100+ Data Science Interview Questions You Must Prepare for 2021

Last updated on Jan 12,2021   *277.9K Views*

Upasana
Research Analyst, Tech Enthusiast, Currently working on Azure IoT & Data Science...

In this Data Science Interview Questions blog, I will introduce you to the most frequently asked questions on Data Science, Analytics and Machine Learning interviews. This blog is the perfect guide for you to learn all the concepts required to clear a Data Science interview. To get in-depth knowledge on Data Science, you can enroll for live **_Data Science Certification Training_** by Edureka with 24/7 support and lifetime access.

Edureka 2019 Tech Career Guide is out! Hottest job roles, precise learning paths, industry outlook & more in the guide. **_Download_** now.

The following are the topics covered in our interview questions:

- **Basic Data Science Interview Questions**

- **Statistics Interview Questions**

- **Data Analysis Interview Questions**

- **Machine Learning Interview Questions**

- **Deep Learning Interview Questions**

Before moving ahead, you may go through the recording of Data Science Interview Questions where our instructor has shared his experience and expertise that will help you to crack any Data Science.

Data Science Interview Questions | Edureka

BASIC DATA SCIENCE INTERVIEW QUESTIONS

Q1. What is Data Science? List the differences between supervised and unsupervised learning.

Data Science is a blend of various tools, algorithms, and machine learning principles with the goal to discover hidden patterns from the raw data. How is this different from what statisticians have been doing for years?

FREE WEBINAR
Time Series Analysis using Python

The answer lies in the difference between explaining and predicting.

Data Product Engineering ⮐

The differences between supervised and unsupervised learning are as follows;

Supervised Learning	Unsupervised Learning
Input data is labelled.	Input data is unlabelled.
Uses a training data set.	Uses the input data set.
Used for prediction.	Used for analysis.
Enables classification and regression.	Enables Classification, Density Estimation, & Dimension Reduction

Q2. What is Selection Bias?

Selection bias is a kind of error that occurs when the researcher decides who is going to be studied. It is usually associated with research where the selection of participants isn't random. It is sometimes referred to as the selection effect. It is the distortion of statistical analysis, resulting from the method of collecting samples. If the selection bias is not taken into account, then some conclusions of the study may not be accurate.

The types of selection bias include:

1. **Sampling bias**: It is a systematic error due to a non-random sample of a population causing some members of the population to be less likely to be included than others resulting in a biased sample.

2. **Time interval**: A trial may be terminated early at an extreme value (often for ethical reasons), but the extreme value is likely to be reached by the variable with the largest variance, even if all variables have a similar mean.

3. **Data**: When specific subsets of data are chosen to support a conclusion or rejection of bad data on arbitrary grounds, instead of according to previously stated or generally agreed criteria.

4. **Attrition**: Attrition bias is a kind of selection bias caused by attrition (loss of participants) discounting trial subjects/tests that did not run to completion.

Q3. What is bias-variance trade-off?

Bias: Bias is an error introduced in your model due to oversimplification of the machine learning algorithm. It can lead to underfitting. When you train your model at that time model makes simplified assumptions to make the target function easier to understand.

Low bias machine learning algorithms — Decision Trees, k-NN and SVM High bias machine learning algorithms — Linear Regression, Logistic Regression

Variance: Variance is error introduced in your model due to complex machine learning algorithm, your model learns noise also from the training data set and performs badly on test data set. It can lead to high sensitivity and overfitting.

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

Total Error

Model Complexity

Bias-Variance trade-off: The goal of any supervised machine learning algorithm is to have low bias and low variance to achieve good prediction performance.

1. The k-nearest neighbour algorithm has low bias and high variance, but the trade-off can be changed by increasing the value of k which increases the number of neighbours that contribute to the prediction and in turn increases the bias of the model.

2. The support vector machine algorithm has low bias and high variance, but the trade-off can be changed by increasing the C parameter that influences the number of violations of the margin allowed in the training data which increases the bias but decreases the variance.

There is no escaping the relationship between bias and variance in machine learning. Increasing the bias will decrease the variance. Increasing the variance will decrease bias.

Q4. What is a confusion matrix?

The confusion matrix is a 2X2 table that contains 4 outputs provided by the **binary classifier**. Various measures, such as error-rate, accuracy, specificity, sensitivity, precision and recall are derived from it. *Confusion Matrix*

Predicted class

|		P	N
Actual Class	P	True Positives (TP)	False Negatives (FN)
	N	False Positives (FP)	True Negatives (TN)

A data set used for performance evaluation is called a **test data set**. It should contain the correct labels and predicted labels.

Two actual classes or observed labels

positive (P)

negative (N)
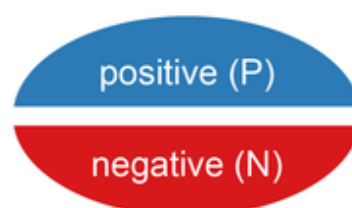
The predicted labels will exactly the same if the performance of a binary classifier is perfect.

Predicted classes of a perfect classifier

Positive prediction

predicted as positive

A classifier

predicted as negative

Negative prediction
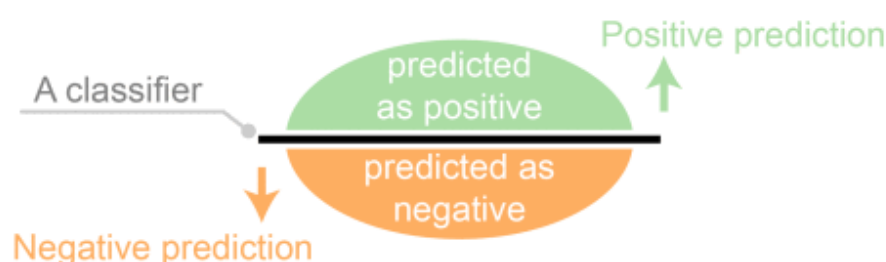
FREE WEBINAR

Time Series Analysis using Python

The predicted labels usually match with part of the observed labels in real-world scenarios.

Predicted classes of a classifier

1. True-positive(TP) — Correct positive prediction

2. False-positive(FP) — Incorrect positive prediction

3. True-negative(TN) — Correct negative prediction

4. False-negative(FN) — Incorrect negative prediction

Four outcomes of a classifier
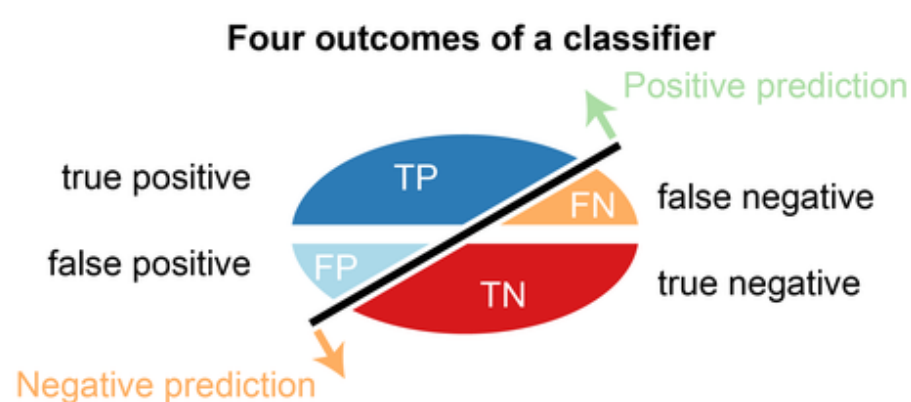
Basic measures derived from the confusion matrix

1. Error Rate = (FP+FN)/(P+N)

2. Accuracy = (TP+TN)/(P+N)

3. Sensitivity(Recall or True positive rate) = TP/P

4. Specificity(True negative rate) = TN/N

5. Precision(Positive predicted value) = TP/(TP+FP)

6. F-Score(Harmonic mean of precision and recall) = (1+b)(PREC.REC)/(b²PREC+REC) where b is commonly 0.5, 1, 2.

STATISTICS INTERVIEW QUESTIONS

Q5. What is the difference between "long" and "wide" format data?

In the **wide-format**, a subject's repeated responses will be in a single row, and each response is in a separate column. In the **long-format**, each row is a one-time point per subject. You can recognize data in wide format by the fact that columns generally represent groups.

Name	Height	Weight
John	160	67
Christopher	182	78

Name	Attribute	Value
John	Height	160
John	Weight	67
Christopher	Height	182
Christopher	Weight	78

Figure: Wide Format   *Figure:* Long Format

Q6. What do you understand by the term Normal Distribution?

Data is usually distributed in different ways with a bias to the left or to the right or it can all be

FREE WEBINAR
Time Series Analysis using Python

However, there are chances that data is distributed around a central value without any bias to the left or right and reaches normal distribution in the form of a bell-shaped curve.

**Figure:** *Normal distribution in a bell curve*

The random variables are distributed in the form of a symmetrical, bell-shaped curve.

Properties of Normal Distribution are as follows;

1. Unimodal -one mode

2. Symmetrical -left and right halves are mirror images

3. Bell-shaped -maximum height (mode) at the mean

4. Mean, Mode, and Median are all located in the center

5. Asymptotic

Q7. What is correlation and covariance in statistics?

Covariance and Correlation are two mathematical concepts; these two approaches are widely used in statistics. Both Correlation and Covariance establish the relationship and also measure the dependency between two random variables. Though the work is similar between these two in mathematical terms, they are different from each other.
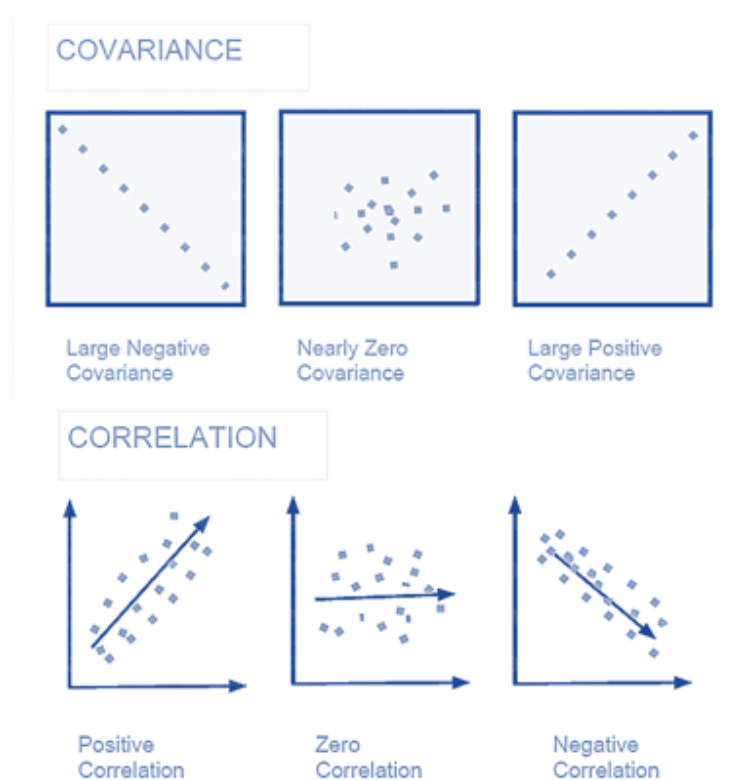
Correlation: Correlation is considered or described as the best technique for measuring and also for estimating the quantitative relationship between two variables. Correlation measures how strongly two variables are related.

Covariance: In covariance two items vary together and it's a measure that indicates the extent to which two random variables change in cycle. It is a statistical term; it explains the systematic relation between a pair of random variables, wherein changes in one variable reciprocal by a corresponding change in another variable.

Q8. What is the difference between Point Estimates and Confidence Interval?

Point Estimation gives us a particular value as an estimate of a population parameter. M Likelihood estimator methods are used to derive Point Estimators for population parameters.

A confidence interval gives us a range of values which is likely to contain the population parameter. The confidence interval is generally preferred, as it tells us how likely this interval is to contain the population parameter. This likeliness or probability is

Q10. What is p-value?

When you perform a hypothesis test in statistics, a p-value can help you determine the strength of your results. p-value is a number between 0 and 1. Based on the value it will denote the strength of the results. The claim which is on trial is called the Null Hypothesis.

Low p-value (≤ 0.05) indicates strength against the null hypothesis which means we can reject the null Hypothesis. High p-value (≥ 0.05) indicates strength for the null hypothesis which means we can accept the null Hypothesis p-value of 0.05 indicates the Hypothesis could go either way. To put it in another way,

High P values: your data are likely with a true null. Low P values: your data are unlikely with a true null.

Q11. In any 15-minute interval, there is a 20% probability that you will see at least one shooting star. What is the probability that you see at least one shooting star in the period of an hour?

Probability of not seeing any shooting star in 15 minutes is

=   1 – P(Seeing one shooting star)
=   1 – 0.2   =   0.8

Probability of not seeing any shooting star in the period of one hour

= $(0.8)^4$ =   0.4096

Probability of seeing at least one shooting star in the one hour

=   1 – P(Not seeing any star)
=   1 – 0.4096   =   0.5904

Q12. How can you generate a random number between 1 – 7 with only a die?

- Any die has six sides from 1-6. There is no way to get seven equal outcomes from a single rolling of a die. If we roll the die twice and consider the event of two rolls, we now have 36 different outcomes.

- To get our 7 equal outcomes we have to reduce this 36 to a number divisible by 7. We can thus consider only 35 outcomes and exclude the other one.

- A simple scenario can be to exclude the combination (6,6), i.e., to roll the die again if 6 appears twice.

- All the remaining combinations from (1,1) till (6,5) can be divided into 7 parts of 5 each. This way all the seven sets of outcomes are equally likely.

Q13. A certain couple tells you that they have two children, at least one of which is a girl. What is the probability that they have two girls?

In the case of two children, there are 4 equally likely possibilities

BB, BG, GB and GG;

where **B** = Boy and **G** = Girl and the first letter denotes the first child.

From the question, we can exclude the first case of BB. Thus from the remaining 3 possibiliti the probability of the case with two girls.

Thus, P(Having two girls given one girl) = **1 / 3**

Q14. A jar has 1000 coins, of which 999 are fair and 1 is double headed. Pick a coin at random, and toss it 10 times. Given

P (B) = 0.001 * 1 = **0.001**
P(A / A + B) = 0.000976 / (0.000976 + 0.001) = **0.4939**
P(B / A + B) = 0.001 / 0.001976 = **0.5061**

Probability of selecting another head = P(A/A+B) * 0.5 + P(B/A+B) * 1 = 0.4939 * 0.5 + 0.5061 = **0.7531**

Q15. What do you understand by statistical power of sensitivity and how do you calculate it?

Sensitivity is commonly used to validate the accuracy of a classifier (Logistic, SVM, Random Forest etc.).

Sensitivity is nothing but "Predicted True events/ Total events". True events here are the events which were true and model also predicted them as true.

Calculation of seasonality is pretty straightforward.

Seasonality = (**True Positives**) / (**Positives in Actual Dependent Variable**)

Q16. Why Is Re-sampling Done?

Resampling is done in any of these cases:

- Estimating the accuracy of sample statistics by using subsets of accessible data or drawing randomly with replacement from a set of data points

- Substituting labels on data points when performing significance tests

- Validating models by using random subsets (bootstrapping, cross-validation)

Q17. What are the differences between over-fitting and under-fitting?

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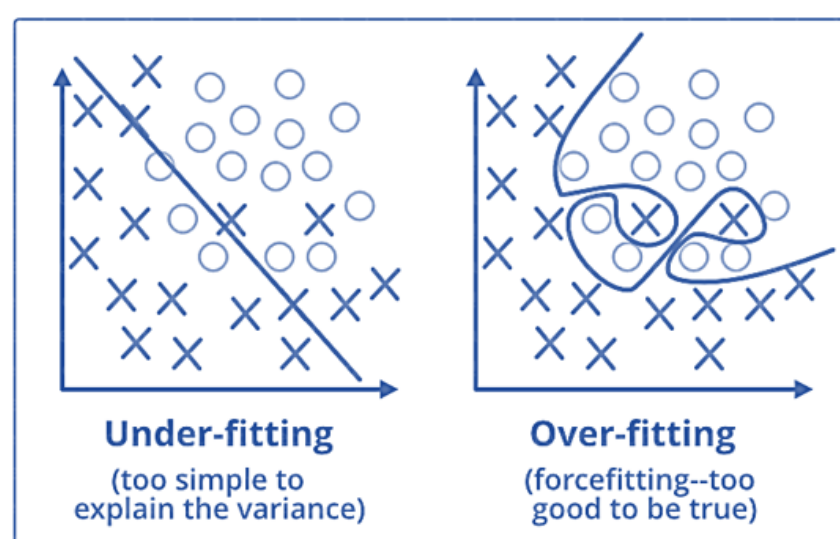
In **overfitting**, a statistical model describes random error or noise instead of the underlying relationship. Overfitting occurs when a model is excessively complex, such as having too many parameters relative to the number of observations. A model that has been overfitted, has poor predictive performance,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

FREE WEBINAR

Time Series Analysis using Python

Q18. How to combat Overfitting and Underfitting?

To combat overfitting and underfitting, you can resample the data to estimate the model accuracy (k-fold cross-validation) and by

Regularisation is the process of adding tuning parameter to a model to induce smoothness in order to prevent overfitting. This is most often done by adding a constant multiple to an existing weight vector. This constant is often the L1(Lasso) or L2(ridge). The model predictions should then minimize the loss function calculated on the regularized training set.

Q20. What Is the Law of Large Numbers?

It is a theorem that describes the result of performing the same experiment a large number of times. This theorem forms the basis of **frequency-style** thinking. It says that the sample means, the sample variance and the sample standard deviation converge to what they are trying to estimate.

Q21. What Are Confounding Variables?

In statistics, a confounder is a variable that influences both the dependent variable and independent variable.

For example, if you are researching whether a lack of exercise leads to weight gain,

lack of exercise = independent variable

weight gain = dependent variable.

A confounding variable here would be any other variable that affects both of these variables, such as the **age of the subject**.

Q22. What Are the Types of Biases That Can Occur During Sampling?

- Selection bias

- Under coverage bias

- Survivorship bias

Q23. What is Survivorship Bias?

It is the logical error of focusing aspects that support surviving some process and casually overlooking those that did not work because of their lack of prominence. This can lead to wrong conclusions in numerous different means.

Q24. What is selection Bias?

Selection bias occurs when the sample obtained is not representative of the population intended to be analysed.

Q25. Explain how a ROC curve works?

The **ROC** curve is a graphical representation of the contrast between true positive rates and false-positive rates at various thresholds. It is often used as a proxy for the trade-off between the sensitivity(true positive rate) and false-positive rate.

FREE WEBINAR
Time Series Analysis using Python
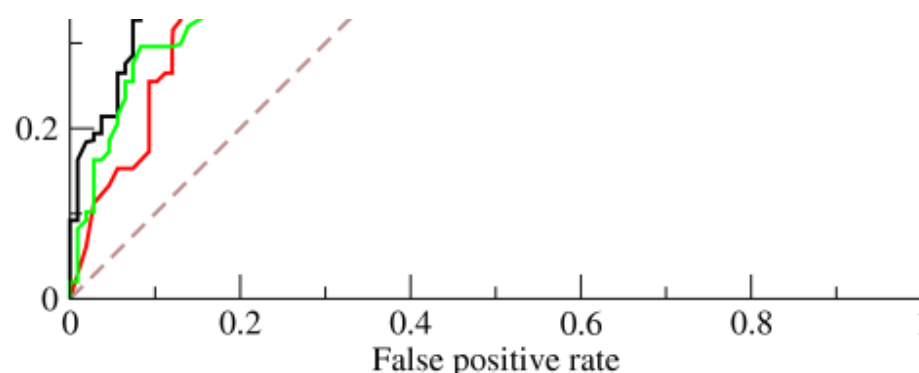
Q26. What is TF/IDF vectorization?

TF–IDF is short for term frequency-inverse document frequency, is a numerical statistic that is intended to reflect how important a word is to a document in a collection or corpus. It is often used as a weighting factor in information retrieval and text mining.

The TF–IDF value increases proportionally to the number of times a word appears in the document but is offset by the frequency of the word in the corpus, which helps to adjust for the fact that some words appear more frequently in general.

Q27. Why we generally use Softmax non-linearity function as last operation in-network?

It is because it takes in a vector of real numbers and returns a probability distribution. Its definition is as follows. Let x be a vector of real numbers (positive, negative, whatever, there are no constraints).

Then the i'th component of Softmax(*x) is* —

$$P(y=j \mid \Theta^{(i)}) = \frac{e^{\Theta^{(i)}}}{\sum_{j=0}^{k} e^{\Theta_k^{(i)}}}$$

Softmax function

$$\text{where } \Theta = w_0 x_0 + w_1 x_1 + \ldots + w_k x_k = \sum_{i=0}^{k} w_i x_i = w^T x$$

It should be clear that the output is a probability distribution: each element is non-negative and the sum over all components is 1.

DATA ANALYSIS INTERVIEW QUESTIONS

Q28. Python or R – Which one would you prefer for text analytics?

We will prefer Python because of the following reasons:

- **Python** would be the best option because it has Pandas library that provides easy performance data analysis tools.

FREE WEBINAR
Time Series Analysis using Python

- **R** is more suitable for machine learning than just text analysis.

- Python performs faster for all types of text analytics

exponentially due to the number of sources and the volume of data generated by these sources.

- It might take up to 80% of the time for just cleaning data making it a critical part of the analysis task.

Q30. Differentiate between univariate, bivariate and multivariate analysis.

Univariate *analyses* are descriptive statistical analysis techniques which can be differentiated based on the number of variables involved at a given point of time. For example, the pie charts of sales based on territory involve only one variable and can the analysis can be referred to as univariate analysis.

The **bivariate** *analysis* attempts to understand the difference between two variables at a time as in a scatterplot. For example, analyzing the volume of sale and spending can be considered as an example of bivariate analysis.

Multivariate analysis deals with the study of more than two variables to understand the effect of variables on the responses.

Q31. Explain Star Schema.

It is a traditional database schema with a central table. Satellite tables map IDs to physical names or descriptions and can be connected to the central fact table using the ID fields; these tables are known as lookup tables and are principally useful in real-time applications, as they save a lot of memory. Sometimes star schemas involve several layers of summarization to recover information faster.

Q32. What is Cluster Sampling?

Cluster sampling is a technique used when it becomes difficult to study the target population spread across a wide area and simple random sampling cannot be applied. Cluster Sample is a probability sample where each sampling unit is a collection or cluster of elements.

For eg., A researcher wants to survey the academic performance of high school students in Japan. He can divide the entire population of Japan into different clusters (cities). Then the researcher selects a number of clusters depending on his research through simple or systematic random sampling.

Let's continue our Data Science Interview Questions blog with some more statistics questions.

Q33. What is Systematic Sampling?

Systematic sampling is a statistical technique where elements are selected from an ordered sampling frame. In systematic sampling, the list is progressed in a circular manner so once you reach the end of the list, it is progressed from the top again. The best example of systematic sampling is equal probability method.

Q34. What are Eigenvectors and Eigenvalues?

Eigenvectors are used for understanding linear transformations. In data analysis, we usually calculate the eigenvectors for a correlation or covariance matrix. 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.

Q35. Can you cite some examples where a false positive is important than a false negative?

Let us first understand what false positives and false negatives are.

- **False Positives** are the cases where you wrongly classified a non-event as an event a.k.a

- **False Negatives** are the cases where you wrongly classify events as non-events, a.k.a Type II error.

Example 1: In the medical field, assume you have to give chemotherapy to patients. Assume a patient comes to that hospital and

Q36. Can you cite some examples where a false negative important than a false positive?

Example 1: Assume there is an airport 'A' which has received high-security threats and based on certain characteristics they identify whether a particular passenger can be a threat or not. Due to a shortage of staff, they decide to scan passengers being predicted as risk positives by their predictive model. What will happen if a true threat customer is being flagged as non-threat by airport model?

Example 2: What if Jury or judge decides to make a criminal go free?

Example 3: What if you rejected to marry a very good person based on your predictive model and you happen to meet him/her after a few years and realize that you had a false negative?

Q37. Can you cite some examples where both false positive and false negatives are equally important?

In the **Banking** industry giving loans is the primary source of making money but at the same time if your repayment rate is not good you will not make any profit, rather you will risk huge losses.

Banks don't want to lose good customers and at the same point in time, they don't want to acquire bad customers. In this scenario, both the false positives and false negatives become very important to measure.

Q38. Can you explain the difference between a Validation Set and a Test Set?

A *Validation set* can be considered as a part of the training set as it is used for parameter selection and to avoid overfitting of the model being built.

On the other hand, a *Test Set* is used for testing or evaluating the performance of a trained machine learning model.

In simple terms, the differences can be summarized as; training set is to fit the parameters i.e. weights and test set is to assess the performance of the model i.e. evaluating the predictive power and generalization.

Q39. Explain cross-validation.

Cross-validation is a model validation technique for evaluating how the outcomes of statistical analysis will **generalize** to an **independent dataset**. Mainly used in backgrounds where the objective is forecast and one wants to estimate how accurately a model will accomplish in practice.

The goal of cross-validation is to term a data set to test the model in the training phase (i.e. validation data set) in order to limit problems like overfitting and get an insight on how the model will generalize to an independent data set.

MACHINE LEARNING INTERVIEW QUESTIONS

Q40. What is Machine Learning?

Machine Learning explores the study and construction of algorithms that can learn from and make predictions on data. Closely related to computational statistics. Used to devise complex models and algorithms that lend themselves to a prediction which in commercial use is known as predictive analytics. Given below, is an image representing the various domains Machine Learning lends itself to.

FREE WEBINAR
Time Series Analysis using Python

Networks

E.g. If you built a fruit classifier, the labels will be "this is an orange, this is an apple and this is a banana", based on showing the classifier examples of apples, oranges and bananas.

Q42. What is Unsupervised learning?

Unsupervised learning is a type of machine learning algorithm used to draw inferences from datasets consisting of input data without labelled responses.

Algorithms: Clustering, Anomaly Detection, Neural Networks and Latent Variable Models

E.g. In the same example, a fruit clustering will categorize as "fruits with soft skin and lots of dimples", "fruits with shiny hard skin" and "elongated yellow fruits".

Q43. What are the various classification algorithms?

The diagram lists the most important **classification algorithms**.

FREE WEBINAR
Time Series Analysis using Python

Q44. What is 'Naive' in a Naive Bayes?

The **Naive Bayes Algorithm** is based on the Bayes Theorem. Bayes' theorem describes the probability of an event, based on prior knowledge of conditions that might be related to the event.

The Algorithm is 'naive' because it makes assumptions that may or may not turn out to be correct.

Q45. Explain SVM algorithm in detail.

SVM stands for support vector machine, it is a supervised machine learning algorithm which can be used for both **Regression and Classification**. If you have n features in your training data set, SVM tries to plot it in n-dimensional space with the value of each feature being the value of a particular coordinate. SVM uses hyperplanes to separate out different classes based on the provided kernel function.

FREE WEBINAR

Time Series Analysis using Python

y

Q46. What are the support vectors in SVM?
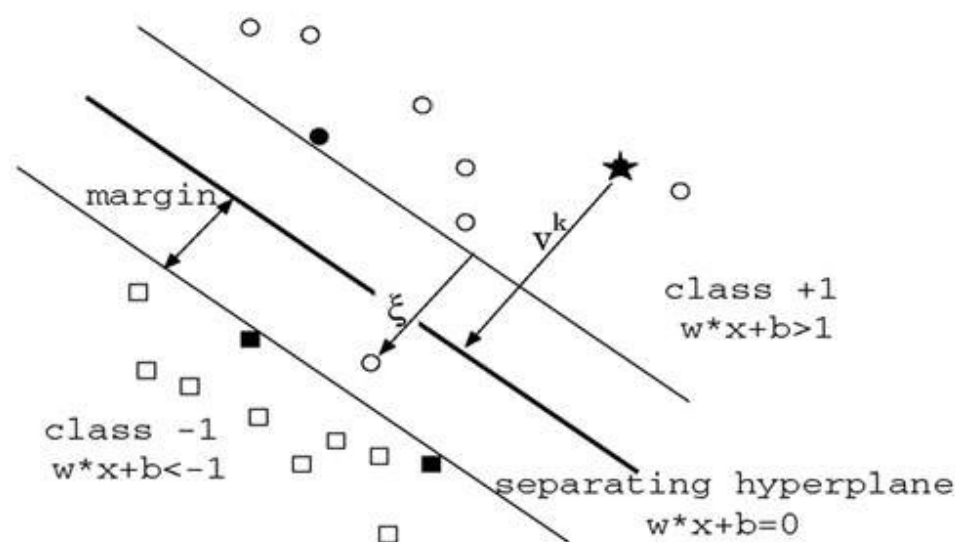
In the diagram, we see that the thinner lines mark the distance from the classifier to the closest data points called the support vectors (darkened data points). The distance between the two thin lines is called the margin.

Q47. What are the different kernels in SVM?

There are four types of kernels in SVM.

1. Linear Kernel

2. Polynomial kernel

3. Radial basis kernel

4. Sigmoid kernel

Q48. Explain Decision Tree algorithm in detail.

A *decision tree* is a supervised machine learning algorithm mainly used for **Regression and Classification**. It breaks down a data set into smaller and smaller subsets while at the same time an associated decision tree is incrementally developed. The final result is a tree with decision nodes and leaf nodes. A decision tree can handle both categorical and numerical data.

Q49. What are Entropy and Information gain in Decision tree algorithm?

The core algorithm for building a decision tree is called **ID3**. **ID3** uses **Entropy** and **Information Gain** *to construct a decision tree.*

Entropy

ta Science Training

A decision tree is built top-down from a root node and involve partitioning of data into homogenious subsets. **ID3** uses enteropy to check the homogeneity of a sample. If the sample is completely homogenious then entropy is zero and if the sample is an equally divided it has entropy of one.

$$Entropy = -p\ log_2 p - q\ log_2 q$$
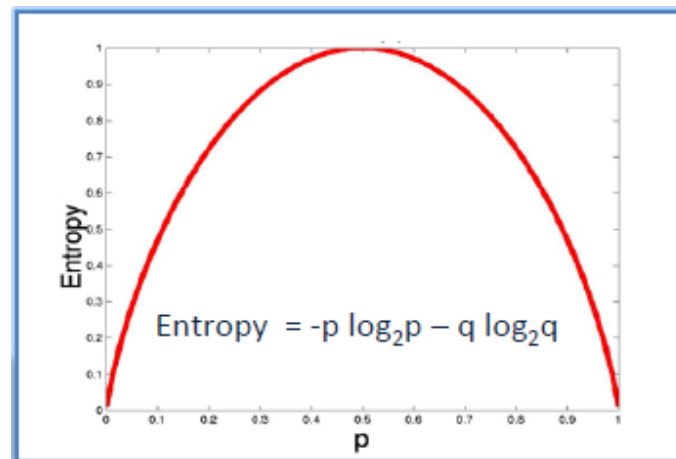
$$Entropy = -0.5\ log_2 0.5 - 0.5\ log_2 0.5 = 1$$

Information Gain

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 attributes that return the highest information gain.

$$Gain(T,X) = Entropy(T) - Entropy(T,X)$$

$$\mathbf{G}(PlayGolf, Outlook) = \mathbf{E}(PlayGolf) - \mathbf{E}(PlayGolf, Outlook)$$
$$= 0.940 - 0.693 = 0.247$$

Q50. What is pruning in Decision Tree?

Pruning is a technique in machine learning and search algorithms that reduces the size of **decision trees** by removing sections of the **tree** that provide little power to classify instances. So, when we remove sub-nodes of a decision node, this process is called *pruning* or opposite process of splitting.

Q51. What is logistic regression? State an example when you have used logistic regression recently.

Logistic Regression often referred to as the logit model is a technique to predict the binary outcome from a linear combination of predictor variables.

FREE WEBINAR

Time Series Analysis using Python

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

D =
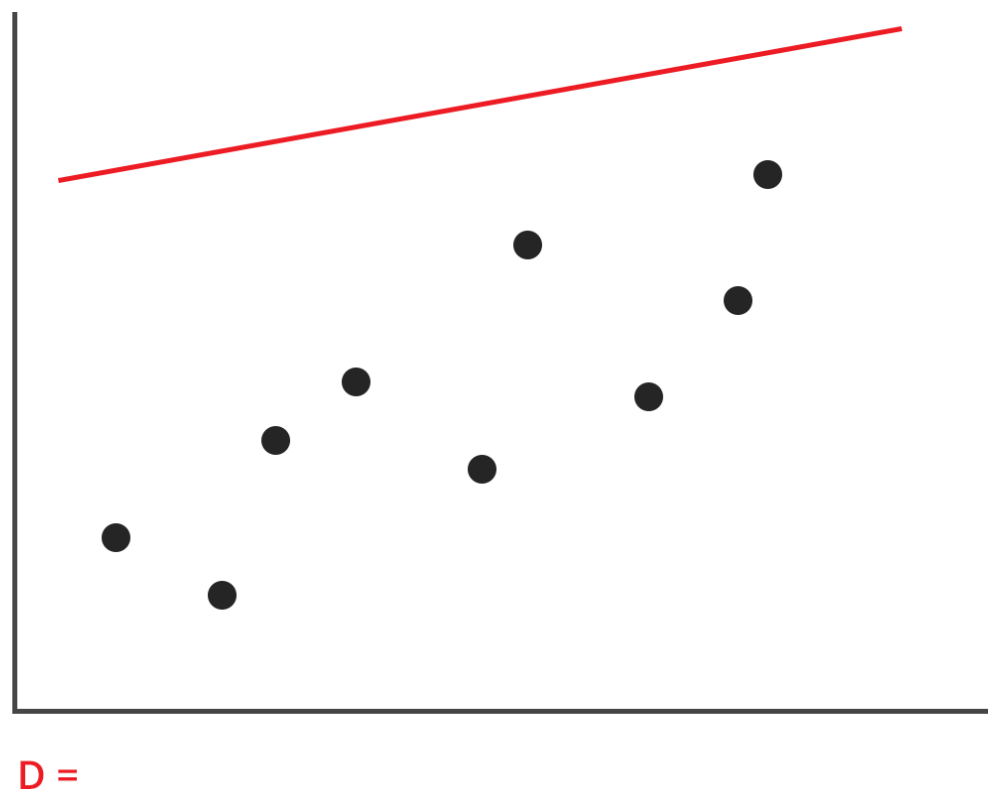
The regression line is the one with the least value of D

Q53. What Are the Drawbacks of the Linear Model?

Some drawbacks of the linear model are:

- The assumption of linearity of the errors.

- It can't be used for count outcomes or binary outcomes

- There are overfitting problems that it can't solve

Q54. What is the difference between Regression and classification ML techniques?

Both Regression and classification machine learning techniques come under **Supervised machine learning algorithms**. In Supervised machine learning algorithm, we have to train the model using labelled data set, While training we have to explicitly provide the correct labels and algorithm tries to learn the pattern from input to output. If our labels are discrete values then it will a classification problem, e.g A,B etc. but if our labels are continuous values then it will be a regression problem, e.g 1.23, 1.333 etc.

Q55. What are Recommender Systems?

Recommender Systems are a subclass of information filtering systems that are meant to predict the preferences or ratings that a user would give to a product. Recommender systems are widely used in movies, news, research articles, products, social tags, music, etc.

Examples include movie recommenders in IMDB, Netflix & BookMyShow, product recommenders in e-commerce sites like Amazon, eBay & Flipkart, YouTube video recommendations and game recommendations in Xbox.

Q56. What is Collaborative filtering?

The process of filtering used by most of the recommender systems to find patterns or information by collaborating viewpoints, various data sources and multiple agents.

FREE WEBINAR
Time Series Analysis using Python

Movie	Alice	Bob	Carol	Dave
Shutter Island	4	3	5	1

Q57. How can outlier values be treated?

Outlier values can be identified by using univariate or any other graphical analysis method. If the number of outlier values is few then they can be assessed individually but for a large number of outliers, the values can be substituted with either the 99th or the 1st percentile values.

All extreme values are not outlier values. The most common ways to treat outlier values

1. To change the value and bring it within a range.

2. To just remove the value.

Q58. What are the various steps involved in an analytics project?

The following are the various **steps involved in an analytics project**:

1. Understand the Business problem

2. Explore the data and become familiar with it.

3. Prepare the data for modelling by detecting outliers, treating missing values, transforming variables, etc.

4. After data preparation, start running the model, analyze the result and tweak the approach. This is an iterative step until the best possible outcome is achieved.

5. Validate the model using a new data set.

6. Start implementing the model and track the result to analyze the performance of the model over the period of time.

Q59. During analysis, how do you treat missing values?

The extent of the missing values is identified after identifying the variables with missing values. If any patterns are identified the analyst has to concentrate on them as it could lead to interesting and meaningful business insights.

If there are no patterns identified, then the missing values can be substituted with mean or median values (imputation) or they can simply be ignored. Assigning a default value which can be mean, minimum or maximum value. Getting into the data is important.

If it is a categorical variable, the default value is assigned. The missing value is assigned a default value. If you have a distribution of data coming, for normal distribution give the mean value.

If 80% of the values for a variable are missing then you can answer that you would be dropping the variable instead of treating the missing values.

Q60. How will you define the number of clusters in a clustering algorithm?

Though the Clustering Algorithm is not specified, this question is mostly in reference to **K-Means clustering** where "K" defines the number of clusters. The objective of clustering is to group similar entities in a way that the entities within a group are similar to each other but the groups are different from each other.

For example, the following image shows three different groups.

FREE WEBINAR
Time Series Analysis using Python

Within Sum of squares is generally used to explain the homogeneity within a cluster. If you plot WSS for a range of number of clusters, you will get the plot shown below.

FREE WEBINAR

Time Series Analysis using Python

- The Graph is generally known as *Elbow Curve*.

- Red circled a point in above graph i.e. *Number of Cluster =6* is the point after which you don't see any decrement in WSS.

- This point is known as the **bending** point and taken as K in K – Means.

This is the widely used approach but few data scientists also use Hierarchical clustering first to create dendrograms and identify the distinct groups from there.

Q61. What is Ensemble Learning?

Ensemble Learning is basically combining a diverse set of learners(Individual models) together to improvise on the stability and predictive power of the model.

Q62. Describe in brief any type of Ensemble Learning?

Ensemble learning has many types but two more popular ensemble learning techniques are mentioned below.

Bagging

Bagging tries to implement similar learners on small sample populations and then takes a mean of all the predictions. In generalised bagging, you can use different learners on different population. As you expect this helps us to reduce the variance error.

FREE WEBINAR

Time Series Analysis using Python
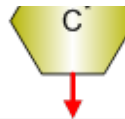
Boosting

Boosting is an iterative technique which adjusts the weight of an observation based on the last classification. If an observation was classified incorrectly, it tries to increase the weight of this observation and vice versa. Boosting in general decreases the bias error and builds strong predictive models. However, they may over fit on the training data.
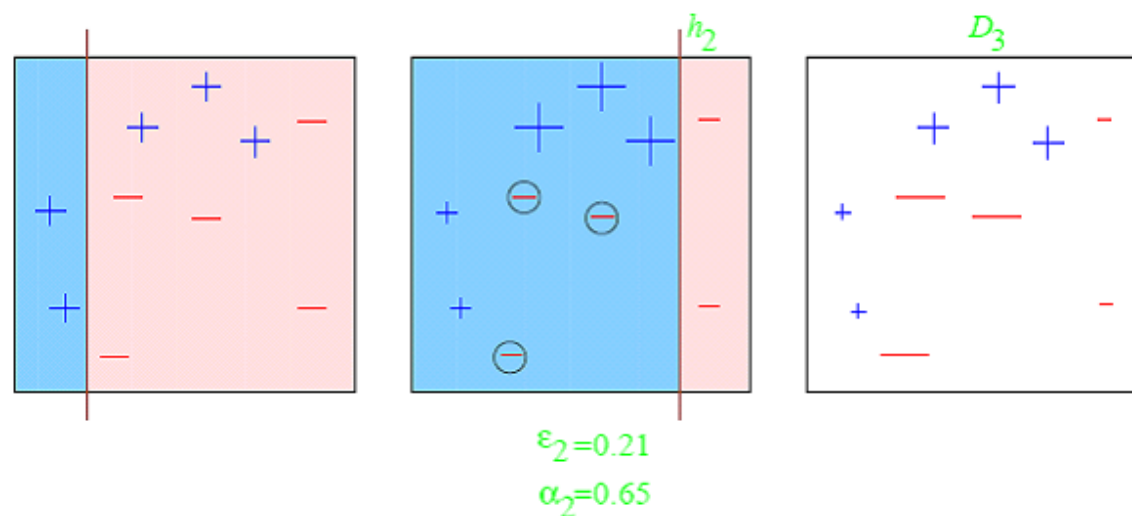
Q63. What is a Random Forest? How does it work?

Random forest is a versatile machine learning method capable of performing both regression and classification tasks. It is also used for dimensionality reduction, treats missing values, outlier values. It is a type of ensemble learning method, where a group of weak models combine to form a powerful model.

In Random Forest, we grow multiple trees as opposed to a single tree. To classify a new object based on attributes, each tree gives a classification. The forest chooses the classification having the most **votes**(Overall the trees in the forest) and in case of regression, it takes the average of outputs by different trees.

Q64. How Do You Work Towards a Random Forest?

The underlying principle of this technique is that several weak learners combined to provide a keen learner. The steps involved are

- Build several decision trees on bootstrapped training samples of data
- On each tree, each time a split is considered, a random sample of mm predictors is chose predictors

- Rule of thumb: At each split m=p√m=p
- Predictions: At the majority rule

fold 1: training[1 2 3], test[4]

fold 1: training[1 2 3 4], test[5]

Q66. What is a Box-Cox Transformation?

The dependent variable for a regression analysis might not satisfy one or more assumptions of an ordinary least squares regression. The residuals could either curve as the prediction increases or follow the skewed distribution. In such scenarios, it is necessary to transform the response variable so that the data meets the required assumptions. A Box cox transformation is a statistical technique to transform non-normal dependent variables into a normal shape. If the given data is not normal then most of the statistical techniques assume normality. Applying a box cox transformation means that you can run a broader number of tests.
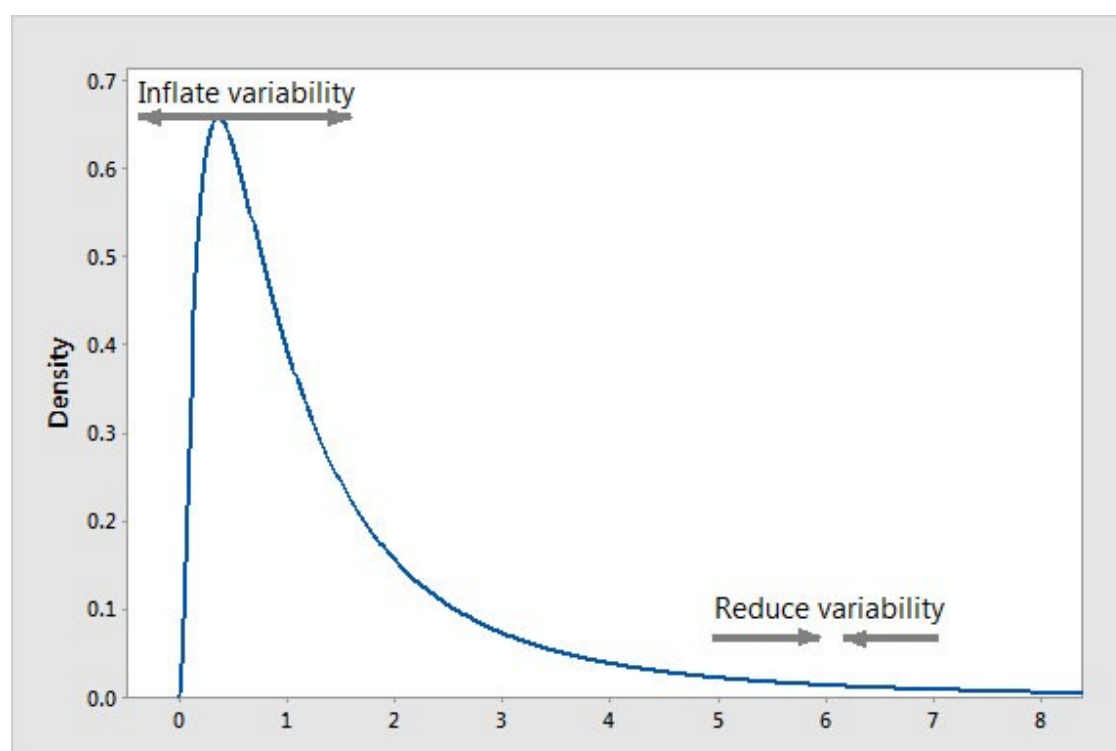
A Box-Cox transformation is a way to transform non-normal dependent variables into a normal shape. Normality is an important assumption for many statistical techniques, if your data isn't normal, applying a Box-Cox means that you are able to run a broader number of tests. The Box-Cox transformation is named after statisticians *George Box* and *Sir David Roxbee Cox* who collaborated on a 1964 paper and developed the technique.

Q67. How Regularly Must an Algorithm be Updated?

You will want to update an algorithm when:

- You want the model to evolve as data streams through infrastructure

- The underlying data source is changing

- There is a case of non-stationarity

- The algorithm underperforms/ results lack accuracy

Q68. If you are having 4GB RAM in your machine and you want to train your model on 10GB data set. How would you go about this problem? Have you ever faced this kind of problem in your machine learning/data science experience so far?

First of all, you have to ask which ML model you want to train.

For Neural networks: Batch size with Numpy array will work.

Steps:

1. Load the whole data in the Numpy array. Numpy array has a property to create a mapping of the complete data set, it doesn't load complete data set in memory.

1. Divide one big data set in small size data sets.

2. Use a partial fit method of SVM, it requires a subset of the complete data set.

3. Repeat step 2 for other subsets.

However, you could actually face such an issue in reality. So, you could check out the **_best laptop for Machine Learning_** to prevent that. Having said that, let's move on to some questions on deep learning.

DEEP LEARNING INTERVIEW QUESTIONS

Q69. What do you mean by Deep Learning?

Deep Learning is nothing but a paradigm of machine learning which has shown incredible promise in recent years. This is because of the fact that Deep Learning shows a great analogy with the functioning of the human brain.

Q70. What is the difference between machine learning and deep learning?

Machine learning is a field of computer science that gives computers the ability to learn without being explicitly programmed. Machine learning can be categorised in the following three categories.

1. Supervised machine learning,

2. Unsupervised machine learning,

3. Reinforcement learning
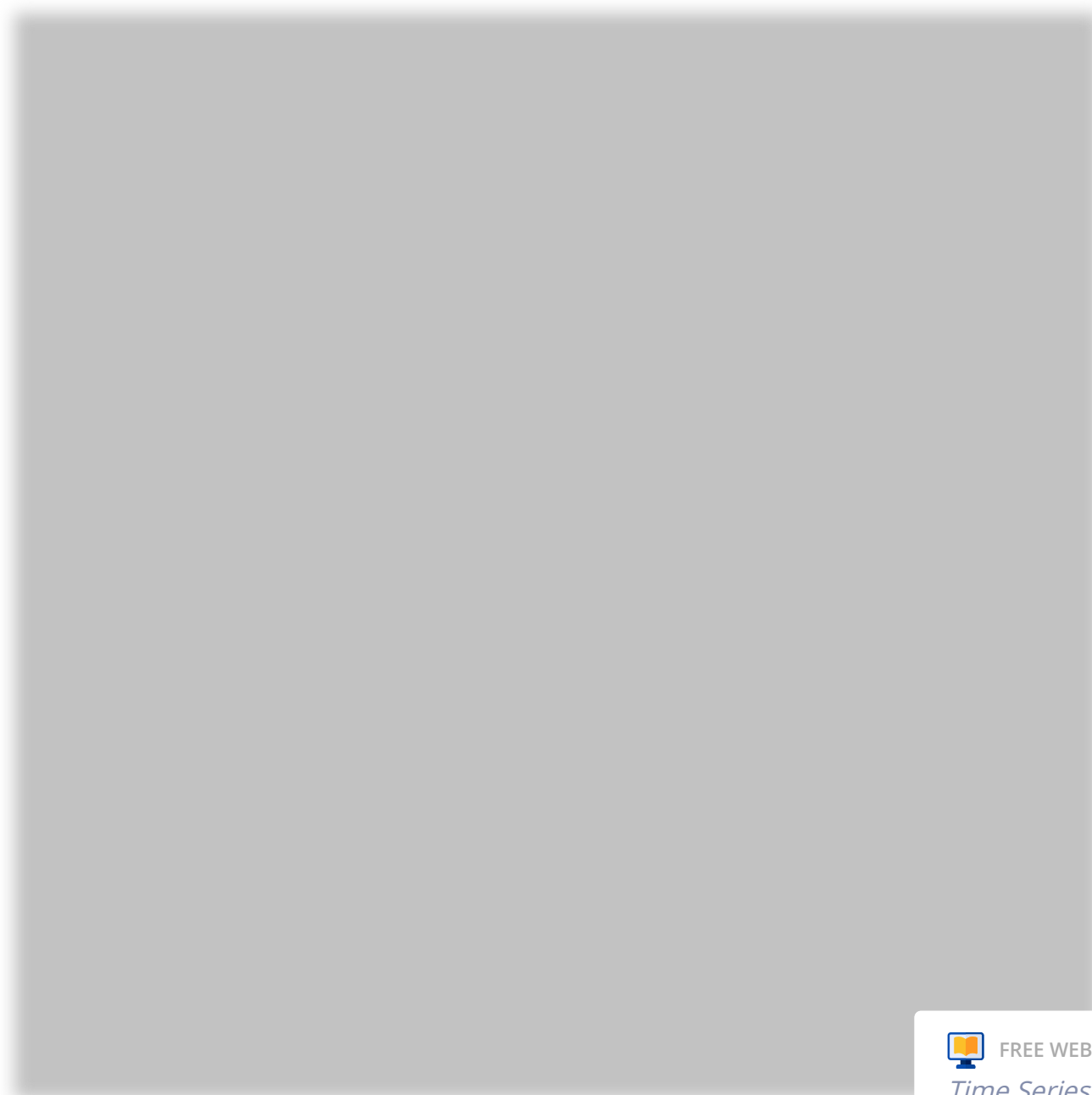
Deep Learning is a subfield of machine learning concerned with algorithms inspired by the structure and function of the brain called artificial neural networks.

## Q72. What is reinforcement learning?

Reinforcement Learning is learning what to do and how to map situations to actions. The end result is to maximise the numerical reward signal. The learner is not told which action to take but instead must discover which action will yield the maximum reward. Reinforcement learning is inspired by the learning of human beings, it is based on the reward/penalty mechanism.

Q73. What are Artificial Neural Networks?

Artificial Neural networks are a specific set of algorithms that have revolutionized machine learning. They are inspired by biological neural networks. **Neural Networks** can adapt to changing the input so the network generates the best possible result without needing to redesign the output criteria.

Q74. Describe the structure of Artificial Neural Networks?

Artificial Neural Networks works on the same principle as a biological Neural Network. It consists of inputs which get processed with weighted sums and Bias, with the help of Activation Functions.
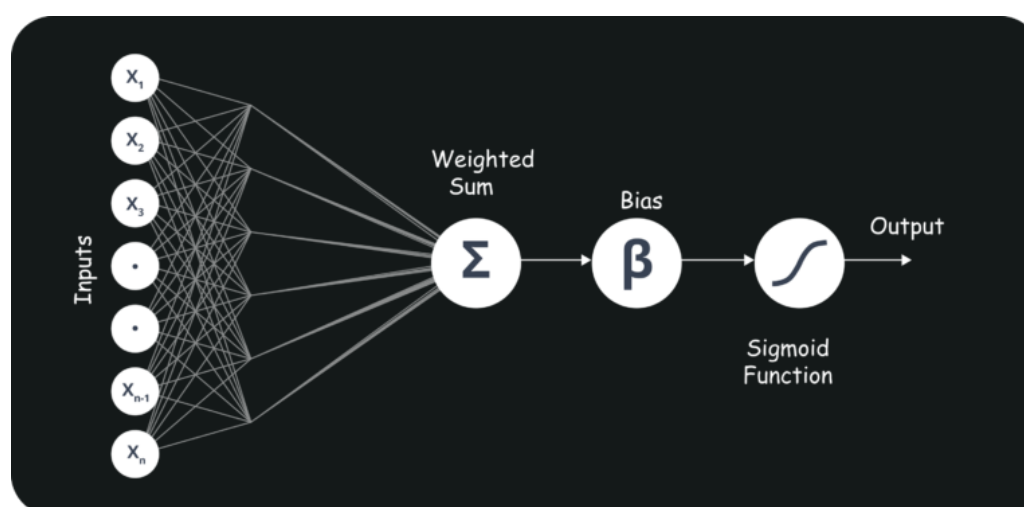
Q75. How Are Weights Initialized in a Network?

There are two methods here: we can either initialize the weights to zero or assign them randomly.

Initializing all weights to 0: This makes your model similar to a linear model. All the neurons and every layer perform the same operation, giving the same output and making the deep net useless.

Initializing all weights randomly: Here, the weights are assigned randomly by initializing them very close to 0. It gives better accuracy to the model since every neuron performs different computations. This is the most commonly used method.

Q76. What Is the Cost Function?

FREE WEBINAR
Time Series Analysis using Python

Also referred to as "loss" or "error," cost function is a measure to evaluate how good your model's performance is. It's used to compute the error of the output layer during backpropagation. We push that error backwards through the neural network and

weights. It will take many updates before reaching the minimum point.

If the learning rate is set too high, this causes undesirable divergent behaviour to the loss function due to drastic updates in weights. It may fail to converge (model can give a good output) or even diverge (data is too chaotic for the network to train).

Q79. What Is the Difference Between Epoch, Batch, and Iteration in Deep Learning?

- Epoch – Represents one iteration over the entire dataset (everything put into the training model).

- Batch – Refers to when we cannot pass the entire dataset into the neural network at once, so we divide the dataset into several batches.

- Iteration – if we have 10,000 images as data and a batch size of 200. then an epoch should run 50 iterations (10,000 divided by 50).

Q80. What Are the Different Layers on CNN?

There are four layers in *CNN*:

1. Convolutional Layer – the layer that performs a convolutional operation, creating several smaller picture windows to go over the data.

2. ReLU Layer – it brings non-linearity to the network and converts all the negative pixels to zero. The output is a rectified feature map.

3. Pooling Layer – pooling is a down-sampling operation that reduces the dimensionality of the feature map.

4. Fully Connected Layer – this layer recognizes and classifies the objects in the image.

| Input | Conv + Maxpool | Conv + Maxpool | Conv + Maxpool | Conv + Maxpool | FC | FC | Output |
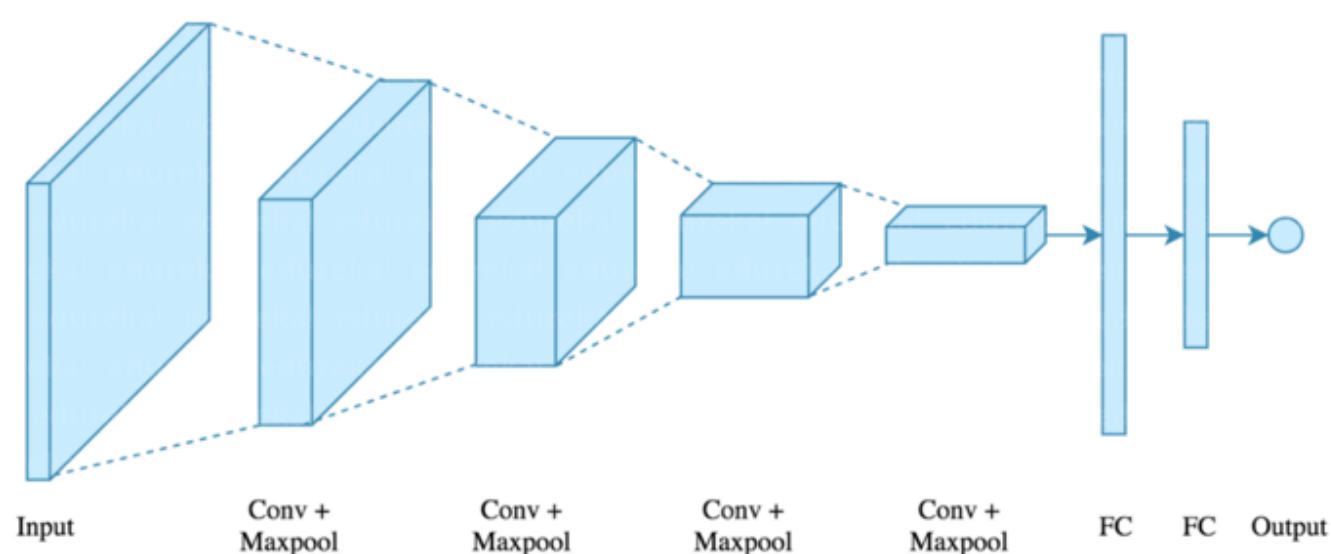
Q81. What Is Pooling on CNN, and How Does It Work?

Pooling is used to reduce the spatial dimensions of a CNN. It performs down-sampling operations to reduce the dimensionality and creates a pooled feature map by sliding a filter matrix over the input matrix.

Q82. What are Recurrent Neural Networks(RNNs)?

RNNs are a type of artificial neural networks designed to recognise the pattern from the sequence of data such as Time series, stock market and government agencies etc. To understand recurrent nets, first, you have to understand the basics of feedforward nets.

FREE WEBINAR
Time Series Analysis using Python

Both these networks RNN and feed-forward named after the way they channel information through a series of mathematical orations performed at the nodes of the network. One feeds information through straight(never touching the same node twice),
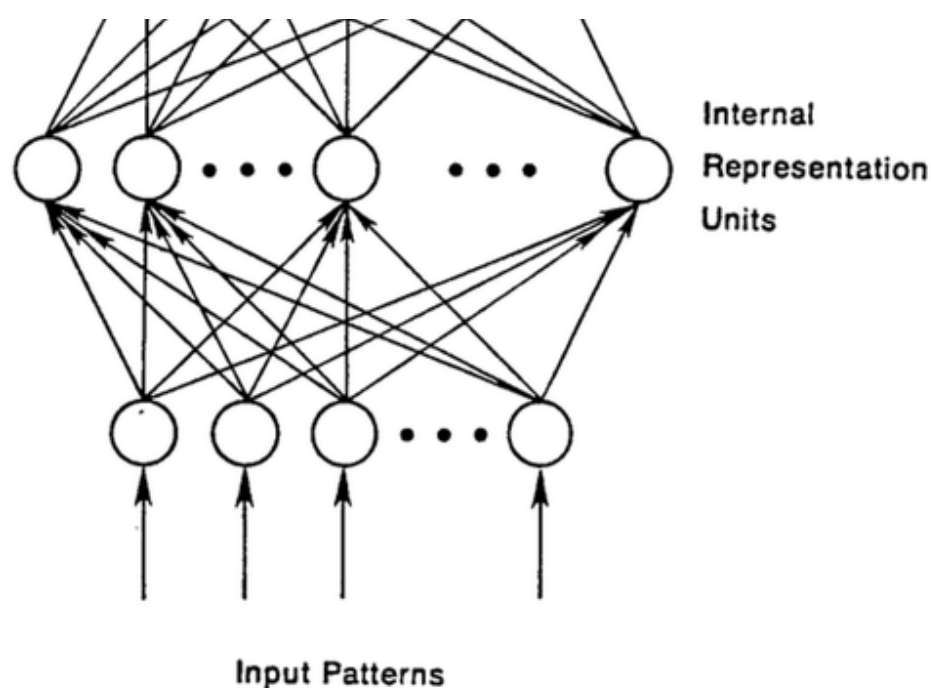
Recurrent networks, on the other hand, take as their input, not just the current input example they see, but also the what they have perceived previously in time.

The decision a recurrent neural network reached at time t-1 affects the decision that it will reach one moment later at time t. So recurrent networks have two sources of input, the present and the recent past, which combine to determine how they respond to new data, much as we do in life.

The error they generate will return via backpropagation and be used to adjust their weights until error can't go any lower. Remember, the purpose of recurrent nets is to accurately classify sequential input. We rely on the backpropagation of error and gradient descent to do so.

Q83. How Does an LSTM Network Work?

Long-Short-Term Memory (LSTM) is a special kind of recurrent neural network capable of learning long-term dependencies, remembering information for long periods as its default behaviour. There are three steps in an LSTM network:

- **Step 1:** The network decides what to forget and what to remember.
- **Step 2:** It selectively updates cell state values.
- **Step 3:** The network decides what part of the current state makes it to the output.

Q84. What Is a Multi-layer Perceptron(MLP)?

As in **Neural Networks, MLPs** have an input layer, a hidden layer, and an output layer. It has the same structure as a single layer **perceptron** with one or more hidden layers. A single layer perceptron can classify only linear separable classes with binary output (0,1), but MLP can classify nonlinear classes.

FREE WEBINAR

*Time Series Analysis using Python*

Except for the input layer, each node in the other layers uses a nonlinear activation function. This means the input layers, the data coming in, and the activation function is based upon all nodes and weights being added together, producing the output. MLP uses a supervised learning method called "backpropagation." In **backpropagation**, the neural network calculates the error with the help of cost function. It propagates this error backward from where it came (adjusts the weights to train the model more accurately).

Q85. Explain Gradient Descent.

To Understand Gradient Descent, Let's understand what is a *Gradient* first.

A **gradient** measures how much the output of a function changes if you change the inputs a little bit. It simply measures the change in all weights with regard to the change in error. You can also think of a gradient as the slope of a function.

Gradient Descent can be thought of climbing down to the bottom of a valley, instead of climbing up a hill. This is because it is a minimization algorithm that minimizes a given function (**Activation Function**).
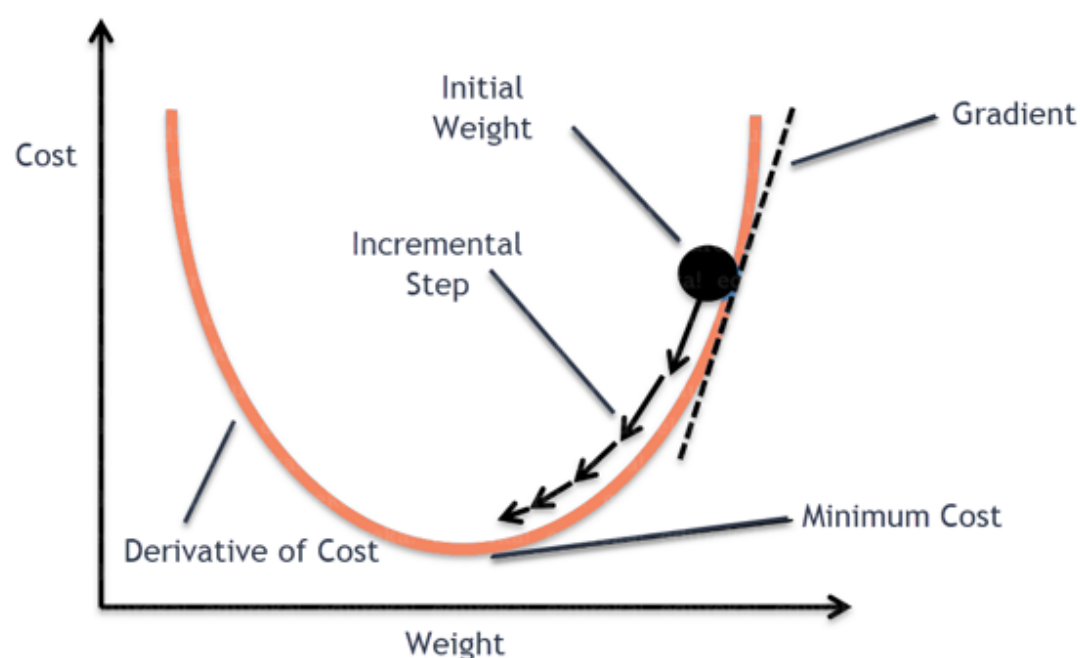
Q86. What is exploding gradients?

While training an RNN, if you see **exponentially growing (very large) error gradients** which accumulate and result in very large updates to neural network model weights during training, they're known as exploding gradients. At an extreme, the values of weights can become so large as to overflow and result in NaN values.

This has the effect of your model is unstable and unable to learn from your training data.

Q87. What is vanishing gradients?

While training an RNN, your slope can become either too small; this makes the training difficult. When the slope is too small, the problem is known as a Vanishing Gradient. It leads to long training times, poor performance, and low accuracy.

Q89. What is Back Propagation and Explain it's Working.

Backpropagation is a training algorithm used for multilayer neural network. In this method, we move the error from an end of the network to all weights inside the network and thus allowing efficient computation of the gradient.

It has the following steps:

FREE WEBINAR
Time Series Analysis using Python

- Forward Propagation of Training Data

- Derivatives are computed using output and target

- **Batch Gradient Descent:** We calculate the gradient for the whole dataset and perform the update at each iteration.

- **Mini-batch Gradient Descent**: It's one of the most popular optimization algorithms. It's a variant of Stochastic Gradient Descent and here instead of single training example, mini-batch of samples is used.

Q91. What are the different *Deep Learning Frameworks*?

- *Pytorch*

- *TensorFlow*

- Microsoft Cognitive Toolkit

- Keras

- Caffe

- Chainer

Q92. What is the role of the Activation Function?

The **Activation function** is used to introduce non-linearity into the neural network helping it to learn more complex function. Without which the neural network would be only able to learn linear function which is a linear combination of its input data. An activation function is a function in an artificial neuron that delivers an output based on inputs.

Q93. Name a few *Machine Learning libraries* for various purposes.

Purpose	Libraries
Scientific Computation	Numpy
Tabular Data	Pandas
Data Modelling & Preprocessing	Scikit Learn
Time-Series Analysis	Statsmodels
Text processing	Regular Expressions, NLTK
Deep Learning	Tensorflow, Pytorch

Q94. What is an Auto-Encoder?

Auto-encoders are simple learning networks that aim to transform inputs into outputs with the minimum possible error. This means that we want the output to be as close to input as possible. We add a couple of layers between the input and the output, and the sizes of these layers are smaller than the input layer. The auto-encoder receives unlabelled input which is then encoded to reconstruct the input.
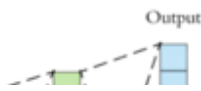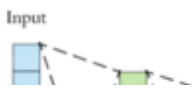
Data Science Masters Program

Weekday / Weekend Batches

See Batch Details

FREE WEBINAR
Time Series Analysis using Python

regularities in the training data. The Boltzmann machine is basically used to optimise the weights and the quantity for the given problem. The learning algorithm is very slow in networks with many layers of feature detectors. "**Restricted Boltzmann Machines**" algorithm has a single layer of feature detectors which makes it faster than the rest.
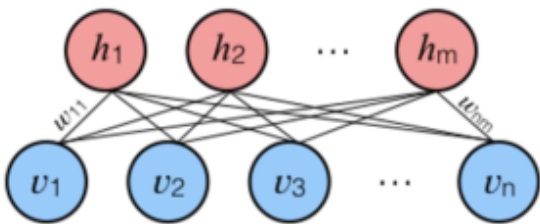
Q96. What Is Dropout and Batch Normalization?

Dropout is a technique of dropping out hidden and visible units of a network randomly to prevent overfitting of data (typically dropping 20 per cent of the nodes). It doubles the number of iterations needed to converge the network.
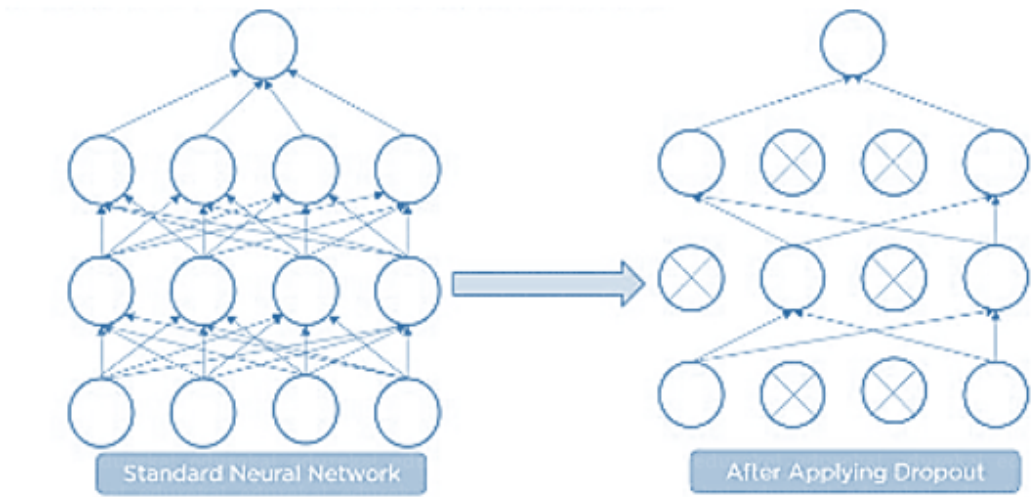
Batch normalization is the technique to improve the performance and stability of neural networks by normalizing the inputs in every layer so that they have mean output activation of zero and standard deviation of one.

### Q97. What Is the Difference Between Batch Gradient Descent and Stochastic Gradient Descent?

Batch Gradient Descent	Stochastic Gradient Descent
The batch gradient computes the gradient using the entire dataset.	The stochastic gradient computes the gradient using a single sample.
It takes time to converge because the volume of data is huge, and weights update slowly.	It converges much faster than the batch gradient because it updates weight more frequently.

Q98. Why Is Tensorflow the Most Preferred Library in Deep Learning?

Tensorflow provides both C++ and Python APIs, making it easier to work on and has a faster compilation time compared to other Deep Learning libraries like Keras and Torch. Tensorflow supports both CPU and GPU computing devices.

Q99. What Do You Mean by Tensor in Tensorflow?

A tensor is a mathematical object represented as arrays of higher dimensions. These arrays of data with different dimensions and ranks fed as input to the neural network are called "***Tensors***."
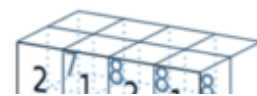
FREE WEBINAR

*Time Series Analysis using Python*

a "DataFlow Graph."

Q101. What is a Generative Adversarial Network?

Suppose there is a wine shop purchasing wine from dealers, which they resell later. But some dealers sell fake wine. In this case, the shop owner should be able to distinguish between fake and authentic wine.

The forger will try different techniques to sell fake wine and make sure specific techniques go past the shop owner's check. The shop owner would probably get some feedback from wine experts that some of the wine is not original. The owner would have to improve how he determines whether a wine is fake or authentic.

The forger's goal is to create wines that are indistinguishable from the authentic ones while the shop owner intends to tell if the wine is real or not accurately

Let us understand this example with the help of an image.
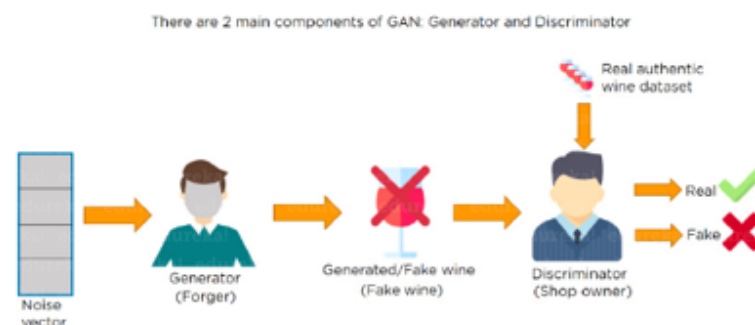
There is a noise vector coming into the forger who is generating fake wine.

Here the forger acts as a Generator.

The shop owner acts as a Discriminator.

The Discriminator gets two inputs; one is the fake wine, while the other is the real authentic wine. The shop owner has to figure out whether it is real or fake.

So, there are two primary components of Generative Adversarial Network (GAN) named:

1. Generator

2. Discriminator

The generator is a CNN that keeps keys producing images and is closer in appearance to the real images while the discriminator tries to determine the difference between real and fake images The ultimate aim is to make the discriminator learn to identify real and fake images.

Apart from the very technical questions, your interviewer could even hit you up with a few simple ones to check your overall confidence, in the likes of the following.

Q102. What are the important skills to have in Python with regard to data analysis?

The following are some of the important skills to possess which will come handy when performing data analysis using Python.

- Good understanding of the built-in data types especially lists, dictionaries, tuples, and sets.

- Mastery of N-dimensional NumPy Arrays.

- Mastery of Pandas dataframes.

FREE WEBINAR

Time Series Analysis using Python

- Ability to perform element-wise vector and matrix operations on NumPy arrays.

- Knowing that you should use the Anaconda distribution and the conda package manager.

I hope this set of Data Science Interview Questions and Answers will help you in preparing for your interviews. All the best!

Got a question for us? Please mention it in the comments section and we will get back to you at the earliest.

Edureka has a specially curated **Data Science course** *which helps you gain expertise in Machine Learning Algorithms like K-Means Clustering, Decision Trees, Random Forest, Naive Bayes. You'll learn the concepts of Statistics, Time Series, Text Mining and an introduction to Deep Learning as well. You'll solve real-life case studies on Media, Healthcare, Social Media, Aviation, HR. New batches for this course are starting soon!!*

Recommended videos for you

▶   ▶   ▶   ▶

Application of Clustering in Data Science Using Real-Time Examples

Linear Regression With R

Data Science : Make Smarter Business Decisions

Python Classes – Pyt Programming Tutori

Watch Now   Watch Now   Watch Now   Watch Now

‹›

Recommended blogs for you

What is the Main Function in Python and how to use it?

Python Career Opportunities: Your Career Guide To Python Programming

How To Implement Classification In Machine Learning?

FIFA World Cup 2018 Analyzing Fifa Datas Python

Read Article   Read Article   Read Article   Read Article

‹›

Comments

1 Comment

Jimmy Vivas says:

Sep 13, 2017 at 9:34 pm GMT

FREE WEBINAR

Time Series Analysis using Python

very well video, thanks to dedicate your time teaching us.

Reply

Trending Courses in Data Science

Data Science Certification Training with Pyth ...

👤 90k Enrolled Learners
📅 Weekend/Weekday
🎥 Live Class

Reviews
⭐⭐⭐⭐⭐ **5** (35650)

Python Programming Certification Training

👤 26k Enrolled Learners
📅 Weekend
🎥 Live Class

Reviews
⭐⭐⭐⭐⭐ **5** (10200)

Machine Learning Certification Training

👤 11k Enrolled Learners
📅 Weekend
🎥 Live Class

Reviews
⭐⭐⭐⭐⭐ **5** (4250)

Data Science Cert Course using R

👤 38k Enrolled Learne
📅 Weekend
🎥 Live Class

Reviews
⭐⭐⭐⭐⭐ **5** (1515

‹›

Browse Categories

Artificial Intelligence | BI and Visualization | Big Data | Blockchain | Cloud Computing | Cyber Security | Data Warehousing and ETL

Databases | DevOps | Digital Marketing | Enterprise | Front End Web Development | Mobile Development | Operating Systems

Programming & Frameworks | Project Management and Methodologies | Robotic Process Automation | Software Testing

Systems & Architecture

edureka!

TRENDING CERTIFICATION COURSES

DevOps Certification Training

AWS Architect Certification Training

Big Data Hadoop Certification Training

Tableau Training & Certification

Python Certification Training for Data Science

TRENDING MASTERS COURSES

Data Scientist Masters Program

DevOps Engineer Masters Program

Cloud Architect Masters Program

Big Data Architect Masters Program

Machine Learning Engineer Masters

FREE WEBINAR

Time Series Analysis using Python

Selenium Certification Training

PMP® Certification Exam Training

Full Stack Web Developer Masters Program

Business Intelligence Masters Program

Subscribe to our Newsletter, and get personalized recommendations.

Sign up with Google

Signup with Facebook

Already have an account? Sign in.

Reviews

Contact us

Blog

Community

Sitemap

Blog Sitemap

Community Sitemap

Webinars

Become an Affiliate

Become a Partner

Hire from Edureka

DOWNLOAD APP

CATEGORIES ⌄

CATEGORIES

Cloud Computing | DevOps | Big Data | Data Science | BI and Visualization | Programming & Frameworks | Software Testing |

Project Management and Methodologies | Robotic Process Automation | Frontend Development | Data Warehousing and ETL | Artificial Intelligence |

Blockchain | Databases | Cyber Security | Mobile Development | Operating Systems | Architecture & Design Patterns | Digital Marketing

TRENDING BLOG ARTICLES ⌄

TRENDING BLOG ARTICLES

Selenium tutorial | Selenium interview questions | Java tutorial | What is HTML | Java interview questions | PHP tutorial | JavaScript interview questions |

Spring tutorial | PHP interview questions | Inheritance in Java | Polymorphism in Java | Spring interview questions | Pointers in C | Linux commands |

Android tutorial | JavaScript tutorial | jQuery tutorial | SQL interview questions | MySQL tutorial | Machine learning tutorial | Python tutorial |

What is machine learning | Ethical hacking tutorial | SQL injection | AWS certification career opportunities | AWS tutorial | What Is cloud computing |

What is blockchain | Hadoop tutorial | What is artificial intelligence | Node Tutorial | Collections in Java | Exception handling in java |

Python Programming Language | Python interview questions | Multithreading in Java | ReactJS Tutorial | Data Science vs Big Data vs Data Analyt… |

Software Testing Interview Questions | R Tutorial | Java Programs | JavaScript Reserved Words and Keywor… | Implement thread.yield() in Java: Exam… |

Implement Optical Character Recogniti… | All you Need to Know About Implemen…

Terms & Conditions

Legal & Privacy

FREE WEBINAR
Time Series Analysis using Python