**Table of Contents**

Statistics

1. Central Limit Theorem
2. Law of Large Numbers
3. Assumption of Normality
4. Normal Distribution
5. Symmetric Distribution
6. Z-score
7. Empirical Rule
8. Hypothesis Testing
9. P-Value
10. Confidence Interval
11. One sample t-test
12. Significance level
13. Point Estimates
14. Standard Deviation
15. Expectation
16. Confidence Level
17. Outliers & Inliers
18. Quartile
19. 5-number summary
20. Skewness
21. Kurtosis
22. Design of Experiments
23. Heteroscedastic Model
24. Autocorrelation
25. Correlation and Covariance
26. Confounding Variable
27. Statistical Interaction
28. Sensitivity
29. Power of Test
30. Cherry Picking
31. P-hacking
32. Significance chasing
33. Confidence test
34. Hypothesis test
35. Residual Analysis
36. Survival Analysis
37. MCMC
38. Causal Inference
39. Meta Analysis
40. Longitudinal Data Analysis
41. Fractal Geometry
42. Chaos Theory
43. Spatial Statistics
44. Cognitive Bias
45. Predictive Analysis
46. Item Response Theory
47. Inferential Statistics

**Statistics**

****

1. **Central limit theorem**

The normal distribution, also known as the bell curve, is a statistical concept used in hypothesis testing and the calculation of confidence intervals. It arises when sample sizes vary without affecting the shape of the population distribution.

To illustrate this, let's consider calculating the average height of people worldwide. Instead of measuring every individual, we take random samples from the population and calculate their mean. By repeating this process multiple times and plotting the means and their frequencies, we can observe a bell-shaped curve closely resembling the original data set.

The Central Limit Theorem states that if an experiment is conducted repeatedly and independently, the average of the results tends to approach the expected value. For this theorem to hold, certain conditions must be met, such as random sampling and independence among the sample values. Additionally, a sample size of at least 30 is typically recommended.

Example, we want to calculate the average height of people in the world, and we take some samples from the general population. Since it is hard or impossible to obtain data regarding the height of every person in the world, we will simply calculate the mean of our sample. By multiplying it several times, we will obtain the mean and their frequencies which we can plot on the graph and create a normal distribution. It will form a bell-shaped curve that will closely resemble the original data set.

<https://github.com/manishmawatwal/statistics_book/blob/main/central_limit_theorem.ipynb>

1. **Law of Large Numbers**

A theorem that describes the result of performing the same experiment very frequently. It states that the sample mean, sample variance, and standard deviation converge to what we are trying to estimate. If an experiment is repeated independently a large number of times, the average of the individual results is close to the expected value. There are two forms of the law of large numbers, the weak law of large numbers states that as ‘n’ increases, the sample statistic of the sequence converges in probability to the population value.   
The strong law of large numbers describes how a sample statistic converges on the population value as the sample size or the number of trials increases. For example, the sample mean will converge on the population mean as the sample size increases.

An increase in the number of trials in an experiment will result in a positive and proportional increase in the results coming closer to the expected value. As an example, let us check the probability of rolling a six-sided dice three times. The expected value obtained is far from the average value. And if we roll a dice a large number of times, we will obtain the average result closer to the expected value (which is 3.5 in this case).

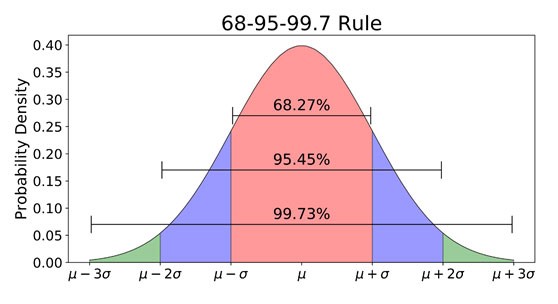
In a financial context, the law of large numbers indicates that a large entity which is growing rapidly cannot maintain that growth pace forever. In business, it suggests that, as a business expands, the percentage rate of growth becomes increasingly difficult to maintain.

<https://github.com/manishmawatwal/statistics_book/blob/main/law_of_large_numbers.ipynb>

1. **Normal Distribution & Empirical Rule & Assumption of Normality**

The normal distribution, also known as the Gaussian distribution or bell-curve distribution, is a statistical distribution commonly observed when data is distributed around a central value, where the mean is equal to the median. This type of distribution is characterized by its lack of bias and follows the empirical rule. It exhibits perfect symmetry, meaning the left and right halves are mirror images of each other.

Empirical rule states that every piece of data in a normal distribution lies within three standard deviations of the mean. It is also known as the 68-95-99.7 rule / three-sigma rule. 68% of values will fall within first standard deviation, 95% will fall within two standard deviations, and 99.75% will fall within three standard deviations (µ ± 3σ).

Examples of phenomena that can be modelled by a normal distribution include the time until the next earthquake. Key characteristics of the normal distribution include its symmetrical nature, unimodal shape (having only one peak or mode), and the fact that the mean, mode, and median all reside at the centre. It is often used to describe central tendency and has asymptotic properties. It is important to note that exponential distributions do not conform to the characteristics of a log-normal or Gaussian distribution. Additionally, qualitative data is not typically expected to exhibit this type of distribution.

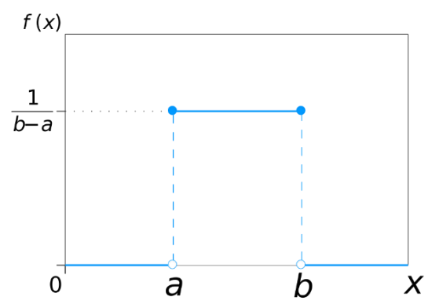
The assumption of normality is a fundamental assumption in many statistical analyses. It assumes that the data being analysed follows a normal distribution or approximately follows a bell-shaped curve. This assumption is important because many statistical tests and models are based on the properties of the normal distribution.

When we assume normality, we expect that the data is symmetrically distributed around the mean, with the majority of the values falling near the centre and fewer values in the tails. It implies that the mean, median, and mode of the data are all approximately equal.

The assumption of normality is particularly relevant in parametric statistical tests, such as t-tests, analysis of variance (ANOVA), and linear regression. These tests rely on the assumption of normality to ensure accurate estimation of parameters and valid hypothesis testing.

1. **Symmetric Distribution**

A symmetric distribution refers to a probability distribution in which the values are evenly distributed around a central point or axis. In other words, if you were to draw a line through the middle of the distribution, the left and right sides would be mirroring images of each other.

In a symmetric distribution, the mean, median, and mode of the data are all equal, and they are located at the centre of the distribution. This means that the distribution has a balanced shape, with values equally likely to occur on both sides of the centre.

Examples of symmetric distributions include the normal distribution (bell-shaped curve), uniform distribution, and triangular distribution with a symmetric shape.

1. **Z-score**

A Z-score is a numerical measurement that describes the relationship of a particular value to the mean of a group of values. It is calculated in terms of standard deviations from the mean.

When the Z-score is 0, it signifies that the data point's score is identical to the mean score. If the Z-score is 1.0, it indicates that the value is one standard deviation away from the mean. Z-scores can be positive or negative, with a positive value suggesting that the score is above the mean and a negative value indicating it is below the mean.

In finance, Z-scores are used as measures of an observation's variability and can be valuable for traders in assessing market volatility. Specifically, in financial analysis, a Z-score below 1.8 suggests that a company may be at risk of facing bankruptcy, while a Z-score closer to 3 implies that a company is in a strong financial position. The Z-score is also employed for standardization purposes, allowing for the scaling down of different features using techniques like Standard Scaler. By using Z-scores, it becomes possible to compare scores between different distributions, facilitating meaningful comparisons and analysis across diverse datasets.

<https://github.com/manishmawatwal/statistics_book/blob/main/z_score.ipynb>

1. **Hypothesis Testing**

Hypothesis testing is a crucial tool for evaluating the meaningfulness of experimental results. It enables us to assess the statistical significance of insights by determining the likelihood of the observed results occurring by chance. In this process, we compare the obtained data with a null hypothesis (H0), which represents the assumption of no effect or no difference. The alternative hypothesis (H1) contradicts the null hypothesis and represents the statement that must be true if the null hypothesis is false.

1. **P-value**

The p-value plays a key role in hypothesis testing, as it quantifies the probability of obtaining results as extreme as the ones observed under the specific hypothesis, assuming the null hypothesis is true. It ranges from 0 to 1, with a lower p-value indicating stronger evidence against the null hypothesis. Typically, a significance level of 0.05 is used as a threshold, where p-values below this threshold suggest rejecting the null hypothesis in favour of the alternative hypothesis. On the other hand, higher p-values indicate that the data is consistent with the null hypothesis. Ultimately, the p-value provides valuable insights into the strength of the evidence and guides decision-making in hypothesis testing scenarios.

With any research study, there is a possibility that the observed differences are a chance event. The only way to know that a difference is really present with certainty, the entire population would need to be studied. The research community and statisticians have to pick a level of uncertainty to which they can agree. This level of uncertainty is called Type 1 error or a false-positive rate (α). This is more commonly called a p-value.

In general, p≤ 0.05 is the agreed upon level.

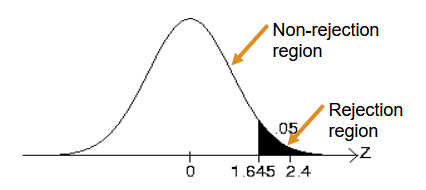
If p-value is less than 1%, there is overwhelming evidence that supports the alternative hypothesis.

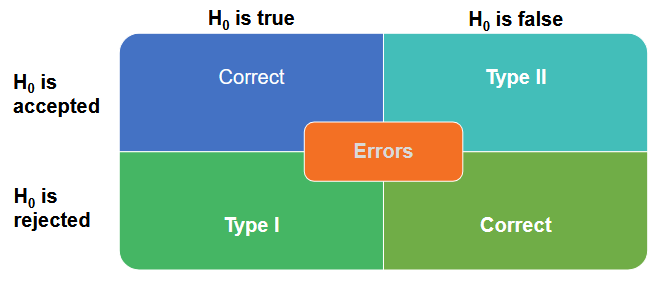
If p-value is between 1% and 5%, there is strong evidence that supports the alternative hypothesis.

If p-value exceeds 10%, there is no evidence that supports the alternative hypothesis.

1. **Hypothesis**

A statement on the parameters which is yet to be proved or established.





Types of hypothesis testing

1. **Confidence interval**

A confidence interval represents the probability that a population parameter will fall within a specific range of values for a certain proportion of times. It is a valuable tool utilized in various statistical analyses, such as hypothesis testing and regression analysis. The confidence interval provides a range of values that is likely to contain the true population parameter.

The confidence interval is determined by the confidence coefficient or confidence level, denoted by . This coefficient indicates the probability or likelihood associated with the interval. For example, a confidence level of 95% corresponds to a confidence coefficient of 0.95, meaning that there is a 95% probability that the true population parameter lies within the calculated interval.

In hypothesis testing, the confidence interval allows researchers to assess the precision and reliability of their estimates. It provides a range within which the population parameter is expected to lie, based on the observed sample data. Similarly, in regression analysis, confidence intervals are employed to gauge the uncertainty associated with the estimated regression coefficients.

The confidence level, often referred to as the level of significance, is denoted by α. It represents the probability of making a Type I error, which is rejecting a true null hypothesis. Typically, a significance level of 0.05 (or 5%) is commonly used, indicating a 5% chance of making a Type I error.

1. **One sample t-test**

In a one-sample t-test, we compare the mean of a sample to a known or hypothesized population mean to determine if there is a significant difference between them. It allows us to assess if there is substantial evidence to support the claim that the sample mean is statistically different from the population mean.

For example, let's consider a scenario where we measure the grams of protein in a sample of energy bars. The labels on these bars claim that they contain 20 grams of protein each. To verify the accuracy of these claims, we can conduct a hypothesis test to determine if the mean protein content of the sampled bars is significantly different from the stated 20 grams. This test will provide us with insights into whether the labels are indeed correct or if there is a significant deviation in the actual protein content.

1. **Significance level and Confidence level**

The significance level is a measure of the probability of obtaining a result that is significantly different from what would be expected under the assumption that the null hypothesis is true. It helps determine the threshold for rejecting the null hypothesis in hypothesis testing. A lower significance level indicates a stricter criterion for rejecting the null hypothesis, as it requires stronger evidence to conclude that the observed results are not due to chance.

On the other hand, the confidence level is a measure of the certainty or reliability with which we can estimate a population parameter based on a sample. It represents the range of values within which the true population parameter is likely to fall. A higher confidence level indicates a greater degree of confidence in the accuracy of the estimated parameter.

The significance level and confidence level are related to each other through a simple formula.

In other words, if we have a confidence level of 95%, the corresponding significance level would be 1 - 0.95 = 0.05. This means that if the null hypothesis is true, there is a 5% probability of obtaining a result that deviates significantly from what is expected.

<https://github.com/manishmawatwal/statistics_book/blob/main/significance_level_confidence_level.ipynb>

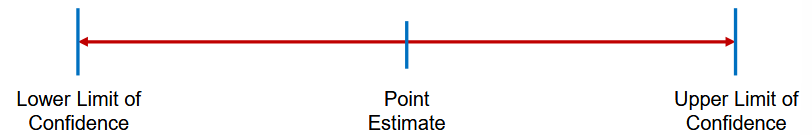
1. **Point Estimates**

Point estimates is a single numeric value that is calculated from sample data to estimate an unknown parameter of a population. These estimates provide an approximation or best guess of the true value of the parameter based on the available data.

For example, if we want to estimate the population mean, we can calculate the sample mean as a point estimate. Similarly, if we want to estimate the population proportion, we can calculate the sample proportion as a point estimate. Other common point estimates include the sample median, sample variance, or sample standard deviation.

It's important to note that point estimates are subject to sampling variability, meaning different samples from the same population may yield different point estimates. Therefore, it's common practice to provide a measure of uncertainty or variability associated with the point estimate, typically in the form of a confidence interval or standard error. These interval estimates or measures of variability provide a range of plausible values for the population parameter.

Whereas Interval estimate gives a measure of certainty, and gives two bounds of the confidence area.



1. **Standard Deviation**

Represents the magnitude of how far the data points are from the mean. A high value indicates that the data is spread to extreme ends, far away from the mean.

1. **Expectation**

Summation or integration of a possible value from a random variable. For a set of ‘n’ events which are mutually exclusive and exhaustive, where for event ‘i’ the expected value is Ei given probability Pi.

E(X)pectation of a Discrete Random Variable

E(X)pectation of a Continuous Random Variable

<https://github.com/manishmawatwal/statistics_book/blob/main/expectation_discrete_continuous.ipynb>

1. **Outliers and Inliers**

Outliers are data points that significantly deviate from the pattern observed in the rest of the dataset. These extreme values can have a detrimental impact on the accuracy and efficiency of a statistical model. When dealing with datasets that contain numerous outliers, it is often preferred to use the median as a measure of central tendency, as it is less influenced by extreme values.

There are two common methods for identifying outliers: the standard deviation or z-score approach, and the interquartile range (IQR). The standard deviation method involves calculating the z-score of each data point, which represents the number of standard deviations it is away from the mean. Data points with z-scores beyond a certain threshold are flagged as outliers. The IQR method involves computing the range between the first quartile (Q1) and the third quartile (Q3), and any values outside this range are considered outliers.

While outliers are typically seen as problematic and are often removed from the dataset, there are certain scenarios where they may be retained. For instance, when analysing critical results or when outliers provide valuable insights into the data, they may be kept. Additionally, in datasets with high skewness, outliers can help provide a more accurate understanding of the underlying distribution.

In contrast to outliers, inliers are data points that exhibit similar characteristics to the rest of the dataset. Identifying inliers can be challenging as it requires external data or a reference point for comparison. Inliers, similar to outliers, can impact model accuracy negatively, and therefore they are typically removed from the dataset as they are considered errors or anomalies.

1. **Quartile**

Quartiles are used to describe the distribution of data by splitting data into three equal portions, and the boundary or edge of these portions are called quartiles. The lower quartile (Q1) is the 25th percentile, the middle quartile (Q2) (also called median) is the 50th percentile, the upper quartile (Q3) is the 75th percentile.

1. **5-number summary**

It is a measure of five entities that cover the entire range of data. Low extreme (Min), First quartile (Q1), Median, Upper quartile (Q3), High extreme (Max).

<https://github.com/manishmawatwal/statistics_book/blob/main/5_point_summary.ipynb>

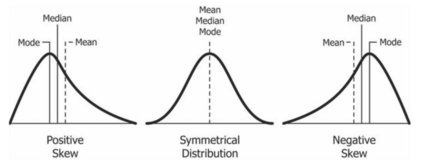
1. **Skewness**

Skewness measures the lack of symmetry in a data distribution. It indicates that there are significant differences between the mean, the mode, and the median of data. Skewed data cannot be used to create a normal distribution. A distribution can exhibit skewness if the tail is longer on one side.

Right skewed = Positive skewed

Left skewed = Negative skewed

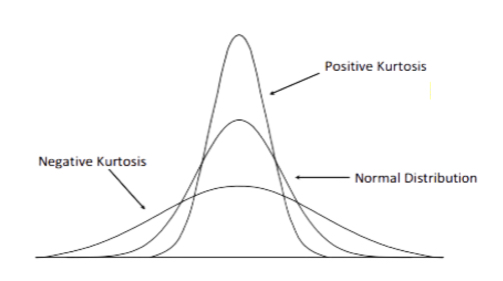
Left-skewed distribution: The distribution of the age of deaths in most populations. Most people live to be between 70 and 80 years old, with fewer and fewer living less than this age. Right-skewed distribution: The distribution of household incomes/wealth distribution.



To transform a Skewed Distribution into a Normal Distribution we apply some linearized function on it. Some common functions that achieve this goal are:

* Logarithmic function: for right-skewed distributions. The only condition is that this function is defined only for strictly positive numbers. f(x) = ln(x)
* Square root transformation: this one has an average effect on distribution shape, it’s weaker than logarithmic transformation, and it’s also used for reducing right-skewed distributions, but is defined only for positive numbers.
* Reciprocal transformation: this one reverses the order among values of the same sign, so large values become smaller, but the negative reciprocal preserves the order among values of the same sign. The only condition is that this function is not defined for zero values.

1. **Kurtosis**

The extreme values present in one tail of distribution or the peaks of frequency distribution versus the other. It is the measure of outliers present in the distribution. A high value of kurtosis represents large amount of outliers in the data. To overcome this, we have to either add more data into the dataset or remove the outliers. The standard normal distribution has a kurtosis of 3 whereas the values of symmetry and kurtosis between -2 and +2 are considered normal and acceptable.

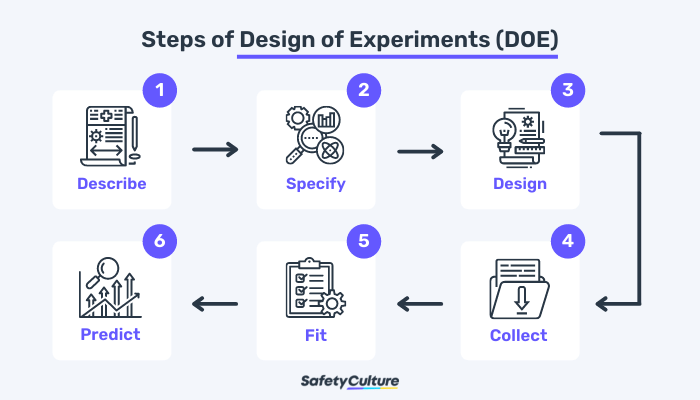
1. **Design of Experiments**

Design of Experiments (DOE) is a statistical approach used to systematically plan, conduct, and analyze experiments in order to understand and optimize processes, products, or systems. It involves carefully selecting and manipulating variables to uncover cause-and-effect relationships and make informed decisions.

For example, let's consider a company that manufactures a new type of light bulb. They want to determine the optimal combination of factors such as bulb material, wattage, and manufacturing process to maximize the bulb's lifespan. In this case, DOE would involve designing a series of experiments where different combinations of these factors are tested and their effects on the bulb's lifespan are measured.

The company would carefully control and vary the levels of each factor while keeping other variables constant, creating different experimental conditions. By comparing the results of these experiments, they can identify which factors have a significant impact on the bulb's lifespan and how different levels of these factors interact with each other.

DOE helps researchers avoid biased conclusions by systematically considering various factors simultaneously and quantifying their effects. It enables them to make data-driven decisions on how to optimize processes, improve product performance, or reduce costs. By efficiently designing and conducting experiments, DOE enables researchers to gain valuable insights and achieve desired outcomes more effectively.



1. **Heteroscedastic Model**

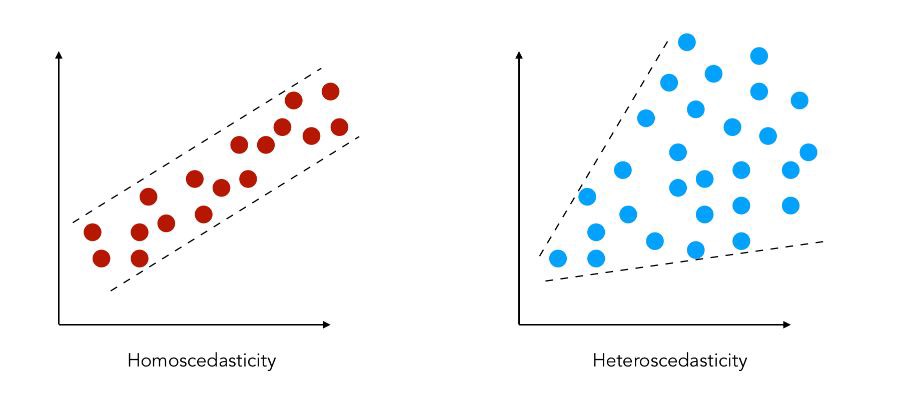
A model whose variation in errors comes out to be inconsistent. It often occurs in two forms – conditional and unconditional.

Heteroscedasticity means unequal scattered distribution. Heteroscedasticity is the systematic change in the spread of the residuals or errors over the range of measured values.

It occurs often in datasets, where we have large range between the largest and the smallest observed values. There are two types of heteroscedasticities

Pure heteroscedasticity – It refers to cases where we specify the correct model and let us observe the non-constant variance in residual plots.

Impure heteroscedasticity – It refers to cases where you incorrectly specify the model, and that causes the non-constant variance.



1. **Autocorrelation**

A representation of the degree of correlation between the two variables in a given time series and a lagged version of itself over successive time intervals. It’s conceptually similar to the correlation between two different time series, but autocorrelation uses the same time series twice: once in its original form and once lagged one or more time periods. The data is correlated in a way that future outcomes are linked to past outcomes. Autocorrelation makes a model less accurate because even errors follow a sequential pattern.

An autocorrelation of +1 represents a perfect positive correlation, while an autocorrelation of -1 represents a perfect negative correlation. Autocorrelation is also known as lagged correlation/serial correlation.

Autocorrelation can be used to determine if a momentum trading strategy makes sense.

For example, if it's rainy today, the data suggests that it's more likely to rain tomorrow than if it's clear today. When it comes to investing, a stock might have a strong positive autocorrelation of returns, suggesting that if it's "up" today, it's more likely to be up tomorrow, too.

1. **Correlation and Covariance**

Correlation is a technique to measure and estimate the quantitative relationship between two variable and is measured in terms of how strong are the variables related. It is dimensionless. The value of correlation between two variables ranges from -1 to +1. The value -1 represents high negative correlation, i.e., if the value in one variable increases, then the value in the other variable will drastically decrease. Similarly, +1 means a positive correlation, an increase in one variable will lead to an increase in the other. Whereas, 0 means there is no correlation.

If two variables are strongly correlated, then they may have a negative impact on the statistical model, and one of them must be dropped.

There are mainly three types of correlation:

* Pearson - Normalized measurement of covariance, assumes both the variables are normally distributed, Measures linear relationship but fail to measure the non-linear relationship between variables.
* Spearman Rank – It is a non-parametric measure, measures both linear and non-linear relationship between two variables.
* Kendall Rank- Non parametric measure for calculating the rank of the correlation coefficient, measures both linear and non-linear relationship between two variables.

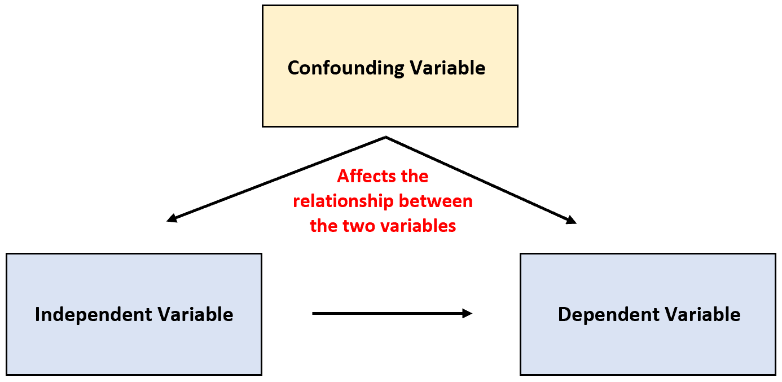
Covariance is a systematic relationship between pair of variables where changes in one affect changes in another variable. The systematic relation is determined between a pair of random variables to see if the change in one will affect the other variable in the pair or not.

Covariance dimension depends on variables. Mathematically, consider 2 random variables, X and Y where the means are represented as and respectively and standard deviations are represented by and respectively and E represents the expected value operator, then:

<https://github.com/manishmawatwal/statistics_book/blob/main/correlation_covariance.ipynb>

1. **Confounding Variable**

A variable that is associated with both the dependent variable and the independent variable, and it can give a wrong estimate that provide useless results.

For example, if we are studying the effect of weight gain, then lack of workout will be the independent variable, and weight gain will be the dependent variable. In this case, the amount of food consumption can be the confounding variable as it will mask or distort the effect of other variables in the study. The effect of weather is another confounding variable.

1. **Statistical Interaction**

When two or more variable interact, which affects a third variable. A real-life example includes the interaction of adding sugar to the stirring of tea. Neither of the two variables has an impact on sweetness, but it is the combination of these two variables that do.

1. **Sensitivity**

Sensitivity refers to the ability of a statistical test or model to correctly identify positive cases or true positives. It is a measure of how well a test or model can detect the presence of a condition or event of interest.

A high sensitivity indicates that the test or model is effective in identifying the condition or event when it is truly present. It means that there are fewer false negatives, which are cases that are mistakenly classified as negative when they should be positive. On the other hand, a low sensitivity implies a higher rate of false negatives, indicating that the test or model may miss a significant number of positive cases.

Sensitivity is particularly important in situations where the consequences of missing positive cases are significant, such as in medical diagnostic tests. For example, in a medical screening test for a disease, a high sensitivity ensures that a greater proportion of individuals with the disease are correctly identified, enabling timely intervention and treatment. Sensitivity is also known as the true positive rate or the recall.

1. **Power of Test**

Probability of rejecting the null hypothesis when it’s false.

To increase the power of the test

* You can increase alpha, but it also increases the chance of type 1 error.
* Increase the sample size, n. This maintains the type 1 error but reduces type 2.

Type 1 error is a false positive conclusion. Type 2 error is a false negative conclusion.

1. **Cherry picking**

Practise of selecting information which supports a certain claim and ignoring any other claim that refutes the desired conclusion. It does not necessarily mean that one side’s information is incorrect, but that a complete picture is not being presented because available evidence to support the other side is not being taken into account.

Cherry picking data happens in politics. For example, in June, 2020, President Trump claimed that the US was doing well in the battle against COVID-19 because the death rate from the disease was declining. This statement totally ignored the information that new records were being set every day for people testing positive for the SARS-CoV-2 virus.

What we have here is a typical case of 'cherry-picking' of data. It isn’t that Trump’s information is incorrect, it just doesn’t present a complete picture of the situation. And politicians are not the only ones guilty of this type of data reporting. Environmentalists, industry representatives, activists of all sorts, and government officials are all in on the cherry-picking harvest.

<https://github.com/manishmawatwal/statistics_book/blob/main/cherry_picking.ipynb>

1. **Significance chasing**

Significance chasing, also known as data dredging or p-hacking or Data Snooping or Data Fishing, refers to the practice of selectively analysing data or conducting multiple statistical tests until a statistically significant result is found, without a strong theoretical or practical justification. It is a problematic behaviour that can lead to false or misleading conclusions and is considered an unethical practice in research.

One example of significance chasing is when a researcher conducts multiple analyses on a dataset, testing various hypotheses and variables, and selectively reports only the statistically significant results while ignoring the non-significant ones. For instance, imagine a study investigating the effects of different variables on academic performance. The researcher might test multiple variables, such as study time, sleep quality, and diet, on a sample of students. By running statistical tests on each variable separately and reporting only the ones that yield significant results, the researcher may inadvertently create a false impression that these variables have a significant impact on academic performance.

Another example is the selective reporting of outcomes within a study. Researchers might conduct various measurements or analyses within a study and choose to report only the outcomes that show significant effects. This cherry-picking of results can lead to a biased interpretation of the data. For instance, consider a clinical trial evaluating the efficacy of a new drug. If the researchers selectively report only the positive outcomes, such as improvements in symptoms, while disregarding any negative or non-significant findings, it can create an inflated perception of the drug's effectiveness.

Significance chasing can also occur in the context of exploring subgroup analyses. Researchers may test multiple subgroups within a study population, searching for significant effects in some subgroups while disregarding others. This can lead to spurious findings and unsubstantiated claims about differential effects. For example, in a study examining the effectiveness of a new therapy for a medical condition, the researchers might conduct subgroup analyses based on age, gender, or other demographic factors. If they selectively report only the subgroups that show significant effects, it can create a false impression that the therapy is more effective for specific groups when, in fact, the results may be due to chance.

In all these examples, significance chasing involves selectively analysing and reporting results that appear statistically significant, while neglecting or downplaying non-significant findings. It is important to recognize that statistical significance alone should not drive the interpretation of research findings. Transparency, rigor, and adherence to sound statistical practices are crucial to ensure the reliability and validity of research results.

1. **Confidence test**

Confidence interval provides a range of values that helps in capturing the unknown parameter. Confidence interval are important in medical research to provide researchers with a strong bias for their estimations.

1. **Hypothesis Test**

Hypothesis testing is used to test an experiment or observation and determine if the results did not occur purely by chance or luck using the below formula where ‘p’ is some parameter.

1. **Residual Analysis**

Residual analysis is a statistical technique used to assess the quality of a regression or predictive model. It involves calculating the differences (residuals) between observed and predicted values. By creating a residual plot, patterns and outliers can be identified, indicating potential problems with the model. Residual analysis helps check assumptions of linearity, constant variance, and independence of errors. A random scatter of residuals around zero is desired, indicating a well-fitted model. It provides insights into the performance and validity of the model's predictions, guiding improvements if needed.

1. **Survival Analysis**

Survival analysis is a statistical method used to analyse time-to-event data, often in medical or social sciences. It focuses on studying the time until an event of interest occurs, such as death, disease recurrence, or failure. Survival analysis considers censoring, where some individuals may not experience the event during the study period. The Kaplan-Meier estimator is commonly used to estimate the survival function. Cox proportional hazards model is a popular regression technique in survival analysis. It helps assess the impact of covariates on survival outcomes. Survival analysis is useful in understanding and predicting the duration until an event happens, providing valuable insights for decision-making and risk assessment.

1. **Markov Chain Monte Carlo (MCMC) Methods**

Markov Chain Monte Carlo (MCMC) methods are a class of algorithms used to approximate complex probability distributions. They involve constructing a Markov chain that samples from the target distribution by iteratively transitioning between states according to certain transition probabilities. MCMC methods are particularly useful when direct sampling from the target distribution is difficult or infeasible. One commonly used MCMC algorithm is the Metropolis-Hastings algorithm, which generates candidate samples and accepts or rejects them based on a certain acceptance criterion. MCMC methods are widely applied in Bayesian statistics, where they allow for posterior inference and estimation of unknown parameters. By leveraging Markov chains, MCMC methods provide a powerful tool for exploring high-dimensional probability spaces and making probabilistic inferences.

1. **Causal Inference**

Causal inference is a field of study that aims to understand cause-and-effect relationships between variables. It involves determining whether an observed association between variables is due to a causal relationship or simply a correlation. Causal inference methods help us make causal claims by considering factors such as confounding variables, counterfactuals, and causal mechanisms. Techniques like randomized controlled trials, propensity score matching, and instrumental variable analysis are commonly used in causal inference. By establishing causality, we can gain insights into how interventions or changes in one variable affect another, enabling us to make informed decisions and predictions based on causal relationships rather than mere associations.

1. **Meta-Analysis**

Meta-analysis is a statistical technique used to combine and analyse data from multiple independent studies on a particular research question or topic. It involves synthesizing the results of these studies to generate a more precise and reliable estimate of the effect or relationship being investigated. By pooling data from various studies, meta-analysis can provide a larger sample size, increased statistical power, and a more comprehensive overview of the research area. It involves systematic literature review, data extraction, and statistical analysis to quantify the overall effect size and assess the consistency or heterogeneity of findings across studies. Meta-analysis is valuable for summarizing existing evidence, identifying patterns or trends, and drawing robust conclusions that can inform decision-making and guide future research in a given field.

1. **Longitudinal Data Analysis**

Longitudinal data analysis refers to the statistical methods and techniques used to analyze data collected over multiple time points from the same individuals or subjects. It focuses on understanding and modeling changes or trends in variables of interest over time. It involves accounting for the correlations and dependencies among repeated measurements on the same subjects and explores how these variables evolve and interact over the course of the study. Longitudinal data analysis techniques include growth curve models, mixed-effects models, generalized estimating equations (GEE), and random-effects models. It allows researchers to investigate individual trajectories, examine the effects of time-varying factors, assess the impact of interventions or treatments, and provide insights into the dynamic nature of the studied phenomenon.

1. **Fractal Geometry**

Fractal geometry is a fascinating area of mathematics that explores the beauty and complexity of irregular shapes found in nature and the world around us. Unlike traditional geometric shapes, like circles or squares, fractals exhibit intricate patterns and structures that repeat at different scales. Imagine zooming in on a fractal shape and discovering that it contains smaller versions of itself, each resembling the overall pattern. This self-similarity is a key feature of fractals. They can be generated using simple mathematical equations or algorithms that iteratively apply transformations. One famous example is the Mandelbrot set, which creates stunningly intricate and infinitely detailed fractal images. Fractal geometry finds applications in various fields, including computer graphics, digital image compression, and understanding complex natural phenomena like coastlines, clouds, and branching structures. Exploring fractals allows us to appreciate the endless beauty and complexity that can emerge from simple mathematical principles.

1. **Chaos Theory**

Chaos theory is a branch of mathematics that studies complex and unpredictable behavior in dynamic systems. It explores how small changes in initial conditions can lead to significant differences in outcomes over time. Imagine a butterfly flapping its wings in one part of the world, and how that tiny action could potentially influence the weather patterns in a distant location. This idea, often referred to as the "butterfly effect," illustrates the sensitivity of chaotic systems to initial conditions. Chaotic systems are characterized by their extreme sensitivity, nonlinearity, and apparent randomness. Despite their unpredictable nature, chaotic systems follow deterministic rules, which means that their behaviour is not random but governed by specific mathematical equations. Chaos theory has applications in various fields, including meteorology, physics, biology, and even social sciences. It helps us understand how seemingly complex and random phenomena can emerge from underlying patterns and interconnectedness, offering insights into the inherent complexity of the world we live in.

1. **Spatial Statistics**

Spatial statistics is a branch of statistics that focuses on analyzing data with a spatial or geographic component. It helps us understand how things are distributed or arranged in space and how they may be related to each other. Imagine you have a map of a city and you want to study the pattern of crime incidents. Spatial statistics allows you to analyze the locations of the incidents and determine if there are any clusters or hotspots where crimes tend to concentrate. It also helps you explore the spatial relationships between different variables, such as the proximity of parks to schools or the correlation between pollution levels and population density in different neighbourhoods. By considering the spatial context of data, spatial statistics provides valuable insights into spatial patterns, trends, and relationships that traditional statistical methods may overlook. It has applications in urban planning, environmental science, epidemiology, and many other fields where spatial data is important for decision-making and understanding spatial processes.

1. **Cognitive Bias**

Cognitive bias refers to the systematic errors in our thinking or decision-making processes that can lead to inaccurate judgments or perceptions. Think of it as a "mental shortcut" that our brains take to simplify complex information and make quick decisions. For example, imagine you're at a party and you meet someone who reminds you of your best friend. You might immediately assume that this person is friendly and trustworthy, simply based on the resemblance. This is called the "halo effect" bias, where your initial positive impression of someone influences your perception of their other qualities.

Cognitive biases can also affect our judgments and beliefs. For instance, confirmation bias occurs when we seek out information that confirms our existing beliefs and ignore or downplay evidence that contradicts them. It's like wearing "blinders" that prevent us from considering alternative perspectives. Another common bias is the availability heuristic, where we overestimate the likelihood of events based on how easily examples come to mind. For instance, if we hear news about a plane crash, we might start believing that flying is dangerous, despite statistical evidence to the contrary.

These biases are part of our cognitive makeup and can sometimes lead us astray. Being aware of them can help us make more informed and objective decisions by actively challenging our assumptions and seeking diverse perspectives.

1. **Predictive Analysis**

Predictive analysis, also known as predictive modeling, is a process of using data and statistical techniques to make predictions about future events or outcomes. It involves analyzing historical data to identify patterns and relationships, and then using those patterns to forecast what might happen in the future.

Imagine you're planning a trip to the beach, and you want to know whether it will rain on a particular day. Predictive analysis would involve gathering data about past weather conditions, such as temperature, humidity, wind speed, and whether it rained or not. By examining this historical data, you can identify patterns and correlations between certain weather conditions and rain. Using this information, you can build a predictive model that takes into account the current weather conditions and predicts the likelihood of rain on your chosen day.

Predictive analysis is widely used in various fields, such as finance, marketing, healthcare, and sports. It helps businesses make informed decisions, such as forecasting customer behaviour, optimizing inventory levels, or identifying potential risks. By leveraging historical data and advanced statistical techniques, predictive analysis enables us to gain insights into the future and make better-informed choices.

1. **Item Response Theory**

Item Response Theory (IRT) is a statistical framework used to analyze responses to test items or questionnaire items. It is particularly useful in educational and psychological assessments.

Imagine a multiple-choice test where each item has different difficulty levels and each respondent has a different level of ability. IRT aims to model the relationship between the respondents' abilities and their responses to the items.

IRT considers two main components: the item characteristics and the individual's ability. The item characteristics include parameters such as item difficulty, discrimination, and guessing probability. The individual's ability is represented on a latent trait continuum, which reflects the underlying construct being measured (e.g., intelligence, proficiency).

IRT models provide valuable insights into the properties of test items, such as their difficulty level and discriminatory power. They also allow us to estimate individuals' abilities more accurately, even when they have not responded to all items. Additionally, IRT models enable the creation of adaptive tests that can dynamically adjust the difficulty of items based on the individual's responses.

By using IRT, researchers and educators can gain a deeper understanding of test items' quality, improve test design, and make more precise interpretations of individuals' abilities. It helps in developing fair and reliable assessments that provide meaningful information about the respondents' skills or attributes.

1. **Order statistics**

Order statistics, in probability theory and statistics, refer to the arrangement of a set of random variables in ascending order. Specifically, given a random sample of size n from a population, the order statistics are the values obtained by sorting the sample in ascending order. Each value in the sorted sample is referred to as the k-th order statistic, where k ranges from 1 to n.

Order statistics provide valuable insights into the distribution of the original random variables and are widely used in various statistical analyses. Here are some key points about order statistics:

The k-th order statistic represents the k-th smallest value in a random sample.

Order statistics allow us to estimate and characterize the underlying population distribution from which the sample is drawn.

The smallest and largest order statistics, referred to as the first and nth order statistics, respectively, provide information about the range of values in the sample.

The distribution of order statistics depends on the distribution of the original random variables and follows specific mathematical properties.

For a random sample from a continuous distribution, the joint distribution of the order statistics can be described by the joint density function.

Order statistics can be used to estimate percentiles, such as the median or quartiles, of the population distribution.

The range between consecutive order statistics provides insights into the spread or variability of the sample.

Order statistics play a crucial role in the construction of order-based nonparametric statistical tests and rank-based inference methods.

The expectation and variance of order statistics can be calculated based on the properties of the underlying distribution.

Order statistics find applications in reliability analysis, extreme value theory, order statistics-based ranking, and selecting the best or worst performers in a group.

Understanding order statistics allows statisticians to extract valuable information from samples and make inferences about the underlying population distribution.

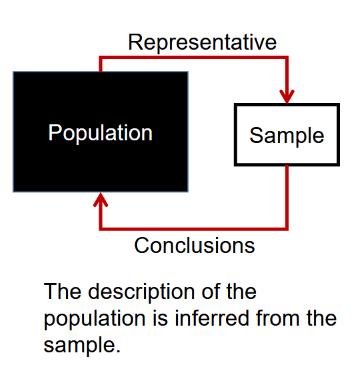
1. **Six sigma**

Six Sigma is a highly effective quality assurance methodology that is widely used in statistical analysis to enhance processes and optimize performance when working with data. It aims to minimize variations and defects in a process by setting a high standard of quality. The term "Six Sigma" refers to the statistical concept of standard deviations, where a process is considered to be at the Six Sigma level when it achieves a remarkable 99.99966% defect-free rate. This means that only 3.4 defects per million opportunities are expected in the outcomes of the process. By implementing Six Sigma principles, organizations can identify and eliminate sources of errors and inefficiencies, leading to improved productivity, customer satisfaction, and overall business success. The methodology utilizes various statistical tools and techniques to measure, analyze, and improve processes, ensuring they operate at the highest level of performance and quality.

1. **Inferential Statistics**

Inferential statistics makes inferences about populations using part of the population. Instead of using the entire population to gather data, the statistician will collect a sample or samples from the millions of data points.

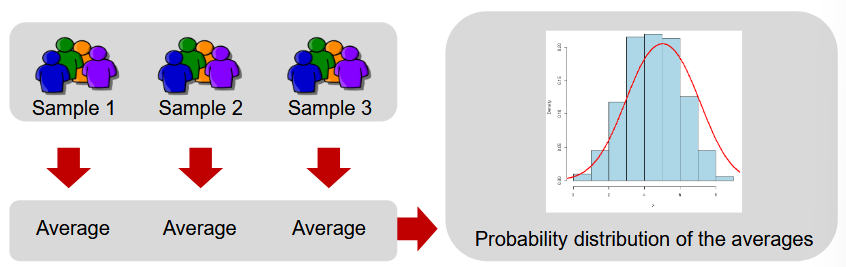
The statistician will then make inferences about the entire population using the samples.



1. **Sampling Distribution**

The probability distribution of a static (sample estimate) is called sampling distribution.

Eg., Suppose we draw all possible samples of size ‘n’ from a given population. We compute a statistic, say mean, of each sample. The probability distribution of this statistic is called a sampling distribution.



1. **Standard Error**

Standard error of the sampling distribution (σx) is mainly determined by the standard deviation of the population and the sample size (n).

1. **Scales of Measurement in Statistics**

1. Nominal – Values assigned to variable represent a descriptive category. However, these values have no inherent numerical value to quantify magnitude. Eg., The variable gender is measured on a nominal scale. Thus, gender only fulfils the identity property.

2. Ordinal – Each value has an ordered relationship to every other value on the scale.

Eg. The results of an Olympic race for the 1st, 2nd and 3rd positions. Stages of cancer, age category.

3. Interval – There is a meaningful and equal distance between two points on the scale. Eg., Temperature. IQ score.

4. Ratio – It has all the four properties of measurement: identity, magnitude, equal intervals, minimum value.

**Probability**

1. **Events**

Events represent outcomes or sets of outcomes from a random experiment. Here are 10 different types of events:

1. Simple Event: An event that consists of a single outcome from a sample space, such as rolling a specific number on a fair six-sided die.
2. Compound Event: An event that consists of multiple outcomes or combinations of outcomes, such as flipping heads twice in a row or drawing two red cards from a deck.
3. Mutually Exclusive Events: Events that cannot occur simultaneously, meaning if one event happens, the other cannot. For example, rolling an odd number and rolling an even number on a die are mutually exclusive events.
4. Independent Events: Events that do not affect each other's probability. The occurrence or non-occurrence of one event has no influence on the other event. Tossing a coin and rolling a die are independent events.
5. Dependent Events: Events that are influenced by each other. The probability of the second event depends on the outcome of the first event. Drawing a card from a deck without replacement is an example of dependent events.
6. Complementary Event: The event that occurs when an event does not happen. If event A represents flipping heads, the complementary event of A is flipping tails.
7. Exhaustive Events: A set of events that together cover all possible outcomes of an experiment. For example, when rolling a fair six-sided die, the events of getting a 1, 2, 3, 4, 5, or 6 are exhaustive events.
8. Impossible Event: An event that cannot occur, having a probability of 0. For example, rolling a 7 on a standard six-sided die is an impossible event.
9. Certain Event: An event that is guaranteed to occur, having a probability of 1. For example, when flipping a fair coin, the event of getting heads or tails is a certain event.
10. Equally Likely Events: Events that have the same probability of occurring. When rolling a fair six-sided die, each face has an equal chance of landing, so all the outcomes are equally likely events.
11. **Likelihood**

It is the process of determining the best data distribution given a specific situation in the data. Likelihood is the probability of classifying a given observant in the presence of some other variable.

1. **Sample Space**

It refers to the set of all possible outcomes of a random experiment or process. For example, when rolling a fair six-sided die, the sample space consists of the numbers {1, 2, 3, 4, 5, 6}. Each outcome in the sample space has an associated probability of occurring. The sample space can be discrete, such as in the case of flipping a coin (where the sample space is {Heads, Tails}), or continuous, such as measuring the height of individuals (where the sample space is an infinite range of real numbers). Understanding the sample space is crucial for defining probabilities and analysing random events. It provides the foundation for calculating probabilities, constructing probability distributions, and making statistical inferences.

1. **Complementary events**

They refer to the occurrence and non-occurrence of a specific outcome. If we have an event A that represents a certain outcome, then the complementary event of A, denoted as A', represents the absence or non-occurrence of that outcome. These two events are mutually exclusive, meaning that only one of them can occur at a time. The probability of event A and its complement A' always adds up to 1, as one of them must happen. If we know the probability of event A, we can easily calculate the probability of its complement A' by subtracting the probability of A from 1. Complementary events are widely used in probability calculations, statistical analysis, and hypothesis testing. They provide a convenient way to reason about the occurrence and non-occurrence of specific outcomes, making them essential in various areas of mathematics, statistics, and decision-making.

1. **Conditional Probability**

Conditional probability is a concept in probability theory that measures the probability of an event occurring given that another event has already occurred. It quantifies the likelihood of an outcome based on some additional information. For example, consider a deck of cards. The probability of drawing a red card from the deck is the unconditional probability. However, if we already know that the first card drawn was a heart, the probability of drawing a red card on the second draw changes. This revised probability is the conditional probability, which takes into account the additional information provided. Conditional probability is calculated by dividing the joint probability of two events (the occurrence of both events) by the probability of the condition (the first event). It plays a crucial role in various fields, including statistics, machine learning, and decision theory, where events are often dependent on each other. Understanding conditional probability is vital for making informed decisions and analysing complex systems.

1. **Joint Probability Distributions**

Joint probability distributions are used to model the simultaneous occurrence of multiple events or random variables. They describe the probability of observing specific combinations of outcomes for these variables. For example, in a two-dice experiment, a joint probability distribution can show the likelihood of rolling a specific number on each die. Joint probability distributions can be represented in various ways, such as tables, matrices, or mathematical equations. They provide valuable information about the relationships and dependencies between variables. Marginal probabilities can be derived from joint probability distributions, which represent the probabilities of individual events regardless of the other variables. Conditional probabilities can also be calculated from joint probability distributions, allowing for the analysis of events given certain conditions. Joint probability distributions are extensively used in fields such as statistics, econometrics, and machine learning for modelling and understanding complex systems with multiple variables.

1. **Prior Probability**

It is the proportion of the dependent variable in the dataset.

Problem: Captain has to decide to bat first

Probability: Only two possibilities

Choose to bat/Doesn’t choose to bat

P (choose to bat) = P (doesn’t choose to bat) = 0.5

Likelihood: Choosing to bat first will depend on

Weather Conditions (rainfall, wind speed), Due pitch, Humidity

1. **PAC Learning**

Probably Approximately Correct is a framework for mathematical analysis. A PAC Learner tries to learn a concept (approximately correct) by selecting a hypothesis from a set of hypotheses that has a low generalization error.

A problem is PAC-learnable if there is an algorithm ‘A’ when given some independently drawn samples, will produce a hypothesis with a small error for any distribution ‘D’ and any concept ‘c’, and that too with a high probability. It may not be possible to find a perfect hypothesis with zero error so the goal is to find a consistent hypothesis that can predict approximately correctly with an upper bound on the error.

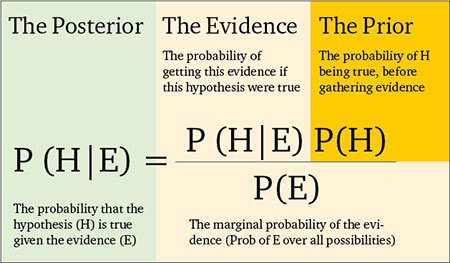
1. **Long tailed distribution**

It is a type of distribution where the tail drops off gradually towards the end of the curve. It is widely used in classification and regression. Pareto principle and the product sales distribution are good examples.

1. **Bayesian and Frequentist**

Bayesian rests on the data which is actually observed in reality and further considers the probability distribution on the hypothesis.

Frequentists rest on the hypothesis of choice and further consider the probability distribution on the data, whether it is observed or not.



1. **Probability Mass Function**

Probability mass functions are used to describe discrete probability distributions and allow us to determine the probability of an observation being exactly equal to a target value.

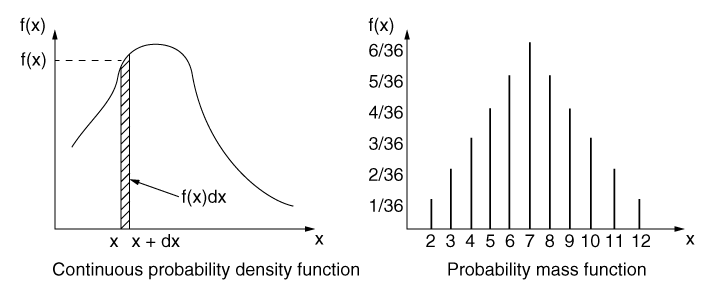
1. **Cumulative Distribution Function**

It can be defined for any kind of random variable, i.e., discrete or continuous, and it tells us the probability that the random variable X takes a value less than or equal to a particular value x:

The CDF is used to determine the probability that an observation will be greater than a certain value, or between two values.

1. **Probability Density Function**

Probability density functions are used to describe continuous probability distributions and allows us to determine the probability of an observation being within a range around our target value by computing the area under the curve for our interval.



1. **Shifted Geometric Distribution**

Probability distribution of the number X of Bernoulli trials needed to get one success, supported on the set {1,2,3,…}. The geometric distribution is denoted by Geo(p) where p>0 is the success probability. The mean of Geo(p) is 1/p

E[X] = 1/p

1. **Bernoulli Trial**

A Bernoulli trial is a random experiment with exactly two possible outcomes, ‘successes and ‘failure’, in which the probability of success is the same every time the experiment is conducted.

1. **Binomial Distribution**

The Binomial distribution describes the probability of obtaining k successes in n Bernoulli experiments, i.e., an experiment which has only two possible outcomes, often call them success and failure. In Binomial Distribution there is a fixed number of trials. Its probability function describes the probability of getting exactly k successes in ‘n’ independent Bernoulli trials:

1. **Geometric Distribution**

The Geometric distribution describes the probability of experiencing a certain amount of failures before experiencing the first success in a series of Bernoulli experiments. In a geometric distribution we are interested in the number of trials required until we obtain a success.

1. **Poisson Distribution**

The Poisson distribution is a discrete probability distribution that models the number of events occurring in a fixed interval of time or space when the events are rare and independent of each other. It is often used to describe the occurrences of rare events, such as the number of customers arriving at a store in a given hour or the number of phone calls received at a call centre in a given minute.

The formula for the Poisson distribution is given as:

Where:

* P(X = k) is the probability of observing k events
* λ (lambda) is the average rate or intensity of events occurring in the given interval
* e is Euler's number (approximately 2.71828)
* k is the number of events

The Poisson distribution is characterized by its mean (λ) and variance (also λ). It is commonly used in various fields, including telecommunications, finance, insurance, and quality control, to analyse and predict the occurrence of rare events.

1. **Gamma Distribution**

The gamma distribution is a continuous probability distribution that is commonly used to model various real-world phenomena, such as the time between events or the distribution of the sum of exponentially distributed random variables. It is a versatile distribution that encompasses several other distributions as special cases, including the exponential distribution and the chi-squared distribution.

The probability density function (PDF) of the gamma distribution is given by the formula:

f(x) = (x^(k-1) \* e^(-x/θ)) / (θ^k \* Γ(k))

Where:

* f(x) is the probability density function at x
* x is a non-negative real number
* k is the shape parameter, often referred to as the "shape" of the distribution
* θ is the scale parameter, often referred to as the "rate" parameter
* Γ(k) is the gamma function, which is a generalization of the factorial function to non-integer values

The shape parameter k determines the shape of the distribution, with higher values resulting in distributions that are more concentrated around the mean. The scale parameter θ determines the rate at which the distribution decays.

The gamma distribution is widely used in fields such as reliability engineering, queueing theory, and actuarial science. It provides a flexible framework for modelling a variety of continuous random variables and has properties that make it mathematically tractable and useful in statistical inference and estimation.

1. **Negative Binomial Distributions**

The negative binomial distribution is a probability distribution that models the number of failures that occur before a specified number of successes in a series of independent and identically distributed Bernoulli trials. The distribution is characterized by two parameters: the number of successes (r) and the probability of success in a single trial (p). The negative binomial distribution can be used to answer questions like "How many failures will occur before obtaining the third success?" or "What is the probability of needing more than five trials to achieve two successes?"

The probability mass function of the negative binomial distribution is given by

P(X = k) = (k+r-1)C(k) \* p^r \* (1-p)^k,

where X is the random variable representing the number of trials needed to obtain r successes, and (k+r-1)C(k) is the binomial coefficient.

The mean of the negative binomial distribution is given by μ = r/p, and the variance is given by σ^2 = r(1-p)/p^2.

The negative binomial distribution is often used in situations where the number of trials needed to achieve a certain number of successes is of interest, such as modeling the number of attempts before making a sale or the number of tosses of a coin before obtaining a specified number of heads.

The negative binomial distribution can be used as an alternative to the geometric distribution when the number of successes is not restricted to being the first success.

1. **Hypergeometric Distributions**

The hypergeometric distribution is a probability distribution that models the probability of obtaining a specific number of successes in a fixed number of draws without replacement from a finite population of two or more distinct types of objects. The distribution is characterized by four parameters: N (total population size), K (number of objects of interest in the population), n (number of draws), and k (number of successes). The hypergeometric distribution is useful when sampling from a finite population without replacement, where the probability of success changes for each draw. It is commonly used in situations like quality control, where a sample is taken from a batch of items to determine the number of defective items. The probability mass function of the hypergeometric distribution is given by

P(X = k) = (K choose k) \* ((N-K) choose (n-k)) / (N choose n),

where X is the random variable representing the number of successes in the sample. The mean of the hypergeometric distribution is given by μ = n \* (K / N), and the variance is given by σ^2 = n \* (K / N) \* ((N - K) / N) \* ((N - n) / (N - 1)). In contrast to the binomial distribution, which models sampling with replacement, the hypergeometric distribution takes into account the finite population size and the fact that each draw changes the probability of success.

1. **Multivariate Distributions**

Multivariate distributions are probability distributions that involve multiple random variables. They provide a framework for modelling and analysing complex systems where multiple variables are interrelated. In a multivariate distribution, each observation consists of a vector of random variables, rather than a single variable. Multivariate distributions allow us to model the joint behaviour and dependencies among multiple variables. Examples of multivariate distributions include the multivariate normal distribution, multivariate t-distribution, and multivariate Poisson distribution. The multivariate normal distribution is one of the most commonly used multivariate distributions and is characterized by its mean vector and covariance matrix. Covariance matrix captures the strength and direction of the linear relationship between pairs of variables in the multivariate distribution. Multivariate distributions can be visualized using scatter plots, contour plots, or 3D plots to show the relationships between variables. Multivariate distributions enable us to perform multivariate statistical analysis, such as multivariate regression and principal component analysis (PCA). Estimating parameters in multivariate distributions requires techniques such as maximum likelihood estimation or Bayesian methods. Multivariate distributions are used in various fields such as finance, economics, genetics, and image processing. Understanding multivariate distributions is essential for analysing and interpreting complex datasets where multiple variables interact and influence each other.

1. **Power Law and Power Law distribution**

Power law (also called the scaling law) states that a relative change in one quantity results in a proportional relative change in another. Example, a square; if you double the length of a side (say, from 2 to 4 inches) then the area will quadruple (from 4 to 16 inches squared). A power law distribution has the form , where:

X and Y are variables of interest, α is the law’s exponent, k is a constant.

Any inverse relationship like is also a power law, because a change in one quantity results in a negative change in another.

Other examples of phenomena with this type of distribution:

Distribution of income, Magnitude of earthquakes, Size of cities according to population, Size of corporations, Trading volumes on the stock market, word frequencies etc.

1. **Disjoint Events**

Disjoint events are events that never occur at the same time. These are also known as mutually exclusive events. These are often visually represented by Venn diagram.

1. **Chebyshev’s Inequality**

Chebyshev's inequality is a powerful theorem in probability theory that provides an upper bound on the probability that a random variable deviate from its mean. It states that for any random variable with a finite variance, the probability that the random variable deviates from its mean by more than k standard deviations is at most 1/k^2, where k is any positive constant greater than 1. This inequality holds regardless of the shape of the probability distribution, making it a versatile tool for bounding probabilities. Chebyshev's inequality is particularly useful when dealing with data that is not normally distributed or when precise information about the distribution is unknown. By applying Chebyshev's inequality, we can make statements about the likelihood of observing extreme values and control the amount of uncertainty associated with such deviations. This inequality has broad applications in statistics, quality control, and risk management, allowing us to establish bounds on the behaviour of random variables and draw meaningful conclusions about their variability.

1. **Stationary and Ergodic Processes**

Stationary and ergodic processes are important concepts in the field of stochastic processes.

1. A stationary process is one whose statistical properties, such as mean and variance, do not change over time. It exhibits a constant behaviour across time, making it useful for modelling and analysis.
2. Stationarity can be defined in two ways: strict stationarity, where the joint distribution of any collection of time points is invariant to shifts in time, and weak stationarity, where the mean, variance, and autocovariance of the process remain constant over time.
3. Stationary processes are commonly encountered in many real-world phenomena, such as stock prices, weather patterns, and signal processing. They provide a convenient framework for analysing and predicting future behaviour.
4. Ergodicity is a property of a stochastic process where its statistical properties can be estimated from a single, long realization of the process. In an ergodic process, the time average and ensemble average of a property converge to the same value.
5. Ergodicity implies that the behaviour observed in a long time series is representative of the overall behaviour of the process. It allows us to make inferences about the process based on a single realization.
6. Many commonly used stochastic processes, such as Markov processes and Gaussian processes, exhibit both stationarity and ergodicity. These properties simplify the analysis and provide meaningful interpretations.
7. Stationary and ergodic processes are often studied using tools like autocovariance and autocorrelation functions, which capture the dependence structure over time.
8. In applications, understanding the stationary and ergodic properties of a process helps in modelling and forecasting. It enables us to estimate parameters, evaluate performance, and assess the reliability of predictions.
9. Stationary and ergodic processes provide a solid foundation for time series analysis, enabling the development of techniques like autoregressive integrated moving average (ARIMA) models and spectral analysis.
10. The concepts of stationarity and ergodicity are fundamental in the study of stochastic processes, providing insights into the behaviour, predictability, and long-term properties of random phenomena.
11. **Martingales**

Martingales are a class of stochastic processes that play a fundamental role in probability theory and mathematical finance. Here are 10 key points about martingales:

1. A martingale is a stochastic process that represents a fair game, where the expected value of the process at any future time is equal to its current value.
2. The concept of a martingale is closely tied to the idea of a random walk, where the process evolves step by step according to random increments.
3. Formally, a martingale is a sequence of random variables that satisfies the property of being "martingale difference," meaning that the expected change in value from the current time to the next is zero.
4. Martingales have numerous applications in finance, such as modelling stock prices, option pricing, and risk management. They provide a framework for understanding the fair value of financial derivatives.
5. Martingales also have applications beyond finance, including in statistics, gambling theory, and signal processing. They provide a powerful tool for analysing and understanding various stochastic phenomena.
6. Martingales possess important properties, such as the martingale convergence theorem, which states that under certain conditions, a martingale will converge almost surely.
7. One of the most well-known martingales is the random walk, where the process moves randomly with equal probability in positive and negative directions.
8. Martingales can be constructed from other processes, such as Brownian motion and Markov chains, by applying certain transformations or conditioning techniques.
9. The theory of martingales has been extended to encompass more general classes of processes, such as submartingales and supermartingales, which relax the fairness assumption.
10. Understanding martingales is crucial for advanced topics in probability theory and stochastic analysis. They provide a rich framework for modelling and analysing random phenomena, and their properties have significant implications in various fields.
11. **Brownian Motion**

Brownian motion is a fundamental concept in physics and mathematics that describes the random movement of particles suspended in a fluid. It was first observed by botanist Robert Brown in 1827 while studying pollen grains under a microscope. In Brownian motion, particles undergo erratic, continuous, and random motion due to the collision with fluid molecules.

The main characteristics of Brownian motion include:

1. Randomness: The path of a particle is unpredictable and exhibits no discernible pattern.
2. Continuity: The motion is continuous, with no sudden jumps or discontinuities.
3. Markov property: The future movement of the particle is independent of its past movement.
4. Gaussian distribution: The displacements of the particle follow a normal distribution.
5. Scaling property: Brownian motion is scale invariant, meaning it looks the same at different time scales.

Brownian motion has significant applications in various fields, including physics, finance, biology, and chemistry. It serves as a mathematical model for diffusive processes, stock price movements, molecular motion, and the behavior of microscopic particles in a fluid. The study of Brownian motion has led to advancements in statistical physics and the development of stochastic calculus.

1. **Stochastic Processes**

Stochastic processes are mathematical models that describe the evolution of random variables over time or space. They are used to analyze and understand systems that involve randomness. In a stochastic process, the future outcomes depend not only on the current state but also on the probabilistic behavior of the system.

For example, consider a stock market. The fluctuations in stock prices can be modeled as a stochastic process, where the price at any given time depends on past prices and other random factors. Stochastic processes are widely used in various fields such as finance, physics, biology, and telecommunications to study phenomena like random walks, queueing systems, birth and death processes, and more. They provide a framework for understanding the behavior and predicting the future of systems affected by uncertainty.

1. **Queuing Theory**

Queuing Theory is a mathematical discipline that focuses on studying waiting lines or queues. It provides a framework for analyzing and optimizing the behavior of systems where customers or entities arrive, wait in line, and are served by one or more service providers. Queuing Theory involves studying key performance measures such as average waiting time, queue length, and system utilization. By modeling arrival and service patterns, queuing models can be used to determine optimal service capacities, staffing levels, and queue management strategies. Queuing Theory finds applications in various fields such as telecommunications, transportation, healthcare, and customer service, where understanding and managing waiting lines is crucial for improving efficiency, minimizing delays, and enhancing customer satisfaction.

1. **Reliability Theory**

Reliability Theory is a branch of probability theory that focuses on studying the reliability and failure characteristics of systems or components over time. It involves analyzing the probability of a system or component functioning without failure for a given period. Reliability measures, such as the mean time between failures (MTBF) and failure rate, are used to quantify the reliability of a system. Reliability Theory considers factors like component failure rates, repair times, and system configurations to assess the overall reliability of complex systems. By modeling failure and repair processes using probabilistic methods, reliability engineers can predict and optimize system performance, determine maintenance strategies, and identify critical components that may require attention. Reliability Theory finds applications in diverse fields, including engineering, manufacturing, transportation, and healthcare, where the reliable functioning of systems is essential for safety, performance, and customer satisfaction.

1. **Renewable Theory**

Renewal theory is a branch of probability theory that deals with the analysis of events or processes that occur repeatedly over time. It is commonly used to model systems with recurring events, such as the renewal of a resource, the arrival of customers to a queue, or the failure and repair of a machine. In renewal theory, the focus is on the interarrival times between consecutive events, which are assumed to be independent and identically distributed random variables. The key concept in renewal theory is the renewal function, which describes the probability that an event occurs within a specific time interval. By studying the properties of the renewal function, one can analyze various aspects of the renewal process, including the expected number of events, the average interarrival time, and the long-term behavior of the system. Renewal theory has applications in diverse fields, including reliability analysis, queueing systems, inventory management, and insurance mathematics.

1. **Decision Theory**

Decision Theory is a powerful framework that provides a systematic approach to decision-making under uncertainty. It recognizes that in real-world scenarios, decisions are often made without complete knowledge of the future outcomes. By incorporating probabilistic reasoning, Decision Theory allows decision makers to assess the potential risks and rewards associated with different courses of action. It involves quantifying uncertainty through probability distributions and evaluating decisions based on their expected utility, which combines the probabilities of outcomes with their desirability. Decision Theory considers individual preferences and trade-offs, allowing decision makers to make choices that align with their risk tolerance and goals. It also emphasizes the value of information, as decision makers can analyze the potential benefits of gathering additional data before making a decision. By applying decision analysis techniques, such as sensitivity analysis and scenario analysis, decision makers can explore the impact of uncertainty and make more informed choices. Bayesian decision theory further enhances decision-making by incorporating prior beliefs and updating them based on observed data. Overall, Decision Theory provides a structured and rational approach to navigate complex and uncertain situations across various domains.

1. **Monte Carlo methods**

Monte Carlo methods are powerful computational techniques used to solve complex problems through random sampling. Here are 10 key points about Monte Carlo methods:

1. Monte Carlo methods rely on the principle of using random sampling to approximate solutions to mathematical or statistical problems.
2. They are particularly useful when analytical or deterministic approaches are impractical or infeasible due to the complexity of the problem.
3. The core idea is to generate a large number of random samples or simulations from a probability distribution that represents the problem at hand.
4. These samples are then used to estimate various quantities of interest, such as the expected value, variance, or probability of certain outcomes.
5. Monte Carlo methods find wide applications in various fields, including physics, engineering, finance, computer science, and optimization.
6. They are especially well-suited for problems involving uncertainty, randomness, or systems with a large number of variables.
7. Monte Carlo simulations can provide insights into the behavior of complex systems and help make informed decisions in the face of uncertainty.
8. One popular application of Monte Carlo methods is in option pricing, where they are used to estimate the value of financial derivatives.
9. Another common use is in integration problems, where the Monte Carlo approach can approximate definite integrals by averaging function values over randomly selected points.
10. Monte Carlo methods have revolutionized computational science and have become a cornerstone of modern scientific and statistical research, enabling the analysis of complex systems and providing approximate solutions to challenging problems.
11. **Failure Mode and Effects Analysis (FMEA)**

A systematic approach to identify and assess potential failure modes in a system, and their impact on system performance. Probability concepts, such as probability of failure and criticality analysis, are used to prioritize and mitigate potential risks.

1. **Extreme Value Theory**

Focuses on the statistical modelling and analysis of extreme events, such as the maximum or minimum values of a random variable. It finds applications in engineering for predicting extreme values in fields like hydrology, structural engineering, and environmental risk assessment.

1. **Percolation Theory**

Studies the behaviour of random systems, such as networks or materials, in which the presence or absence of connections between elements follows a random pattern. It has applications in physics, computer science, and the study of complex systems.

1. **Information Theory**

Concerned with quantifying and measuring the amount of information in a random variable or a communication system. It provides insights into data compression, coding theory, and the limits of data transmission.

1. **Simpson’s Paradox**

Occurs when a trend or association observed in several groups of data disappears or reverses when the groups are combined. It highlights the importance of considering subgroup effects and the potential for misleading conclusions when data is aggregated.

**Machine Learning**

1. **Data Mining**

Application of machine learning methods to large databases.

1. **Data Modeling**

First step towards the design of a database. Data modelling creates a conceptual model based on the relationship between various data models. The process involves moving from the conceptual stage to the logical model to the physical schema. It involves the systematic method of applying data modelling techniques.

1. **Database Design**

Process of designing the database. The database design creates an output which is a detailed data model of the database.

1. **Data Warehouse**

Central repository of information that can be analysed to make more informed decisions. Data flows into a data warehouse from transactional systems, relational databases, and other sources, typically on a regular cadence.

1. **Data Privacy**

* Federated Learning/ Multi Party Computation - It is possible for a group of separate entities to collectively train a Machine Learning model by pooling their data but without explicitly having to share it with each other. Simply you can pool your data without sharing with each other. This also known as secure computation, which is a subfield of cryptography with the aim of creating ways for parties to jointly compute a function over their inputs while keeping those inputs private.
* Homomorphic Encryption - In this technique it is possible to do Machine Learning on data that are encrypted and that stay encrypted throughout. This technique allows computation on cipher texts, generating an encrypted outcome which, when decrypted, equals the result of the operations as if they had been performed on the plaintext.
* Differential Privacy - One of the most promising approaches within privacy-preserving Machine Learning is Differential Privacy. This technique allows you to collect personal data with quantifiable privacy protections in such a way that the output cannot be tied back to the presence or absence of any individual in that dataset. In general terms, an algorithm is differentially private if an observer examining the output is not able to determine whether a specific individual’s information was used in the computation.

To protect individual privacy, random noise is generated according to a carefully chosen distribution, which will lead to perturbation of the true answer so that the true answer plus noise is returned to the user. Differential privacy is mostly compatible with, or even beneficial to, meaningful data analysis despite its protective strength. This also offers protection from over fitting where its benefits thus go even beyond data security.

1. **Entropy in ML**

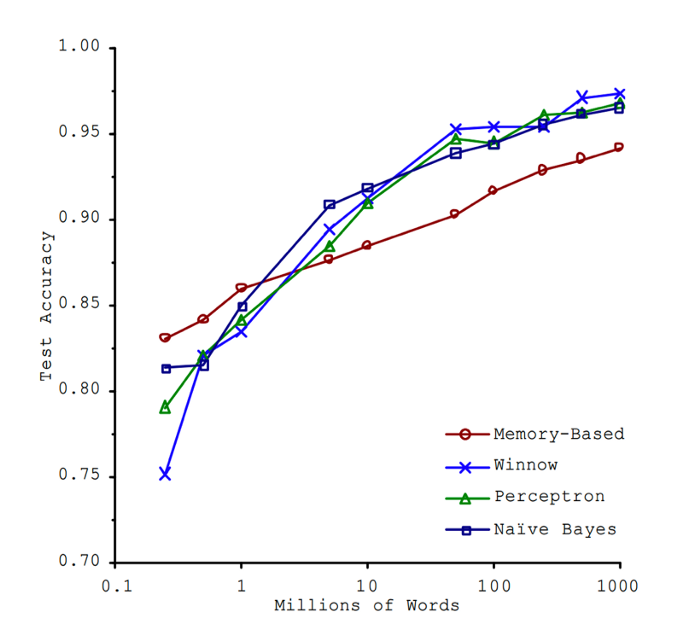
Measure of impurity and randomness in a DT.

The entropy of a given dataset tells us how pure or impure the values of the dataset are.

For example, suppose we are given a box with 10 blue marbles. Then, the entropy of the box is 0 as it contains marbles of the same colour, i.e., there is no impurity. If we need to draw a marble from the box, the probability of it being blue will be 1.0. However, if we replace 4 of the blue marbles with 4 red marbles in the box, then the entropy increases to 0.4 for drawing blue marbles.

1. **Unreasonable effectiveness of data**

The size of the dataset used to train the model mattered far more than the choice of ML approach. And, the performance differences between the models became very small as the dataset grew large. More data is almost always better.



1. **No Free Lunch Theorem**

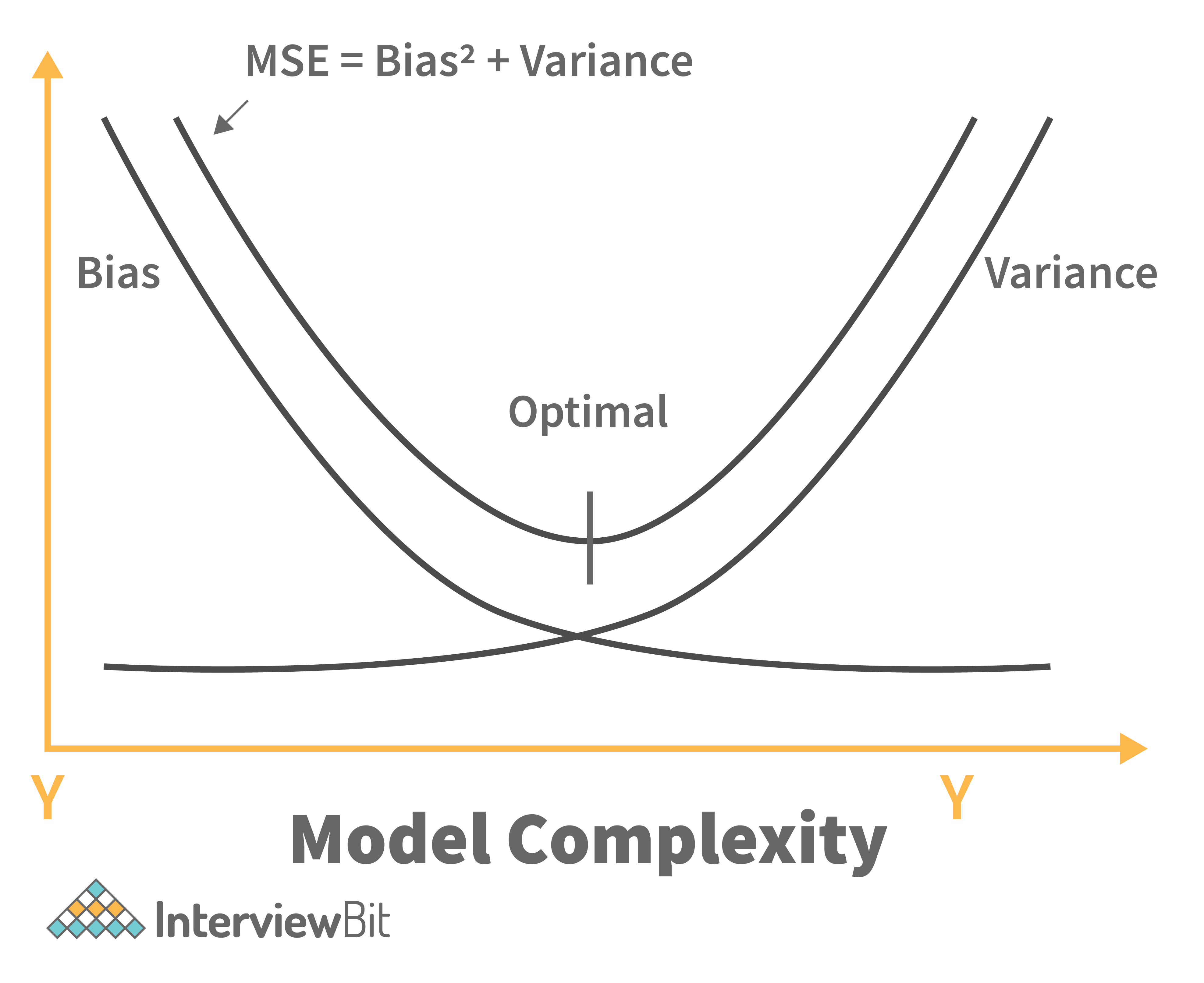
A theoretical finding that suggests all optimisation algorithms perform equally well when their performance is averaged over all possible objective functions.

1. **Variance**

Variance is a type of error that occurs when the model ends up being too complex and learns features from the data along with the noise that exists in it. This kind of error can occur if the algorithm used to train the model has high complexity, even though the data and the underlying patterns and trends are quite easy to discover. This makes the model very sensitive, that performs well on the training dataset but poorly on the testing dataset. Variance generally leads to poor accuracy in testing and results in overfitting.

1. **Bias Variance Tradeoff**

Optimum balance between bias and variance. If you try to decrease bias, the variance will increase and vice-versa. Normally, as you increase the complexity of your model, you will see a reduction in error due to lower bias in the model. However, this happens only until a particular point. As you continue to make your model more complex, you end up over-fitting your model and hence your model will start suffering from high variance.



1. **Confusion Matrix**

Table used to estimate the performance of a model.

True Positive (TP): Positive Prediction is correct/ Observation is positive, and is predicted to be positive.

False Positive (FP): Positive Prediction is incorrect/ Observation is negative, but is predicted positive.

True Negative (TN): Negative Prediction is correct/ Observation is negative, and is predicted to be negative.

False Negative (FN): Negative Prediction is incorrect/ Observation is positive, but is predicted negative.

FP is Type 1 error, which means that a positive result was predicted but the actual result is negative.

FN is Type 2 error, which means that a negative result was predicted but the actual result is positive.

Misclassification rate/Error rate

Specificity is the measure of true negative rate.

Recall/Sensitivity/True Positive Rate

Precision is the measure of a positive predicted value.

F-score is the harmonic mean of precision and recall.

High recall indicates the class is correctly recognized (small number of FN).

Low recall indicates the class is incorrectly recognized (large number of FN).

High recall, low precision means that most of the positive examples are correctly recognized (low FN), but there are a lot of false positives.

Low recall, high precision shows that we miss a lot of positive examples (high FN), but those we predict as positive are indeed positive (low FP).

1. **R squared and Adjusted R-squared**

R-squared is expressed as a percentage between 0 and 100, with 100 signalling perfect correlation and zero no correlation. One misconception about regression analysis is that a low R-squared value is always a bad thing.

R-squared (R2) is a statistical measure that explains to what extent the variance of one variable explains the variance of the second variable. So, if the R2 of a model is 0.50, then approximately half of the observed variation can be explained by the model's inputs.

Adjusted R-squared is a modified version of R-squared that has been adjusted for the number of predictors in the model. The adjusted R-squared increases when the new term improves the model more than would be expected by chance. It decreases when a predictor improves the model by less than expected. Typically, the adjusted R-squared is positive, not negative. It is always lower than the R-squared.

Adding more independent variables or predictors to a regression model tends to increase the R-squared value, which tempts makers of the model to add even more variables. This is called overfitting and can return an unwarranted high R-squared value. Adjusted R-squared is used to determine how reliable the correlation is and how much it is determined by the addition of independent variables.

1. **Cost Function**

Cost functions are a tool to evaluate how good the model performs. It takes into consideration the errors and losses made in the output layer during the back propagation process.

1. **Loss Function**

Loss function captures the difference between the actual and predicted values for a single record. Most commonly used loss functions are Mean-squared error and Hinge Loss. Loss function is used as a way to measure how well the model is performing.

1. **Cross Validation**

Statistical model validation technique used for improving model’s performance. Model will be trained and tested with rotation using different samples of the training dataset to ensure that the model performs well for unknown data. Most commonly used techniques are:

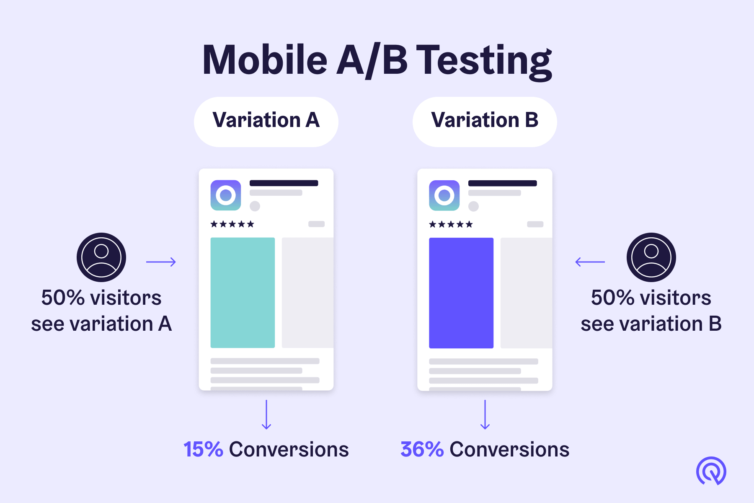
* K-fold method
* Leave p-out method
* Leave-one-out method
* Holdout method



1. **A/B Testing**

The goal of A/B testing is to pick the best variant among two hypotheses, the use cases of this kind of testing could be a web page or application responsiveness, landing page redesign, banner testing, marketing campaign performance etc. The first step is to confirm a conversion goal, and then statistical analysis is used to understand which alternative performs better for the given conversion goal.

This is a statistical hypothesis testing for randomized experiments with two variables, A and B. The objective of A/B Testing is to detect any changes to a web page to maximize or increase the outcome of a strategy. A/B testing is used when we wish to test a new feature in a product.

In the A/B test, we give users two variants of the product, and we label these variants as A and B. The A variant can be the product with the new feature added, and the B variant can be the product without the new feature. After users use these two products, we capture their ratings for the product. If the rating of product variant A is statistically and significantly higher, then the new feature is considered an improvement and useful and is accepted. Otherwise, the new feature is removed from the product.

1. **Semi supervised ML**

With more common supervised machine learning methods, you train a machine learning algorithm on a ‘labelled’ dataset in which each record includes the outcome information. This allows the algorithm to deduce patterns and identify relationships between your target variable and the rest of the dataset based on information it already has.

When you don’t have enough labelled data to produce an accurate model and you don’t have the ability or resources to get more data, you can use semi-supervised techniques to increase the size of your training data.

For example, imagine you are developing a model intended to detect fraud for a large bank. Some fraud you know about, but other instances of fraud are slipping by without your knowledge. You can label the dataset with the fraud instances you’re aware of, but the rest of your data will remain unlabelled. You can use a semi-supervised learning algorithm to label the data, and retrain the model with the newly labelled dataset. Then, you apply the retrained model to new data, more accurately identifying fraud using supervised learning techniques. However, there is no way to verify that the algorithm has produced labels that are 100% accurate, resulting in less trustworthy outcomes than traditional supervised techniques.

1. **Clustering**

Dividing data points into a number of groups. The division is done in a way that all the data points in the same group are more similar to each other than the data points in other groups. Example: Fuzzy clustering, K-means clustering, Density-based clustering etc.

K-means clustering – Unsupervised ML method. It is the technique of classifying data using a certain set of clusters which is called K clusters. The two methods to calculate the optimal value of k in k-means are:

* Silhouette score method
* Elbow method: We need to calculate the Within-Cluster-Dum of Squared Errors (WSS) for different k values. The WSS is described as the sum of the squares of the distance between each data value and its centroid. Select the k for which the WSS error starts to become negligible.



1. **Dimensionality Reduction**

Reducing the number of features in a dataset to avoid overfitting and reduce the variance. It reduces the dimensions and size of the dataset.

Reduces the storage space and time for model execution.

Removes the issue of multi-collinearity thereby improving the parameter interpretation of the ML model.

Easier to visualize data when the dimensions are reduced.

Avoids the curse of increased dimensionality.

1. **K-nearest neighbors**

Supervised learning classification/regression algorithm which uses proximity to make classifications or predictions. ‘k’ in KNN is a parameter that refers to the number of nearest neighbours to include in the majority of the voting process.

For classification problems, a class label is assigned on the basis of a majority vote-i.e., the label that is most frequently around a given data point is used. In regression problem, the average of k nearest neighbours is taken to make a prediction.

KNN requires standardisation. KNN is affected by outliers. Since it heavily relies on memory to store all its training data, it is also referred to as instance-based or memory-based learning method.

Limitations

* Curse of dimensionality – Refers to a set of problems that arise when working with high-dimensional data. The dimension of a dataset corresponds to the number of attributes/features that exist in a dataset. A dataset with a large number of attributes, generally of the order of a hundred or more, is referred to as high dimensional data. Some of the difficulties that come with high dimensional data manifest during analysing or visualizing the data to identify patterns, and some manifest while training machine learning models. The difficulties related to training machine learning models due to high dimensional data are referred to as the ‘Curse of Dimensionality’.
* Calculate the distance of the test case from all training cases.

These distance metrics help to form decision boundaries, which partitions query points into different regions.

* Euclidean distance – most commonly used
* Manhattan distance
* Minkowski distance
* Hamming distance

1. **Naïve Bayes**

Bayes theorem finds the probability of an event occurring given the probability of another event has already occurred.



Conditions for Naïve Bayes Classification:

* Features are assumed to be independent (hence naïve) - This is a strong assumption and unrealistic for real data; however, the technique is very effective on a large range of complex problems.
* Each feature is given the same weight. None of the attributes is irrelevant and assumed to be contributing equally to the outcome.

1. **Gaussian Naïve Bayes**

Continuous values associated with each feature are assumed to be distributed according to a Gaussian distribution/Normal distribution. Standard Naïve Bayes only supports categorical features, while Gaussian Naïve Bayes only supports continuously valued features.

1. **SVM Classifier**

Support vector machine, used for classification and prediction tasks. SVM consists of a separating plane that discriminates between the two classes of variables. This separating plane is known as hyperplane.

Hyperparametrs of SVM are: ‘C’, ‘gamma’, ‘kernel’

‘C’ adds penalty to each misclassified point.

‘gamma’ essentially controls the distance of influence of a single training point.

There are two types of SVM classifiers:

* Linear SVM – SVM algorithm predicts a straight hyperplane dividing the two classes. The hyperplane is also called as maximum margin hyperplane.
* Nonlinear SVM

Some of the kernels used in SVM are

* Linear Kernel, Gaussian Kernel, Polynomial Kernel
* Laplace RBF Kernel
* Hyperbolic Kernel
* Radial Basis Kernel
* Sigmoid Kernel

Advantages of SVM classifier

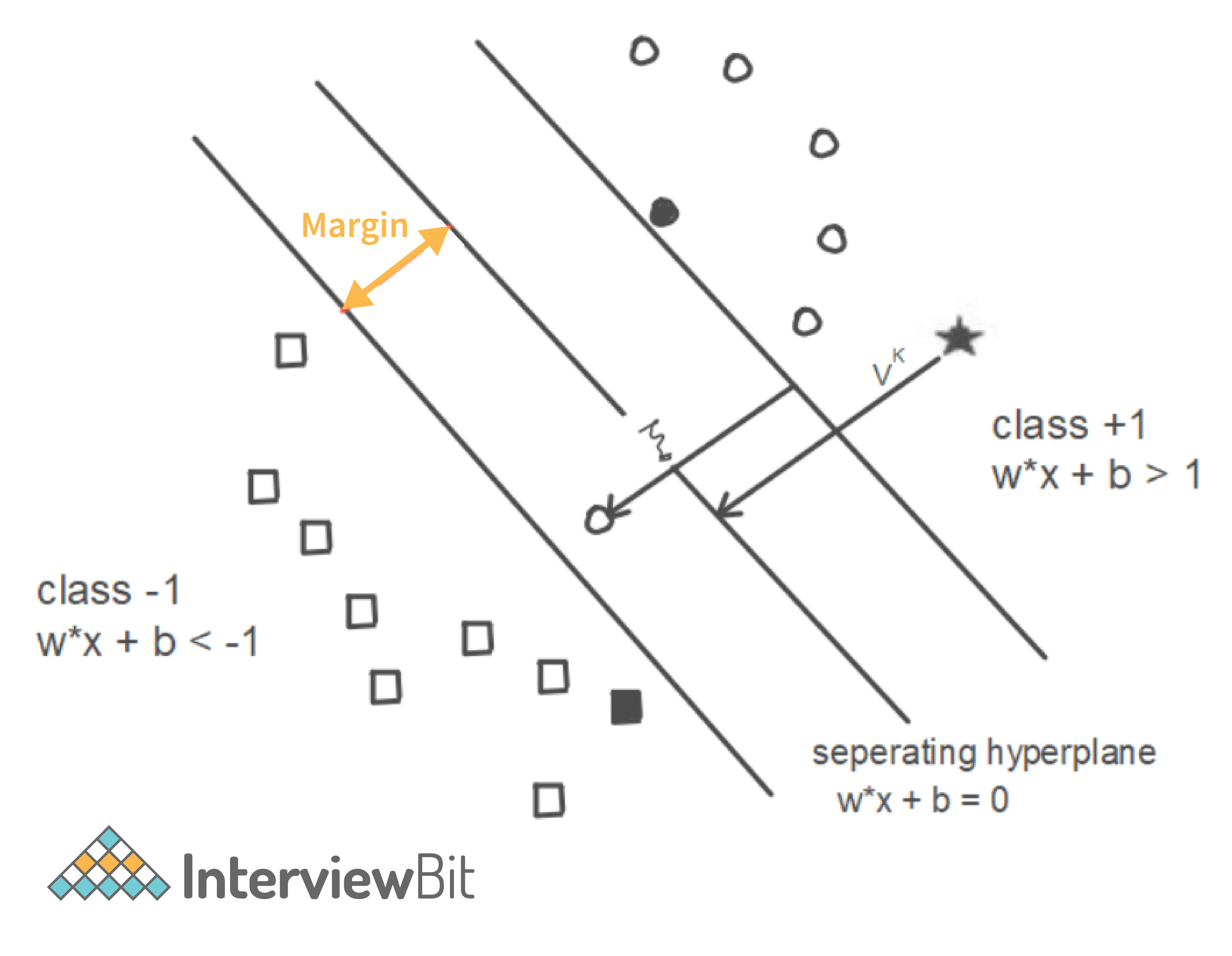
* SVMs are effective when the number of features is quite large.
* Nonlinear data can also be classified using customized hyperplanes built by using kernel trick

Disadvantages of SVM classifier

* SVM have high algorithmic complexity and extensive memory requirements due to the use of quadratic programming.
* SVMs have good generalization performance, but they can be extremely slow in test phase.

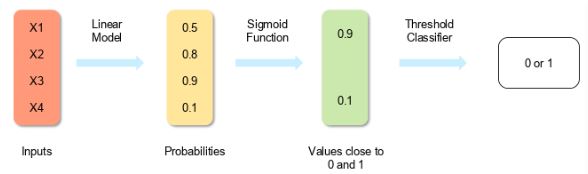
Support Vectors are data points that are nearest to the hyperplane.

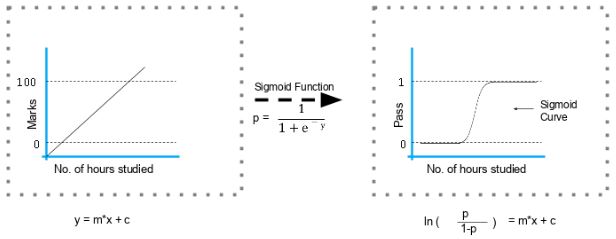
There are two types of SVMs based on hardness-

Hard margin SVMs work only if the data is linearly separable and these types of SVMs are quite sensitive to the outliers. But our main objective is to find a good balance between keeping the margins as large as possible and limiting the margin violation i.e., instances that end up in the middle of margin or even on the wrong side, and this method is called soft margin SVM.

1. **Logistic Regression**

Also known as Logit model. It measures the relationship between the dependent variable and one or more independent variables by estimating probabilities using its underlying logistic function (sigmoid).





For example, if you want to predict whether a particular political leader will win the election or not. In this case, the outcome of prediction is binary i.e., 0 or 1 (Win/Lose). The predictor variables here would be the amount of money spent for campaigning of a particular candidate, the amount of time spent in campaigning, etc.

1. **Decision Tree**

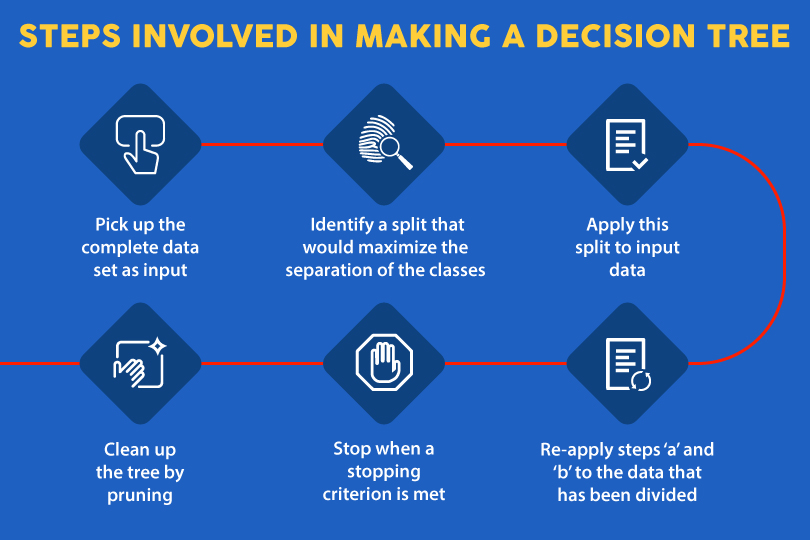
Supervised learning algorithm for both classification and regression. Dependent variable can be both a numerical value and a categorical value. Each node denotes the test on an attribute, each edge denotes the outcome of that attribute, each leaf note holds the class label.

Pruning is the process of removing the sections of the tree that are not necessary or are redundant. It helps to make smaller DT, which performs better, give high accuracy and speed.

Information gain depends on the decrease in entropy after the dataset is split on an attribute. Constructing a DT is about finding the attributes that return highest information gain.

Steps for making a DT

* Take the entire dataset as input.
* Calculate entropy of the target variable, as well as the predictor attributes.
* Calculate your information gain of all attributes.
* Choose the attribute with the highest information gain as the root node.
* Repeat the same procedure on every branch until the decision node of each branch is finalized.



1. **Random Forest**

Combination of multiple DT. To classify a new object based on attributes, each tree gives a classification. The forest chooses the classification having the most votes and in regression, it takes the average of outputs by different trees. RF is based on bagging.

Steps in making a Random Forest

Step 1: Randomly select *k* features from a total of *m* features where k << m

Step 2: Among the *k* features, calculate the node D using the best split point.

Step 3: Split the node into daughter nodes using the best split.

Step 4: Repeat steps two and three until leaf nodes are finalized.

Build forest by repeating steps 1-4 for ‘n’ times to create ‘*n’* number of trees.



In general, RF is always better than DT. But there are few scenarios in which DT is better than RF

* Explain-ability
* Computation
* Features

1. **Ensemble Learning**

Sometimes the datasets are very complex, and it is difficult for one model to be able to grasps the underlying trends in these datasets. In such situations, we combine several individual models together to improve performance. This is called ensemble learning.

1. **Bagging and Boosting**

Bagging (bootstrap aggregating) is an ensemble learning method. In this technique, we generate some data using the bootstrap method, in which we use an already existing dataset and generate multiple samples of the N size. This bootstrapped data is then used to train multiple models in parallel, which makes the bagging model more robust than a simple model.

Once all the models are trained, when it’s time to make a prediction, we make predictions using all the trained models and then average the result in the case of regression, and for classification, we choose the result, generated by models, that have the highest frequency.

Boosting is an ensemble learning method. In boosting, we create multiple models and sequentially train them by combining weak models iteratively in a way that training a new model depends on the models trained before it.

In doing so, we take the patterns learned by a previous model and test them on a dataset when training the new model. In each iteration, we give more importance to observations in the dataset that are incorrectly handled or predicted by previous models. Boosting is useful in reducing bias in models as well. Different types of boosting algorithms are

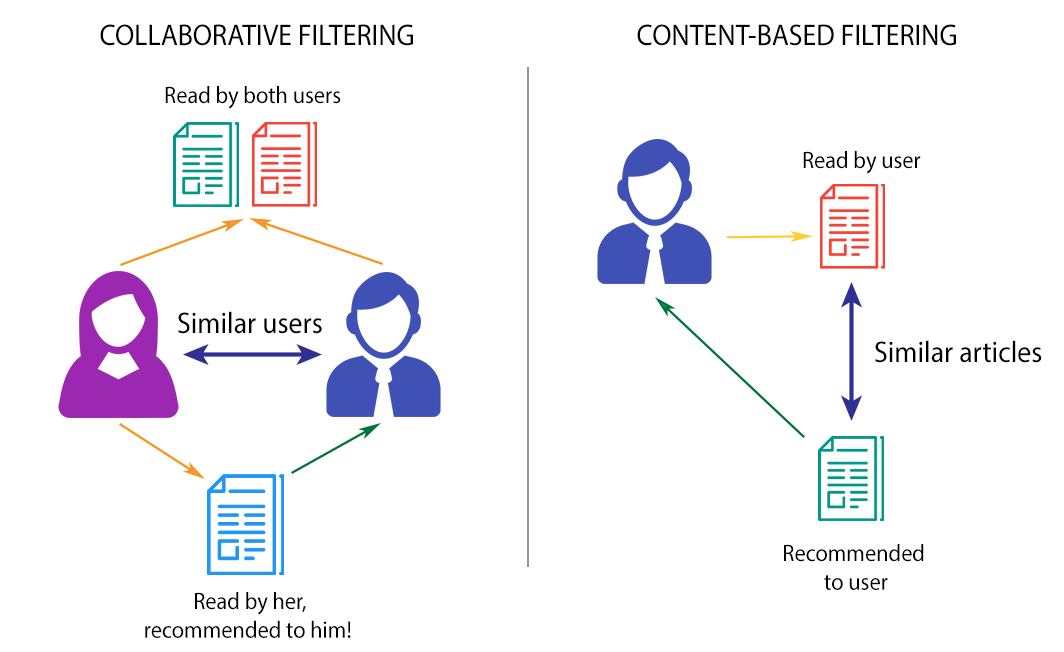
* AdaBoost
* Gradient Boosting
* XGBoost

1. **Stacking**

Stacking is an ensemble learning method. We can combine weak models that use different learning algorithms. Stacking works by training multiple weak models and then using them together by training another model, called a meta-model, to make predictions based on the multiple outputs of predictions returned by these multiple weak models.

1. **Recommendation System**

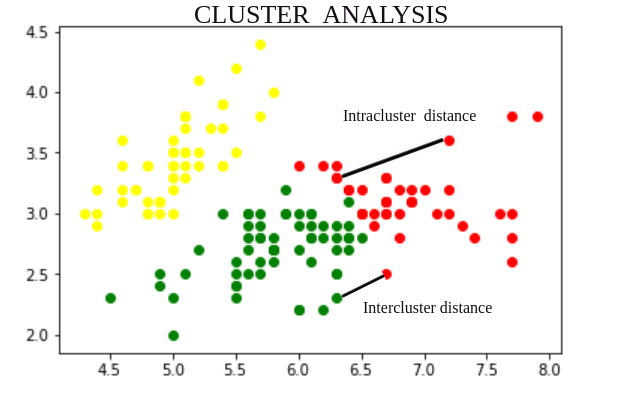
Systems that generate recommendations based on what they know about the user’s tastes from their activities on the platform. These recommendations can also be generated based on what users with a similar taste like watching. Generally, content-based filtering is better than collaborative filtering.

* Collaborative Filtering: Explains the behaviour of other users and their purchase history in terms of ratings, selection etc. The engine makes predictions on what might interest a person based on the preferences of other users. For example, If User A, similar to User B, watched and liked a movie, then that movie will be recommended to User B, and similarly, if User B watched and liked a movie, then that would be recommended to User A.
* Content-based Filtering: Recommendations are generated by making use of the properties of the content that a user is interested in. For example, if a user is watching movies belonging to the action and mystery genre and giving them good ratings, it is a clear indication that the user likes movies of this kind. If shown movies of a similar genre as recommendations, there is a higher probability that the user would like those recommendations as well.

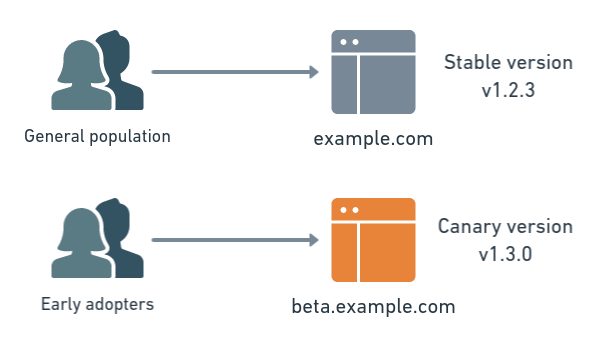
1. **Market Basket Analysis**

Market Basket Analysis is a modelling technique based upon the theory that if you buy a certain group of items, you are more (or less) likely to buy another group of items. Understanding the relationships and the strength of those relationships is valuable information that can be used to make recommendations, cross-sell, up-sell, offer coupons, etc. Market Basket Analysis is one of the key techniques used by large retailers to uncover associations between items. It works by looking for combinations of items that occur together frequently in transactions. It allows retailers to identify relationships between the items that people buy.

1. **Intercluster and Intracluster**

The aim of the clustering process is to discover overall distribution patterns and interesting correlations among the data attributes. It is the task of grouping a set of objects in such a way that objects in the same group are more similar to each other than to those in other groups. Cluster analysis itself is not one specific algorithm, but the general task to be solved. It can be achieved by various algorithms that differ significantly in their understanding of what constitutes a cluster and how to efficiently find them. Popular notions of clusters include groups with small distances between cluster members, dense areas of the data space, intervals or particular statistical distributions.

1. **Canary Deployment**

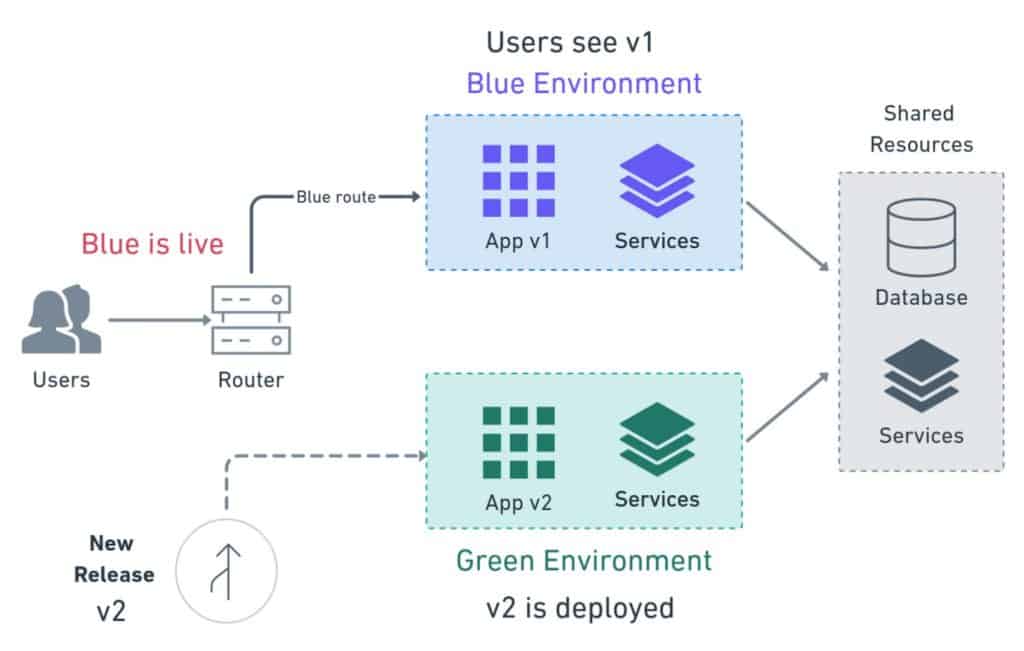
Canary release allows you to rollout your feature to only a subset of users as an initial test to make sure nothing else in your system broke. When you deploy the new software version, you shift some percentage – say, 10% - of your user base to the new version while maintaining 90% of users on the old version. If that 10% reports no errors, you can roll it out to gradually more users, until the new version is being used by everyone. If the 10% has problems, though, you can roll it right back, and 90% of your users will have never seen the problem. Canary deployment benefits include zero downtime, easy rollout and quick rollback – plus the added safety from gradual rollout process. Drawbacks include – the expense of maintaining multiple server instances, the difficult clone-or-don’t clone database decision.

1. **Blue Green Deployment**

It is a technique that reduces downtime and risk by running two identical production environments called Blue and Green.

At any time, only one of the environments is live, with the live environment serving all production traffic. For this example, Blue is currently live, and Green is idle. As you prepare a new version of your model, deployment and the final stage of testing takes place in the environment that is not live, in this example, Green. Once you have deployed and fully tested the model in Green, you switch the router, so all incoming requests now go to Green instead of Blue. Green is now live, and Blue is idle.

If something unexpected happens with your new version on Green, you can immediately roll back to the last version by switching back to Blue.



**Deep Learning**

1. **Genetic Algorithm**

It is used to solve complicated problems with a greater number of variables & possible outcomes/solutions. The combinations of different solutions are passed through the Darwinian based algorithm to find the best solutions. The poorer solutions are then replaced with the offspring of good solutions. Robotics is a common field.

Five phases are considered in a genetic algorithm:

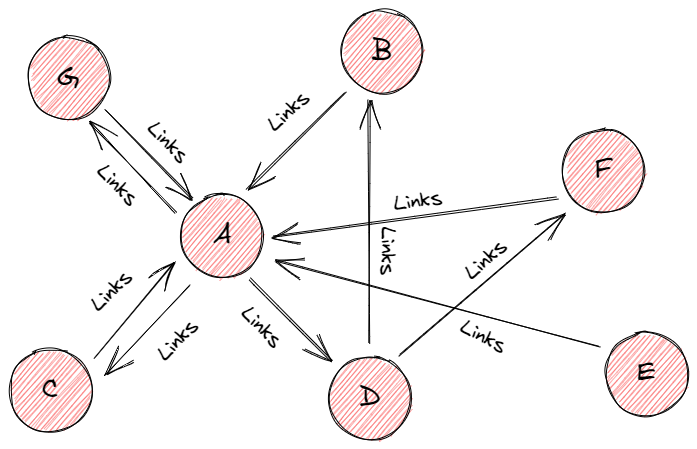
* Initial population
* Fitness function
* Selection
* Crossover – most significant phase in GA.
* Mutation

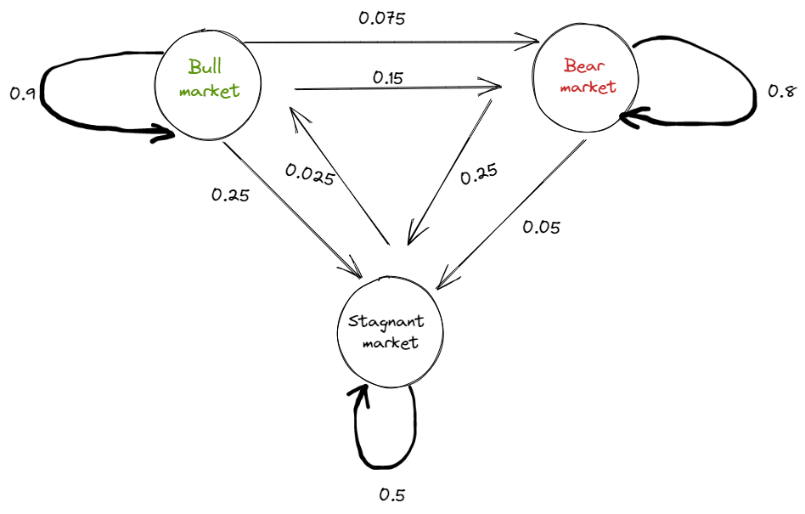
1. **Markov Chains**

Markov Chains defines that a state’s future probability depends only on its current state. Markov Chain is used in word recommendation. In this system, the model recognizes and recommends the next word based on the immediate previous word.

Markov chains are used to calculate the probability of an event occurring by considering it as a state transitioning to another state or a state transitioning to the same state as before.

* PageRank used by Google – To use the PageRank algorithm, we assume the web to be a directed graph, with web pages acting as nodes and hyperlinks acting as edges. PageRank assigns a value to a page depending on the number of backlinks referring to it. A page that is connected to many other pages earns a higher rank.
* Predicting Market Trends – Markov chains and their associated diagrams may be used to estimate the probability of various financial market climates and so forecast the likelihood of future market circumstances.





1. **Exploding gradients**

**Exploding Gradients** are exponentially growing error gradients that update the NN weights to a great extent, causes an overflow, and results in ‘Nan’ values. Exploding Gradients is the problematic scenario where large error gradients accumulate to result in very large updates to the weights of neural network models in the training stage. In an extreme case, the value of weights can overflow and result in ‘Nan’ values. Hence the model becomes unstable and is unable to learn from the training data.

1. **Vanishing Gradients**

Vanishing Gradients causes increase in training time and poor performance and low accuracy. A condition when the slope is too small during the training.

1. **Overfitting and Underfitting**

Overfitting happens when the model is too robust, **low bias** and **high variance**.

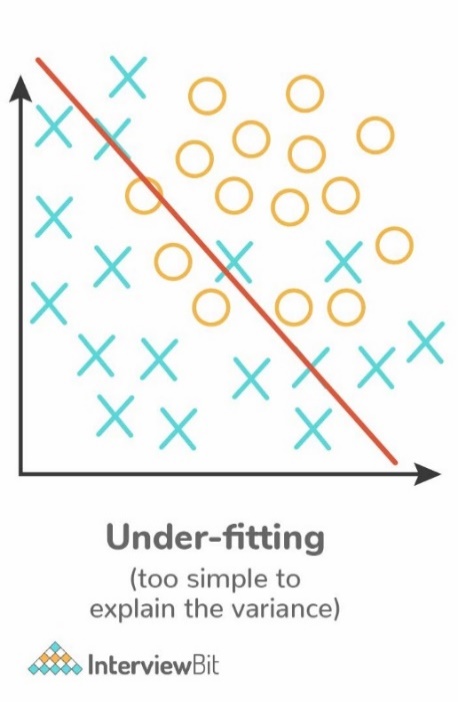
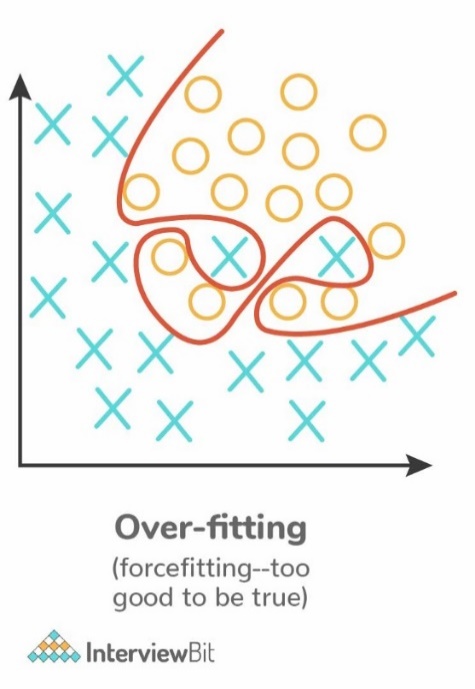
Underfitting happens when the model is too simple, **high bias** and **low variance**.

Techniques to reduce Overfitting

* Keep the model simple – take fewer variables into account, thereby removing some of the noise in the training data
* Use **cross-validation**
* **Stop early** while training
* Use random **dropouts**
* Increase training data
* Use **regularization** such as LASSO, that penalize certain model parameters

Techniques to reduce Underfitting

* Increase model **complexity**
* Increase the number of **features**
* Remove **noise** from the data
* Increase the number of training **epochs**



1. **Regularisation**

Regularization entails addition of penalties to different parameters in the ML model for reducing the freedom of the model to avoid the issue of overfitting. E.g., Linear model regularization, Lasso/L1 regularization etc. The model predictions should then minimize the loss function calculated on the regularized training set.

Linear model regularization applies penalty over coefficients that multiplies the predictors. Lasso/L1 regularization shrinks some coefficients to zero, thereby removing it from the model.

1. **Hyperparameters**

Hyperparameters are parameter whose value is set before the learning process so that the network training requirements can be identified and the structure of the network is improved.

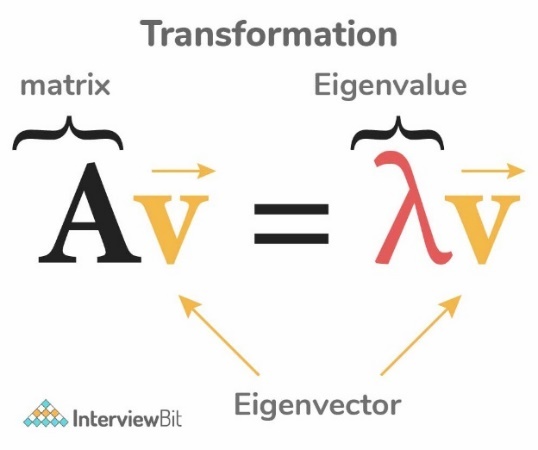
1. **Eigenvectors and Eigenvalues**

Eigenvectors are column vectors or unit vectors whose length/magnitude is equal to 1. These are used in PCA for gathering valuable insights from the given matrix.

Eigenvectors depict the direction in which a linear transformation moves and acts by compressing, flipping, or stretching. Used to understand linear transformations and are calculated for a correlation or covariance matrix.

Eigenvalue is the strength of the transformation in the direction of the eigenvector.   
Eigenvector’s direction remains unchanged when a linear transformation is applied to it.

Eigenvalues are coefficients that are applied on eigenvectors which give these vectors different values for length or magnitude. A matrix can be decomposed into Eigenvectors and Eigenvalues and this process is called Eigen decomposition.



1. **Principal Component Analysis**

It is a statistical procedure that allows you to summarize the information content in large data tables by means of a smaller set of “summary indices” that can be more easily visualized and analysed. In PCA the number of input dimensions is equal to principal components. PCA is unsupervised ML algorithm. PCA simplifies the complexity in high-dimensional data while retaining trends and patterns. It does this by transforming the data into fewer dimensions, which act as summaries of features.

1. **Multicollinearity**

Multicollinearity occurs when two or more independent variables (also known as predictor) are highly correlated with one another in a regression model.

This means that an independent variable can be predicted from another independent variable in a regression model.

Solutions for Multicollinearity-

* Drop the variables causing the problem. If using a large number of X-variables, a stepwise regression could be used to determine which of the variables to drop. Removing collinear X-variables is the simplest method of solving the multicollinearity problem.
* If all the X-variables are retained, then avoid making inferences about the individual parameters. Also, restrict inferences about the mean value of Y of values to X that lie in the experimental region.
* Re-code the form of the independent variables. For example, if x1 and x2 are collinear, you might try using x1 and the ratio x2/x1 instead.
* Ridge and Lasso Regression
* By standardizing the variables i.e., by subtracting the mean value or taking the deviated forms of the variables.
* Increase in sample size may sometimes solve the problem of multicollinearity.

1. **Feature Vectors**

It is an n-dimensional vector of numerical features that represent an object. In ML, feature vectors are used to represent numeric or symbolic characteristics/features of an object in a mathematical way.

1. **Affine Transformation**

Affine transformation is a linear mapping method that preserves points, straight lines, and planes. Sets of parallel lines remain parallel after an affine transformation.

The affine transformation technique is typically used to correct for geometric distortions or deformations that occur with non-ideal camera angles. For example, satellite imagery uses affine transformations to correct for wide angle lens distortion, panorama stitching, and image registration. Transforming and fusing the images to a large, flat coordinate system is desirable to eliminate distortion. This enables easier interactions and calculations that don’t require accounting for image distortion.

1. **Batch Normalisation**

A technique which attempts to improve the performance and stability of the neural network. It is done by normalizing the inputs in each layer so that the mean output activation remains 0 with the standard deviation at 1. The model becomes less sensitive to hyperparameter tuning. Weight initialization becomes an easy task. It introduces mild regularisation in the network.

1. **Batch Learning**

The model is trained with data in only a single batch, this is known as batch learning or offline learning.

1. **Batch Gradient Descent vs Stochastic Gradient Descent**

|  |  |
| --- | --- |
| **Batch Gradient Descent** | **Stochastic Gradient Descent** |
| Helps in computing the gradient using the complete data set | Helps in computing the gradient using only a single sample |
| Takes time to converge | Takes less time to converge |
| Volume is large for analysis purposes | Volume is low |
| Update weights infrequently | Update weights frequently |

1. **Backpropagation**

Backpropagation, short for "backward propagation of errors," is an algorithm for supervised learning of artificial neural networks using gradient descent. Given an ANN and an error function, the method calculates the gradient of the error function with respect to the neural network's weights.

The "backwards" part of the name stems from the fact that calculation of the gradient proceeds backwards through the network, with the gradient of the final layer of weights being calculated first and the gradient of the first layer of weights being calculated last. Partial computations of the gradient from one layer are reused in the computation of the gradient for the previous layer. This backwards flow of the error information allows for efficient computation of the gradient at each layer versus the naive approach of calculating the gradient of each layer separately.

Backpropagation law is also known as generalized Delta rule.

Limitations

* Slow convergence
* Scaling
* Local minima problem

1. **Perceptron**

Simplest NN that contains a single neuron which performs 2 functions. First function is to perform the weighted sum of all the inputs and the second function is an activation function.

1. **Radial Basis Function**

Particular type of ANN used for function approximation problems. RBF Networks have three-layer architecture (input layer, a hidden layer, and the output layer), universal approximation, and faster learning speed. RBF are feed-forward neural networks. The input layer receives input data and passes it into the hidden layer, where the computation occurs. The hidden layer is the most powerful. The output layer is designated for prediction tasks like classification or regression.

1. **Recurrent Neural Network**

RNN are an ANN that works on a sequence of data, time series and others.

1. **LSTM**

Long Short-Term Memory is a RNN that is capable of learning long term dependencies and recalling information for a longer period as part of its default behaviour.

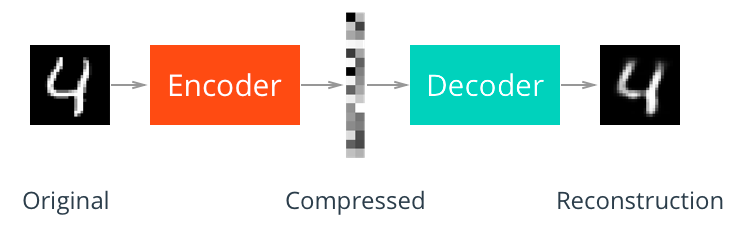
1. **Boltzmann Machine**

It is an algorithm that discover fascinating features representing complex regularities present in the data. It optimizes the quantity and weight.

1. **Autoencoders**

Transform input into output with minimal possible error. An autoencoder receives unlabelled input that is encoded for reconstructing the output. It is a kind of ANN. Autoencoder also tries to generate a representation as close as possible to its original input from the reduced encoding. These models are trained as supervised machine learning models and during inference, they work as unsupervised models that’s why they are called self-supervised models. Application of Autoencoders include:

* File Compression: Autoencoders reduce the dimensionality of input data which is referred to as file compression. Autoencoders work with all kinds of data like images, videos, and audio, this helps in sharing and viewing data faster than we could do with its original file size.
* Image de-noising: Autoencoders does not require any human interaction, once trained on any kind of data it can reproduce that data with less noise than the original image.
* Image Transformation: Autoencoders can transform B/W images to coloured one and vice versa, we can up-sample and down-sample the input data, etc.



1. **Lazy learning algorithms**

Lazy learning is a learning method in which generalization of the training data is, in theory, delayed until a query is made to the system, as opposed to eager learning, where the system tries to generalize the training data before receiving queries. E.g., KNN.

Training is fast, but prediction is slow.

Also known as instance-based learning.

Lazy classifiers are very useful when working with large datasets that have a few attributes.

1. **Instance based learning**

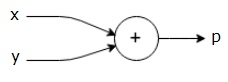
Instance-based learning refers to a family of techniques for classification and regression, which produce a class label/predication based on the similarity of the query to its nearest neighbour in the training set. E.g., KNN, Self-Organizing Map.

1. **Computational Graph**

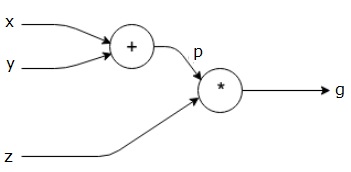
Directed Graph which has a network of nodes where nodes represent mathematical operations and the edges represent tensors. Computational graphs are a way of expressing and evaluating a mathematical expression.

Example, for a mathematical equation p = x+ y

The computational graph will be



Another example, g = (x + y) \* z



1. **GraphML**

Graph ML will introduce you to a set of tools used for processing network data and leveraging the power of the relation between entities that can be used for prediction, modelling, and analytics tasks.

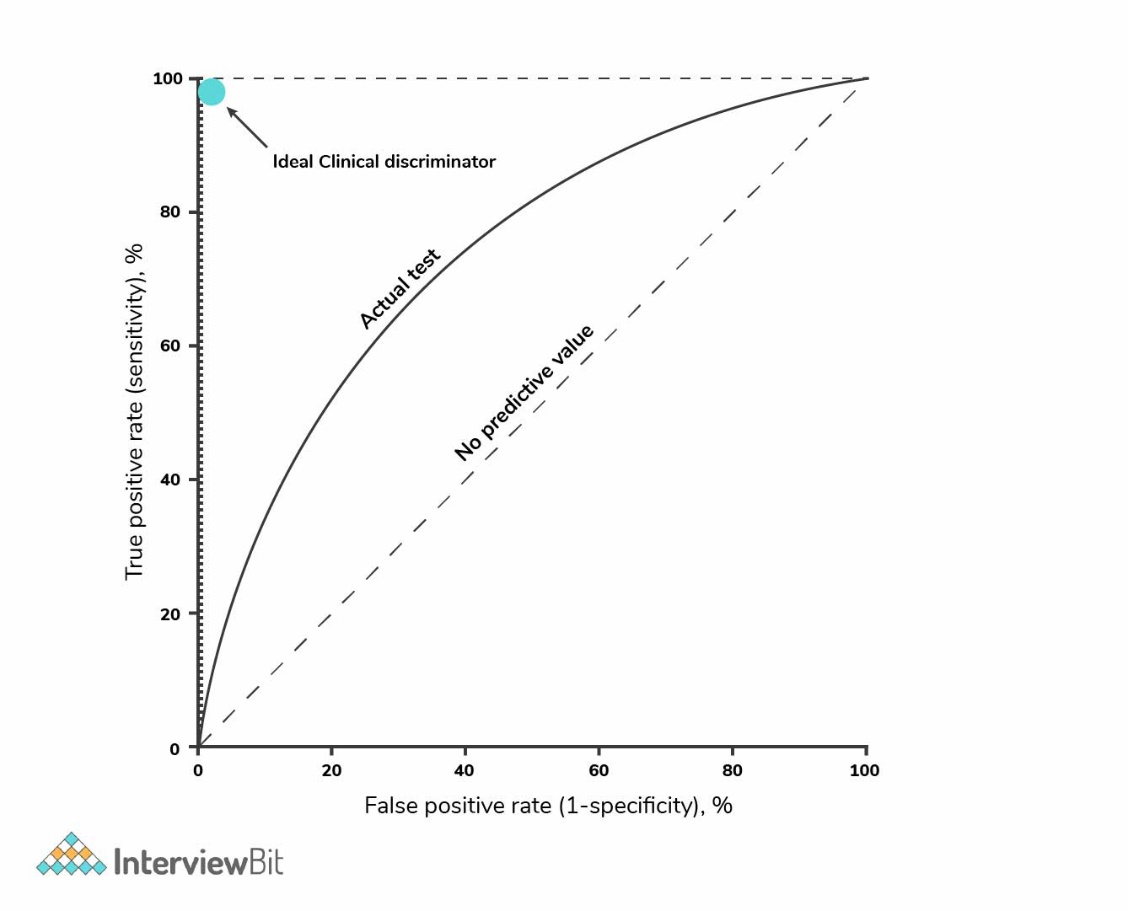
Representing data structures as graphs allow us to discover relationships and patterns which could have been ignored if we model our data around isolated data points.

1. **ROC Curve**

Receiver Operating Characteristic curve is a graphical representation between false-positive rates and true positive rates at different thresholds. It is a proxy for trade-off between sensitivity and specificity.

TPR represents the proportion of observations correctly predicted as positive out of overall positive observations.

FPR represents the proportion of observations incorrectly predicted out of overall negative observations.

****A completely random model is represented by a straight line, having a 0.5 ROC. The amount of deviation a ROC has from this straight line denotes the efficiency of the model.