Top 45 Data Science Interview Questions You Must Prepare In 2019

edureka.co/blog/interview-questions/data-science-interview-questions

Last updated on Jun 19,2019*101.5K Views*

Sandeep Dayananda

3 / 3 Blog from Interview Questions

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:

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.

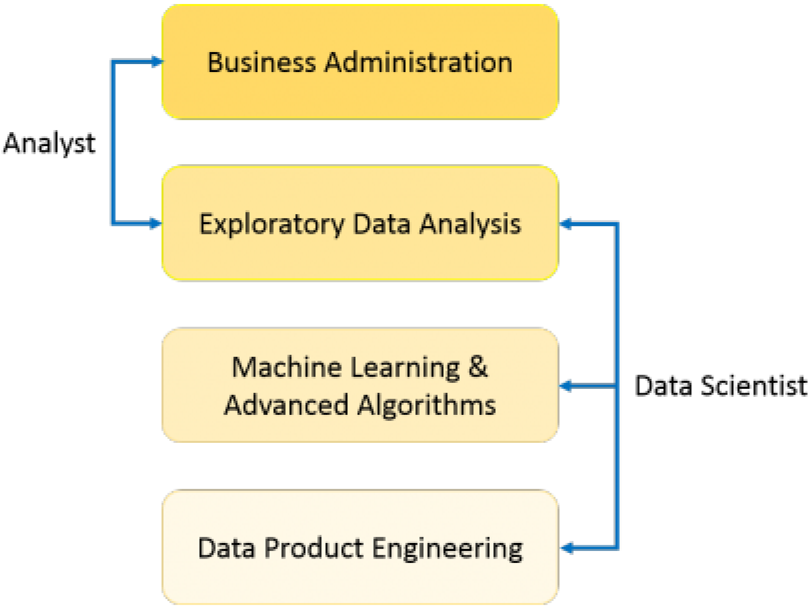
Data Science Interview Questions | Edureka

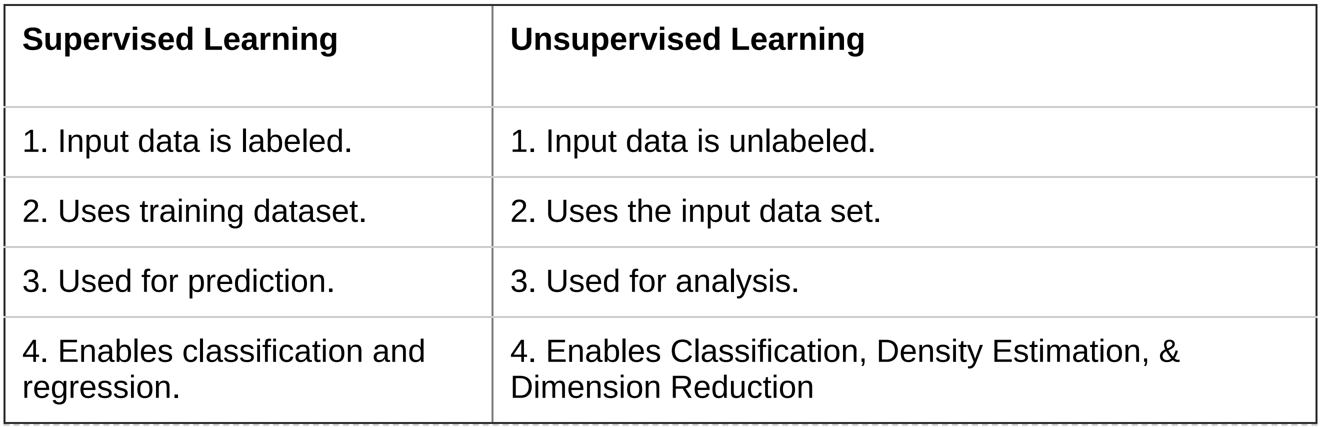
BASIC DATA SCIENCE INTERVIEW QUESTIONS

1. What is Data Science? Also, 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?

The answer lies in the difference between explaining and predicting.





Supervised Learning	Unsupervised Learning
1. Input data is labeled.	1. Input data is unlabeled.
2. Uses training dataset.	2. Uses the input data set.
3. Used for prediction.	3. Used for analysis.
4. Enables classification and regression.	4. Enables Classification, Density Estimation, & Dimension Reduction

Supervised Learning vs Unsupervised Learning

2. 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.
- 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.
- Familiarity with Scikit-learn. ****Scikit-Learn Cheat Sheet****
- Ability to write efficient list comprehensions instead of traditional for loops.
- Ability to write small, clean functions (important for any developer), preferably pure functions that don't alter objects.
- Knowing how to profile the performance of a Python script and how to optimize bottlenecks.

The following will help to tackle any problem in data analytics and machine learning.

3. 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.

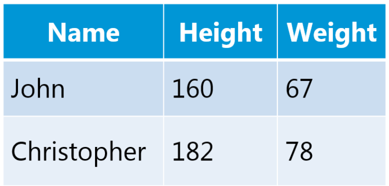
The types of selection bias include:

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.

STATISTICS INTERVIEW QUESTIONS

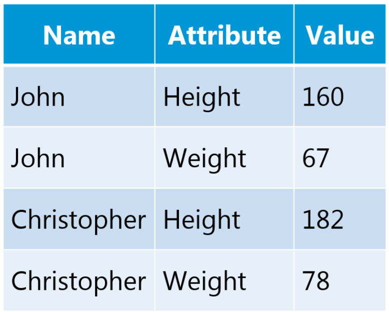
4. What is the difference between "long" and "wide" format data?

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

Figure: *Wide Format*



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

Figure: *Long Format*

5. 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 jumbled up.

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.

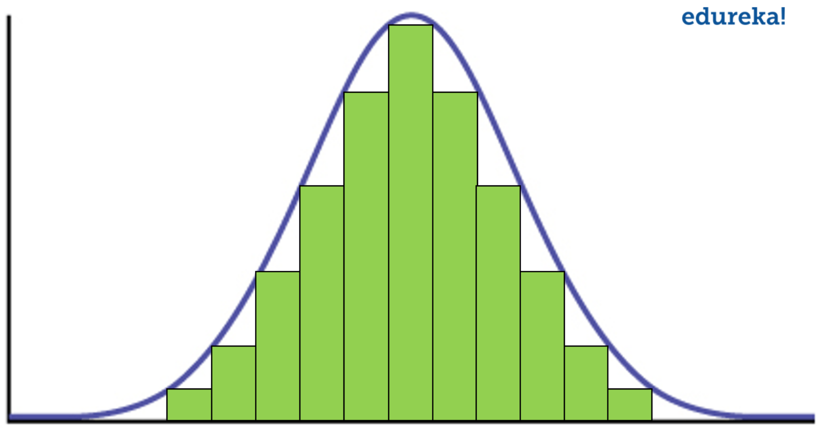


Figure: *Normal distribution in a bell curve*

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

Properties of Nornal Distribution:

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

6. What is the goal of A/B Testing?

It is a statistical hypothesis testing for a randomized experiment with two variables A and B.

The goal of A/B Testing is to identify any changes to the web page to maximize or increase the outcome of an interest. A/B testing is a fantastic method for figuring out the best online promotional and marketing strategies for your business. It can be used to test everything from website copy to sales emails to search ads

An example of this could be identifying the click-through rate for a banner ad.

7. 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**)

**where true positives are positive events which are correctly classified as positives*.

8. What are the differences between overfitting and underfitting?

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

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

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

DATA ANALYSIS INTERVIEW QUESTIONS

9. 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 to use data structures and high-performance data analysis tools.
- R is more suitable for machine learning than just text analysis.
- Python performs faster for all types of text analytics.

Python vs R

10. How does data cleaning plays a vital role in analysis?

Data cleaning can help in analysis because:

- Cleaning data from multiple sources helps to transform it into a format that data analysts or data scientists can work with.
- Data Cleaning helps to increase the accuracy of the model in machine learning.
- It is a cumbersome process because as the number of data sources increases, the time taken to clean the data increases 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 analysis task.

11. 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.

12. 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.

13. 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.

14. 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.

15. 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 Type I error.
- **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 he is tested positive for cancer, based on the lab prediction but he actually doesn’t have cancer. This is a case of false positive. Here it is of utmost danger to start chemotherapy on this patient when he actually does not have cancer. In the absence of cancerous cell, chemotherapy will do certain damage to his normal healthy cells and might lead to severe diseases, even cancer.

Example 2: Let's say an e-commerce company decided to give $1000 Gift voucher to the customers whom they assume to purchase at least $10,000 worth of items. They send free voucher mail directly to 100 customers without any minimum purchase condition because they assume to make at least 20% profit on sold items above $10,000. Now the issue is if we send the $1000 gift vouchers to customers who have not actually purchased anything but are marked as having made $10,000 worth of purchase.

16. 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?

17. 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.

18. 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.

19. 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

20. 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.

Speech Recognition *Face Recognition* *Anti Virus* *Weather Prediction*

Figure: *Applications of Machine Learning*

21. What is the Supervised Learning?

Supervised learning is the machine learning task of inferring a function from labeled training data. The training data consist of a set of training examples.

Algorithms: Support Vector Machines, Regression, Naive Bayes, Decision Trees, K-nearest Neighbor Algorithm and Neural 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.

22. What is Unsupervised learning?

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

23. What are the various classification algorithms?

The below diagram lists the most important classification algorithms.

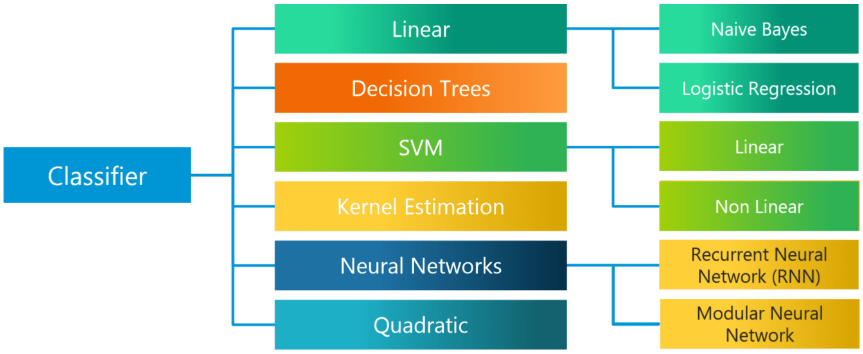


Figure: *Various Classification algorithms*

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

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

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 campaigning of a particular candidate, the amount of time spent in campaigning, etc.

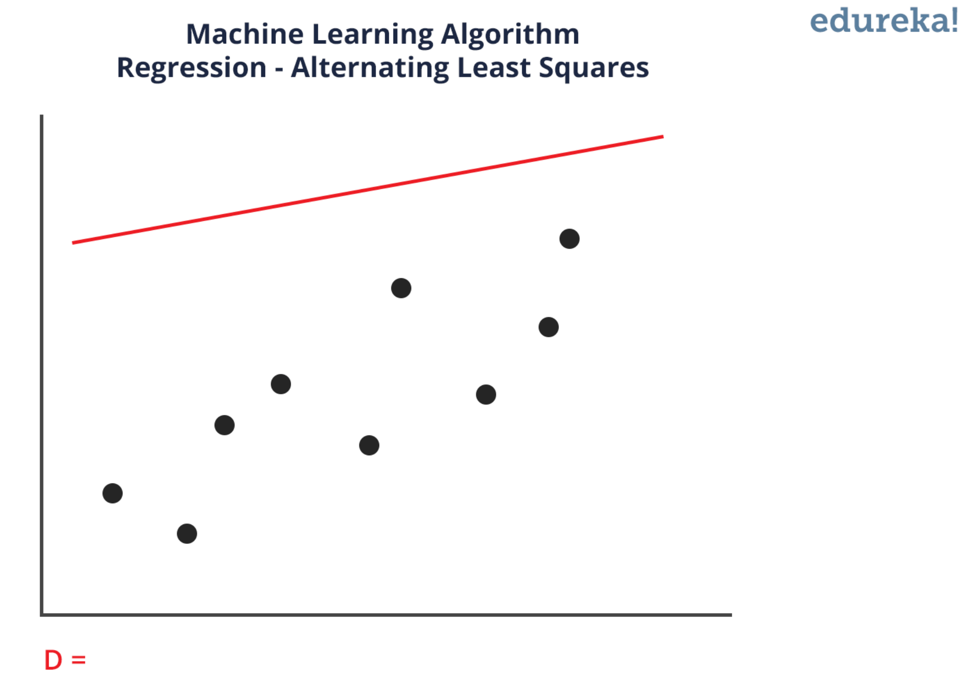
25. What are Recommender Systems?

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.

26. What is Linear Regression?

Linear regression is a statistical technique where the score of a variable Y is predicted from the score of a second variable X. X is referred to as the predictor variable and Y as the criterion variable.

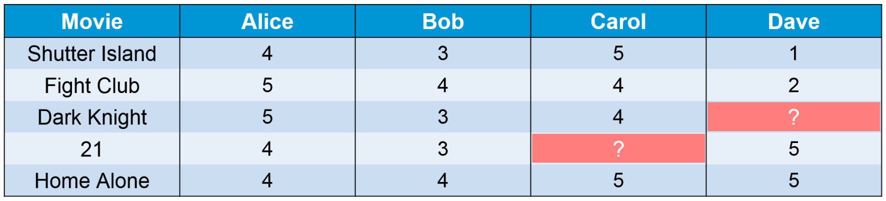


The regression line is the one with the least value of D

27. 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.

An example of collaborative filtering can be to predict the rating of a particular user based on his/her ratings for other movies and others’ ratings for all movies. This concept is widely used in recommending movies in IMDB, Netflix & BookMyShow, product recommenders in e-commerce sites like Amazon, eBay & Flipkart, YouTube video recommendations and game recommendations in Xbox.



Movie	Alice	Bob	Carol	Dave
Shutter Island	4	3	5	1
Fight Club	5	4	4	2
Dark Knight	5	3	4	?
21	4	3	?	5
Home Alone	4	4	5	5

Figure: *Predicting the rating of Dave for Dark Knight and Carol for 21 using Collaborative Filtering*

28. 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 in within a range.
2. To just remove the value.

29. 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 modeling 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.

30. 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.

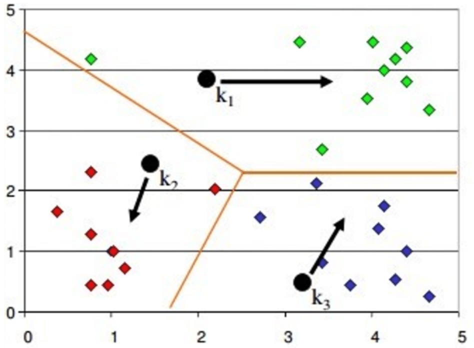
If it is a categorical variable, the default value is assigned. The missing value is assigned a default value. If you have a distribution of data coming, for normal distribution give the mean value.

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.

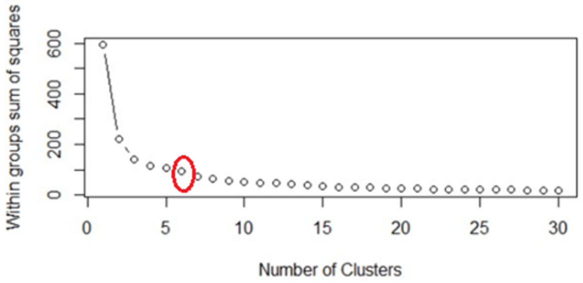
31. How will you define the number of clusters in a clustering algorithm?

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.



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.



- The Graph is generally known as Elbow Curve.
- Red circled 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.

Now that we have seen the Machine Learning Questions, Let's continue our Data Science Interview Questions blog with some Probability questions.

PROBABILITY INTERVIEW QUESTIONS

32. 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

33. 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.

34. 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 possibilities of **BG**, **GB** & **BB**, we have to find the probability of the case with two girls.

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

35. 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 that you see 10 heads, what is the probability that the next toss of that coin is also a head?

There are two ways of choosing the coin. One is to pick a fair coin and the other is to pick the one with two heads.

Probability of selecting fair coin = 999/1000 = **0.999**
Probability of selecting unfair coin = 1/1000 = **0.001**

Selecting 10 heads in a row = Selecting fair coin * Getting 10 heads + Selecting an unfair coin

P (A) = 0.999 * (1/2)^5 = 0.999 * (1/1024) = **0.000976**
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**

DEEP LEARNING INTERVIEW QUESTIONS

36. What do you mean by Deep Learning and Why has it become popular now?

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.

Now although Deep Learning has been around for many years, the major breakthroughs from these techniques came just in recent years. This is because of two main reasons:

- The increase in the amount of data generated through various sources
- The growth in hardware resources required to run these models

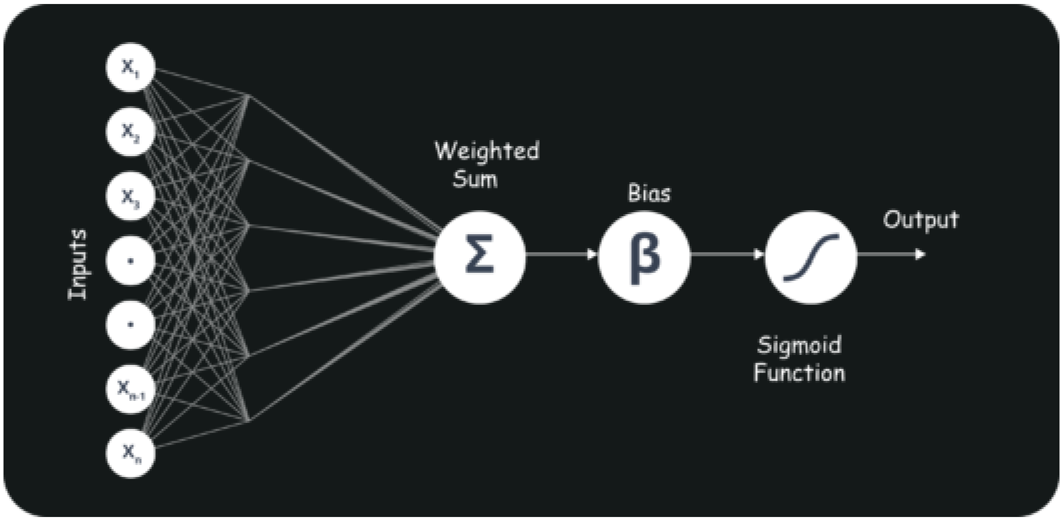
GPUs are multiple times faster and they help us build bigger and deeper deep learning models in comparatively less time than we required previously

37. 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 input so the network generates the best possible result without needing to redesign the output criteria.

38. 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.



39. 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**).

40. 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:

- Forward Propagation of Training Data
- Derivatives are computed using output and target
- Back Propagate for computing derivative of error wrt output activation
- Using previously calculated derivatives for output
- Update the Weights

41. What are the variants of Back Propagation?

- **Stochastic Gradient Descent:** We use only single training example for calculation of gradient and update parameters.
- **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.

42. What are the different Deep Learning Frameworks?

- Pytorch
- TensorFlow
- Microsoft Cognitive Toolkit
- Keras
- Caffe
- Chainer

43. What is the role of 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

44. What is an Auto-Encoder?

Autoencoders 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 autoencoder receives unlabeled input which is then encoded to reconstruct the input.

45. What is a Boltzmann Machine?

Boltzmann machines have a simple learning algorithm that allows them to discover interesting features that represent complex regularities in the training data. The Boltzmann machine is basically used to optimize 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.

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!!*