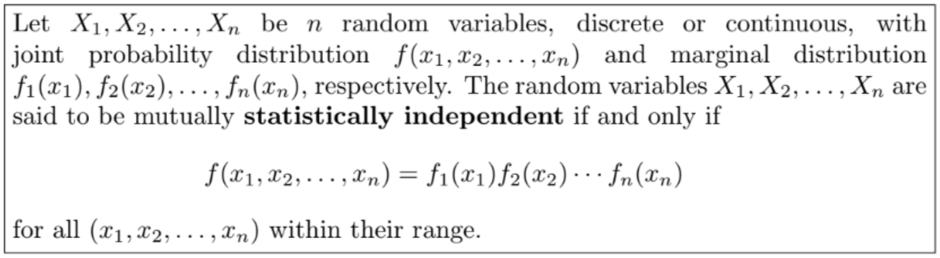


Let X and Y be two random variables, discrete or continuous. The **conditional distribution** of the random variable Y given that $X = x$ is

$$f(y|x) = \frac{f(x,y)}{g(x)}, \text{ provided } g(x) > 0.$$

Similarly, the conditional distribution of X given that $Y = y$ is

$$f(x|y) = \frac{f(x,y)}{h(y)}, \text{ provided } h(y) > 0.$$

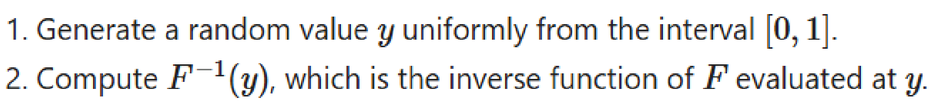


Let $X_1, X_2, \ldots, X_n$ be n random variables, discrete or continuous, with joint probability distribution $f(x_1, x_2, \ldots, x_n)$ and marginal distribution $f_1(x_1), f_2(x_2), \ldots, f_n(x_n)$, respectively. The random variables $X_1, X_2, \ldots, X_n$ are said to be mutually **statistically independent** if and only if

$$f(x_1, x_2, \ldots, x_n) = f_1(x_1)f_2(x_2)\cdots f_n(x_n)$$

for all $(x_1, x_2, \ldots, x_n)$ within their range.

If X is a random variable with CDF- F, then $F(X)$ follows a uniform distribution between 0 and 1. This opens up the possibility of generating artificial data with any desired distribution, given that we know F. The process is as follows:



1. Generate a random value y uniformly from the interval $[0, 1]$.
2. Compute $F^{-1}(y)$, which is the inverse function of F evaluated at y.

There are three common probability distributions that can be used to generate data:

- Uniform

	● Binomial ● Gaussian (*Data generation using any one of above method followed by Naive Bayes)
Input/Datasets/Test Cases:	Size of data to be generated and range.
Results:	Print histogram of data generated. Write result values in table
Analysis and conclusion:	Write your own analysis of output and conclusion(Minimum 1 statement Analysis, Minimum 1 Statement Conclusion)
References:	Reference Links(Any 2)