TOP 30

Data Science

INTERVIEW QUESTION

Created by- TOPPERWORLD

Q 1. What is Data Science?

Ans: An interdisciplinary field that constitutes various scientific processes, algorithms, tools, and machine learning techniques working to help find common patterns and gather sensible insights from the given raw input data using statistical and mathematical analysis is called Data Science.

1. It starts with gathering the business requirements and relevant data.
2. Once the data is acquired, it is maintained by performing data cleaning, data warehousing, data staging, and data architecture.
3. Data processing does the task of exploring the data, mining it, and analyzing it which can be finally used to generate the summary of the insights extracted from the data.
4. Once the exploratory steps are completed, the cleansed data is subjected to various algorithms like predictive analysis, regression, text mining, recognition patterns, etc depending on the requirements.
5. In the final stage, the results are communicated to the business in a visually appealing manner. This is where the skill of data visualization, reporting, and different business intelligence tools come into the picture.

Q 2.What is the difference between data analytics and data science?

Ans:

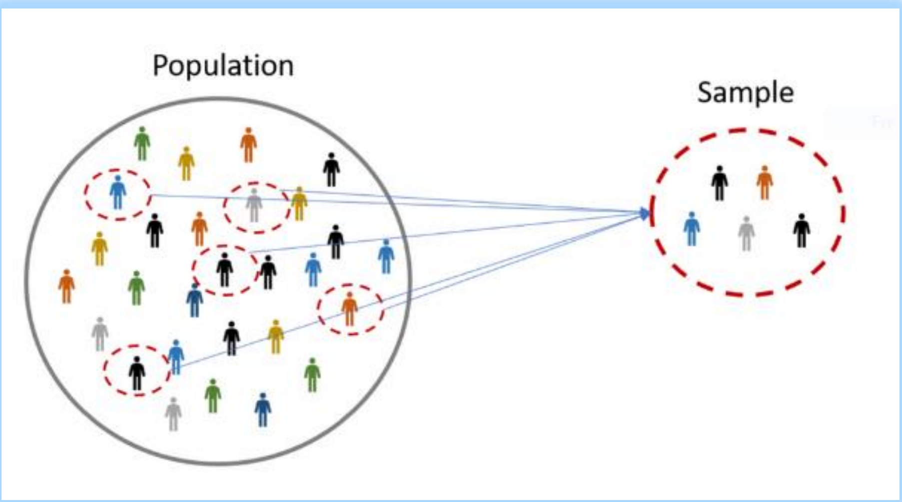
- Data science involves the task of transforming data by using various technical analysis methods to extract meaningful insights using which a data analyst can apply to their business scenarios.

- Data analytics deals with checking the existing hypothesis and information and answers questions for a better and effective business-related decision-making process.
- Data Science drives innovation by answering questions that build connections and answers for futuristic problems. Data analytics focuses on getting present meaning from existing historical context whereas data science focuses on predictive modeling.

- Data Science can be considered as a broad subject that makes use of various mathematical and scientific tools and algorithms for solving complex problems whereas data analytics can be considered as a specific field dealing with specific concentrated problems using fewer tools of statistics and visualization.

Q 3. What are some of the techniques used for sampling? What is the main advantage of sampling?

Ans: Data analysis can not be done on a whole volume of data at a time especially when it involves larger datasets. It becomes crucial to take some data samples that can be used for representing the whole population and then perform analysis on it. While doing this, it is very much necessary to carefully take sample data out of the huge data that truly represents the entire dataset.



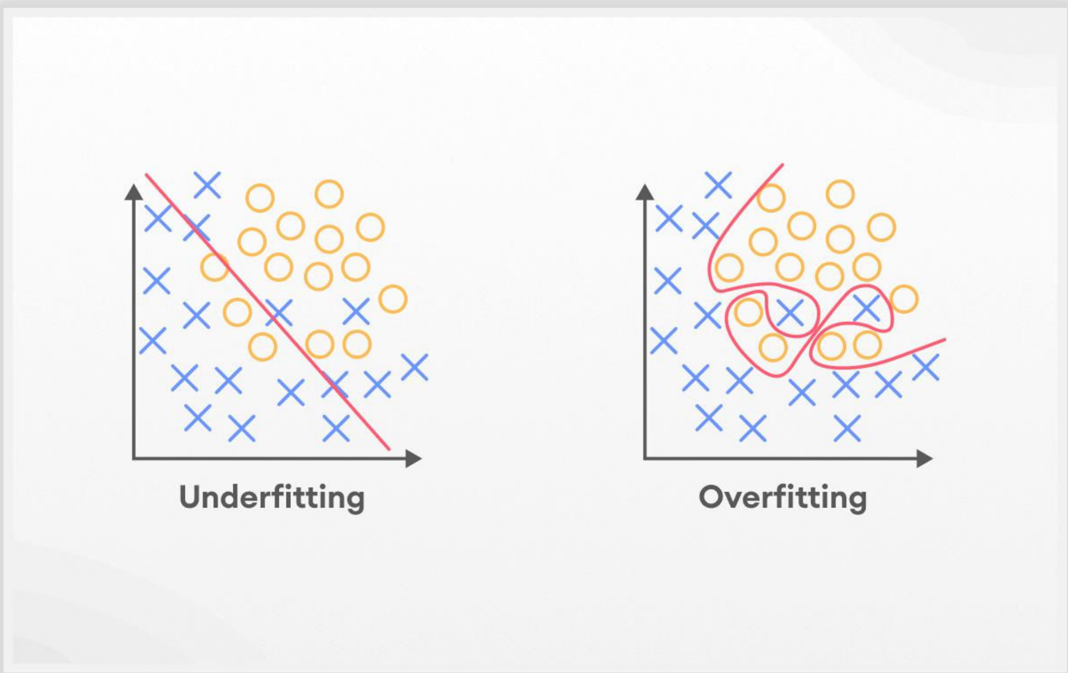
There are majorly two categories of sampling techniques based on the usage of statistics, they are:

1) **Probability Sampling techniques:** Clustered sampling, Simple random sampling, Stratified sampling.
2) **Non-Probability Sampling techniques:** Quota sampling, Convenience sampling, snowball sampling, etc.

Q 4. List down the conditions for Overfitting and Underfitting.

Ans:

- **Overfitting:** The model performs well only for the sample training data. If any new data is given as input to the model, it fails to provide any result. These conditions occur due to low bias and high variance in the model. Decision trees are more prone to overfitting.
- **Underfitting:** Here, the model is so simple that it is not able to identify the correct relationship in the data, and hence it does not perform well even on the test data. This can happen due to high bias and low variance. Linear regression is more prone to Underfitting.



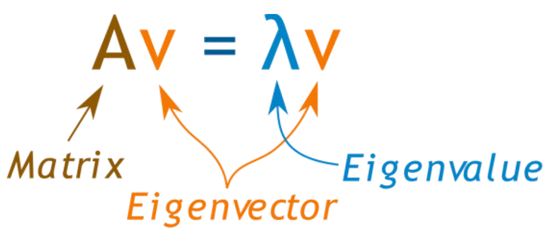
Q 5. Differentiate between the long and wide format data.

Ans:

Long format Data	Wide-Format Data
Here, each row of the data represents the one-time information of a subject. Each subject would have its data in different/ multiple rows.	Here, the repeated responses of a subject are part of separate columns.
The data can be recognized by considering rows as groups.	The data can be recognized by considering columns as groups.
This data format is most commonly used in R analyses and to write into log files after each trial.	This data format is rarely used in R analyses and most commonly used in stats packages for repeated measures ANOVAs.

Q 6. What are Eigenvectors and Eigenvalues?

Ans: Eigenvectors are column vectors or unit vectors whose length/magnitude is equal to 1. They are also called right vectors. Eigenvalues are coefficients that are applied on eigenvectors which give these vectors different values for length or magnitude.



A matrix can be decomposed into Eigenvectors and Eigenvalues and this process is called Eigen decomposition. These are then eventually used in machine learning methods like PCA (Principal Component Analysis) for gathering valuable insights from the given matrix.

Q 7. What does it mean when the p-values are high and low?

Ans: A p-value is the measure of the probability of having results equal to or more than the results achieved under a specific hypothesis assuming that the null hypothesis is correct. This represents the probability that the observed difference occurred randomly by chance.

- Low p-value which means values ≤ 0.05 means that the null hypothesis can be rejected and the data is unlikely with true null.
- High p-value, i.e values ≥ 0.05 indicates the strength in favor of the null hypothesis. It means that the data is like with true null.
- p-value = 0.05 means that the hypothesis can go either way.

Q 8. When is resampling done?

Ans: Resampling is a methodology used to sample data for improving accuracy and quantify the uncertainty of population parameters.

It is done to ensure the model is good enough by training the model on different patterns of a dataset to ensure variations are handled.

It is also done in the cases where models need to be validated using random subsets or when substituting labels on data points while performing tests.

Q 9. What do you understand by Imbalanced Data?

Ans: Data is said to be highly imbalanced if it is distributed unequally across different categories. These datasets result in an error in model performance and result in inaccuracy.

Q 10. Are there any differences between the expected value and mean value?

Ans: There are not many differences between these two, but it is to be noted that these are used in different contexts. The mean value generally refers to the probability distribution whereas the expected value is referred to in the contexts involving random variables.

Q 11. What do you understand by Survivorship Bias?

Ans: This bias refers to the logical error while focusing on aspects that survived some process and overlooking those that did not work due to lack of prominence. This bias can lead to deriving wrong conclusions.

Q 12.Define the terms KPI, lift, model fitting, robustness and DOE.

Ans:

- **KPI:** KPI stands for Key Performance Indicator that measures how well the business achieves its objectives.
- **Lift:** This is a performance measure of the target model measured against a random choice model. Lift indicates how good the model is at prediction versus if there was no model.
- **Model fitting:** This indicates how well the model under consideration fits given observations.
- **Robustness:** This represents the system ’ s capability to handle differences and variances effectively.
- **DOE:** stands for the design of experiments, which represents the task design aiming to describe and explain information variation under hypothesized conditions to reflect variables.

Q 13. Define confounding variables.

Ans: Confounding variables are also known as confounders. These variables are a type of extraneous variables that influence both independent and dependent variables causing spurious association and mathematical relationships between those variables that are associated but are not casually related to each other.

Q 14. Define and explain selection bias?

Ans: The selection bias occurs in the case when the researcher has to make a decision on which participant to study. The selection bias is associated with those researches when the participant selection is not random. The selection bias is also called the selection effect. The selection bias is caused by as a result of the method of sample collection.

Four types of selection bias are explained below:

1. **Sampling Bias:** As a result of a population that is not random at all, some members of a population have fewer chances of getting included than others, resulting in a biased sample. This causes a systematic error known as sampling bias.
2. **Time interval:** Trials may be stopped early if we reach any extreme value but if all variables are similar invariance, the variables with the highest variance have a higher chance of achieving the extreme value.
3. **Data:** It is when specific data is selected arbitrarily and the generally agreed criteria are not followed.
4. **Attrition:** Attrition in this context means the loss of the participants. It is the discounting of those subjects that did not complete the trial.

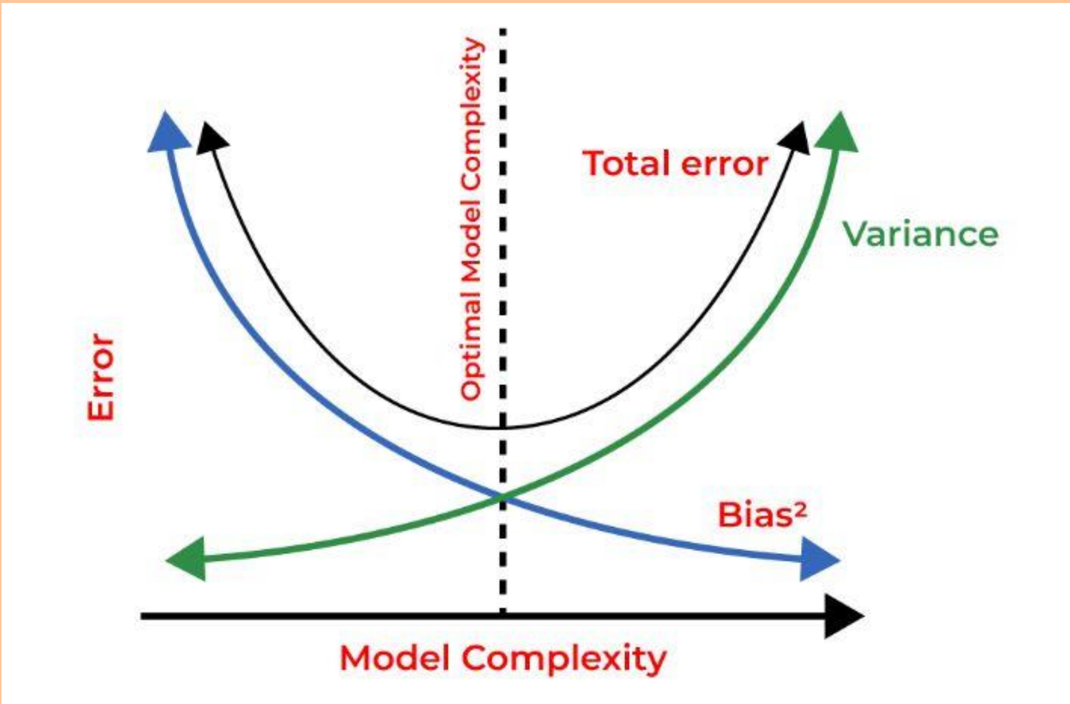
Q 15. Define bias-variance trade-off?

Ans:

- **Bias:** It is a kind of error in a machine learning model when an ML Algorithm is oversimplified. When a model is trained, at that time it makes simplified assumptions so that it can easily understand the target function. Some algorithms that have low bias are Decision Trees, SVM, etc. On the other hand, logistic and linear regression algorithms are the ones with a high bias.
- **Variance:** Variance is also a kind of error. It is introduced into an ML Model when an ML algorithm is made highly complex. This model also learns noise from the data set that is meant for training. It further

performs badly on the test data set. This may lead to over lifting as well as high sensitivity.

When the complexity of a model is increased, a reduction in the error is seen. This is caused by the lower bias in the model. But, this does not happen always till we reach a particular point called the optimal point. After this point, if we keep on increasing the complexity of the model, it will be over lifted and will suffer from the problem of high variance. We can represent this situation with the help of a graph as shown below:



As you can see from the image above, before the optimal point, increasing the complexity of the model reduces the error (bias). However, after the optimal point, we see that the increase in the complexity of the machine learning model increases the variance.

- **Trade-off Of Bias And Variance:** So, as we know that bias and variance, both are errors in machine learning models, it is very essential that any machine learning model has low variance as well as a low bias so that it can achieve good performance.

Let us see some examples. The **K-Nearest Neighbor Algorithm** is a good example of an algorithm with low bias and high variance. This trade-off can easily be reversed by increasing the k value which in turn results in increasing the number of neighbours. This, in turn, results in increasing the bias and reducing the variance.

Another example can be the algorithm of a support vector machine. This algorithm also has a high variance and obviously, a low bias and we can reverse the trade-off by increasing the value of parameter C. Thus, increasing the C parameter increases the bias and decreases the variance.

So, the trade-off is simple. If we increase the bias, the variance will decrease and vice versa.

Q 16. Define the confusion matrix?

Ans: It is a matrix that has 2 rows and 2 columns. It has 4 outputs that a binary classifier provides to it. It is used to derive various measures like specificity, error rate, accuracy, precision, sensitivity, and recall.

		Predicted	
		0	1
Actual	0	TN	FP
	1	FN	TP

$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$

The test data set should contain the correct and predicted labels. The labels depend upon the performance. For instance, the predicted labels are the same if the binary classifier performs perfectly. Also, they match the part of observed labels in real-world scenarios. The four outcomes shown above in the confusion matrix mean the following:

1) **True Positive:** This means that the positive prediction is correct.
2) **False Positive:** This means that the positive prediction is incorrect.
3) **True Negative:** This means that the negative prediction is correct.
4) **False Negative:** This means that the negative prediction is incorrect.

The formulas for calculating basic measures that comes from the confusion matrix are:

1. **Error rate**: (FP + FN)/(P + N)
2. **Accuracy**: (TP + TN)/(P + N)
3. **Sensitivity** = TP/P
4. **Specificity** = TN/N
5. **Precision** = TP/(TP + FP)
6. **F-Score** = (1 + b)(PREC.REC)/(b2 PREC + REC) Here, b is mostly 0.5 or 1 or 2.

In these formulas:

FP = false positive
FN = false negative
TP = true positive
RN = true negative

Also,

Sensitivity is the measure of the True Positive Rate. It is also called recall.
Specificity is the measure of the true negative rate.
Precision is the measure of a positive predicted value.
F-score is the harmonic mean of precision and recall.

Q 17. What is logistic regression? State an example where you have recently used logistic regression.

Ans: Logistic Regression is also known as the logit model. It is a technique to predict the binary outcome from a linear combination of variables (called the predictor variables).

For example, let us say that we want to predict the outcome of elections for a particular political leader. So, we want to find out whether this leader is going to win the election or not. So, the result is binary i.e. win (1) or loss (0). However, the input is a combination of linear variables like the money spent on advertising, the past work done by the leader and the party, etc.

Q 18. What is Linear Regression? What are some of the major drawbacks of the linear model?

Ans: Linear regression is a technique in which the score of a variable Y is predicted using the score of a predictor variable X. Y is called the criterion variable. Some of the drawbacks of Linear Regression are as follows:

- The assumption of linearity of errors is a major drawback.
- It cannot be used for binary outcomes. We have Logistic Regression for that.
- Overfitting problems are there that can’t be solved.

Q 19. What is a random forest? Explain it’s working.

Ans: Classification is very important in machine learning. It is very important to know to which class does an observation belongs. Hence, we have various classification algorithms in machine learning like logistic regression, support vector machine, decision trees, Naive Bayes classifier, etc.

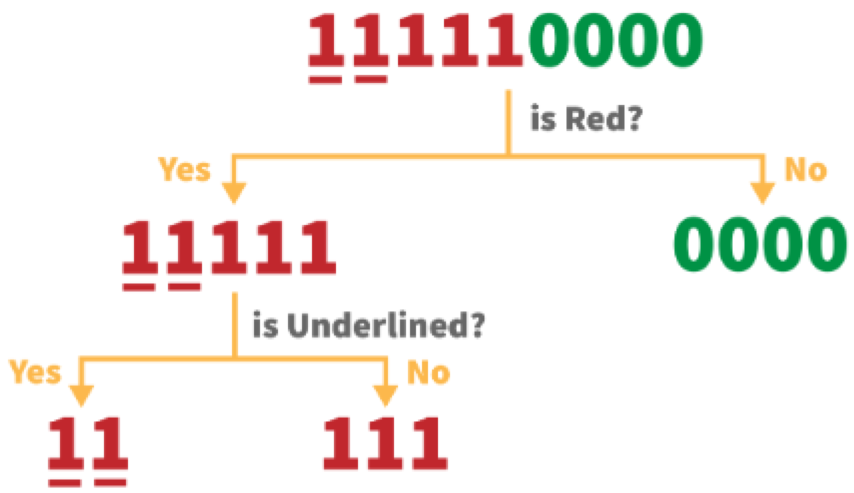
One such classification technique that is near the top of the classification hierarchy is the **random forest** classifier.

So, firstly we need to understand a decision tree before we can understand the random forest classifier and its works. So, let us say that we have a string as given below:

So, we have the string with 5 ones and 4 zeroes and we want to classify the characters of this string using their features.

These features are colour (red or green in this case) and whether the observation (i.e. character) is underlined or not. Now, let us say that we are only interested in red and underlined observations. S

o, the decision tree would look something like this:



So, we started with the colour first as we are only interested in the red observations and we separated the red and the green-coloured characters. After that, the "No" branch i.e. the branch that had all the green coloured characters was not expanded further as we want only red-underlined characters. So, we expanded the "Yes" branch and we again got a "Yes" and a "No" branch based on the fact whether the characters were underlined or not.

So, this is how we draw a typical decision tree. However, the data in real life is not this clean but this was just to give an idea about the working of the decision trees.

➢ Random Forest

It consists of a large number of decision trees that operate as an ensemble. Basically, each tree in the forest gives a class prediction and the one with the maximum number of votes becomes the prediction of our model. For instance, in the example shown below, 4 decision trees predict 1, and 2 predict 0. Hence, prediction 1 will be considered.

The underlying principle of a random forest is that several weak learners combine to form a keen learner. The steps to build a random forest are as follows:

- Build several decision trees on the samples of data and record their predictions.

- Each time a split is considered for a tree, choose a random sample of mm predictors as the split candidates out of all the pp predictors. This happens to every tree in the random forest.
- Apply the rule of thumb i.e. at each split m = p √ m = p.
- Apply the predictions to the majority rule.

Q 20. In a time interval of 15-minutes, the probability that you may see a shooting star or a bunch of them is 0.2. What is the percentage chance of you seeing at least one star shooting from the sky if you are under it for about an hour?

Ans: Let us say that Prob is the probability that we may see a minimum of one shooting star in 15 minutes.

So, Prob = 0.2

Now, the probability that we may not see any shooting star in the time duration of 15 minutes is = 1 - Prob

1-0.2 = 0.8

The probability that we may not see any shooting star for an hour is:

$$= (1-Prob)(1-Prob)(1-Prob)*(1-Prob)$$
$$= 0.8 * 0.8 * 0.8 * 0.8 = (0.8)^4$$
$$\approx 0.40$$

So, the probability that we will see one shooting star in the time interval of an hour is = 1-0.4 = 0.6

So, there are approximately 60% chances that we may see a shooting star in the time span of an hour.

Q 21. What is deep learning? What is the difference between deep learning and machine learning?

Ans: Deep learning is a paradigm of machine learning. In deep learning, multiple layers of processing are involved in order to extract high features from the data. The neural networks are designed in such a way that they try to simulate the human brain.

Deep learning has shown incredible performance in recent years because of the fact that it shows great analogy with the human brain.

The difference between machine learning and deep learning is that deep learning is a paradigm or a part of machine learning that is inspired by the structure and functions of the human brain called the artificial neural networks.

Q 22. What is a Gradient and Gradient Descent?

Ans:

- **Gradient:** Gradient is the measure of a property that how much the output has changed with respect to a little change in the input. In other words, we can say that it is a measure of change in the weights with respect to the change in error.
- **Gradient Descent:** Gradient descent is a minimization algorithm that minimizes the Activation function. Well, it can minimize any function given to it but it is usually provided with the activation function only.

Gradient descent, as the name suggests means descent or a decrease in something. The analogy of gradient descent is often taken as a person climbing down a hill/mountain. The following is the equation describing what gradient descent means:

So, if a person is climbing down the hill, the next position that the climber has to come to is denoted by "b" in this equation. Then, there is a minus sign because it denotes the minimization (as gradient descent is a minimization algorithm). The Gamma is called a waiting factor and the remaining term which is the Gradient term itself shows the direction of the steepest descent.

Q23.How are the time series problems different from other regression problems?

Ans:

- Time series data can be thought of as an extension to linear regression which uses terms like autocorrelation, movement of averages for summarizing historical data of y-axis variables for predicting a better future.
- Forecasting and prediction is the main goal of time series problems where accurate predictions can be made but sometimes the underlying reasons might not be known.
- Having Time in the problem does not necessarily mean it becomes a time series problem. There should be a relationship between target and time for a problem to become a time series problem.
- The observations close to one another in time are expected to be similar to the ones far away which provide accountability for seasonality. For instance, today's weather would be similar to tomorrow's weather but not similar to weather from 4 months from today. Hence, weather prediction based on past data becomes a time series problem.

Q 24. What are RMSE and MSE in a linear regression model?

Ans:

- **RMSE:** RMSE stands for Root Mean Square Error. In a linear regression model, RMSE is used to test the performance of the machine learning model. It is used to evaluate the data spread around the **line of best** fit. So, in simple words, it is used to measure the deviation of the residuals.

RMSE is calculated using the formula:

$$RMSE = \sqrt{\frac{\Sigma(y_i - \hat{y}_i)^2}{N - P}}$$

- **Yi** is the actual value of the output variable.
- **Y(Cap)** is the predicted value and,
- **N** is the number of data points.

- **MSE:** Mean Squared Error is used to find how close is the line to the actual data. So, we make the difference in the distance of the data points from the line and the difference is squared. This is done for all the data points and the submission of the squared difference divided by the total number of data points gives us the Mean Squared Error (MSE).

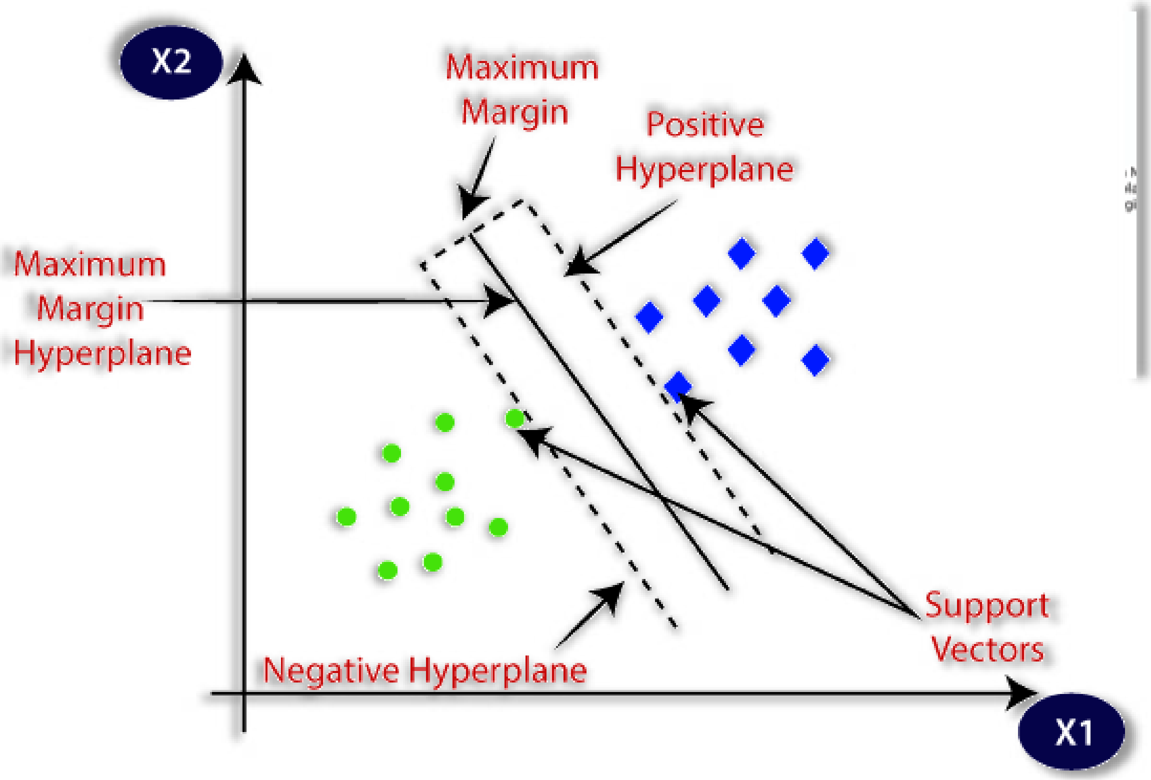
So, if we are taking the squared difference of N data points and dividing the sum by N, what does it mean? Yes, it represents the average of the squared difference of a data point from the line i.e. the average of the squared difference between the actual and the predicted values. The formula for finding MSE is given below:

$$\mathrm{MSE} = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2$$

- **Yi** is the actual value of the output variable (the ith data point)
- **Y(cap)** is the predicted value and,
- **N** is the total number of data points.

So, **RMSE** is the square root of **MSE**.

Q 25. What are Support Vectors in SVM (Support Vector Machine)?



In the above diagram, we can see that the thin lines mark the distance from the classifier to the closest data points (darkened data points). These are called support vectors. So, we can define the support vectors as the data points or vectors that are nearest (closest) to the hyperplane. They affect the position of the hyperplane. Since they support the hyperplane, they are known as support vectors.

Q 26. So, you have done some projects in machine learning and data science and we see you are a bit experienced in the field. Let’s say your laptop’s RAM is only 4GB and you want to train your model on 10GB data set.

What will you do? Have you experienced such an issue before?

Ans: In such types of questions, we first need to ask what ML model we have to train. After that, it depends on whether we have to train a model based on Neural Networks or SVM.

The steps for Neural Networks are given below:

1. The Numpy array can be used to load the entire data. It will never store the entire data, rather just create a mapping of the data.
2. Now, in order to get some desired data, pass the index into the NumPy Array.
3. This data can be used to pass as an input to the neural network maintaining a small batch size.

The steps for SVM are given below:

1. For SVM, small data sets can be obtained. This can be done by dividing the big data set.
2. The subset of the data set can be obtained as an input if using the partial fit function.
3. Repeat the step of using the partial fit method for other subsets as well.

Now, you may describe the situation if you have faced such an issue in your projects or working in machine learning/ data science.

Q 27. Explain Neural Network Fundamentals.

Ans: In the human brain, different neurons are present. These neurons combine and perform various tasks. The Neural Network in deep learning tries to imitate human brain neurons. The neural network learns the patterns from the data and uses the knowledge that it gains from various patterns to predict the output for new data, without any human assistance.

A perceptron is the simplest neural network that contains a single neuron that performs 2 functions. The first function is to perform the weighted sum of all the inputs and the second is an activation function.

There are some other neural networks that are more complicated. Such networks consist of the following three layers:

1) **Input Layer:** The neural network has the input layer to receive the input.
2) **Hidden Layer:** There can be multiple hidden layers between the input layer and the output layer. The initially hidden layers are used for detecting the low-level patterns whereas the further layers are responsible for combining output from previous layers to find more patterns.
3) **Output Layer:** This layer outputs the prediction.

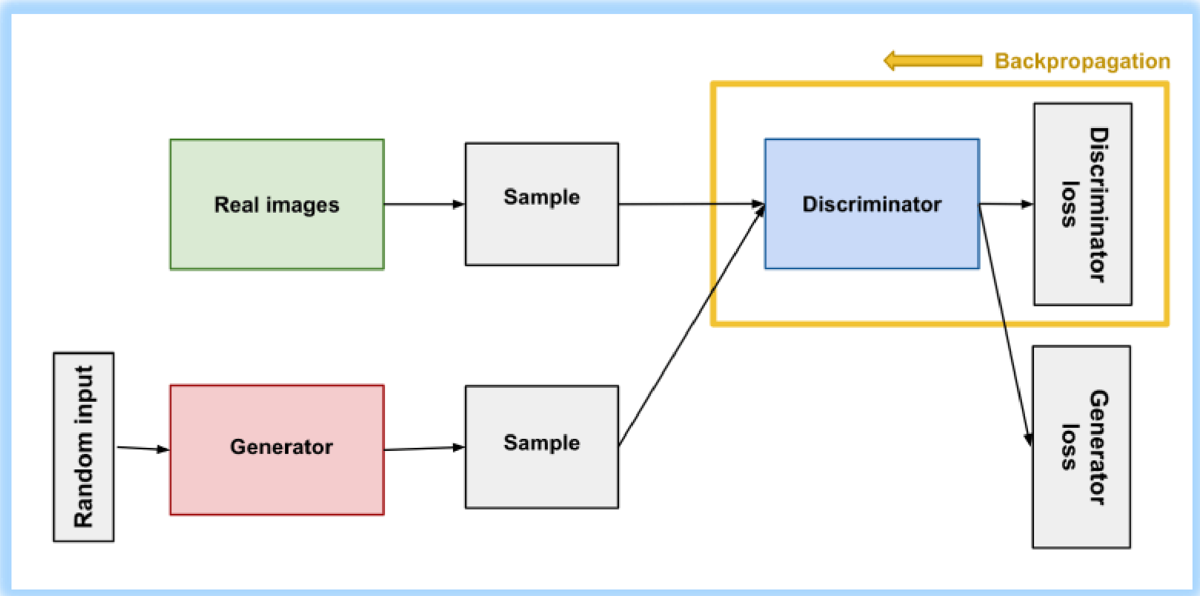
Q 28. What is Generative Adversarial Network?

Ans: This approach can be understood with the famous example of the wine seller. Let us say that there is a wine seller who has his own shop. This wine seller purchases wine from the dealers who sell him the wine at a low cost so that he can sell the wine at a high cost to the customers.

Now, let us say that the dealers whom he is purchasing the wine from, are selling him fake wine. They do this as the fake wine costs way less than the original wine and the fake and the real wine are indistinguishable to a normal consumer (customer in this case).

The shop owner has some friends who are wine experts and he sends his wine to them every time before keeping the stock for sale in his shop. So, his friends, the wine experts, give him feedback that the wine is probably fake. Since the wine seller has been purchasing the wine for a long time from the same dealers, he wants to make sure that their feedback is right before he complains to the dealers about it. Now, let us say that the dealers also have got a tip from somewhere that the wine seller is suspicious of them.

So, in this situation, the dealers will try their best to sell the fake wine whereas the wine seller will try his best to identify the fake wine. Let us see this with the help of a diagram shown below:



From the image above, it is clear that a noise vector is entering the generator (dealer) and he generates the fake wine and the discriminator has to distinguish between the fake wine and real wine. This is a **Generative Adversarial Network** (GAN).

In a GAN, there are 2 main components viz. Generator and Discrminator. So, the generator is a CNN that keeps producing images and the discriminator tries to identify the real images from the fake ones.

Q 29. What is a computational graph?

Ans: A computational graph is also known as a "Dataflow Graph" . Everything in the famous deep learning library TensorFlow is based on the computational graph. The computational graph in Tensorflow has a network of nodes where each node operates. The nodes of this graph represent operations and the edges represent tensors.

Q 30. What are auto-encoders?

Ans: Auto-encoders are learning networks. They transform inputs into outputs with minimum possible errors. So, basically, this means that the output that we want should be almost equal to or as close as to input as follows.

Multiple layers are added between the input and the output layer and the layers that are in between the input and the output layer are smaller than the input layer. It received unlabelled input. This input is encoded to reconstruct the input later.

ABOUT US

At TopperWorld, we are on a mission to empower college students with the knowledge, tools, and resources they need to succeed in their academic journey and beyond.

➢ Our Vision

- ❖ Our vision is to create a world where every college student can easily access high-quality educational content, connect with peers, and achieve their academic goals.
- ❖ We believe that education should be accessible, affordable, and engaging, and that's exactly what we strive to offer through our platform.

➢ Unleash Your Potential

- ❖ In an ever-evolving world, the pursuit of knowledge is essential. TopperWorld serves as your virtual campus, where you can explore a diverse array of online resources tailored to your specific college curriculum.
- ❖ Whether you're studying science, arts, engineering, or any other discipline, we've got you covered.
- ❖ Our platform hosts a vast library of e-books, quizzes, and interactive study tools to ensure you have the best resources at your fingertips.

➢ The TopperWorld Community

- ❖ Education is not just about textbooks and lectures; it's also about forming connections and growing together.

- TopperWorld encourages you to engage with your fellow students, ask questions, and share your knowledge.
- We believe that collaborative learning is the key to academic success.

Start Your Journey with TopperWorld

- Your journey to becoming a top-performing college student begins with TopperWorld.
- Join us today and experience a world of endless learning possibilities.
- Together, we'll help you reach your full academic potential and pave the way for a brighter future.
- Join us on this exciting journey, and let's make academic success a reality for every college student.