Probability and Statistics for Machine Learning - By Suraj

Table of Contents

S.No. Topic

1. Fundamentals of Probability

2. Descriptive & Inferential Statistics

Table of Contents

23 Markov Chains

24 Monte Carlo Methods

25 Stochastic Processes

26 Conditional Probability Distribution

27 Probability Density Function

28 Cumulative Distribution Function

29 Moment and Moment Generating Function

30 Skewness

31 Kurtosis

32 Covariance

33 Independence

34 Conditional Independence

35 Stationary Distribution

36 Bootstrap Sampling

37 Cross Validation

38 Expectation Maximisation Algorithm

39 Kullback-Leibler Divergence

40 Mutual Information

41 Singular Value Decomposition

42 Power Law

43 Heavy Tailed Distribution

44 Entropy

45 p-value

46 Markov chain Monte Carlo

47 Empirical Distribution Function

48 Kernel Density Function

49 Copula

50 Extreme Value Theory

51 Brownian Motion

52 Information Theory

53 Shannon Entropy

54 Bias-Variance Tradeoff

55 T-Statistic and margin of error.

56 Binomial Distribution.

A. Fundamentals of probability

1. Experiment

An experiment is any process that generates an outcome. In probability theory, we are interested in the possible outcomes of an experiment and the likelihood of each outcome occurring.

2. Sample Space

Sample Space

Sample space is a term used in probability theory and statistics to refer to the set of all possible outcomes or events that can occur in a given experiment or situation. It is denoted by the symbol "S".

For example, if you flip a coin, the sample space is {Heads, Tails}. If you roll a dice, the sample space is {1, 2, 3, 4, 5, 6}. If you draw a card from a standard deck of 52 cards, the sample space is {Ace of Spades, Ace of Hearts, Ace of Diamonds, Ace of Clubs, 2 of Spades, 2 of Hearts, 2 of Diamonds, 2 of Clubs, 3 of Spades, 3 of Hearts, ..., King of Diamonds, King of Clubs}.

3. Event

In probability theory and statistics, an event is a subset of the sample space, which is the set of all possible outcomes of an experiment. An event is said to occur if any one of the outcomes in the subset occurs. Events are often denoted by capital letters, such as A, B, C, etc.

Let's consider the experiment of rolling two dice. The sample space for this experiment is:

S = {(1, 1), (1, 2), (1, 3), ..., (6, 4), (6, 5), (6, 6)} where each element of the sample space is an ordered pair representing the numbers rolled on the two dice.

Now, let's define some events for this experiment. For example, we could define event A as the event that the sum of the two dice is even:

A = {(1, 1), (1, 3), (1, 5), (2, 2), (2, 4), ..., (6, 2), (6, 4), (6, 6)}

We can also define another event, event B, as the event that the first die rolls a 4

B = {(4, 1), (4, 2), (4, 3), (4, 4), (4, 5), (4, 6)}

4. Probability

Probability is a measure of the likelihood or chance of an event occurring. It is expressed as a number between 0 and 1, with 0 indicating that the event is impossible and 1 indicating that the event is certain to occur.

Let's consider the experiment of flipping a fair coin. The sample space for this experiment is: S = {H, T}

Where "H" represents Heads and "T" represents Tails. Since the coin is fair, the probability of flipping Heads is 0.5 and the probability of flipping Tails is also 0.5

Probability of an event

P(A) = n(A) / n(S)

Where P(A) is the probability of event A, n(A) is the number of outcomes in event A, and n(S) is the total number of outcomes in the sample space.

For example, let's consider the experiment of rolling a fair six-sided die. The sample space for this experiment is:

S = {1, 2, 3, 4, 5, 6}

Let's define event A as the event that the outcome is an even number:

A = {2, 4, 6}

The probability of event A can be calculated as

P(A) = n(A)/n(S) = 3/6 = 0.5

5. Complement

In probability theory and statistics, the complement of an event A is the event that A does not occur. It is denoted by A' or Ac, and consists of all outcomes in the sample space that are not in A.

The probability of the complement of an event A is given by: P(A') = 1 - P(A)

where P(A) is the probability of event A.

Let's consider the experiment of rolling a six-sided die. The sample space for this experiment is:

S = {1, 2, 3, 4, 5, 6}

Let's define event A as the event that the outcome is an even number: A = {2, 4, 6}

The complement of event A, denoted by A', consists of all outcomes in the sample space that are not in A:

A' = {1, 3, 5}

The probability of event A can be calculated as:

P(A) = n(A) / n(S) = 3 / 6 = 0.5

The probability of the complement of event A can be calculated using the formula:

P(A') = 1 - P(A) = 1 - 0.5 = 0.5

6. Union and Interesection

In set theory, Union and Intersection are two fundamental operations that can be performed on sets.

The Union of two sets A and B is the set of all elements that belong to either A or B or both. This can be represented symbolically as A ∪ B.

The Intersection of two sets A and B is the set of all elements that belong to both A and B. This can be represented symbolically as A ∩ B.

7. Conditional prob.

Conditional probability is a fundamental concept in probability theory, which describes the probability of an event A given that another event B has occurred. It is denoted by P(A | B), and it represents the probability of A occurring, given that B has already occurred.

Mathematically, the conditional probability of A given B can be calculated as:

P(A | B) = P(A ∩ B) / P(B)

where P(A ∩ B) represents the probability of both A and B occurring together, and P(B) represents the probability of B occurring.

8. Bayes theorem

Bayes Theorem is a mathematical formula used to calculate conditional probabilities. It is used to find the probability of an event occurring given that another event has already occurred. Bayes theorem is based on the idea of conditional probability, which is the probability of an event happening given that another event has already occurred.

The formula for Bayes Theorem is as follows:

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

where $P(A|B)$ is the probability of A given that B has occurred, $P(B|A)$ is the probability of B given that A has occurred, $P(A)$ is the prior probability of A occurring, and $P(B)$ is the prior probability of B occurring.

9. Random Variable

A random variable is a variable whose value is determined by a random process, such as the outcome of a coin toss or the roll of a die.

In probability theory, random variables are used to model uncertainty and to calculate the probabilities of different outcomes.

It is typically denoted by a capital letter, such as X, and can take on a range of possible values.

Random variables can be classified as discrete or continuous, depending on the possible values they can take on.

A discrete random variable is one that takes on a countable number of possible values, such as the number of heads obtained in a series of coin flips or the number of people in a room. The probability distribution of a discrete random variable is often represented by a probability mass function (PMF), which gives the probability of each possible value.

A continuous random variable is one that takes on an infinite number of possible values within a certain range, such as the height of a person or the time it takes to complete a task. The probability distribution of a continuous random variable is often represented by a probability density function (PDF), which gives the probability of a value falling within a certain range.

10. Probability Distribution

A probability distribution function (PDF) is a function that describes the probability of a random variable taking on a certain value or range of values. It can be used to model the likelihood of different outcomes in a variety of fields, including statistics, physics, and finance.

In statistics, there are many different probability distributions that can be used to model data, such as the normal distribution, binomial distribution, and Poisson distribution. The choice of distribution depends on the nature of the data and the research question being addressed.

PDFS can be used to calculate various statistical properties of the data, such as the mean, variance, and standard deviation. They can also be used to perform hypothesis testing, make predictions about future events, and analyze the probability of different outcomes in a variety of fields.

II, Expected Value

Expected value, also known as the mean or average value, is a measure of central tendency that represents the average outcome of a probability distribution.

Expected value is a concept that is commonly used in probability theory and statistics. It represents the average value of a random variable over many repeated trials. The expected value is calculated by multiplying each possible outcome of the random variable by its corresponding probability and summing the products.

The expected value is a useful tool for decision-making in situations where there is uncertainty. It has many real world examples, few of which are quoted below

Insurance - Insurance orgs use expected values to calculate premiums for insurance policies. They estimate the prob of an event occuring(car accident) and the expected cost of the event. The premium charged is often based on expected cost of the event.

Finance - Investors use expected values to calculate the expected return on an investment. They estimate the prob of different outcomes(stock up or down) and the expected return for each outcome. The expected return is used to evaluate the risk and reward of the investment.

Manufacturing - They use expected values to control the quality of their products. They estimate the prob. of defects occuring during the production and expected cost of each defect. The expected cost is used to determine the acceptable level of defects and to prioritise quality

control measures.

Healthcare - Professionals in this domain use expected value to estimate the effectiveness of treatments. They estimate the prob. of different outcomes(recovery etc) and the expected benefit of each outcome. The benefit in turn, is used to make decisions about treatments.

12. Variance and Std. dev

Variance and Standard deviation are measures of the spread or dispersion of a set of data. They are commonly used in statistics to describe the variability of a dataset.

Variance is a measure of how far a set of data points are spread out from their mean(avg) value. It is calculated by taking the avg. of the sqrd. differences between each data point and mean.

Std. Deviation is the square root of the variance. It is a measure of the spread or dispersion of the dataset. It tells us how much the data deviates from the mean on average.

Use cases of variance and Std. deviation in real world

Finance - Std. deviation is used to measure the risk associated with an investment. The higher the Std. deviation, The riskier is the investment considered to be.

Quality control - variance and Std. deviation are used to measure the variability of a prod. process. A higher variance or Std. deviation indicates that the process is not consistent and needs to be improved.

Health - Std. deviation is used to measure the variability of bio. data such as blood pressure or glucose levels. A high Std. dev may indicate that a patient's health is not stable and needs to be monitored.

13. Joint distribution

Joint distribution is a statistical concept that describes the distribution of two or more random variables together.

It shows the probability of all possible outcomes for multiple variables simultaneously.

They are used in data analysis, ML and other fields.

Joint distribution of two R.VS X and Y can be represented by a PDF(Probability density function) or PMF(Probability mass function), depending upon whether X and Y are continuos or discrete variables.

Joint distribution PDF or PMF is represented by f(x,y) and satisfies the prop that f(x,y) is non -neg for all x and y.

The total prob of all possible outcomes is 1 ie summation(summation(f(x,y))) =1

Use cases of joint distribution in real world are

Medical diagnosis - It is used to estimate the prob of a patient having multiple symptoms or diseases simultaneously. For eg - A doctor can use the joint distribution of blood pressure and cholestrol levels to diagnose a patient with high bp and high cholestrol.

Stock market analysis - In finance, it can be used to model the relationships between diff stocks or assets to estimate the prob. of a particular comb of returns.

Customer segmentation - In marketing, Joint distribution can be used to segment customers based on multiple variables like age, income and purchasing patterns. Analysts can use the joint dist. of these variables to identify diff customer segments and create targeted marketing campaigns.

14. Marginal Probability

It is a statistical concept that describes the probability distribution of a single random variable from a joint probability distribution.

It shows the prob. of one variable independently, without considering other variables.

It is obtained by summing(discrete) or integrating(continuos) the joint prob distribution over all possible values of other variable

Use cases in real world

Medical diagnosis - It is used to estimate the prob. of a patient having a particular symptom or disease independently. For ex - Doctor can use marginal prob of blood pressure to diagnose a patient with high bp without considering other diseases.

Insurance claims - It is used to estimate the probability of a single event occuring such as prob of a car accident or a house fire.

Customer analysis - It is used to analyse customer behaviour by considering the prob of single event, such as prob of customer making a purchase or clicking on an adv.

15. Independence

Independence is a statistical concept that describes the relationship between two random variables. Two random variables X and Y are said to be independent if the occurence of one event does not affect the prob of the occurence of the other.

$P(X=x, Y=y) = P(X=x) * P(Y=y)$

Use cases

Coin toss - The toss of a coin : Prob of getting heads on the first toss doesn't depend on the outcome of the second toss and vice versa.

Medical tests - Results of multiple tests are often used to make a diagnosis. If the tests are independent, The prob of a disease can be calculated as prod of prob of each test being positive.

Weather - In forecasting, The temp and precipitation are two indep. variables. The occurence of rain doesn't depend on temp and vice versa.

Example

Survey to check if their is a relationship between gender and those who like dogs.

Using chi-square test for independence to determine if there is a significant association between gender and liking dogs.

for ex-

Chi-square statistic: 0.0

p-value: 1.0

Degrees of freedom: 1

Expected values: [[1.5 1.5]

[1.5 1.5]]

The p-value is 1.0, which is greater than the significance level of 0.05. This means that we fail to reject the null hypothesis of independence and conclude that there is not enough evidence to suggest that there is a significant association between gender and liking dogs.

16. Conditional independence

It is a concept in probability theory that describes the relationship between two random variables, given the value of a third random variable.

It is a form of independence that occurs when the occurence of one event has no effect on the probability of another event, given the occ. of third event.

Two R.VS X and Y are said to be conditionally independent given a third random variable Z, iff

$P(X,Y|Z) = P(X|Z) * P(Y|Z)$

It is useful concept in many areas, including machine learning where it is often used in bayesian networks to simplify the represetntation and computation of complex prob. models. It can also be used to make inferences about causal relationships between variables.

A causal relationship between variable referes to a relationship in which one variable, known as the cause or independent variable, directly affects other variable, known as the effect or dependent variable. Like amount of fertilizer and plant length growth.

Example

Suppose a dataset of medical records exist that includes information about patient's age, smoking status and whether they have lung cancer or not. We are interested in understanding the relationship between age and lung cancer, and we want to know whether smoking affects this relationship.

We use conditional independence to determine whether age and lung cancer are indep, given smoking status. If they are cond. independent, then smoking status doesn't affect the relationship between age and lung canecer. If they are not, then smoking does affect this relationship.

For same, We calculate the following probab

P(Age,lung_cancer| smoking_status)

P(Age| smoking_status)

P(lung_cancer| smoking_status)

If it satisfies the equation above then, we can say that age and lung cancer are cond. independent given smoking status

17. Law of large numbers

It is a fundamental concept in probability theory that states that as the number of independent, identically distributed(IID) random variable increases, Their sample mean shall converge to their expected value.v

In simpler terms, if one repeatedly flips a fair coin, The law of large numbers tell us that as you keep flipping it more and more times, the proportion of times you get heads will approach 0.5. This is because the prob of getting heads is 0.5 and as one flips the coin more and more times, The sample mean of the proportion of heads shall converge to its true value.

It is an important concept in statistics, finance and other fields that deal with random variables. The law of large numbers helps to explain why we can rely on statistical measures such as

means, variances and other summary statistics as well as predict future outcomes with greater accuracy based on past data.

A use case of law of large numbers is in monte carlo simulations, Where a large number of random samples are generated to estimate the value of complex mathematical functions or to simulate real world scenarios. The law of large number ensures that as we increase the number of simulations, The estimate becomes more accurate.

10. Central Limit Theorem

It is a fundamental concept in statistics that describes the behaviour of the mean of a random sample of independent observations drawn from any distribution, as the sample size becomes sufficiently large.

It states that regardless of the shape of the original population distribution, The distribution of the sample means approaches a normal distribution, as the sample size increases.

This is important because it allows us to make inferences about the population mean using a sample mean, and to use statistical tests that assume normality.

Real world use cases include

Survey sampling - When conducting a survey, It is usually not pract. to collect data from the entire population. Instead, a sample is drawn from the population and statistics are computed based on the sample. The CLT allows us to make inferences about the population mean based on the sample mean.

Quality control - In manufacturing, It is imp. to ensure that products meet certain quality stds. Samples of the products are tested and the mean of the sample is used to estimate the mean of the population. The CLT helps us to determine how many samples are needed to estimate the population mean with a given level of accuracy.

Finance - It is used to analyze stock returns. By assumption, stock returns are normally

distributed, finance analysts can make predictions about future returns and estimate the probability of different outcomes.

9. Hypothesis Testing

It is a statistical method used to determine whether a hypothesis about a population is supported by the data.

It involves two types of hypothesis -

Null hypothesis - It is the default hypothesis, which assumes that theere is no significant difference between two or more groups.

Alternative hypothesis - It is the one we want to test, which assumes that there is a significant difference between two or more groups.

For example - Let's say we want to test whether the avg. height of men and women is significantly different. The null hypothesis would be that there is no significant diff in the avg height between men and women, while the alternative hypothesis would be that there is a significant difference.

Steps to conduct a hypothesis test

State the null and alternate hypotheis

Null hypothesis(H0) - The two groups are not significantly different.

Alternative hypothesis(Ha) - The two groups are significantly different.

Choose a significance level(alpha) to determine the test. Common alpha values are 0.05,0.01 and 0.001.

Calculate data from the sample.

Calculate the test statistics - The test statistic depends on the type of test being performed(t-test, z-test or chi-squared) and the data being analysed.

Determine the p -value - The p-value is the probability of obtaining a test statistic as extreme as the one observed, assuming that the null hypo. is true.

6/ Compare the p-value with chosen significance level - If the p-value is less than or equal to the significance level, we reject the null hypothesis and accept the alternative hypothesis. If the p-value is greater than the significance level, we fail to reject the null hypothesis.

T-test statistic

A t-test statistic is a mesaure used in statistical hypothesis testing to determine if there is a significant difference betweent he mean of two groups. It is used when the pop stand dev is not known and the sample size is rel. small (less than 30)

T-test compares the difference between means of two groups to the variations within the groups. It measures the difference between the sample means divided by the std. error of the diff bet. the means. The resulting value is t-value.

The t-value tells us how many std. error the diff between the two means is away from zero. The larger the t-value, The greater the evidence against the null hypothesis ie the means of the two groups are not sig. different. The t-value is compared to a critical value from a t-distribution with degrees of freedom equal to the sample size minus 2 and the p-value is calculated as prob of obt. a t value as extreme or more extreme than the observed value.

In summary, The t-test statistics is a measure used to det. the significance of the diff bet the means of two groups, taking into account the variation within the groups.

2.0. Maximum likelihood estimation

Maximum likelihood estimate(MLE) is a statistical method that is commonly used to estimate the parameters of a probability distribution by maximising the likelihood function.

The likelihood function measures how likely it is to observe the data given a particular set of parameter values.

For ex - If we havve a set of data points that we believe come from a normal distribution but we don't know the mean and std. deviation of that distribution, we can use MLE to estimate those params

L(mu,sigma_sq) = prod(1 to n)(1/sq.root(2pisigma_sq) *exp(-(x-u)_sq/2_sigma_sq))

21 Confidence interval

Confidence interval

A confidence interval is a range of values that is likely to contain an unknown population parameter with a certain level of confidence. It is often used in statistics to estimate the true value of a population parameter based on a sample from that population.

T-statistic and margin of error

The t-statistic is a measure of how many standard errors a sample mean is away from the true population mean, given a certain sample size and a certain level of confidence. It is calculated as the difference between the sample mean and the hypothesized population mean divided by the standard error of the mean.

The margin of error is a measure of the amount of error that can be expected in an estimate of a population parameter based on a sample from that population. It is calculated as the product of the t-statistic, the standard deviation of the sample, and a factor that depends on the sample size and the level of confidence.

22. Bayesian Inference

Bayesian inference is a method of statistical inference in which the probability of a hypothesis is updated based on new evidence, using Bayes' theorem. It is a way to quantify how confident we are in a hypothesis, given some prior belief about its probability and some observed evidence. Bayesian inference can be used to make predictions, estimate parameters, and test hypotheses.

Steps of Bayesian Inference are as follows

Specify a prior prob. distribution for the hypothesis of interest.

Collect new evidence and update the prior prob. using bayes theorem to obtain a posterior prob. distribution.

Use the posterior prob. distribution to make predictions or estimate params.

In Bayesian inference, the likelihood function plays a crucial role in calculating the posterior distribution. The posterior distribution is proportional to the product of the likelihood function and the prior distribution. In other words, the likelihood function informs us how likely the observed data is, given the values of the parameters, and the prior distribution informs us about our belief or uncertainty about the values of the parameters before observing the data.

Some use cases of Bayesian inference include:

Medical diagnosis: Bayesian inference can be used to estimate the probability of a patient having a certain disease, given some observed symptoms and prior knowledge about the prevalence of the disease.

Customer segmentation: Bayesian inference can be used to segment customers based on their behavior and preferences, and to estimate the probability of a customer belonging to a certain segment.

Forecasting: Bayesian inference can be used to make predictions about future events, such as stock prices or weather patterns, using observed data and prior knowledge about the underlying processes.

23 - Markov chains

Markov Chains are a mathematical framework for modeling systems that transition between different states over time. It is a stochastic process that satisfies the Markov property, which means that the probability of transitioning to the next state depends only on the current state, and not on any past states.

Markov Chains have a wide range of applications, such as in weather forecasting, finance, biology,

and computer science. For example, Markov Chains can be used to model the weather by representing the weather as different states (e.g., sunny, cloudy, rainy), and modeling the transitions between these states based on historical weather data.

Suppose we want to model the weather using a Markov Chain with three states: sunny, cloudy, and rainy. We can represent the transitions between these states using a transition probability matrix.

24. Monte carlo methods

Monte Carlo methods are a class of computational algorithms that rely on repeated random sampling to obtain numerical results. These methods are widely used in various fields such as physics, engineering, finance, and computer graphics.

The basic idea behind Monte Carlo methods is to use random sampling to approximate complex mathematical problems that may not have a closed-form analytical solution. Monte Carlo methods work by generating a large number of random samples from a probability distribution that represents the problem being solved. These samples are then used to estimate the expected value or probability of some outcome.

One of the most common applications of Monte Carlo methods is in the estimation of integrals.

We can use Monte Carlo methods to approximate the area of a unit circle by randomly sampling points within a square that encloses the circle, and then calculating the proportion of points that fall within the circle. The expected value of this proportion is equal to the area of the circle divided by the area of the square, which is equal to $\pi/4$. Therefore, we can estimate the value of π by multiplying the proportion by 4.

25. Stochastic process

Stochastic Processes refer to mathematical models that describe the evolution of random

variables over time or space. These models are used to analyze and predict the behavior of systems that exhibit random behavior or noise. The variables in these models are often referred to as stochastic variables and their behavior is governed by a set of probabilistic rules.

Use cases

Stock Price Modeling: Stochastic processes are often used to model stock prices in finance. The models help investors to make informed decisions about buying or selling stocks based on the estimated future price movements.

Signal Processing: Stochastic processes are used in signal processing to model the noise present in signals. By modeling and analyzing the noise, signal processing techniques can remove noise from signals to improve their quality.

Epidemic/Pandemic prediction: Stochastic processes are used in epidemiology to model the spread of diseases. By analyzing the stochastic behavior of the disease, researchers can predict the likelihood of an epidemic and design effective control strategies.

Suppose we have a system that can be in one of two states, A or B. The system switches between the two states randomly, and the time spent in each state is also random. We can model this system using a Markov chain, where the states represent A and B, and the transition probabilities represent the likelihood of switching between the states.

Transition prob

P(A -> B) = 0.3 P(A -> A) = 0.7 P(B -> A) = 0.5 P(B -> B) = 0.5

26. Conditional prob. distribution

Conditional Probability Distribution is a probability distribution for a random variable when one or more other random variables are given or held constant. It gives the probability distribution of one

variable based on the knowledge of the distribution of another variable.

Use cases -

Machine learning- Conditional probability distribution is used in the context of modeling the probability distribution of a target variable given input features. For instance, in a spam classification problem, we can model the probability of an email being spam given some input features such as the email content, sender information, etc.

Finance - It is used to model the returns on investments.

Example -

Suppose we have a dataset of the heights and weights of a group of people. We want to know the probability distribution of the weight of a person given that their height is greater than 6 feet.

To calculate this, we can use Bayes' theorem which states:

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

27. Probability density function

A probability density function (PDF) is a function that describes the relative likelihood for a random variable to take on a given value. In other words, it provides the probability of a continuous random variable falling within a particular range of values, rather than taking on any one specific value.

The PDF is defined as the derivative of the cumulative distribution function (CDF), which describes the probability that a random variable X is less than or equal to a certain value x.

The PDF is useful in a wide range of applications, including statistical analysis, machine learning, and signal processing. It is commonly used to model the distribution of data in scientfic experiments, to

estimate the likelihood of future events, and to generate synthetic data for simulations and testing.

Example - Analyze a dataset of daily temperatures in a city.

By computing the PDF of the temperature data, we can gain insights into the typical range of temperatures experienced by residents of the city, as well as the likelihood of extreme hot or cold temperatures occurring. This information could be used to guide decisions around infrastructure planning, emergency preparedness, and other areas where temperature is a critical factor.

28. Cumulative distribution function

The Cumulative Distribution Function (CDF) of a random variable is a function that gives the probability that the variable takes a value less than or equal to a given value. In other words, it is the probability distribution of the random variable X.

Use cases

Probability calculations: The CDF can be used to calculate the probability of an event occurring between two values.

Hypothesis testing: The CDF can be used to calculate critical values for hypothesis testing.

Data visualization: The CDF can be used to visualize the distribution of a random variable.

The CDF is defined mathematically as follows:

$F(x) = P(X \leq x)$

29. Moment and Moment Generating function

The moment generating function (MGF) is a tool used in probability theory to find moments of probability distributions. It is defined as the expected value of the exponential of a random variable multiplied by a parameter t.

The moment generating function of a random variable X is given by:

$M_X(t) = E[e^{(tX)}]$

The MGF is used to derive moments of the distribution of the random variable X. Specifically, the nth moment of X can be obtained by taking the nth derivative of the MGF and evaluating it at t=0:

$M_X^n(0) = E[X^n]$

The MGF can be used to find the mean, variance, and other moments of a probability distribution.

One of the important use cases of the moment generating function is that it can be used to determine the distribution of a sum of independent random variables. Specifically, if X and Y are independent random variables with MGFs $M_X(t)$ and $M_Y(t)$, then the MGF of their sum X+Y is the product of their MGFs:

$M_{(X+Y)}(t) = M_X(t) * M_Y(t)$

30. Skewness

Skewness is a measure of the asymmetry of a probability distribution. It is a measure of the degree to which the tails of the distribution differ from the normal distribution. A positive skewness indicates that the distribution has a longer tail on the right side, while a negative skewness indicates that the distribution has a longer tail on the left side.

Skewness is commonly used in statistical analysis to understand the shape of a distribution and to identify any outliers in the dataset. It can be useful in detecting problems with data collection

or data entry.

31. Kurtosis

Kurtosis is a measure of the peakedness and tails of a probability distribution. It is a measure of the degree of outlier influence on the distribution. The kurtosis of a distribution can be positive, zero, or negative.

Positive kurtosis indicates that a distribution has fatter tails and a sharper peak than a normal distribution. Negative kurtosis indicates that a distribution has thinner tails and a flatter peak than a normal distribution.

Zero kurtosis indicates that a distribution has a shape that is approximately equal to a normal distribution.

A high kurtosis value can indicate that a dataset has many outliers, which may skew the analysis. A low kurtosis value can indicate that a dataset has a high concentration of values around the mean.

32. Covariance

Covariance is a statistical measure that indicates the degree to which two random variables are related. Specifically, it measures how much two variables change together. If the covariance is positive, it indicates that the two variables tend to increase or decrease together, whereas if it is negative, it indicates that they tend to move in opposite directions. A covariance of zero indicates that the two variables are not related.

Covariance can be used to determine the relationship between two variables, such as the relationship between the height and weight of a group of people

Suppose we have a dataset containing the height and weight of a group of people. We want to calculate the covariance between these two variables.

33. Cond Independence

Conditional independence refers to a situation where two events A and B are independent of each other given some third event C

One practical example of conditional independence is in medical diagnosis. Suppose we want to diagnose a patient with a certain disease based on a set of symptoms. If we know that the patient has a family history of the disease, then the probability of the patient having the disease may be higher, making the symptoms less informative in the diagnosis. Thus, in this case, the occurrence of the family history event renders the symptom and the disease occurrence independent of each other.

Correlation is close to zero, suggests that height and weight are independent.

35. Stationary distribution

Stationary distribution refers to a probability distribution that remains constant over time, even though the underlying process generating the data may be changing

An example of a stationary process is a random walk, where the distribution of the variable remains the same at each step. Consider a simple random walk model, where a coin is flipped to determine whether the next step is up or down. The probability of moving up or down remains constant, and the distribution of the number of steps in either direction is stationary. In this case, we can calculate the stationary distribution by solving the equation:

$P = pP + qP$

where P is the stationary distribution, p is the probability of moving up, q is the probability of moving down, and $p+q=1$

36. Bootstrap sampling

Bootstrap sampling is a statistical technique used for estimating the sampling distribution of a

statistic. It involves repeatedly sampling from the original dataset with replacement and using those samples to estimate the variablity of the statistic of interest. This technique is particularly useful when the underlying population distribution is unknown or when the sample size is small.

One of the most common use cases of bootstrap sampling is to estimate the confidence interval of a statistic. For example, suppose we have a sample of 100 observations and we want to estimate the 95% confidence interval of the mean. We can use bootstrap sampling to estimate the distribution of the sample mean and then use the percentiles of that distribution to construct the confidence interval.

37. Cross val

Cross-validation is a technique used in machine learning to evaluate the performance of a model by splitting a dataset into subsets, training the model on one subset, and testing it on another subset. The main purpose of cross-validation is to estimate the performance of a model on an independent dataset. Cross-validation is widely used in machine learning, as it allows for a more accurate estimate of a model's performance and can help prevent overfitting.

38. Expectation maximisation algorithm

The Expectation Maximization (EM) algorithm is an iterative method for maximum likelihood estimation in statistical models. It is particularly useful when dealing with incomplete data or hidden variables. The algorithm consists of two main steps: the Expectation (E) step, where we estimate the expected value of the latent variables given the observed data, and the Maximization (M) step, where we maximize the likelihood function with respect to the model parameters given the estimated expected values.

The EM algorithm has many use cases, including clustering, mixture models, latent variable models, and more. For example, it can be used to cluster data into different groups based on similarities in the data, or to estimate parameters in a model that involves hidden variables.

39. Kullback leibler divergence

Kullback-Leibler (KL) divergence is a measure of how different two probability distributions are from each other. It measures the amount of information lost when approximating one probability distribution with another.

$KL(P \parallel Q) = \sum P(x) * \log(P(x) / Q(x))$

where:

$KL(P \parallel Q)$ represents the KL divergence between probability distributions P and Q.

$P(x)$ is the probability of an event x occurring in distribution P.

$Q(x)$ is the probability of the same event x occurring in distribution Q.

$\sum$ denotes the summation symbol, indicating that the formula is a sum over all possible events x.

log is the natural logarithm.

The KL divergence is non-negative, and it is equal to zero if and only if P and Q are identical.

Use cases -

Model selection: KL divergence can be used to compare the performance of different models. For example, if we have two models that predict the same output, we can use KL divergence to measure the difference between the distributions of the predicted values.

Feature selection: KL divergence can be used to measure the information gain of adding a new feature to a model. For example, if we have a classification problem and we want to add a new feature to our model, we can use KL divergence to measure the difference between the class distributions with and without the new feature.

Optimization: KL divergence can be used as a loss function in optimization problems. For example, in some unsupervised learning problems, we want to find a distribution that is similar to the data distribution. We can use KL divergence to measure the difference between the data distribution and the model distribution and minimize it using gradient descent.

40. Mutual information

Mutual information is a measure of the amount of information that two random variables share. It measures how much knowing one variable can tell us about the other variable.

The mutual information between two random variables X and Y is defined as follows:

$I(X; Y) = H(X) - H(X|Y)$

where $H(X)$ is the entropy of X and $H(X|Y)$ is the conditional entropy of X given Y.

One use case of mutual information is feature selection in machine learning. It can be used to identify the most informative features in a dataset. By calculating the mutual information between each feature and the target variable, we can select the features that have the highest mutual information and use them for prediction.

41. SVD

Singular Value Decomposition (SVD) is a matrix factorization method that decomposes a matrix into three matrices: U, Σ, and V, where U and V are orthogonal matrices and Σ is a diagonal matrix containing the singular values of the original matrix.

SVD is commonly used in machine learning for dimensionality reduction, noise reduction, and data compression. It can also be used for image compression, recommendation systems, and natural language processing.

42. Power law

Power law, also known as the Pareto principle or the 80/20 rule, is a phenomenon where a small number of events or individuals have a disproportionately large impact. It is characterized by a heavy-tailed distribution where the frequency of occurrences decreases rapidly as the value increases.

Power law is commonly observed in many real-world phenomena such as the distribution of wealth, the frequency of word usage in language, the popularity of websites, and the number of connections in social networks.

One of the most popular use cases of power law is in the analysis of networks, where it is often used to model the degree distribution of nodes in a network. The degree of a node is defined as the number of edges connecting to it, and the degree distribution is the frequency distribution of nodes with a certain degree.

43. Heavy tailed distribution

Power law, also known as the Pareto principle or the 80/20 rule, is a phenomenon where a small number of events or individuals have a disproportionately large impact. It is characterized by a heavy-tailed distribution where the frequency of occurrences decreases rapidly as the value increases.

Power law is commonly observed in many real-world phenomena such as the distribution of wealth, the frequency of word usage in language, the popularity of websites, and the number of connections in social networks.

One of the most popular use cases of power law is in the analysis of networks, where it is often used to model the degree distribution of nodes in a network. The degree of a node is defined as the number of edges connecting to it, and the degree distribution is the frequency distribution of nodes with a certain degree.

44. Entropy

Entropy is a measure of the uncertainty or randomness of a system or a random variable. In information theory, entropy is used to quantify the amount of information contained in a message or signal.

The entropy of a discrete random variable X with probability mass function $p(x)$ is defined as:

$H(x) = -\sum p(x) \log p(x)$

Use cases include Data compression, Cryptography, Image processing and ML

45. P-value

p-value is a statistical measure used to determine the significance of results obtained in a statistical hypothesis test. It is the probability of obtaining a result as extreme or more extreme than the one observed, assuming the null hypothesis is true. If the p-value is below a certain significance level (usually 0.05), we reject the null hypothesis and conclude that there is a statistically significant difference or relationship between variables.

The use cases of p-value include:

Hypothesis testing: p-value is commonly used to test hypotheses in fields such as biology, psychology, economics, and engineering.

Model selection: p-value can be used to compare the performance of different models and select the one with the best fit to the data.

Feature selection: p-value can be used to identify the most significant features in a dataset and select the ones that are most relevant for a particular task.

Quality control: p-value can be used to monitor the quality of a manufacturing process by testing whether the product meets certain specifications.

Risk assessment: p-value can be used to assess the risk of certain events, such as a disease outbreak or a financial crisis

46. Markov chain Monte carlo

Markov Chain Monte Carlo (MCMC) is a computational method that uses random sampling to

estimate complex probability distributions. It is particularly useful in Bayesian statistics, where it is used to sample from the posterior distribution of a parameter of interest. The basic idea behind MCMC is to use a Markov chain to generate a sequence of samples from the desired distribution, and then to use these samples to estimate its properties.

One common use case for MCMC is in Bayesian inference, where it is used to estimate the posterior distribution of a parameter given some data. For example, suppose we are interested in estimating the proportion of defective items in a production process. We might model this as a binomial distribution with an unknown parameter p, and use MCMC to estimate the posterior distribution of p based on some observed data.

47. Empirical distribution

Empirical distribution is a type of probability distribution that is estimated based on observed data. It is a non-parametric approach to modeling data, which means that it does not assume any specific functional form for the underlying probability distribution. Instead, the empirical distribution function is constructed by plotting the sorted observed data as the horizontal axis and their corresponding probabilities as the vertical axis.

48. Kernel density function

Kernel Density Estimation (KDE) is a non-parametric method for estimating the probability density function of a given dataset. The KDE is a way to estimate the probability density function of a random variable. It estimates the density function using a kernel function and a bandwidth parameter. The kernel function is a non-negative function that integrates to one and is used to smooth the data. The bandwidth parameter controls the amount of smoothing applied to the data. The kernel function is centered at each point in the dataset, and the kernel functions are summed to create the final estimate of the density function.

The KDE is useful in cases where we have a dataset but do not know the underlying distribution. KDE is often used for data visualization, to smooth histograms and to estimate the density of a dataset.

49. Copula

A copula is a mathematical function that describes the dependence between random variables. It provides a way to model the joint distribution of two or more random variables, by specifying the dependence structure separately from the marginal distributions of the variables.

A simple example of a copula is the Gaussian copula, which is based on the assumption that the marginal distributions of the variables are Gaussian and the dependence structure is captured by the correlation matrix.

50. Extreme value theory

Extreme Value Theory (EVT) is a branch of statistics that deals with extreme events or outliers that occur rarely but have a significant impact. EVT is concerned with the statistical modeling of extreme values or tail events, which are rare events that fall outside the range of normal expectations.

51. Brownian motion

Brownian Motion, also known as Wiener process, is a stochastic process used in probability theory and statistics to model random movements or fluctuations of particles in a fluid or gas.

The main characteristic of Brownian Motion is its ability to simulate random and continuous movements with no predictable patterns.

Mathematically, Brownian Motion is a continuous-time stochastic process, usually denoted by $W(t)$, where t represents time, and $W(t)$ denotes the position of the particle at time t. The process is characterized by the following properties:

The increment $W(t2) - W(t1)$ is normally distributed with mean zero and variance $(t2 - t1)$. The increments $W(t2) - W(t1)$ are independent for disjoint time intervals.

52. Information theory

Information theory is a branch of mathematics and computer science that deals with the quantification, storage, and communication of information. It is concerned with the fundamental limits of communication and the amount of information that can be transmitted through a noisy channel. Information theory is used in various fields such as signal processing, data compression, cryptography, and machine learning.

The central concept in information theory is entropy, which measures the amount of uncertainty or randomness in a system. The entropy of a random variable X with probability distribution $P(x)$ is defined as:

$$H(x) = -\sum P(x) \log P(x)$$

One of the main use cases of information theory is data compression. Data compression is the process of reducing the size of a data file by encoding it in a more compact form.

Another use case of information theory is in cryptography. Cryptography is the practice of secure communication in the presence of adversaries.

53. Shannon entropy

Shannon entropy is a measure of the uncertainty or unpredictability of information content. It was introduced by Claude Shannon in 1948 in his paper "A Mathematical Theory of Communication." The Shannon entropy of a message is calculated based on the probability of each possible symbol in the message, and it provides an upper bound on the average length of a lossless compression algorithm that can be used to represent the message.

The Shannon entropy of a discrete random variable X with n possible outcomes and probability distribution $p(x)$ is given by:

$H(x) = -\sum p(x) \log_2 p(x)$

where log2 is the base-2 logarithm.

Use cases

Information theory: Shannon entropy is a fundamental concept in information theory, and it is used to study the properties of communication channels and encoding schemes.

Cryptography: Shannon entropy is used to measure the strength of encryption keys and to generate random numbers.

Machine learning: Shannon entropy can be used as a measure of the purity of a node in a decision tree or the amount of information gained by splitting a node.

Example

Suppose we have a sequence of symbols: "ABACADABRA". We can calculate the Shannon entropy of this sequence as follows:

Count the number of occurrences of each symbol: A: 5 B: 2 C: 1 D: 1 R: 2

Calculate the probability of each symbol by dividing the count by the total number of symbols: P(A) = 5/11 = 0.45 P(B) = 2/11 = 0.18 P(C) = 1/11 = 0.09 P(D) = 1/11 = 0.09 P(R) = 2/11 = 0.18

Calculate the Shannon entropy as the sum of the products of each symbol's probability and its logarithm (base 2): $H(x) = -(0.45\log_2(0.45) + 0.18\log_2(0.18) + 0.09\log_2(0.09) + 0.09\log_2(0.09) + 0.18*\log_2(0.18)) = 2.334$

Therefore, the Shannon entropy of the sequence "ABACADABRA" is 2.334 bits per symbol. This means that on average, it takes at least 2.334 bits to represent each symbol in the sequence using a lossless compression algorithm.

54. Bias Variance Tradeoff

The bias-variance tradeoff is a fundamental concept in machine learning that describes the relationship between the complexity of a model and its ability to generalize to new data. It refers to the tradeoff between the model's ability to fit the training data (bias) and its ability to generalize to new data (variance).

Bias refers to the error that is introduced by approximating a real-life problem with a simplified model. A high bias model is one that is too simple to capture the underlying patterns in the data and tends to underfit. On the other hand, variance refers to the error that is introduced by the model's sensitivity to small fluctuations in the training data. A high variance model is one that is too complex and tends to overfit the training data.

55. T-Statistic margin error

t-statistic is a measure of the difference between two sample means that determines whether they are statistically significant or not. It is commonly used in hypothesis testing to determine whether the difference between two groups is meaningful or just due to random chance.

The t-statistic takes into account the size of the difference between two sample means, the variance of each group, and the size of the sample. In general, a larger t-statistic indicates a more significant difference between the two groups.

The margin of error in t-statistic refers to the range of values within which the true population mean is likely to fall, based on the sample data. The margin of error is affected by the size of the sample, the level of confidence desired, and the standard deviation of the population.

A larger margin of error indicates that the sample mean is less precise and more likely to be further from the true population mean. On the other hand, a smaller margin of error indicates that the sample mean is more precise and closer to the true population mean.

56. Binomial Distribution

In probability theory and statistics, a binomial distribution is a probability distribution that describes the number of successes in a fixed number of independent trials, where each trial can have one of two possible outcomes (usually referred to as "success" or "failure"), and the probability of success is constant throughout the trials.

The binomial distribution is characterized by two parameters: n, the number of trials, and p, the probability of success on each trial. The probability of obtaining exactly k successes in n trials is given by the binomial probability mass function:

$P(k;n,p) = (n \text{ choose } k)\, p^k\, (1-p)^{(n-k)}$

where "n choose k" represents the number of ways to choose k items from a set of n items, and can be computed as $n!/(k!(n-k)!)$.

it can be used to model the probability of a certain number of defective items in a batch of products, or the probability of a certain number of mutations occurring in a population of organisms.