Company: Google

Role: Data Scientist

**1.Why do you use feature selection?**

We use Feature selection because It is desirable to reduce the number of input variables to both reduce the computational cost of modelling and, in some cases, to improve the performance of the model. Feature selection is the process of reducing the number of input variables when developing a predictive model.

**2.What is the effect on the coefficients of logistic regression if two 3. predictors are highly correlated?**

When predictor variables are correlated, the precision of the estimated regression coefficients decreases as more predictor variables are added to the model.

**3.What are the confidence intervals of the coefficients?**

The coefficient confidence intervals provide a measure of precision for linear regression coefficient estimates. A 100(1–α)% confidence interval gives the range that the corresponding regression coefficient will be in with 100(1–α)% confidence.

**4.What’s the difference between Gaussian Mixture Model and K-Means?**

Gaussian mixture models can be used to cluster unlabelled data in much the same way as k-means. There are, however, a couple of advantages to using Gaussian mixture models over k-means.

First and foremost, k-means does not account for variance. By variance, we are referring to the width of the bell shape curve.

In two dimensions, variance (covariance to be exact) determines the shape of the distribution. One way to think about the k-means model is that it places a circle (or, in higher dimensions, a hyper-sphere) at the center of each cluster, with a radius defined by the most distant point in the cluster. The second difference between k-means and Gaussian mixture models is that the former performs hard classification whereas the latter performs soft classification. In other words, k-means tells us what data point belong to which cluster but won’t provide us with the probabilities that a given data point belongs to each of the possible clusters.

**5.How do you pick k for K-Means?**

The optimal K value usually found is **the square root of N**, where N is the total number of samples. Use an error plot or accuracy plot to find the most favourable K value. KNN performs well with multi-label classes, but you must be aware of the outliers.

**6.How do you know when Gaussian Mixture Model is applicable?**

An approach is **to find the clusters using soft clustering methods** and then see if they are gaussian. If they are then you can apply a GMM model which represents the whole dataset.

**7.Assuming a clustering model’s labels are known, how do you evaluate the performance of the model?**

Three important factors by which clustering can be evaluated are

(a) Clustering tendency (b) Number of clusters, **k**(c) Clustering quality

**Company: Uber**

**Role: Data Scientist**

**1.Pick any product or app that you really like and describe how you would improve it.**

**2.How would you find an anomaly in a distribution?**

The simplest approach to identifying irregularities in data is to flag the data points that deviate from common statistical properties of a distribution, including mean, median, mode, and quantiles.

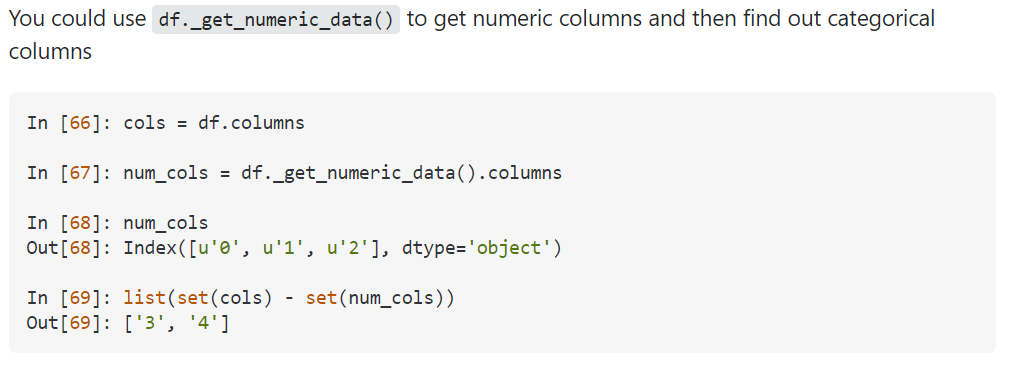
**Company: TCS**

**Role: Data Scientist**

1. Explain about Time series models you have used?

Moving Average (MA) method is the simplest and most basic of all the time series forecasting models. This model is used for a univariate (one variable) time series. In a MA model, the output (or future) variable is assumed to have a linear dependence on the current and past values. Thus, the new series is created from the average of the past values. MA model is suitable for identifying and highlighting trends and trend cycles.

1. SQL Questions - Group by Top 2 Salaries for Employees - use Row num and Partition
2. Pandas find Numeric and Categorical Columns. For Numeric columns in Data frame, find the mean of the entire column and add that mean value to each row of those numeric columns.



Pandas dataframe.mean() function return the mean of the values for the requested axis.

1. What is Gradient Descent? What is Learning Rate and Why we need to reduce or increase?

Gradient Descent is **an optimization algorithm for finding a local minimum of a differentiable function**. Gradient descent is simply used in machine learning to find the values of a function's parameters (coefficients) that minimize a cost function as far as possible.

In machine learning and statistics, the learning rate is a tuning parameter in an optimization algorithm that determines the step size at each iteration while moving toward a minimum of a loss function.

Generally, a large learning rate allows the model to learn faster, at the cost of arriving on a sub-optimal final set of weights. A smaller learning rate may allow the model to learn a more optimal or even globally optimal set of weights but may take significantly longer to train.

1. What is Log-Loss and ROC-AUC?

AUC - ROC curve is a **performance measurement for the classification problems at various threshold settings**. ROC is a probability curve and AUC represents the degree or measure of separability. It tells how much the model is capable of distinguishing between classes.

Log-loss is **indicative of how close the prediction probability is to the corresponding actual/true value** (0 or 1 in case of binary classification). The more the predicted probability diverges from the actual value, the higher is the log-loss value.

1. What is multi-collinearity? How will you choose one features if there are 2 highly correlated features? Give Examples with the techniques used.

Multicollinearity is the occurrence of high intercorrelations among two or more independent variables in a multiple regression model.

**The potential solutions include the following:**

1. Remove some of the highly correlated independent variables.
2. Linearly combine the independent variables, such as adding them together.
3. Perform an analysis designed for highly correlated variables, such as principal components analysis or partial least squares regression
4. VIF – Variance Inflation Factor – Explain.

Variance inflation factor measures how much the behaviour (variance) of an independent variable is influenced, or inflated, by its interaction/correlation with the other independent variables. Variance inflation factors allow a quick measure of how much a variable is contributing to the standard error in the regression.

1. Do you know to use Amazon Sage Maker for MLOPS?

Amazon Sage Maker **helps data scientists and developers to prepare, build, train, and deploy high-quality machine learning (ML) models quickly** by bringing together a broad set of capabilities purpose-built for ML.

1. Explain your Projects end to end (15-20mins).

Company: Capital One

Role: Data Scientist

1. How would you build a model to predict credit card fraud?

We can perform in 5 steps:

Exploratory Data Analysis.

Train-test split.

Modelling.

Hyperparameter Tuning.

Evaluating Final Model Performance.

1. How do you handle missing or bad data?
2. How would you derive new features from features that already exist?

Binning, (also called banding or discretisation), can be used to create new categorical features that group individuals based on the value ranges of existing features. You can use binning to create new target features you want to predict or new input features.

1. If you’re attempting to predict a customer’s gender, and you only have 100 data points, what problems could arise?

Overfitting. We might learn too much into some particular patterns within this small sample set so we lose generalization abilities on other datasets.

1. Suppose you were given two years of transaction history. What features would you use to predict credit risk?

* Transaction amount,
* Transaction count,
* Transaction frequency,
* transaction category: bar, grocery, jwery etc.
* transaction channels: credit card, debit card, international wire transfer etc.
* distance between transaction address and mailing address,
* fraud/ risk score.

1. Design an AI program for Tic-tac-toe
2. Explain overfitting and what steps you can take to prevent it.
3. Why does SVM need to maximize the margin between support vectors?

Maximizing the margin seems good because **points near the decision surface represent very uncertain classification decisions**: there is almost a 50% chance of the classifier deciding either way. By construction, an SVM classifier insists on a large margin around the decision boundary.

Company: Latentview Analytics

Role: Data Scientist

Experience: 2 years

1. What is mean and median

Mean

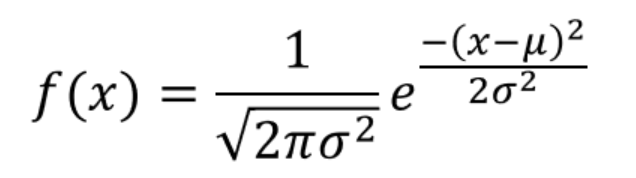
The mean value is the average value. To calculate the mean, find the sum of all values, and divide the sum by the number of values:

Median

The median value is the value in the middle, after you have sorted all the values:

1. Difference between normal and gaussian distribution

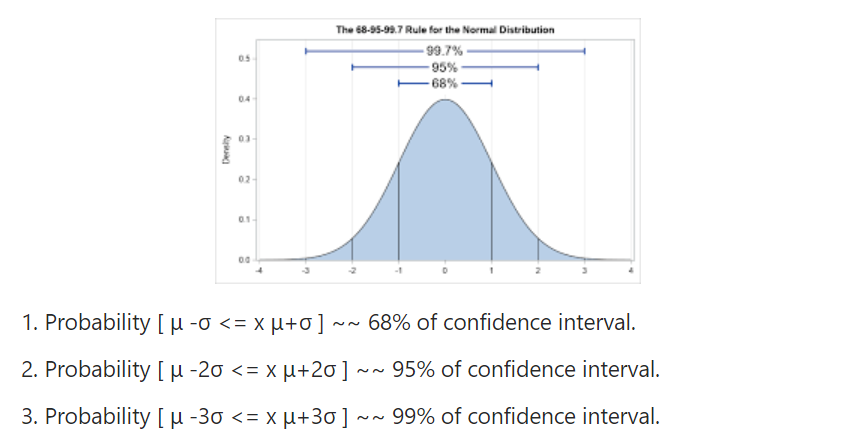
A gaussian and normal distribution is the same in statistics theory. Gaussian distribution is also known as a normal distribution. The curve is made with the help of probability density function with the random values. F(x) is the PDF function and x is the value of gaussian & used to represent the real values of random variables having unknown distribution. The normal distribution contains the curve between the x values and corresponding to the y values but the gaussian distribution made the curve with the x random variables and corresponding the PDF values.



Where σ is the standard deviation and μ is the mean value and x is the random variables. It is also known as the Normal distribution with mean 0 and the standard deviation is 1.

There is no difference between Gaussian and Normal distribution. We can find the p-value and z-value with the help of normal distribution.

Gaussian distribution has an empirical property that checks that in which confidence interval the data points come into these are called as an empirical property of the normal distribution:



It is used for continuous values of the unknown distribution. The mean, median and mode are the same in any distribution when we plot a gaussian bell curve for that distribution.

1. What is central limit theorem

The central limit theorem (CLT) is simple. It just says that with a large sample size, sample means are normally distributed. ... Putting it all together the CLT just says that when you have roughly 30 or more observations in your sample, the average of those numbers is part of a bell-shaped curve.

1. What is null hypothesis

A null hypothesis is a type of hypothesis used in statistics that proposes that there is no difference between certain characteristics of a population (or data-generating process)

1. What is covariance and correlation and how will you interpret it.

In simple words, both the terms measure the relationship and the dependency between two variables. “Covariance” indicates the direction of the linear relationship between variables. “Correlation” on the other hand measures both the strength and direction of the linear relationship between two variables.

1. How will you find out the outliers in the dataset and is it always to remove outliers?
2. Explain about Machine Learning

Machine learning is a branch of artificial intelligence (AI) and computer science which focuses on the use of data and algorithms to imitate the way that humans learn, gradually improving its accuracy.

1. Explain the algorithm of your choice
2. Different methods of missing values imputation

* Mean imputation. Simply calculate the mean of the observed values for that variable for all individuals who are non-missing.
* Substitution.
* Hot deck imputation.
* Cold deck imputation.
* Regression imputation.
* Stochastic regression imputation.
* Interpolation and extrapolation.

Company: Verizon

Role: Data Scientist

1. How many cars are there in Chennai? How do u structurally approach coming up with that number?

As a general framework for the problems of this nature, this is the approach one must take. This is a market-estimate/ sizing problem and a very popular one at that in consulting interviews. Some other variants of the problem are

a. Estimate the number of cricket bats in a country / city

b. Estimate the number of wheels in a city

c. Estimate the size of Wine market in a city/state

Potential Catch

a. Avoid bottom-up approaches for these problems.

b. Try not to rush into micro segments to soon.

c. Identify and choose your segments carefully.

d. Make your assumptions based on real life estimates.

General Framework for such market sizing problems

1. Start with the total population of the city.

2. Divide the population based on Genders. Male and Female.

3. Then make various segments and allocate proportionally.

Either go with this Segmentation and further drill down into

Middle class 2. Upper Middle Class 3. Lower Middle.

OR

Further think of micro segments to divide the car users or usage of the cars.

Public transport users, Income groups, Taxis, Private vehicles,

Families with more than one vehicle, Cars sold and not sold.

Cars registered in city. Cars from outside, Cars in govt duty etc.

Kindly Note:

These problems are meant to see your structured thinking and problem-solving approaches. You must show your working very clearly.

Always be confident about the nature of the problem. Read and read again to be sure that its a market sizing problem.

Must give a number in the end carefully calculated based on the assumptions made at every step.

1. Multiple Linear Regression?

Multiple linear regression (MLR), also known simply as multiple regression, is a statistical technique that uses several explanatory variables to predict the outcome of a response variable. The goal of multiple linear regression (MLR) is to model the linear relationship between the explanatory (independent) variables and response (dependent) variable.

1. OLS vs MLE?

“OLS” stands for “ordinary least squares” while “**MLE**” stands for “maximum likelihood estimation.” Maximum likelihood estimation, or MLE, is a method used in estimating the parameters of a statistical model and for fitting a statistical model to data. The ordinary least squares, or OLS, can also be called the linear least squares. This is a method for approximately determining the unknown parameters located in a linear regression model.

1. R2 vs Adjusted R2? During Model Development which one do we consider?

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. ... 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. Lift chart, drift chart

A lift chart graphically represents the improvement that a mining model provides when compared against a random guess, and measures the change in terms of a lift score. By comparing the lift scores for different models, you can determine which model is best. You can also determine the point at which the model's predictions become less useful. For example, by reviewing the lift chart, you might realize that a promotional campaign is likely to be effective against only 30% of your customers, and use that figure to limit the scope of the campaign.

The Drift Charts window is consisted from three diagrams. The first two are the raw data / time series for the channels in which the Drift Analysis is performed. The third diagram displays the actual Drift Time Series. The red horizontal lines are specifying the defined Threshold.

1. Sigmoid Function in Logistic regression

**In order to map predicted values to probabilities**, we use the Sigmoid function. The function maps any real value into another value between 0 and 1. In machine learning, we use sigmoid to map predictions to probabilities.

1. ROC what is it? AUC and Differentiation?

AUC - ROC curve is a performance measurement for the classification problems at various threshold settings. ROC is a probability curve and AUC represents the degree or measure of separability. In Machine Learning, performance measurement is an essential task. So, when it comes to a classification problem, we can count on an AUC - ROC Curve. When we need to check or visualize the performance of the multi - class classification problem, we use AUC (Area Under the Curve) ROC (Receiver Operating Characteristics) curve. It is one of the most important evaluation metrics for checking any classification model’s performance. It is also written as AUROC (Area Under the Receiver Operating Characteristics).

An excellent model has AUC near to the 1 which means it has good measure of separability. A poor model has AUC near to the 0 which means it has worst measure of separability. In fact, it means it is reciprocating the result. It is predicting 0s as 1s and 1s as 0s. And when AUC is 0.5, it means model has no class separation capacity whatsoever.

1. Linear Regression from Multiple Linear Regression

Multiple linear regression (MLR), also known simply as multiple regression, is a statistical technique that uses several explanatory variables to predict the outcome of a response variable. Multiple regression is an extension of linear (OLS) regression that uses just one explanatory variable.

1. P-Value what is it and its significance? What does P in P-Value stand for? What is Hypothesis Testing? Null hypothesis vs Alternate Hypothesis?

The p-value is the probability that the null hypothesis is true. (1 – the p-value) is the probability that the alternative hypothesis is true. A low p-value shows that the results are replicable. A low p-value shows that the effect is large or that the result is of major theoretical, clinical or practical importance.

n statistics, the p-value is the probability of obtaining results at least as extreme as the observed results of a statistical hypothesis test, assuming that the null hypothesis is correct. A smaller p-value means that there is stronger evidence in favour of the alternative hypothesis.

H0: The null hypothesis: It is a statement of no difference between sample means or proportions or no difference between a sample mean or proportion and a population mean or proportion. In other words, the difference equals 0.

Ha: The alternative hypothesis: It is a claim about the population that is contradictory to H0 and what we conclude when we reject H0.Since the null and alternative hypotheses are contradictory, you must examine evidence to decide if you have enough evidence to reject the null hypothesis or not. The evidence is in the form of sample data.After you have determined which hypothesis the sample supports, you make a decision. There are two options for a decision. They are “reject H0” if the sample information favours the alternative hypothesis or “do not reject H0” or “decline to reject H0” if the sample information is insufficient to reject the null hypothesis.

1. Bias Variance Trade off?

Bias is the simplifying assumptions made by the model to make the target function easier to approximate. Variance is the amount that the estimate of the target function will change given different training data. Trade-off is tension between the error introduced by the bias and the variance.

1. Over fitting vs Underfitting in Machine learning?

In statistics and machine learning, one of the most common tasks is to fit a model to a set of training data, so as to be able to make reliable predictions on general untrained data.

In overfitting, a statistical model describes random error or noise instead of the underlying relationship. Overfitting occurs when a model is excessively complex, such as having too many parameters relative to the number of observations. A model that has been overfit has poor predictive performance, as it overreacts to minor fluctuations in the training data.

Underfitting occurs when a statistical model or machine learning algorithm cannot capture the underlying trend of the data. Underfitting would occur, for example, when fitting a linear model to non-linear data. Such a model too would have poor predictive performance.

1. Estimation of Multiple Linear Regression

The least squares method is the most widely used procedure for developing estimates of the model parameters. For simple linear regression, the least squares estimate of the model parameters β0 and β1 are denoted b0 and b1. Using these estimates, an estimated regression equation is constructed: ŷ = b0 + b1x .

1. Forecasting vs Prediction difference? Regression vs Time Series?

Prediction is concerned with estimating the outcomes for unseen data. Forecasting is a sub-discipline of prediction in which we are making predictions about the future, on the basis of time-series data. Thus, the only difference between prediction and forecasting is that we consider the temporal dimension

A regression will analyze the mean of the dependent variable in relation to changes in the independent variables. Time Series: A time series measures data over a specific period of time. Data points will typically be plotted in charts for further analysis.

1. p,d,q values in ARIMA models

A nonseasonal ARIMA model is classified as an "ARIMA(p,d,q)" model, where: p is the number of autoregressive terms, d is the number of nonseasonal differences needed for stationarity, and. q is the number of lagged forecast errors in the prediction equation.

Company: Fractal

Role: Data Scientist

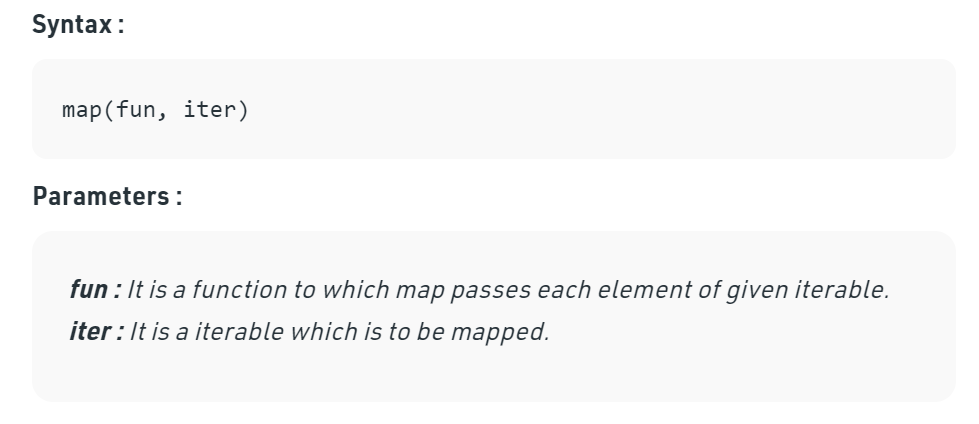
1.Difference between array and list

A list in Python is a collection of items which can contain elements of multiple data types, which may be either numeric, character logical values, etc. It is an ordered collection supporting negative indexing. A list can be created using [] containing data values.

An array is a vector containing homogeneous elements i.e. belonging to the same data type. Elements are allocated with contiguous memory locations allowing easy modification, that is, addition, deletion, accessing of elements. In Python, we have to use the array module to declare arrays. If the elements of an array belong to different data types, an exception “Incompatible data types” is thrown.

2.Map function

map () function returns a map object (which is an iterator) of the results after applying the given function to each item of a given iterable (list, tuple etc.)



3.Scenario, if coupon distributed randomly to customers of swiggy, how to check there buying behaviour. Use segmenting customers Compare customers who got coupon and who did not

4.Which is faster dictionary or list for look up

The reason is because a dictionary is a lookup, while a list is an iteration.

Dictionary uses a hash lookup, while your list requires walking through the list until it finds the result from beginning to the result each time.

to put it another way. The list will be faster than the dictionary on the first item, because there's nothing to look up. it's the first item, boom. it's done. but the second time the list has to look through the first item, then the second item. The third time through it has to look through the first item, then the second item, then the third item.etc.So each iteration the lookup takes more and more time. The larger the list, the longer it takes. While the dictionary is always a more or less fixed lookup time (it also increases as the dictionary gets larger, but at a much slower pace, so by comparison it's almost fixed).

5.How to merge two arrays

You can use either the spread operator [...array1, ...array2], or a functional way []. concat(array1, array2) to merge 2 or more arrays. These approaches are immutable because the merge result is stored in a new array.

If you’d like to perform a mutable merge, i.e. merge into an array without creating a new one, then you can use array1.push(...array2) approach.

6How much time svm takes to complete if 1 iteration takes 10sec for 1st class. And there are 4 classes.

7.Kernals in svm, there difference

**Company name: Infosys**

**Role: Data scientist**

1) curse of dimensionality? How would you handle it?

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 is referred to as ‘Curse of Dimensionality’. The popular aspects of the curse of dimensionality; ‘data sparsity’ and ‘distance concentration’ are discussed in the following sections.

To overcome the issue of the curse of dimensionality, Dimensionality Reduction is used to reduce the feature space with consideration by a set of principal features.

2) How to find the multi collinearity in the data set

Multicollinearity can be detected via various methods. In this article, we will focus on the most common one – VIF (Variable Inflation Factors).” VIF determines the strength of the correlation between the independent variables. It is predicted by taking a variable and regressing it against every other variable.

3)Explain the difference ways to treat multi collinearity!

Multicollinearity can be detected via various methods. In this article, we will focus on the most common one – VIF (Variable Inflation Factors). ” VIF determines the strength of the correlation between the independent variables. It is predicted by taking a variable and regressing it against every other variable.

4) How you decide which feature to keep and which feature to eliminate after performing multi collinearity test?

5)Explain logistic regression

Logistic regression is a statistical analysis method used to predict a data value based on prior observations of a data set. A logistic regression model predicts a dependent data variable by analysing the relationship between one or more existing independent variables.

6)we have sigmoid function which gives us the probability between 0-1 then what is the need of logloss in logistic regression?

Log Loss is the most important classification metric based on probabilities. It’s hard to interpret raw log-loss values, but log-loss is still a good metric for comparing models. For any given problem, a lower log loss value means better predictions.

7) P value and its significance in statistical testing?

In statistics, the p-value is the probability of obtaining results at least as extreme as the observed results of a statistical hypothesis test, assuming that the null hypothesis is correct. A smaller p-value means that there is stronger evidence in favour of the alternative hypothesis.

8) How do you split the time series data and evaluation metrics for time series data

9) How did you deploy your model in production? How often do you retrain it?

Company: Wipro

Role: Data Scientist

1. Difference between WHERE and HAVING in SQL

WHERE Clause is used to filter the records from the table based on the specified condition.

HAVING Clause is used to filter record from the groups based on the specified condition.

1. Basics of Logistics Regression

Logistic regression is a statistical analysis method used to predict a data value based on prior observations of a data set. Logistic regression has become an important tool in the discipline of machine learning. The approach allows an algorithm being used in a machine learning application to classify incoming data based on historical data. As more relevant data comes in, the algorithm should get better at predicting classifications within data sets. Logistic regression can also play a role in data preparation activities by allowing data sets to be put into specifically predefined buckets during the extract, transform, load (ETL) process in order to stage the information for analysis.A logistic regression model predicts a dependent data variable by analyzing the relationship between one or more existing independent variables. For example, a logistic regression could be used to predict whether a political candidate will win or lose an election or whether a high school student will be admitted to a particular college.

The resulting analytical model can take into consideration multiple input criteria. In the case of college acceptance, the model could consider factors such as the student’s grade point average, SAT score and number of extracurricular activities. Based on historical data about earlier outcomes involving the same input criteria, it then scores new cases on their probability of falling into a particular outcome category.

1. How do you treat outliers?
2. Explain confusion matrix?

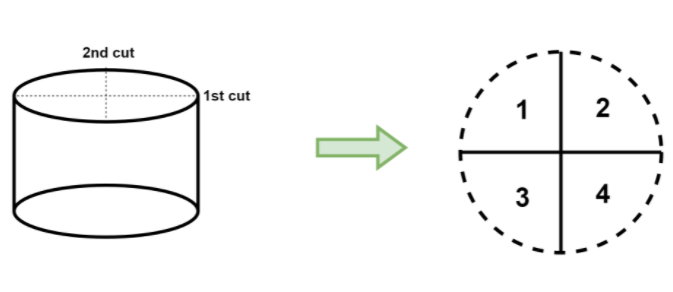
A confusion matrix is a summary of prediction results on a classification problem. The number of correct and incorrect predictions are summarized with count values and broken down by each class. This is the key to the confusion matrix. The confusion matrix shows the ways in which your classification model.

1. Explain PCA (Wanted me to explain the co-variance matrix and eigen vectors and values and the mathematical expression and mathematical derivation for co-variance matrix)

So, in order to identify these correlations, we compute the covariance matrix. The covariance matrix is a p × p symmetric matrix (where p is the number of dimensions) that has as entries the covariances associated with all possible pairs of the initial variables. they provide an estimate of the variance in individual random variables and also measure whether variables are correlated.

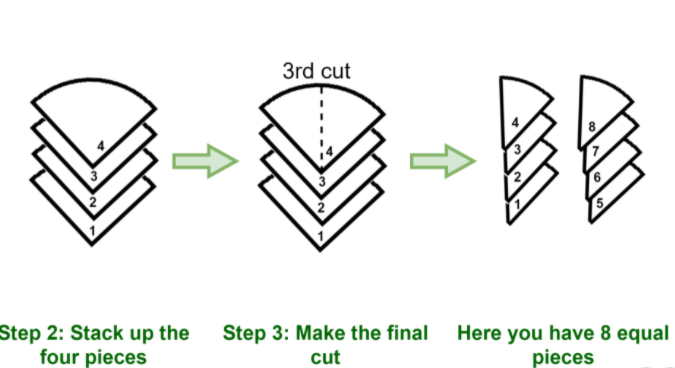
1. How do you cut a cake into 8 equal parts using only 3 straight cuts?

Step 1: Cut the cake into quarters (4 pieces) using 2 of the cuts – one horizontally down the centre of the cake and the other vertically down the centre of the cake. This will leave you with 4 pieces (or slices) of cake.



Step 2: Then take all 4 pieces and arrange them in a stack that is 4 pieces high.

Step 3: Finally, you can just cut that stack of 4 pieces in half – using your third and final cut – and then you will end up with 8 pieces of cake!

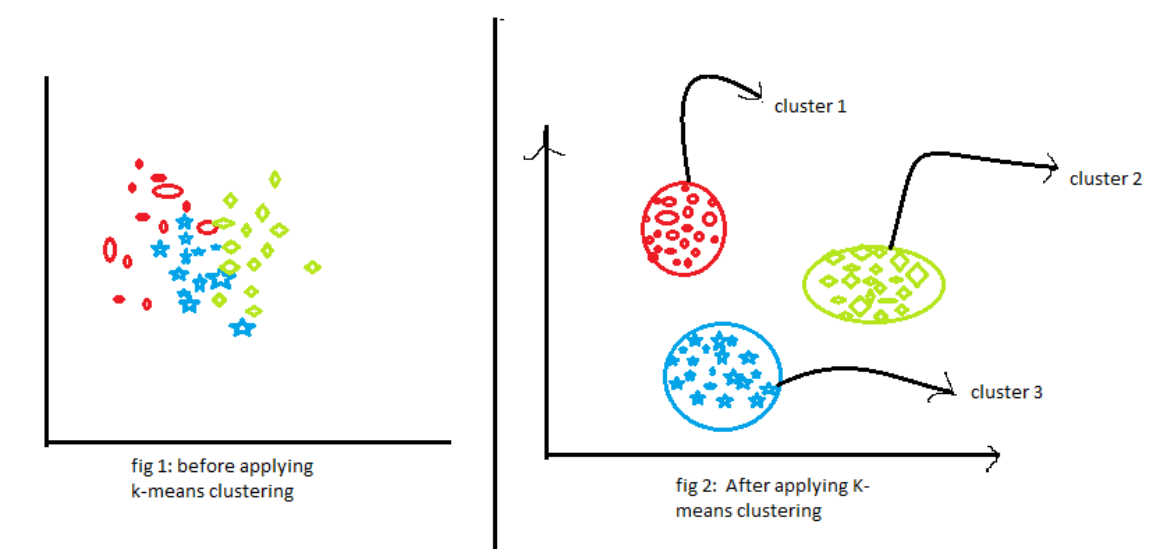


1. Explain kmeans clustering

K-means clustering is a very famous and powerful unsupervised machine learning algorithm. It is used to solve many complex unsupervised machine learning problems. Before we start let’s take a look at the points which we are going to understand.

A K-means clustering algorithm tries to group similar items in the form of clusters. The number of groups is represented by K.

Let’s take an example. Suppose you went to a vegetable shop to buy some vegetables. There you will see different kinds of vegetables. The one thing you will notice there that the vegetables will be arranged in a group of their types. Like all the carrots will be kept in one place, potatoes will be kept with their kinds and so on. If you will notice here then you will find that they are forming a group or cluster, where each of the vegetables is kept within their kind of group forming the clusters.



ow, look at the above two figures. what did you observe? Let us talk about the first figure. The first figure shows the data before applying the k-means clustering algorithm. Here all three different categories are messed up. When you will see such data in the real world, you will not able to figure out the different categories.

Now, look at the second figure(fig 2). This shows the data after applying the K-means clustering algorithm. you can see that all three different items are classified into three different categories which are called clusters.

How Does the K-means clustering algorithm work?

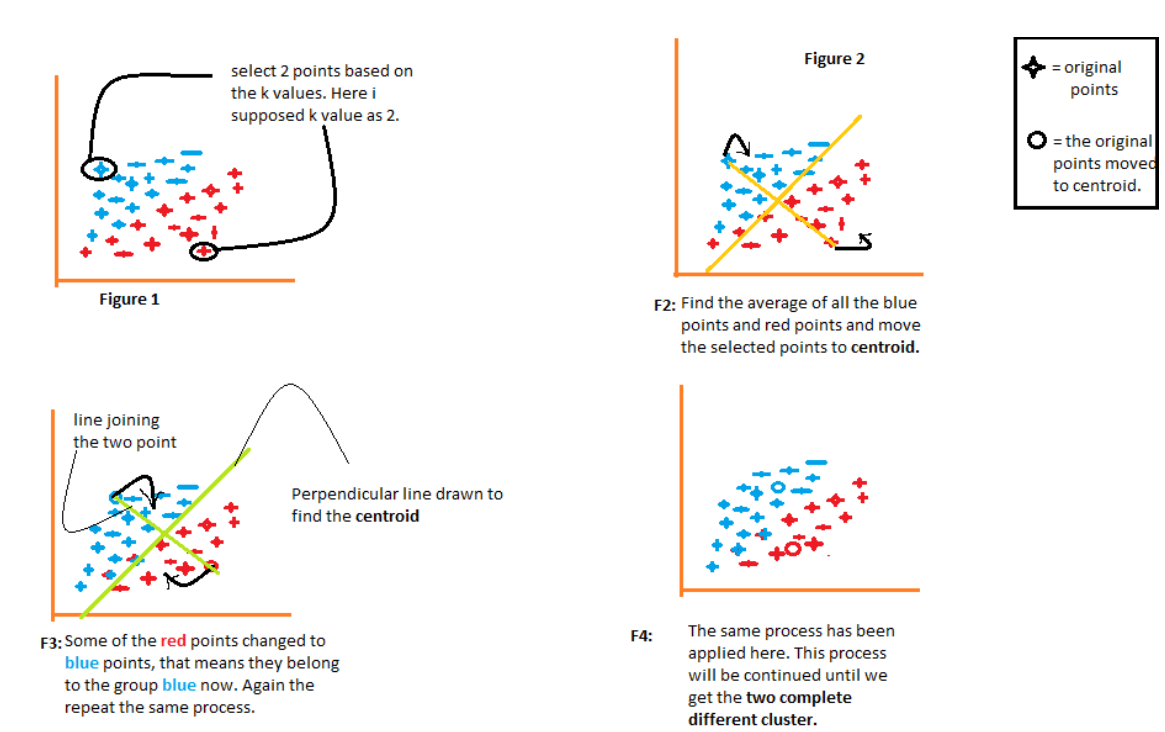
k-means clustering tries to group similar kinds of items in form of clusters. It finds the similarity between the items and groups them into the clusters. K-means clustering algorithm works in three steps. Let’s see what are these three steps.

1.Select the k values.

2.Initialize the centroids.

Select the group and find the average.

Let us understand the above steps with the help of the figure because a good picture is better than the thousands of words.



We will understand each figure one by one.

* Figure 1 shows the representation of data of two different items. the first item has shown in blue color and the second item has shown in red color. Here I am choosing the value of K randomly as 2. There are different methods by which we can choose the right k values.
* In figure 2, Join the two selected points. Now to find out centroid, we will draw a perpendicular line to that line. The points will move to their centroid. If you will notice there, then you will see that some of the red points are now moved to the blue points. Now, these points belong to the group of blue color items.
* The same process will continue in figure 3. we will join the two points and draw a perpendicular line to that and find out the centroid. Now the two points will move to its centroid and again some of the red points get converted to blue points.
* The same process is happening in figure 4. This process will be continued until and unless we get two completely different clusters of these groups.

1. How is KNN different from k-means clustering?

K-means clustering represents an unsupervised algorithm, mainly used for clustering, while KNN is a supervised learning algorithm used for classification.

1. What would be your strategy to handle a situation indicating an imbalanced dataset?

**7 Techniques to Handle Imbalanced Data**

1. Use the right evaluation metrics.
2. Resample the training set. ...
3. Use K-fold Cross-Validation in the right way.
4. Ensemble different resampled datasets.
5. Resample with different ratios.
6. Cluster the abundant class.
7. Design your own models.

10.Stock market prediction: You would like to predict whether or not a certain company will declare bankruptcy within the next 7 days (by training on data of similar companies that had previously been at risk of bankruptcy). Would you treat this as a classification or a regression problem?

Classification

Company: Accenture

Role: Data Scientist

1. What is difference between K-NN and K-Means clustering?

K-means clustering represents an unsupervised algorithm, mainly used for clustering, while KNN is a supervised learning algorithm used for classification.

1. How to handle missing data? What imputation techniques can be used?
2. Explain topic modelling in NLP and various methods in performing topic modeling.

Topic modelling is a method in natural language processing (NLP) used to train machine learning models. It refers to the process of logically selecting words that belong to a certain topic from within a document. From a business standpoint, topic modeling provides great time- and effort-saving benefits.

**The three most common techniques of topic modeling are:**

1.Latent Semantic Analysis (LSA) Latent semantic analysis (LSA) aims to leverage the context around the words in order to capture hidden concepts or topics. ...

2.Probabilistic Latent Semantic Analysis (pLSA)

3.Latent Dirichlet Allocation (LDA)

1. Explain how you would find and tackle an outlier in the dataset.
2. Follow up: What about inlier?
3. Explain back propagation in few words and its variants?

Backpropagation is an essential mechanism by which neural networks get trained. It is a mechanism used to fine-tune the weights of a neural network (otherwise referred to as a model in this article) in regards to the error rate produced in the previous iteration. It is similar to a messenger telling the model if the net made a mistake or not as soon as it predicted.

1. Is interpretability important for machine learning model? If so, ways to achieve interpretability for a machine learning models?

Interpretability is important to different people for different reasons: Data scientists want to build models with high accuracy. They want to understand the details to find out how they can pick the best model and improve that model.

Fairness: if we ensure our predictions are unbiased, we prevent discrimination against under-represented groups.

Robustness: we need to be confident the model works in every setting, and that small changes in input don't cause large or unexpected changes in output.

1. Is interpretability important for machine learning model? If so, ways to achieve interpretability for a machine learning models?
2. How would you design a data science pipeline?

Generally, the primary processes of a data science pipeline are:

* Data engineering (including collection, cleansing, and preparation)
* Machine learning (model learning and model validation)
* Output (model deployment and data visualization)

But the first step in deploying a data science pipeline is identifying the business problem you need the data to address and the data science workflow. Formulate questions you need answers to — that will direct the machine learning and other algorithms to provide solutions you can use.

Once that’s done, the steps for a data science pipeline are:

1. Data collection, including the identification of data sources and extraction of data from sources into usable formats
2. Data preparation, which may include ETL
3. Data modeling and model validation, in which machine learning is used to find patterns and apply rules to the data via algorithms and then tested on sample data
4. Model deployment, applying the model to the existing and new data
5. Reviewing and updating the model based on changing business requirements
6. Explain bias - variance trade off. How does this affect the model?

In statistics and machine learning, the bias–variance trade-off is the property of a model that the variance of the parameter estimates across samples can be reduced by increasing the bias in the estimated parameters.The variance is an error from sensitivity to small fluctuations in the training set.

1. What does a statistical test do?

A statistical test provides a mechanism for making quantitative decisions about a process or processes. The intent is to determine whether there is enough evidence to "reject" a conjecture or hypothesis about the process. The conjecture is called the null hypothesis.

1. How to determine if a coin is biased? Hint: Hypothesis testing

If you were testing H0: coin is fair (p=0.5) against the alternative hypothesis Ha: coin is biased toward tails (p<0.5), you would only reject the null hypothesis in favour of the alternative hypothesis if the number of heads was some number less than 5.

Company: Tiger Analytics

Role: Senior Analyst

1.What is deep learning, and how does it contrast with other machine learning algorithms?

Deep learning is a type of machine learning, which is a subset of artificial intelligence. Machine learning is about computers being able to think and act with less human intervention; deep learning is about computers learning to think using structures modelled on the human brain.

2.When should you use classification over regression?

The main difference between Regression and Classification algorithms that Regression algorithms are used to predict the continuous values such as price, salary, age, etc. and Classification algorithms are used to predict/Classify the discrete values such as Male or Female, True or False, Spam or Not Spam, etc.

3.Using Python how do you find Rank, linear and tensor equations for an given array of elements? Explain your approach.

4.What exactly do you know about Bias-Variance decomposition?

The bias–variance decomposition is a way of analysing a learning algorithm's expected generalization error with respect to a particular problem as a sum of three terms, the bias, variance, and a quantity called the irreducible error, resulting from noise in the problem itself.

5.What is the best recommendation technique you have learnt and what type of recommendation technique helps to predict ratings?

Content-based technique is a domain-dependent algorithm and it emphasizes more on the analysis of the attributes of items in order to generate predictions

Two ways to calculate similarity are Pearson Correlation and Cosine Similarity. Basically, the idea is to find the most similar users to your target user (nearest neighbours) and weight their ratings of an item as the prediction of the rating of this item for target user.

6.How can you assess a good logistic model?

**Measuring the performance of Logistic Regression**

1. One can evaluate it by looking at the confusion matrix and count the misclassifications (when using some probability value as the cut-off) or.
2. One can evaluate it by looking at statistical tests such as the Deviance or individual Z-scores.

7.How to you read the text from an image? Explain?

OCR is a tool to allow computers to recognize the text from physical documents to be interpreted as data.Some OCR programs will add the text recognized from a scanned document as metadata to the file, allowing certain programs to search for the document using any text contained within the document.

8.What are all the options to convert speech to text? Explain and name few available tools to implement the same?

company Name : Tata IQ

Role: Data Analyst

Why data science as a career?

It has already been declared as the hottest job, data scientist brings in skill sets and knowledge from various backgrounds such as mathematics, statistics, Analytics, modeling, and business acumen. These skills help them to identify patterns which can help the organization to recognize new market opportunities.

Stats:

What is p value?

A p-value is a measure of the probability that an observed difference could have occurred just by random chance.

What is histograms?

A histogram is a graphical representation that organizes a group of data points into user-specified ranges. Similar in appearance to a bar graph, the histogram condenses a data series into an easily interpreted visual by taking many data points and grouping them into logical ranges or bins.

What is confidence interval?

A confidence interval, in statistics, refers to the probability that a population parameter will fall between a set of values for a certain proportion of times.

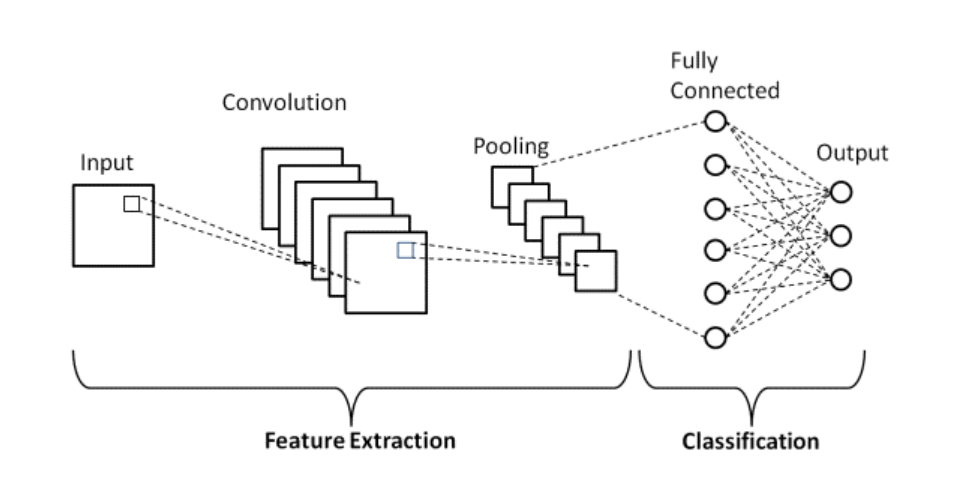
Role: Junior Data Scientist

1) Explain the architecture of CNN

**Basic Architecture**

There are two main parts to a CNN architecture

* A convolution tool that separates and identifies the various features of the image for analysis in a process called as Feature Extraction
* A fully connected layer that utilizes the output from the convolution process and predicts the class of the image based on the features extracted in previous stages.



**Convolution Layers**

There are three types of layers that make up the CNN which are the convolutional layers, pooling layers, and fully-connected (FC) layers. When these layers are stacked, a CNN architecture will be formed. In addition to these three layers, there are two more important parameters which are the dropout layer and the activation function which are defined below.

### 1. Convolutional Layer

This layer is the first layer that is used to extract the various features from the input images. In this layer, the mathematical operation of convolution is performed between the input image and a filter of a particular size MxM. By sliding the filter over the input image, the dot product is taken between the filter and the parts of the input image with respect to the size of the filter (MxM).

The output is termed as the Feature map which gives us information about the image such as the corners and edges. Later, this feature map is fed to other layers to learn several other features of the input image.

### 2. Pooling Layer

In most cases, a Convolutional Layer is followed by a Pooling Layer. The primary aim of this layer is to decrease the size of the convolved feature map to reduce the computational costs. This is performed by decreasing the connections between layers and independently operates on each feature map. Depending upon method used, there are several types of Pooling operations.

In Max Pooling, the largest element is taken from feature map. Average Pooling calculates the average of the elements in a predefined sized Image section. The total sum of the elements in the predefined section is computed in Sum Pooling. The Pooling Layer usually serves as a bridge between the Convolutional Layer and the FC Layer

### 3. Fully Connected Layer

The Fully Connected (FC) layer consists of the weights and biases along with the neurons and is used to connect the neurons between two different layers. These layers are usually placed before the output layer and form the last few layers of a CNN Architecture.

In this, the input image from the previous layers are flattened and fed to the FC layer. The flattened vector then undergoes few more FC layers where the mathematical functions operations usually take place. In this stage, the classification process begins to take place.

### 4. Dropout

Usually, when all the features are connected to the FC layer, it can cause overfitting in the training dataset. Overfitting occurs when a particular model works so well on the training data causing a negative impact in the model’s performance when used on a new data.

To overcome this problem, a dropout layer is utilised wherein a few neurons are dropped from the neural network during training process resulting in reduced size of the model. On passing a dropout of 0.3, 30% of the nodes are dropped out randomly from the neural network.

### 5. Activation Functions

Finally, one of the most important parameters of the CNN model is the activation function. They are used to learn and approximate any kind of continuous and complex relationship between variables of the network. In simple words, it decides which information of the model should fire in the forward direction and which ones should not at the end of the network.

It adds non-linearity to the network. There are several commonly used activation functions such as the ReLU, Softmax, tanH and the Sigmoid functions. Each of these functions have a specific usage. For a binary classification CNN model, sigmoid and softmax functions are preferred an for a multi-class classification, generally softmax us used.

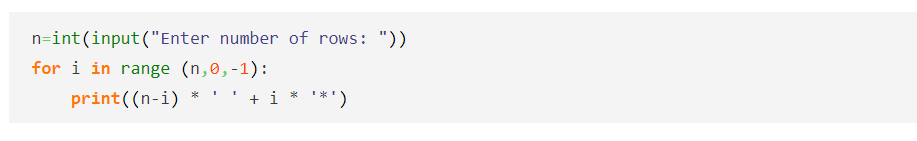
2)If we put a 3×3 filter over 6×6 image what will be the size of the output image

we get 4 x 4 image

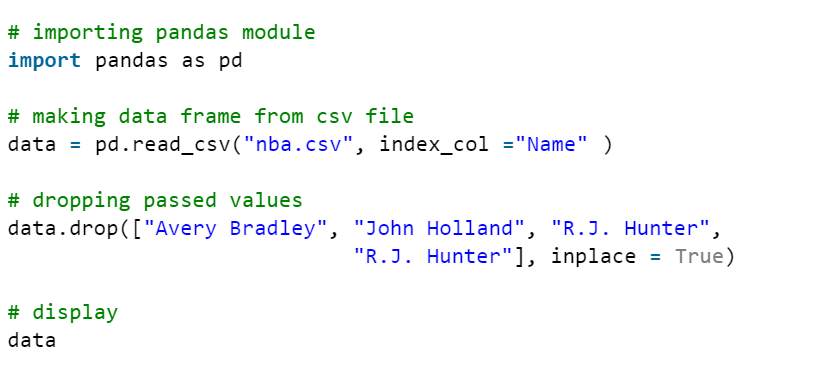
3) What will you do to reduce overfitting In deep learning models

we can reduce the complexity of a neural network to reduce overfitting in one of two ways: Change network complexity by changing the network structure (number of weights). Change network complexity by changing the network parameters (values of weights)

4) Can you write a program for inverted star program in python



5)Write a program to create a dataframe and remove elements from it



Company: Mindtree

Role: Data Scientist

1. What is central tendency

Central tendency is a descriptive summary of a dataset through a single value that reflects the center of the data distribution. Along with the variability (dispersion) of a dataset, central tendency is a branch of descriptive statistics.

1. Which central tendency method is used If there exists any outliers

The median is the most informative measure of central tendency for skewed distributions or distributions with outliers. For example, the median is often used as a measure of central tendency for income distributions, which are generally highly skewed.

1. Central limit theorem

The central limit theorem states that if you have a population with mean μ and standard deviation σ and take sufficiently large random samples from the population with replacement, then the distribution of the sample means will be approximately normally distributed.

1. Chi-Square test

A chi-square test is a statistical test used to compare observed results with expected results. The purpose of this test is to determine if a difference between observed data and expected data is due to chance, or if it is due to a relationship between the variables you are studying.

1. A/B testing

A/B testing is a user experience research methodology. A/B tests consist of a randomized experiment with two variants, A and B. It includes application of statistical hypothesis testing or "two-sample hypothesis testing" as used in the field of statistics.

1. Difference between Z and t distribution (Linked to A/B testing)

Z Test is the statistical hypothesis which is used in order to determine that whether the two samples means calculated are different in case the standard deviation is available and sample is large whereas the T test is used in order to determine a how averages of different data sets differs from each other in case

1. Outlier treatment method

**Some of the most popular methods for outlier detection are:**

1. Z-Score or Extreme Value Analysis (parametric)
2. Probabilistic and Statistical Modeling (parametric)
3. Linear Regression Models (PCA, LMS)
4. Proximity Based Models (non-parametric)
5. Information Theory Models.
6. ANOVA test

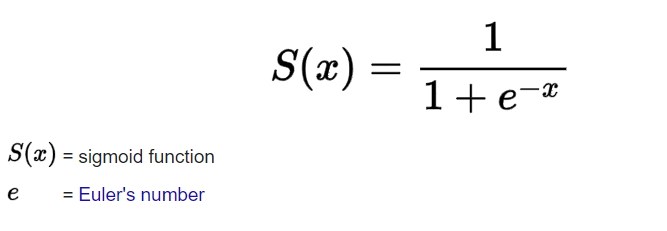
Analysis of variance (ANOVA) is a statistical technique that is used to check if the means of two or more groups are significantly different from each other. ANOVA checks the impact of one or more factors by comparing the means of different samples. Another measure to compare the samples is called a t-test.

1. Cross validation

Cross-validation is a resampling procedure used to evaluate machine learning models on a limited data sample. The procedure has a single parameter called k that refers to the number of groups that a given data sample is to be split into. As such, the procedure is often called k-fold cross-validation.

1. How will you work in a machine learning project if there is a huge imbalance in the data

11.Formula of sigmoid function



12.Can we use sigmoid function in case of multiple classification

We usually use softmax function at the end of the neural network when dealing with multiclass classification to get the output in a probabilistic shape. It's more convenient to see how confident the model is. Yes you can, but i recommend that you use sigmoid when your data can belong to more then 1 class at a time.

13.What is Area under the curve

The Area Under the Curve (AUC) is the measure of the ability of a classifier to distinguish between classes and is used as a summary of the ROC curve. The higher the AUC, the better the performance of the model at distinguishing between the positive and negative classes.

14.Which metric is used to split a node in Decision Tree

In the decision tree chart, each internal node has a decision rule that splits the data. Gini referred to as the Gini ratio, which measures the impurity of the node. You can say a node is pure when all of its records belong to the same class, such nodes known as the leaf node.

**Company: Genpact**

**Role: Data Scientist**

1. Why do we select validation data other than test data?

The validation dataset is different from the test dataset that is also held back from the training of the model, but is instead used to give an unbiased estimate of the skill of the final tuned model when comparing or selecting between final models.

1. Difference between linear logistic regression?

The Differences between Linear Regression and Logistic Regression. Linear Regression is used to handle regression problems whereas Logistic regression is used to handle the classification problems. Linear regression provides a continuous output but Logistic regression provides discreet output.

1. Why do we take such a complex cost function for logistic?

You need a function that measures the performance of a Machine Learning model for given data. Cost Function quantifies the error between predicted values and expected values. `If you can't measure it, you can't improve it.

1. Difference between random forest and decision tree?

Decision trees are very easy as compared to the random forest. A decision tree combines some decisions, whereas a random forest combines several decision trees. Thus, it is a long process, yet slow. Whereas, a decision tree is fast and operates easily on large data sets, especially the linear one.

1. How would you decide when to stop splitting the tree?

Stop splitting the current node if it does not improve the entropy by at least some pre-set(threshold) value. Stop partitioning if the number of datapoints are less then some preset(Threshold) values. Restricting the depth of the tree to some pre-set(Threshold) value.

1. Measures of central tendency

There are three main measures of central tendency: the mode, the median and the mean. Each of these measures describes a different indication of the typical or central value in the distribution.

1. What is the requirement of k means algorithm

Every data point is allocated to each of the clusters through reducing the in-cluster sum of squares. In other words, the K-means algorithm identifies k number of centroids, and then allocates every data point to the nearest cluster, while keeping the centroids as small as possible.

1. Which clustering technique uses combining of clusters

Hierarchical clustering, as the name suggests is an algorithm that builds hierarchy of clusters. This algorithm starts with all the data points assigned to a cluster of their own. Then two nearest clusters are merged into the same cluster.

1. Which is the oldest probability distribution

The binomial distribution is one of the oldest known probability distributions. It was discovered by Bernoulli, J. in his work entitled Ars Conjectandi (1713).

**Company: Ford**

**Role: Data Scientist**

1. How would you check if the model is suffering from multi Collinearity?

The best way to identify the multicollinearity is to calculate the Variance Inflation Factor (VIF) corresponding to every independent Variable in the Dataset. VIF tells us about how well an independent variable is predictable using the other independent variables.

1. What is transfer learning? Steps you would take to perform transfer learning.

Transfer learning is the reuse of a pre-trained model on a new problem. It's currently very popular in deep learning because it can train deep neural networks with comparatively little data. This is very useful in the data science field since most real-world problems typically do not have millions of labelled data points to train such complex models.

You can use transfer learning on your own [predictive modeling](https://machinelearningmastery.com/gentle-introduction-to-predictive-modeling/) problems.

Two common approaches are as follows:

1. Develop Model Approach
2. Pre-trained Model Approach

### Develop Model Approach

1. **Select Source Task**. You must select a related predictive modeling problem with an abundance of data where there is some relationship in the input data, output data, and/or concepts learned during the mapping from input to output data.
2. **Develop Source Model**. Next, you must develop a skillful model for this first task. The model must be better than a naive model to ensure that some feature learning has been performed.
3. **Reuse Model**. The model fit on the source task can then be used as the starting point for a model on the second task of interest. This may involve using all or parts of the model, depending on the modeling technique used.
4. **Tune Model**. Optionally, the model may need to be adapted or refined on the input-output pair data available for the task of interest.

### Pre-trained Model Approach

1. **Select Source Model**. A pre-trained source model is chosen from available models. Many research institutions release models on large and challenging datasets that may be included in the pool of candidate models from which to choose from.
2. **Reuse Model**. The model pre-trained model can then be used as the starting point for a model on the second task of interest. This may involve using all or parts of the model, depending on the modeling technique used.
3. **Tune Model**. Optionally, the model may need to be adapted or refined on the input-output pair data available for the task of interest.

This second type of transfer learning is common in the field of deep learning.

1. Why is CNN architecture suitable for image classification? Not an RNN?

While simple neural networks have some success in classifying basic binary images, they can't handle complex images with pixel dependencies. They also don't have the computational power which is needed to handle images with large pixels, which is exactly where CNNs come in.

1. What are the approaches for solving class imbalance problem?
2. When sampling what types of biases can be inflected? How to control the biases?
3. Explain concepts of epoch, batch, iteration in machine learning.

Iterations is the number of batches of data the algorithm has seen (or simply the number of passes the algorithm has done on the dataset). Epochs is the number of times a learning algorithm sees the complete dataset.

1. What type of performance metrics would you choose to evaluate the different classification models and why?

We can use classification performance metrics such as Log-Loss, Accuracy, AUC(Area under Curve) etc. Another example of metric for evaluation of machine learning algorithms is precision, recall, which can be used for sorting algorithms primarily used by search engines.

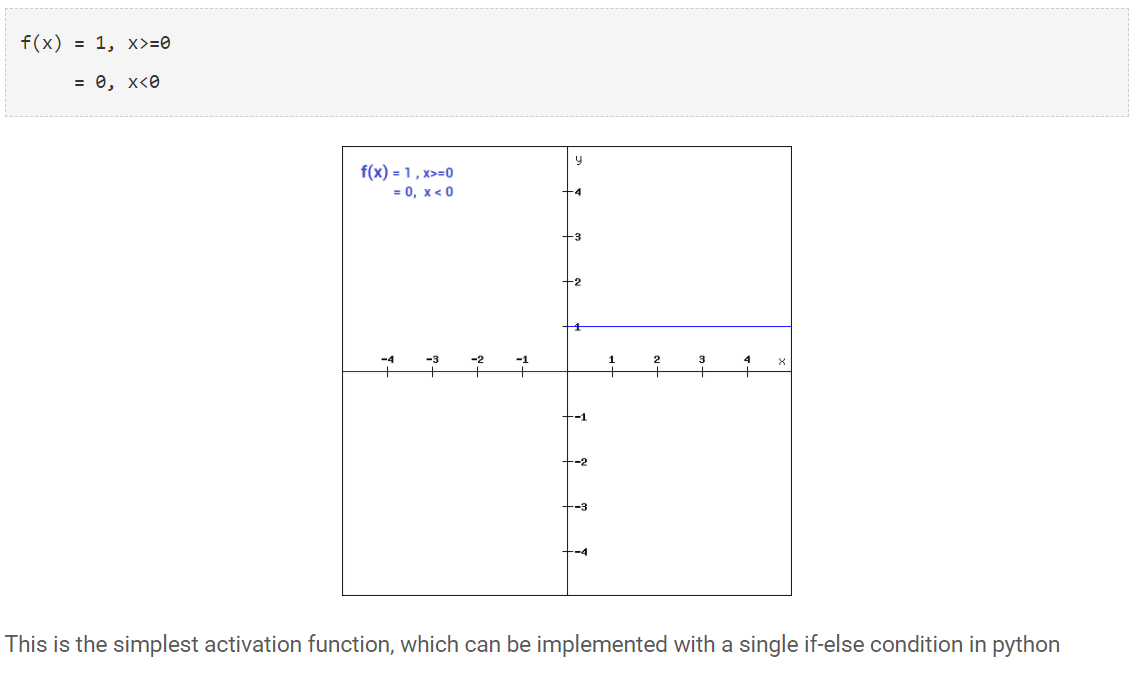
1. What are some of the types of activation functions and specifically when to use them?

## Popular types of activation functions and when to use them

### 1. Binary Step Function

The first thing that comes to our mind when we have an activation function would be a threshold based classifier i.e. whether or not the neuron should be activated based on the value from the linear transformation.

In other words, if the input to the activation function is greater than a threshold, then the neuron is activated, else it is deactivated, i.e. its output is not considered for the next hidden layer. Let us look at it mathematically-

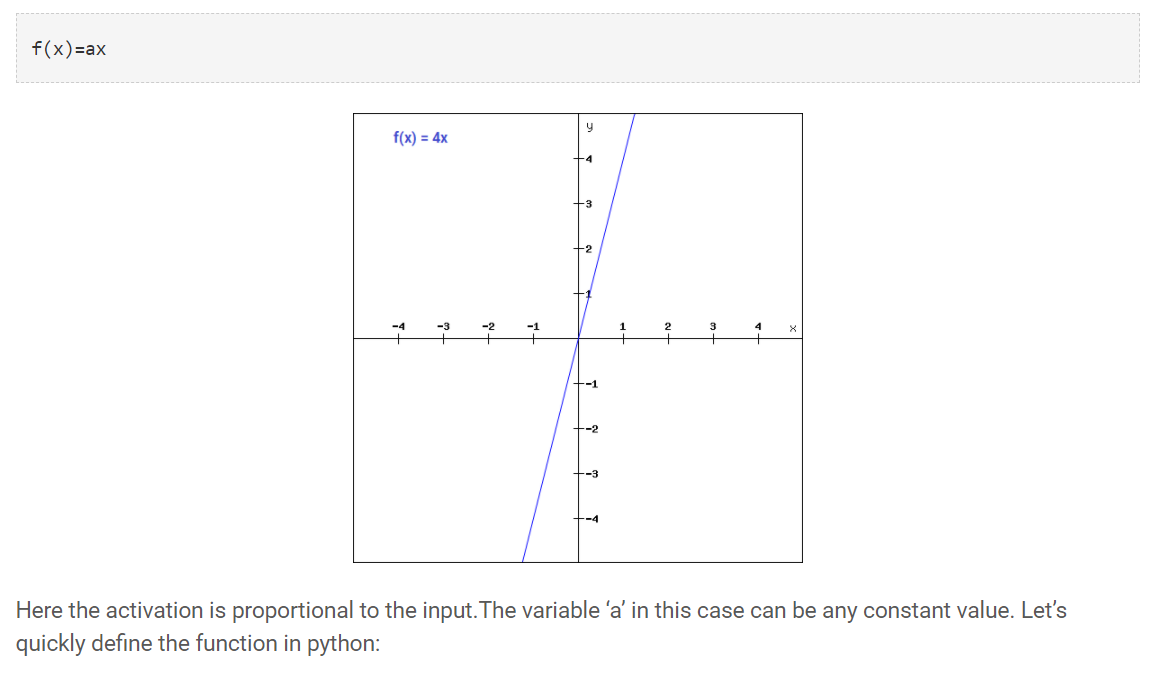


The binary step function can be used as an activation function while creating a binary classifier. As you can imagine, this function will not be useful when there are multiple classes in the target variable. That is one of the limitations of binary step function.

Moreover, the gradient of the step function is zero which causes a hindrance in the back propagation process. That is, if you calculate the derivative of f(x) with respect to x, it comes out to be 0.

### 2. Linear Function

We saw the problem with the step function, the gradient of the function became zero. This is because there is no component of x in the binary step function. Instead of a binary function, we can use a linear function. We can define the function as-

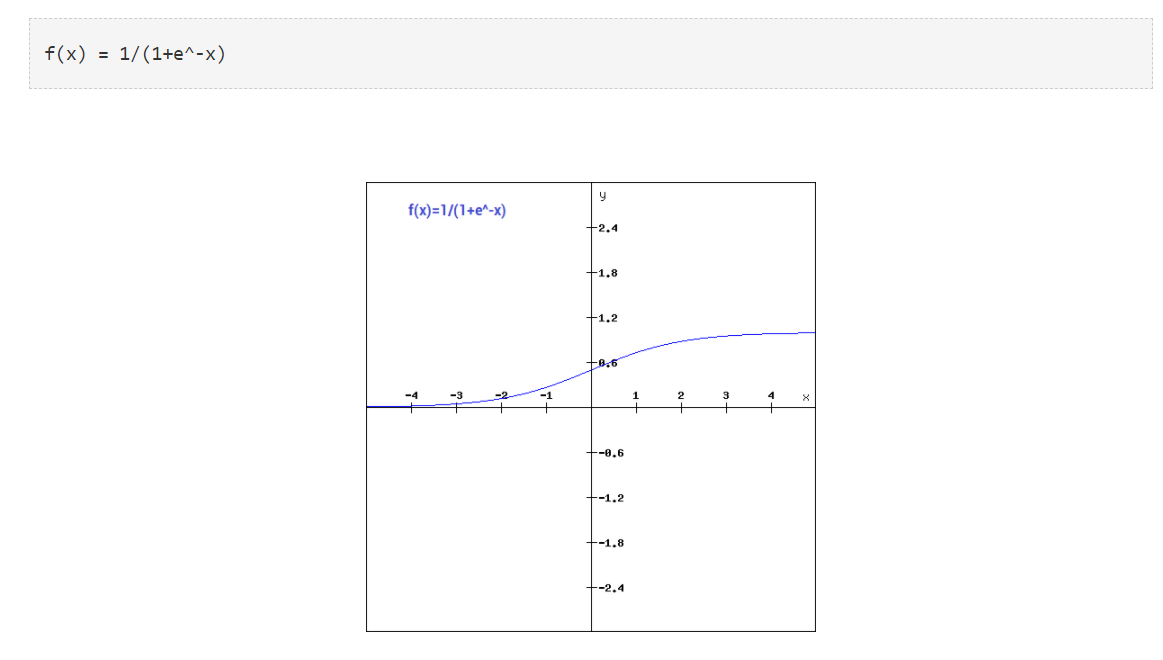


Although the gradient here does not become zero, but it is a constant which does not depend upon the input value x at all. This implies that the weights and biases will be updated during the backpropagation process but the updating factor would be the same.

In this scenario, the neural network will not really improve the error since the gradient is the same for every iteration. The network will not be able to train well and capture the complex patterns from the data. Hence, linear function might be ideal for simple tasks where interpretability is highly desired.

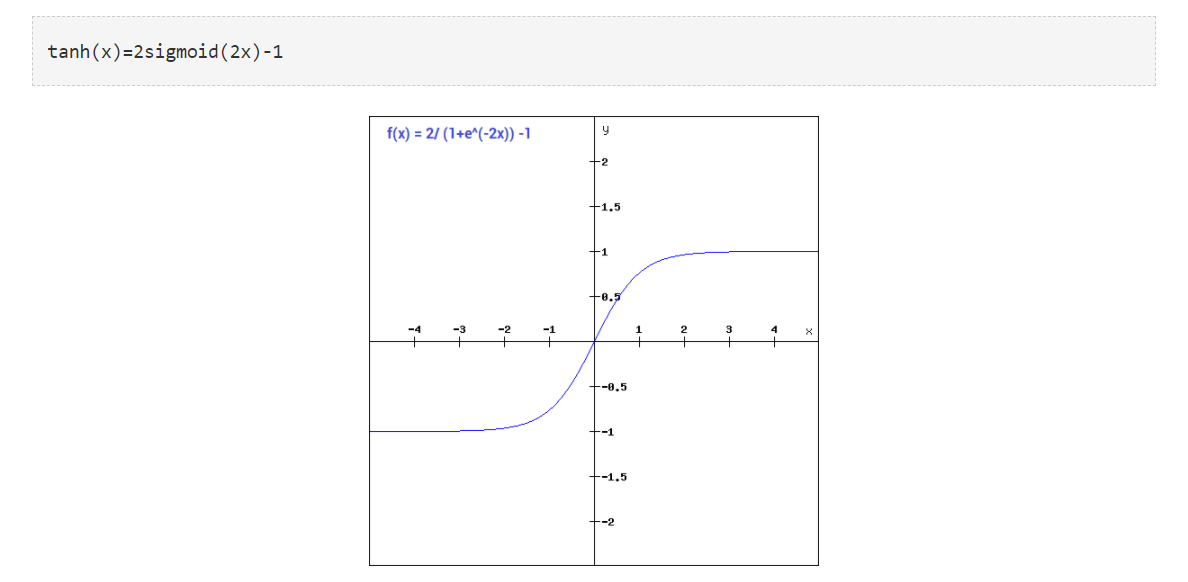
### 3. Sigmoid

The next activation function that we are going to look at is the Sigmoid function. It is one of the most widely used non-linear activation function. Sigmoid transforms the values between the range 0 and 1. Here is the mathematical expression for sigmoid-



### 4. Tanh

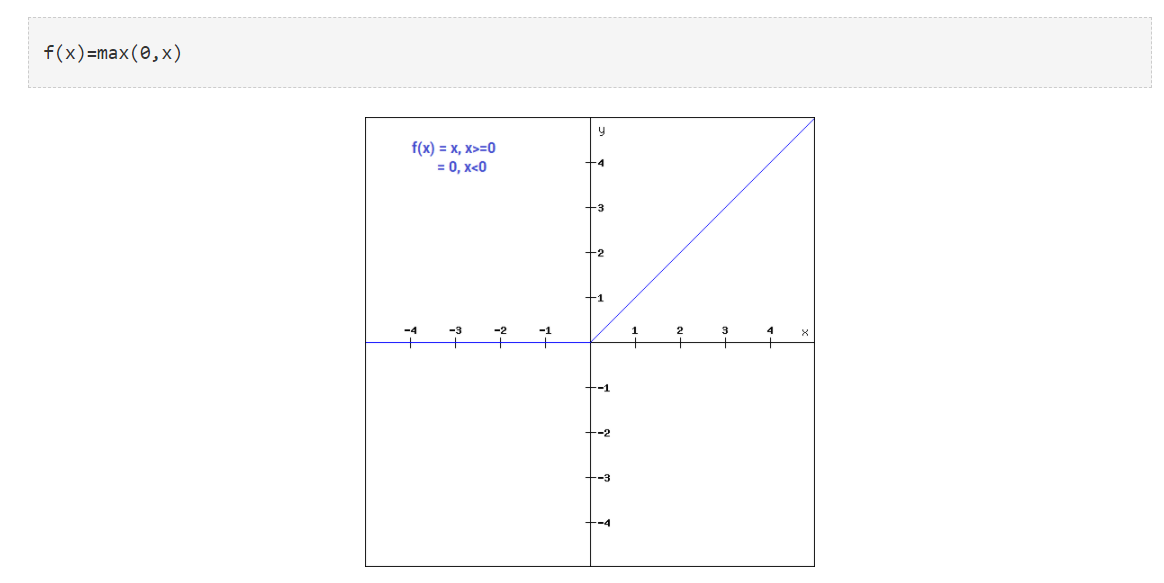
The tanh function is very similar to the sigmoid function. The only difference is that it is symmetric around the origin. The range of values in this case is from -1 to 1. Thus the inputs to the next layers will not always be of the same sign. The tanh function is defined as-



### 5. ReLU

The ReLU function is another non-linear activation function that has gained popularity in the deep learning domain. ReLU stands for Rectified Linear Unit. The main advantage of using the ReLU function over other activation functions is that it does not activate all the neurons at the same time.

This means that the neurons will only be deactivated if the output of the linear transformation is less than 0. The plot below will help you understand this better-

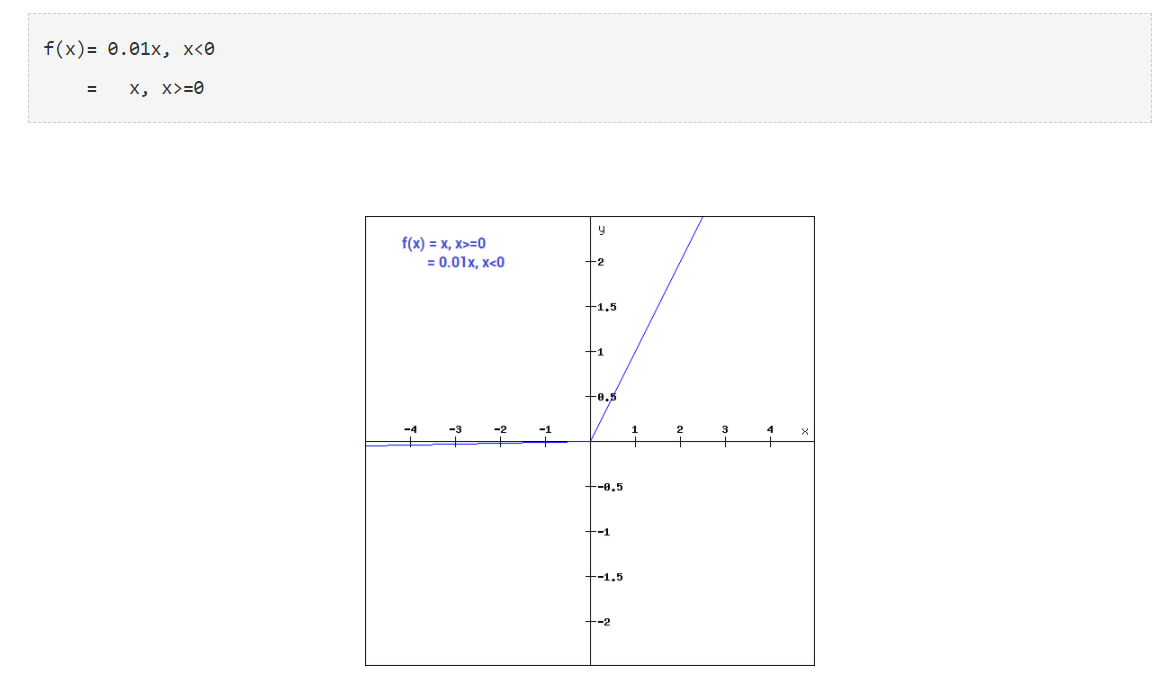


For the negative input values, the result is zero, that means the neuron does not get activated. Since only a certain number of neurons are activated, the ReLU function is far more computationally efficient when compared to the sigmoid and tanh function.

### 6. Leaky ReLU

Leaky ReLU function is nothing but an improved version of the ReLU function. As we saw that for the ReLU function, the gradient is 0 for x<0, which would deactivate the neurons in that region.

Leaky ReLU is defined to address this problem. Instead of defining the Relu function as 0 for negative values of x, we define it as an extremely small linear component of x. Here is the mathematical expression-



1. What are the conditions that should be satisfied for a time series to be stationary?

**When the following conditions are satisfied then a time series is stationary.**

* Mean is constant and does not depend on time.
* Autocovariance function depends on s and t only through their difference |s-t| (where t and s are moments in time)
* The time series under considerations is a finite variance process.

1. What is the difference between Batch and Stochastic Gradient Descent?

Batch Gradient Descent: Batch Gradient Descent involves calculations over the full training set at each step as a result of which it is very slow on very large training data. Thus, it becomes very computationally expensive to do Batch GD. However, this is great for convex or relatively smooth error manifolds. Also, Batch GD scales well with the number of features.

SGD tries to solve the main problem in Batch Gradient descent which is the usage of whole training data to calculate gradients as each step. SGD is stochastic in nature i.e it picks up a “random” instance of training data at each step and then computes the gradient making it much faster as there is much fewer data to manipulate at a single time, unlike Batch GD.

**Company: Quantiphi**

**Role: Machine Learning Engineer**

1. What happens when neural nets are too small?

Their gradient tends to get smaller as we move backward through the hidden layers, which means that neurons in the earlier layers learn much more slowly than neurons in later layers. This causes minor weight updates

1. Why do we need pooling layer in CNN? Common pooling methods?

Pooling layers are used to reduce the dimensions of the feature maps. Thus, it reduces the number of parameters to learn and the amount of computation performed in the network. The pooling layer summarises the features present in a region of the feature map generated by a convolution layer.

Pooling layers provide an approach to down sampling feature maps by summarizing the presence of features in patches of the feature map. Two common pooling methods are average pooling and max pooling that summarize the average presence of a feature and the most activated presence of a feature respectively.

1. Are ensemble models better than individual models? Why/why - not?

Ensembles are used to achieve better predictive performance on a predictive modeling problem than a single predictive model. The way this is achieved can be understood as the model reducing the variance component of the prediction error by adding bias (i.e. in the context of the bias-variance trade-off).

1. In brief, how would you perform the task of sentiment analysis?

Sentiment Analysis is a procedure used to determine if a chunk of text is positive, negative or neutral. In text analytics, natural language processing (NLP) and machine learning (ML) techniques are combined to assign sentiment scores to the topics, categories or entities within a phrase.

**How to Perform Sentiment Analysis?**

1. Step 1: Crawl Tweets Against Hash Tags.
2. Analyzing Tweets for Sentiment.
3. Step 3: Visualizing the Results.
4. Step 1: Training the Classifiers.
5. Step 2: Preprocess Tweets.
6. Step 3: Extract Feature Vectors.
7. How should brands use Sentiment Analysis?

**Company: Cognizant**

**Role: Data Scientist**

1. SQL question on inner join and cross join

Inner Join clause in SQL Server creates a new table (not physical) by combining rows that have matching values in two or more tables. This join is based on a logical relationship (or a common field) between the tables and is used to retrieve data that appears in both tables.

The CROSS JOIN is used to generate a paired combination of each row of the first table with each row of the second table. This join type is also known as cartesian join. Suppose that we are sitting in a coffee shop and we decide to order breakfast.

1. SQL question on group-by

The GROUP BY Statement in SQL is used to arrange identical data into groups with the help of some functions. i.e if a particular column has same values in different rows then it will arrange these rows in a group. Important Points: GROUP BY clause is used with the SELECT statement.

1. What is the difference between gradient and slope, differentiation and integration?

A derivative of a function is a representation of the rate of change of one variable in relation to another at a given point on a function. The slope describes the steepness of a line as a relationship between the change in y-values for a change in the x-values. ... A function's derivative is a function in and of itself.

1. What are vanishing and exploding gradients in neural networks?

In machine learning, the vanishing gradient problem is encountered when training artificial neural networks with gradient-based learning methods and backpropagation. The problem is that in **some cases, the gradient will be vanishingly small**, effectively preventing the weight from changing its value.

Exploding gradients are a problem where large error gradients accumulate and result in very large updates to neural network model weights during training. This has the effect of your model being unstable and unable to learn from your training data.

**Company: Husqvarna Group**

**Role: Data Scientist**

1. Data Pre-Processing Steps used.

To make the process easier, data preprocessing is divided into four stages: data cleaning, data integration, data reduction, and data transformation.

1. Sales forecasting how is it done using Statistical vs DL models – Efficiency

simple moving average, weighted moving average, exponential smoothing, and single regression analysis), the weighted moving average is the most accurate, since specific weights can be placed in accordance with their importance.

1. What are the Evaluation Metric parameters for testing Logistic Regression?

You can evaluate a logistic regression model using accuracy score, which is the overall accuracy of the model. If you want to look at how the classifier does within a certain class (positive and negative prediction power), you may use other metrics such as precision, Recall, confusion matrix, etc.

1. What packages in Python can be used for ML? Why do we prefer one over another?

Skikit-learn was built on top of two Python libraries – NumPy and SciPy and has become the most popular Python machine learning library for developing machine learning algorithms. Scikit-learn has a wide range of supervised and unsupervised learning algorithms that works on a consistent interface in Python.

1. Numpy vs Pandas basic difference.

NumPy library provides objects for multi-dimensional arrays, whereas Pandas is capable of offering an in-memory 2d table object called Data Frame. NumPy consumes less memory as compared to Pandas. Indexing of the Series objects is quite slow as compared to NumPy arrays.

1. Feature on which this Imputation was done, and which method did we use there?
2. Tuple vs Dictionary. Where do we use them?

Tuples are used to store multiple items in a single variable. Tuple is one of 4 built-in data types in Python used to store collections of data, the other 3 are List, Set, and Dictionary, all with different qualities and usage. A tuple is a collection which is ordered and unchangeable.

Dictionary in Python is an ordered collection of data values, used to store data values like a map, which, unlike other Data Types that hold only a single value as an element, Dictionary holds key: value pair. Key-value is provided in the dictionary to make it more optimized.

1. What is NER - Named Entity Recognition?

Named Entity Recognition is the process of NLP which deals with identifying and classifying named entities. The raw and structured text is taken and named entities are classified into persons, organizations, places, money, time, etc.

**Company: Deloitte**

**Role: Data Scientist**

1. Conditional Probability

Conditional probability is defined as the likelihood of an event or outcome occurring, based on the occurrence of a previous event or outcome. Conditional probability is calculated by multiplying the probability of the preceding event by the updated probability of the succeeding, or conditional, event.

1. Can Linear Regression be used for Classification? If Yes, why if No why?

Linear regression is suitable for predicting output that is continuous value, such as predicting the price of a property. Its prediction output can be any real number, range from negative infinity to infinity. Whereas logistic regression is for classification problems, which predicts a probability range between 0 to 1.

1. Hypothesis Testing. Null and Alternate hypothesis

A null hypothesis is a type of conjecture used in statistics that proposes that there is no difference between certain characteristics of a population or data-generating process. The alternative hypothesis proposes that there is a difference.

Hypothesis testing is an act in statistics whereby an analyst tests an assumption regarding a population parameter. The methodology employed by the analyst depends on the nature of the data used and the reason for the analysis.

Hypothesis testing is used to assess the plausibility of a hypothesis by using sample data. Such data may come from a larger population, or from a data-generating process. The word "population" will be used for both of these cases in the following descriptions.

1. Why use Decision Trees?

A decision tree is a decision support tool that uses a tree-like model of decisions and their possible consequences, including chance event outcomes, resource costs, and utility. It is one way to display an algorithm that only contains conditional control statements.

1. PCA Advantages and Disadvantages?

PCA pumps not only control pain but also have other benefits. People feel less anxious and depressed. They are not as sleepy, because they use less medicine. Often they are able to move around more.

Furthermore, if w decreases with non-negligible ratio as z does, then PCA fails to reproduce the original behaviour of w. ... Also, time varying w can be confused with the incorrect value of constant one when the decreasing (or increasing) ratio of w is small but not negligible.

1. What is Naive Bayes Theorem? Multinomial, Bernoulli, Gaussian Naive Bayes.

Naïve Bayes Classifier is one of the simple and most effective Classification algorithms which helps in building the fast machine learning models that can make quick predictions. It is a probabilistic classifier, which means it predicts on the basis of the probability of an object.

Company: Axtria

------------

1.RNN, NN and CNN difference.

The main difference between CNN and RNN is the ability to process temporal information or data that comes in sequences, such as a sentence for example.Whereas, RNNs reuse activation functions from other data points in the sequence to generate the next output in a series.

1. Supervised, unsupervised and reinforcement learning with their algo example.

## Supervised Learning

Consider yourself as a student sitting in a classroom wherein your teacher is supervising you, **“how you can solve the problem”** or **“whether you are doing correctly or not”**. Likewise, in Supervised Learning input is provided as a **labelled dataset,** a model can learn from it to provide the result of the problem easily.

### Types of Problems

Supervised Learning deals with two types of problem- **classification problems** and **regression problems**.

#### **Classification problem**s

This algorithm helps to predict a discrete value. It can be thought, the input data as a member of a particular class or group. For instance, taking up the photos of the fruit dataset, each photo has been labelled as a mango, an apple, etc. Here, the algorithm has to **classify** the new images into any of these categories. Examples:

* Naive Bayes Classifier
* Support Vector Machines
* Logistic Regression

#### Regression problems

These problems are used for continuous data. For example, predicting the price of a piece of land in a city, given the area, location, number of rooms, etc. And then the input is sent to the machine for calculating the price of the land according to previous examples. Examples-

* Linear Regression
* Nonlinear Regression
* Bayesian Linear Regression

## Unsupervised Learning

This learning algorithm is completely opposite to Supervised Learning. In short, there is **no complete and clean labelled dataset in unsupervised learning**. Unsupervised learning is self-organized learning. Its main aim is to explore the underlying patterns and predicts the output.  Here we basically provide the machine with data and ask to look for hidden features and cluster the data in a way that makes sense. Example

* K – Means clustering
* Neural Networks
* Principal Component Analysis

## Reinforcement Learning

It is neither based on supervised learning nor unsupervised learning. Moreover, here the algorithms learn to react to an environment on their own. It is rapidly growing and moreover producing a variety of learning algorithms. These algorithms are useful in the field of Robotics, Gaming etc.

For a learning agent, there is always a start state and an end state. However, to reach the end state, there might be a different path. In Reinforcement**Learning Problem** an **agent** tries to manipulate the **environment**. The agent travels from one **state**to another. The agent gets the **reward**(appreciation) on success but will **not receive any reward**or appreciation on failure. In this way, the agent learns from the environment.

### Key Differences Between Supervised vs Unsupervised Learning vs Reinforcement Learning

1. Supervised Learning deals with two main tasks Regression and Classification. Unsupervised Learning deals with clustering and associative rule mining problems. Whereas Reinforcement Learning deals with exploitation or exploration, Markov’s decision processes, Policy Learning, Deep Learning and value learning.
2. Supervised Learning works with the labelled data and here the output data patterns are known to the system. But, the unsupervised learning deals with unlabeled data where the output is based on the collection of perceptions. Whereas in Reinforcement Learning Markov’s Decision process- the agent interacts with the environment in discrete steps.
3. The name itself says, Supervised Learning is highly supervised. And Unsupervised Learning is not supervised. As against, Reinforcement Learning is less supervised which depends on the agent in determining the output.
4. The input data in Supervised Learning in labelled data. Whereas, in Unsupervised Learning the data is unlabelled. The data is not predefined in Reinforcement Learning.
5. Supervised Learning predicts based on a class type. Unsupervised Learning discovers underlying patterns. And in Reinforcement Learning, the learning agent works as a reward and action system.
6. Supervised learning maps labelled data to known output. Whereas, Unsupervised Learning explore patterns and predict the output. Reinforcement Learning follows a trial and error method.
7. To sum up, in Supervised Learning, the goal is to generate formula based on input and output values. In Unsupervised Learning, we find an association between input values and group them. In Reinforcement Learning an agent learn through delayed feedback by interacting with the environment.
8. Difference between ai, ml and dl

AI is an umbrella discipline that covers everything related to making machines smarter. ML refers to an AI system that can self-learn based on the algorithm. Systems that get smarter and smarter over time without human intervention is ML. Deep Learning (DL) is a machine learning (ML) applied to large data sets.

1. How u do dimensionality reduction.

What are the steps in dimensionality reduction?

**Seven Techniques for Data Dimensionality Reduction**

1. Missing Values Ratio. ...
2. Low Variance Filter. ...
3. High Correlation Filter. ...
4. Random Forests / Ensemble Trees. ...
5. Principal Component Analysis (**PCA**). ...
6. Backward Feature Elimination. ...
7. Forward Feature Construction.
8. What is Multicollinearity

Multicollinearity refers to a situation in which more than two explanatory variables in a multiple regression model are highly linearly related. We have perfect multicollinearity if, for example as in the equation above, the correlation between two independent variables is equal to 1 or −1.

1. Parameters of random forest

(The parameters of a random forest are the variables and thresholds used to split each node learned during training). Scikit-Learn implements a set of sensible default hyperparameters for all models, but these are not guaranteed to be optimal for a problem.

7 Parameters of deep learning algos

Parameters are key to machine learning algorithms.  
  
**Some examples of model parameters include:**

* The weights in an artificial neural network.
* The support vectors in a support vector machine.
* The coefficients in a linear regression or logistic regression.

**Company: Bridgei2i**

**Role: Senior Analytics Consultant**

1)    What is the difference between Cluster and Systematic Sampling?

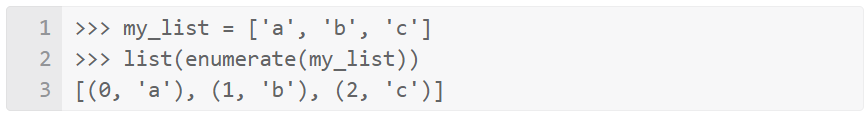
While systematic sampling uses fixed intervals from the larger population to create the sample, cluster sampling breaks the population down into different clusters. ... Cluster sampling divides the population into clusters and then takes a simple random sample from each cluster.

2)    Differentiate between a multi-label classification problem and a multi-class classification problem.

Multi-label classification is a generalization of multiclass classification, which is the single-label problem of categorizing instances into precisely one of more than two classes; in the multi-label problem there is no constraint on how many of the classes the instance can be assigned to.

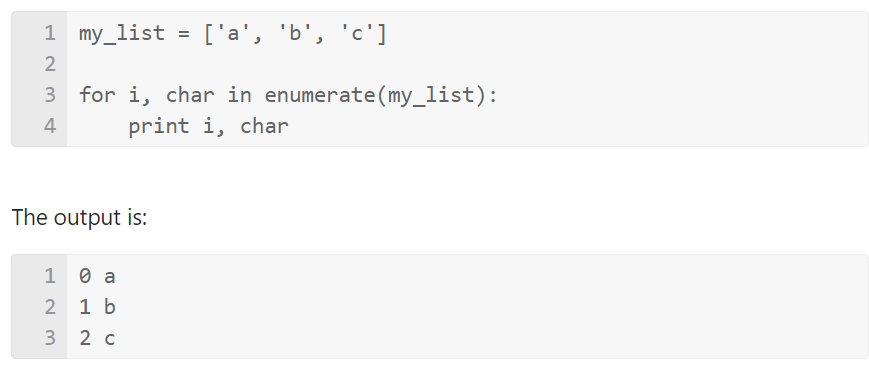
3)    How can you iterate over a list and also retrieve element indices at the same time?

enumerate function. It takes each element in a sequence (like a list) and sticks it's location right before it. For example:



Note that enumerate () returns an object to be iterated over, so wrapping it in list () just helps us see what enumerate () produces.

An example that directly relates to your question is this:



4)    What is Regularization and what kind of problems does regularization solve?

Regularization is a technique used for tuning the function by adding an additional penalty term in the error function. The additional term controls the excessively fluctuating function such that the coefficients don't take extreme values.

Overfitting is a phenomenon that occurs when a Machine Learning model is constraint to training set and not able to perform well on unseen data. Regularization is a technique used to reduce the errors by fitting the function appropriately on the given training set and avoid overfitting

7)    Can you cite some examples where a false positive is important than a false negative?

A false positive is where you receive a positive result for a test, when you should have received a negative result. ... Some examples of false positives: A pregnancy test is positive, when in fact you aren't pregnant. A cancer screening test comes back positive, but you don't have the disease.

8)    What is the advantage of performing dimensionality reduction before fitting an SVM?

* Dimensionality Reduction helps in data compression, and hence reduced storage space.
* It reduces computation time.
* It also helps remove redundant features, if any.
* Dimensionality Reduction helps in data compressing and reducing the storage space required.

9)    How will you find the correlation between a categorical variable and a continuous variable ?

**Point biserial Correlation**

The point biserial correlation coefficient is a special case of Pearson’s correlation coefficient. I am not going to go in the mathematical details of how it is calculated, but you can read more about it [here](https://en.wikipedia.org/wiki/Point-biserial_correlation_coefficient). I will highlight three important points to keep in mind though:

* Similar to the Pearson coefficient, the point biserial correlation can range from -1 to +1.
* The point biserial calculation assumes that the continuous variable is normally distributed and homoscedastic.
* If the dichotomous variable is artificially binarized, i.e. there is likely continuous data underlying it, biserial correlation is a more apt measurement of similarity. There is a simple formula to calculate the biserial correlation from point biserial correlation, but nonetheless this is an important point to keep in mind.

10)   How will you calculate the accuracy of a model using a confusion matrix?

Classification accuracy is the ratio of correct predictions to total predictions made. classification accuracy = correct predictions / total predictions. 1. classification accuracy = correct predictions / total predictions. It is often presented as a percentage by multiplying the result by 100.

13)    What do you understand by statistical power of sensitivity and how do you calculate it?

Statistical power is the probability of a hypothesis test of finding an effect if there is an effect to be found. A power analysis can be used to estimate the minimum sample size required for an experiment, given a desired significance level, effect size, and statistical power.

14)    What is pruning, entropy and information gain in decision tree algorithm?

The information gain is based on the decrease in entropy after a dataset is split on an attribute. Constructing a decision tree is all about finding attribute that returns the highest information gain (i.e., the most homogeneous branches).The result is the Information Gain, or decrease in entropy.

15)    What are the types of biases that can occur during sampling?

**Types of Sampling Bias**

* Observer Bias. Observer bias occurs when researchers subconsciously project their expectations on the research. ...
* Self-Selection/Voluntary Response Bias. ...
* Survivorship Bias. ...
* Recall Bias.

**Company: Deloitte**

**Role: Data Scientist**

1. Conditional Probability

Conditional probability is defined as the likelihood of an event or outcome occurring, based on the occurrence of a previous event or outcome. Conditional probability is calculated by multiplying the probability of the preceding event by the updated probability of the succeeding, or conditional, event.

1. Why Bayes theorem? DB Bayes and Naïve Bayes Theorem?

Bayes' theorem provides a way to revise existing predictions or theories (update probabilities) given new or additional evidence. In finance, Bayes' theorem can be used to rate the risk of lending money to potential borrowers.

It is a classification technique based on Bayes' Theorem with an assumption of independence among predictors. In simple terms, a Naive Bayes classifier assumes that the presence of a particular feature in a class is unrelated to the presence of any other feature.

Data Science Interview Questions:

1. What is the Central Limit Theorem and why is it important?

In probability theory, the central limit theorem (CLT) states that the distribution of a sample variable approximates a normal distribution (i.e., a “bell curve”) as the sample size becomes larger, assuming that all samples are identical in size, and regardless of the population's actual distribution shape.

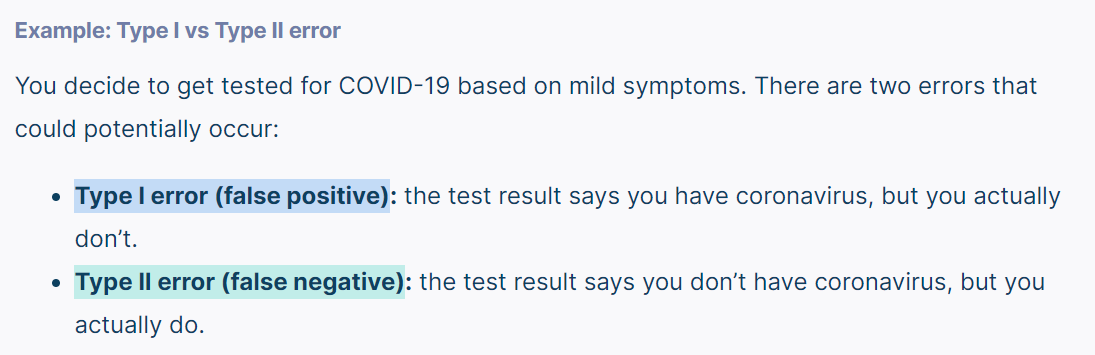
The Central Limit Theorem is important for statistics because it allows us to safely assume that the sampling distribution of the mean will be normal in most cases. This means that we can take advantage of statistical techniques that assume a normal distribution, as we will see in the next section.

2.What is the difference between type I vs type II error?

In statistics, a **Type I error** is a false positive conclusion, while a **Type II error** is a false negative conclusion.

Making a statistical decision always involves uncertainties, so the risks of making these errors are unavoidable in hypothesis testing.

The probability of making a Type I error is the significance level, or alpha (α), while the probability of making a Type II error is beta (β). These risks can be minimized through careful planning in your study design.



1. Explain the 80/20 rule, and tell me about its importance in model validation.

The 80-20 rule maintains that 80% of outcomes (outputs) come from 20% of causes (inputs). In the 80-20 rule, you prioritize the 20% of factors that will produce the best results. A principle of the 80-20 rule is to identify an entity's best assets and use them efficiently to create maximum value.

1. Is it better to spend five days developing a 90-percent accurate solution or 10 days for 100-percent accuracy?

Depends on the context- Is error acceptable? Fraud detection, quality assurance

1. Most common characteristics used in descriptive statistics?

Descriptive statistics are broken down into measures of central tendency and measures of variability (spread). Measures of central tendency include the mean, median, and mode, while measures of variability include standard deviation, variance, minimum and maximum variables, kurtosis, and skewness.

1. What do you mean by degree of freedom?

Degrees of Freedom refers to the maximum number of logically independent values, which are values that have the freedom to vary, in the data sample.Calculating Degrees of Freedom is key when trying to understand the importance of a Chi-Square statistic and the validity of the null hypothesis.

1. Why is the t-value same for 90% two tail and 95% one tail test?

because they answer different questions, one being more concrete than the other. The one-tailed question limits the values we are interested in, so the same statistic now has a different inferential meaning, resulting in lower error probability, hence higher observed significance.

1. What does it mean if a model is heteroscedastic? what about homoscedastic?

heteroscedasticity refers to cases where you incorrectly specify the model, and that causes the non-constant variance. When you leave an important variable out of a model, the omitted effect is absorbed into the error term.

Homoscedastic (also spelled "homoscedastic") refers to a condition in which the variance of the residual, or error term, in a regression model is constant. That is, the error term does not vary much as the value of the predictor variable changes.

1. You roll a biased coin (p(head)=0.8) five times. What’s the probability of getting three or more heads?

To start off the question, we need 3, 4, or 5 heads to satisfy the cases. 5 heads: All heads, so (4/5)^5=1024/3125

4 heads: All heads but 1. There are 5 ways to organize this, and then a (4/5)^4\*(1/5)^1=256/3125. Since there are 5 cases, we have 1280/3125.

3 heads: All heads but 2. There are 10 ways to organize this, and then a (4/5)^3\*(1/5)^2=64/3125. Since there are 10 cases, we have 640/3125 .We sum all these cases up to get (1024+1280+640)/3125=2944/3125. We have a 2944/3125 or 0.94208 probability to get 3 or more heads.

1. What does interpolation and extrapolation mean? Which is generally more accurate?

Extrapolation is an estimation of a value based on extending a known sequence of values or facts beyond the area that is certainly known. Interpolation is an estimation of a value within two known values in a sequence of values. Polynomial interpolation is a method of estimating values between known data points.

Data Science Interview Questions:

1. What the aim of conducting A/B Testing?

A/B testing allows individuals, teams and companies to make careful changes to their user experiences while collecting data on the results. This allows them to construct hypotheses and to learn why certain elements of their experiences impact user behavior.

1. What does the term Variance Inflation Factor mean?

Variance inflation factor measures how much the behaviour (variance) of an independent variable is influenced, or inflated, by its interaction/correlation with the other independent variables. Variance inflation factors allow a quick measure of how much a variable is contributing to the standard error in the regression.

1. What is the significance of Gamma and Regularization in SVM?

gamma is a parameter for non linear hyperplanes. The higher the gamma value it tries to exactly fit the training data set gammas = [0.1, 1, 10, 100]for gamma in gammas: svc = svm.SVC(kernel='rbf', gamma=gamma).fit(X, y)

The regularization parameter (lambda) serves as a degree of importance that is given to miss-classifications. SVM pose a quadratic optimization problem that looks for maximizing the margin between both classes and minimizing the amount of miss-classifications.For non-linear-kernel SVM the idea is the similar.

Data Science Interview questions:

How will you calculate the Sensitivity of machine learning models?

Sensitivity is a measure of the proportion of actual positive cases that got predicted as positive (or true positive). Sensitivity is also termed as Recall.

What do you mean by cluster sampling and systematic sampling?

Cluster sampling is a probability sampling method in which you divide a population into clusters, such as districts or schools, and then randomly select some of these clusters as your sample. In single-stage sampling, you collect data from every unit within the selected clusters.

Systematic sampling is a type of probability sampling method in which sample members from a larger population are selected according to a random starting point but with a fixed, periodic interval. This interval, called the sampling interval, is calculated by dividing the population size by the desired sample size.

Explain Eigenvectors and Eigenvalues.

Eigenvectors are the directions along which a particular linear transformation acts by flipping, compressing or stretching. Eigenvalue can be referred to as the strength of the transformation in the direction of eigenvector or the factor by which the compression occurs.

Explain Gradient Descent.

Gradient Descent is an optimization algorithm for finding a local minimum of a differentiable function. Gradient descent is simply used in machine learning to find the values of a function's parameters (coefficients) that minimize a cost function as far as possible.

How does Backpropagation work? Also, it states its various variants.

Back-propagation is just a way of propagating the total loss back into the neural network to know how much of the loss every node is responsible for, and subsequently updating the weights in such a way that minimizes the loss by giving the nodes with higher error rates lower weights and vice versa.

What do you know about Autoencoders?

Autoencoders are artificial neural networks that can learn from an unlabeled training set. This may be dubbed as unsupervised deep learning. They can be used for either dimensionality reduction or as a generative model, meaning that they can generate new data from input data

What is Dropout in Neural Networks?

Dropout is a technique where randomly selected neurons are ignored during training. They are “dropped-out” randomly. This means that their contribution to the activation of downstream neurons is temporally removed on the forward pass and any weight updates are not applied to the neuron on the backward pass.

What is the difference between Batch and Stochastic Gradient Descent?

Batch gradient descent, at all steps, takes the steepest route to reach the true input distribution. SGD, on the other hand, chooses a random point within the shaded area, and takes the steepest route towards this point. At each iteration, though, it chooses a new point.

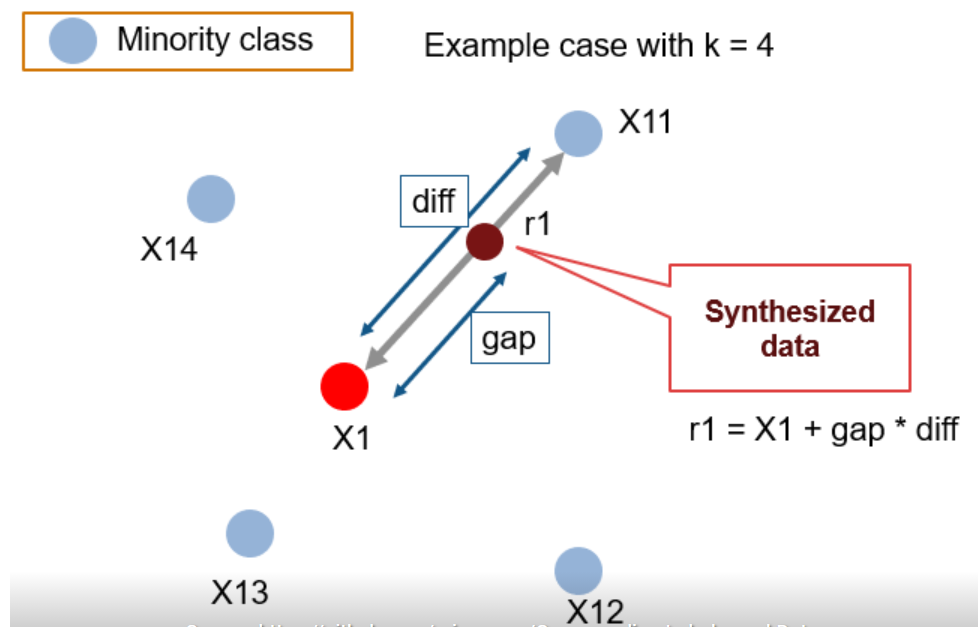
**Company: Latentview.**

1. Explain the SMOTE

SMOTE is an oversampling technique where the synthetic samples are generated for the minority class. This algorithm helps to overcome the overfitting problem posed by random oversampling. It focuses on the feature space to generate new instances with the help of interpolation between the positive instances that lie together.

Working Procedure:

At first the total no. of oversampling observations, N is set up. Generally, it is selected such that the binary class distribution is 1:1. But that could be tuned down based on need. Then the iteration starts by first selecting a positive class instance at random. Next, the KNN’s (by default 5) for that instance is obtained. At last, N of these K instances is chosen to interpolate new synthetic instances. To do that, using any distance metric the difference in distance between the feature vector and its neighbors is calculated. Now, this difference is multiplied by any random value in (0,1] and is added to the previous feature vector. This is pictorially represented below:



1. What is stratified sampling technique

Stratified sampling is a type of sampling method in which the total population is divided into smaller groups or strata to complete the sampling process. After dividing the population into strata, the researcher randomly selects the sample proportionally.

1. Explain the working of random forest and xgboost

random forest builds multiple decision trees and merges them together to get a more accurate and stable prediction. Random forest has nearly the same hyperparameters as a decision tree or a bagging classifier. Random forest adds additional randomness to the model, while growing the trees.

XGBoost is a decision-tree-based ensemble Machine Learning algorithm that **uses a gradient boosting framework**. In prediction problems involving unstructured data (images, text, etc.) A wide range of applications: Can be used to solve regression, classification, ranking, and user-defined prediction problems.

1. How do you optimise the Recall of your output?

Generally, if you want higher precision you need to restrict the positive predictions to those with highest certainty in your model, which means predicting fewer positives overall (which, in turn, usually results in lower recall).

1. What are chisquare and ANOVA test

The chi-square is used to investigate whether the distribution of classes and is compatible with a distribution model (often equal distribution, but not always), while ANOVA is used to investigate whether differences in means between samples are significant or not.

Analysis of variance, or ANOVA, is a statistical method that separates observed variance data into different components to use for additional tests. A one-way ANOVA is used for three or more groups of data, to gain information about the relationship between the dependent and independent variables.

1. In python they asked for LOC,ILOC, how do you remove duplicate,How to unique values in column,

**The main distinction between loc and iloc is:**

* loc is label-based, which means that you have to specify rows and columns based on their row and column labels.
* iloc is integer position-based, so you have to specify rows and columns by their integer position values (0-based integer position).

**Pandas drop\_duplicates() method helps in removing duplicates from the data frame.**

1. Syntax: DataFrame.drop\_duplicates(subset=None, keep='first', inplace=False)
2. Parameters:
3. subset: Subset takes a column or list of column label. It's default value is none. ...
4. keep: keep is to control how to consider duplicate value.

Company: Enquero Global

Role: Data Scientist

1.Previous job role and responsibilities

Tell about your previous job role and responsibilities

2.Problem statement of your project and how do you overcome challenges

Explain about your project in details and mention how did you overcome those challenges

3.How do you handle feature which had many categories?

Perform feature engineering. The importance of feature selection can best be recognized when you are dealing with a dataset that contains a vast number of features. This type of dataset is often referred to as a high dimensional dataset. Now, with this high dimensionality, comes a lot of problems such as - this high dimensionality will significantly increase the training time of your machine learning model, it can make your model very complicated which in turn may lead to Overfitting.

Often in a high dimensional feature set, there remain several features which are redundant meaning these features are nothing but extensions of the other essential features. These redundant features do not effectively contribute to the model training as well. So, clearly, there is a need to extract the most important and the most relevant features for a dataset in order to get the most effective predictive modelling performance.

4.When to use precision and recall.

1. Precision: This tells when you predict something positive, how many times they were actually positive. whereas,
2. Recall: This tells out of actual positive data, how many times you predicted correctly.

5.What are outliers & how do you handle them

Outliers is the observation that lies an abnormal distance from other values in a random sample from a population.

3 different methods of dealing with outliers:

1. *Univariate method:* This method looks for data points with extreme values on one variable.
2. *Multivariate method:* Here we look for unusual combinations on all the variables.
3. *Minkowski error:* This method reduces the contribution of potential outliers in the training process.