CLOUDYML

50+ Important Data Science Interview Questions & Answers

(Save It Now)

www.cloudyml.com

1. How is Data modeling different from Database design?

Data Modeling: It can be considered as the first step towards the design of a database. Data modeling creates a conceptual model based on the relationship between various data models. The process involves moving from the conceptual stage to the logical model to the physical schema. It involves the systematic method of applying data modeling techniques.

Database Design: This is the process of designing the database. The database design creates an output which is a detailed data model of the database. Strictly speaking, database design includes the detailed logical model of a database but it can also include physical design choices and storage parameters.

2. What is the benefit of dimensionality reduction?

Dimensionality reduction reduces the dimensions and size of the entire dataset. It drops unnecessary features while retaining the overall information in the data intact. Reduction in dimensions leads to faster processing of the data.
The reason why data with high dimensions is considered so difficult to deal with is that it leads to high time consumption while processing the data and training a model on it. Reducing dimensions speeds up this process, removes noise, and also leads to better model accuracy.

3. Explain stacking in Data Science.

Just like bagging and boosting, stacking is also an ensemble learning method. In bagging and boosting, we could only combine weak models that used the same learning algorithms, e.g., logistic regression. These models are called homogeneous learners.
However, in stacking, we can combine weak models that use different learning algorithms as well. These learners are called heterogeneous learners. Stacking works by training multiple (and different) weak models or learners and then using them together by training another model, called a meta-model, to make predictions based on the multiple outputs of predictions returned by these multiple weak models.

4. What are Loss Function and Cost Functions? Explain the key Difference Between them?

When calculating loss we consider only a single data point, then we use the term loss function.

Whereas, when calculating the sum of error for multiple data then we use the cost function. There is no major difference.

In other words, the loss function is to capture the difference between the actual and predicted values for a single record

whereas cost functions aggregate the difference for the entire training dataset.

The Most commonly used loss functions are Mean-squared error and Hinge loss.

Mean-Squared Error(MSE): In simple words, we can say how our model predicted values against the actual values.

MSE = summation(predicted value - actual value)**2/n(no of data points)
Hinge loss: It is used to train the machine learning classifier, which is

L(y) = max(0,1- y_true*y_pred)

Where y = -1 or 1 indicating two classes and y represents the output form of the classifier. The most common cost function represents the total cost as the sum of the fixed costs and the variable costs in the equation y = mx + b

5. What is SVM? Can you name some kernels used in SVM?

SVM stands for support vector machine. They are used for classification and prediction tasks. SVM consists of a separating plane that discriminates between the two classes of variables.

This separating plane is known as hyperplane. Some of the kernels used in SVM are –

Polynomial Kernel
Gaussian Kernel
Laplace RBF Kernel
Sigmoid Kernel
Hyperbolic Kernel

6. What are the subsets of SQL?

The following are the four significant subsets of the SQL:

Data definition language (DDL): It defines the data structure that consists of commands like CREATE, ALTER, DROP, etc.

Data manipulation language (DML): It is used to manipulate existing data in the database. The commands in this category are SELECT, UPDATE, INSERT, etc.

Data control language (DCL): It controls access to the data stored in the database. The commands in this category include GRANT and REVOKE.

Transaction Control Language (TCL): It is used to deal with the transaction operations in the database. The commands in this

category are COMMIT, ROLLBACK, SET TRANSACTION, SAVEPOINT, etc.

7. Difference between a shallow and a deep copy?

It's faster to do shallow repetitions. It does, however, handle pointers and references in a "lazy" manner. It just copies over the pointer price rather than producing a current copy of the specific knowledge the pointer links to. As a result,

each of the initial and subsequent copies can have pointers that relate to the same underlying knowledge. Deep repetition clones the underlying data completely. It is not shared by the first and, as a result, by the copy.

8. Which of the following machine learning algorithms can be used for inputting missing values of both categorical and continuous variables?

K-means clustering
Linear regression
K-NN (k-nearest neighbor)
Decision trees

The K nearest neighbor algorithm can be used because it can compute the nearest neighbor and if it doesn't have a value, it just computes the nearest neighbor based on all the other features.

When you're dealing with K-means clustering or linear regression, you need to do that in your pre-processing, otherwise, they'll crash. Decision trees also have the same problem, although there is some variance

9. What is an RNN (recurrent neural network)?

RNN is an algorithm that uses sequential data. RNN is used in language translation, voice recognition, image capturing etc. There are different types of RNN networks such as one-to-one, one-to-many, many-to-one and many-to-many. RNN is used in Google’s Voice search and Apple’s Siri.

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

This is statistical hypothesis testing for randomized experiments with two variables, A and B. The objective of A/B testing is to detect any changes to a web page to maximize or increase the outcome of a strategy.

CLOUDYML

SWITCH YOUR CAREER TO

DATA SCIENCE & ANALYTICS

"Make your career in this fastest growing Data Science industry without paying your lakh of rupees."

Features of this course :-

- ✅Get Hands-on Practical Learning Experience
- ✅Topic Wise Structured Tutorial Videos
- ✅Guided Practice Assignments
- ✅Capstone End-to-End Projects
- ✅1-1 Doubt Clearance Support Everyday
- ✅One Month Internship Opportunity
- ✅Interview QnA PDF Collection
- ✅Course Completion Certificate
- ✅Lifetime Course Content Access
- ✅No Prior Coding Experience Required to Join
- ✅Resume Review Feature
- ✅Daily Interview QnA Mail Everyday
- ✅Job Opening Mail & More.

Python
Macine Learning
Deep Learning
SQL
Maths & Stats.
Tablue
PowerBI
Excel
AWS Sagemaker
Google Data Studio
AWS QuickSight

Visit Our Website & Enroll Today

Chosen & Trusted by 8000+ Happy Learners

www.cloudyml.com

11. What is a star schema?

Star schema is the fundamental schema among the data mart schema and it is simplest. It is said to be star as its physical
model resembles to the star shape having a fact table at its center and the dimension tables at its peripheral representing the star's points.

12. What is batch normalization?

Batch normalization is a technique through which attempts could be made to improve the performance and stability of the neural network. This can be done by normalizing the inputs in each layer so that the mean output activation remains 0 with the standard deviation at 1.

13. What is GAN?

The Generative Adversarial Network takes inputs from the noise vector and sends them forward to the Generator, and then to Discriminator, to identify and differentiate unique and fake inputs.

14. What is the case when in SQL Server?

The CASE statement is used to construct logic in which one column's value is determined by the values of other columns.
At least one set of WHEN and THEN commands makes up the SQL Server CASE Statement. The condition to be tested is specified by the WHEN statement. If the WHEN condition returns TRUE, the THEN sentence explains what to do.

When none of the WHEN conditions return true, the ELSE statement is executed. The END keyword brings the CASE statement to a close.

15. What is pickling and unpickling?

Pickle module accepts any Python object and converts it into a string representation and dumps it into a file by using dump function, this process is called pickling. While the process of retrieving original Python objects from the stored string representation is called unpickling.

16. What do you understand by a random forest model?

It combines multiple models together to get the final output or, to be more precise, it combines multiple decision trees together to get the final output. So, decision trees are the building blocks of the random forest model.

17. How are Data Science and Machine Learning related to each other?

Data Science and Machine Learning are two terms that are closely related but are often misunderstood. Both of them deal with data. Data Science is a broad field that deals with large volumes of data and allows us to draw insights out of this voluminous data. Machine Learning, on the other hand, can be thought of as a sub-field of Data Science. It also deals with data, but here, we are solely focused on learning how to convert the processed data into a functional model, which can be used to

map inputs to outputs, e.g., a model that can expect an image as an input and tell us if that image contains a flower as an output.

18. What is a kernel function in SVM?

In the SVM algorithm, a kernel function is a special mathematical function. In simple terms, a kernel function takes data as input and converts it into a required form. This transformation of the data is based on something called a kernel trick, which is what gives the kernel function its name. Using the kernel function, we can transform the data that is not linearly separable (cannot be separated using a straight line) into one that is linearly separable.

19. Explain TF/IDF vectorization.

The expression ‘TF/IDF’ stands for Term Frequency–Inverse Document Frequency. It is a numerical measure that allows us to determine how important a word is to a document in a collection of documents called a corpus. TF/IDF is used often in text mining and information retrieval.

20. What are exploding gradients?

Exploding Gradients is the problematic scenario where large error gradients accumulate to result in very large updates to the weights of neural network models in the training stage. In an extreme case, the value of weights can overflow and result in NaN values. Hence the model becomes unstable and is unable to learn from the training data.

21. What is systematic sampling and cluster sampling ?

Systematic sampling is a type of probability sampling method. The sample members are selected from a larger population with a random starting point but a fixed periodic interval. This interval is known as the sampling interval. The sampling interval is calculated by dividing the population size by the desired sample size.

Cluster sampling involves dividing the sample population into separate groups, called clusters. Then, a simple random sample of clusters is selected from the population. Analysis is conducted on data from the sampled clusters.

22. What is macro in excel?

Macro refers to an algorithm or a set of actions that help automate a task in Excel by recording and playing back the steps taken to complete that task. Once the steps are stored, you create a Macro, and it can be edited and played back as many times as the user wants.

Macro is great for repetitive tasks and also eliminates errors. For example, suppose an account manager has to share reports regarding the company employees for non-payment of dues. In that case, it can be automated using a Macro and doing minor changes every month, as needed.

23. What does KPI stand for in statistics?

A KPI is a quantifiable measure to evaluate whether the objectives are being met or not.
It is a reliable metric to measure the performance level of an organisation or individual.
An example of a KPI in an organisation such as the expense ratio.

In terms of performance, KPIs are an effective way of measuring whether an organisation or individual is meeting expectations.

24. What is the difference between Deep Learning and Machine Learning?

Deep Learning allows machines to make various business-related decisions using artificial neural networks that simulate the human brain, which is one of the reasons why it needs a vast amount of data for training. Machine Learning gives machines the ability to make business decisions without any external help, using the knowledge gained from past data. Machine Learning systems require relatively small amounts of data to train themselves, and most of the features need to be manually coded and understood in advance.

25. What is Cross-validation in Machine Learning?

Cross-validation allows a system to increase the performance of the given Machine Learning algorithm. This sampling process is done to break the dataset into smaller parts that have the same number of rows, out of which a random part is selected as a test set and the rest of the parts are kept as train sets. Cross-validation consists of the following techniques:

•Holdout method
•K-fold cross-validation
•Stratified k-fold cross-validation
•Leave p-out cross-validation

26. What is Epoch in Machine Learning?

Epoch in Machine Learning is used to indicate the count of passes in a given training dataset where the Machine Learning algorithm has done its job. Generally, when there is a large chunk of data, it is grouped into several batches. All these batches go through the given model, and this process is referred to as iteration. Now, if the batch size comprises the complete training dataset, then the count of iterations is the same as that of epochs.

27. What is Dimensionality Reduction?

In the real world, Machine Learning models are built on top of features and parameters. These features can be multidimensional and large in number. Sometimes, the features may be irrelevant and it becomes a difficult task to visualize them. This is where dimensionality reduction is used to cut down irrelevant and redundant features with the help of principal variables. These principal variables conserve the features, and are a subgroup, of the parent variables.

28. What is p-value in hypothesis testing?

If the p-value is more than then critical value, then we fail to reject the H0
If p-value = 0.015 (critical value = 0.05) – strong evidence

If p-value = 0.055 (critical value = 0.05) – weak evidence
If the p-value is less than the critical value, then we reject the H0
If p-value = 0.055 (critical value = 0.05) – weak evidence
If p-value = 0.005 (critical value = 0.05) – strong evidence

CLOUDYML

SWITCH YOUR CAREER TO
DATA SCIENCE & ANALYTICS

"Make your career in this fastest growing Data Science industry without paying your lakh of rupees."

Features of this course :-

- ✅Get Hands-on Practical Learning Experience
- ✅Topic Wise Structured Tutorial Videos
- ✅Guided Practice Assignments
- ✅Capstone End-to-End Projects
- ✅1-1 Doubt Clearance Support Everyday
- ✅One Month Internship Opportunity
- ✅Interview QnA PDF Collection
- ✅Course Completion Certificate
- ✅Lifetime Course Content Access
- ✅No Prior Coding Experience Required to Join
- ✅Resume Review Feature
- ✅Daily Interview QnA Mail Everyday
- ✅Job Opening Mail & More.

Python
Macine Learning
Deep Learning
SQL
Maths & Stats.
Tablue
PowerBI
Excel
AWS Sagemaker
Google Data Studio
AWS QuickSight

Visit Our Website & Enroll Today

Chosen & Trusted by 8000+ Happy Learners

www.cloudyml.com

29. What is the difference between one tail and two tail hypothesis testing?

2-tail test: Critical region is on both sides of the distribution
H0: x = μ
H1: x <> μ
1-tail test: Critical region is on one side of the distribution
H1: x <= μ
H1: x > μ

30. What is the Six sigma in statistic?

In quality control, an error-free data set is generated using six sigma statistics. σ is known as standard deviation. The lower the standard deviation, the less likely that a process performs accurately and commits errors. If a process delivers 99.99966% error-free results, it is said to be six sigma. A six sigma model is one that outperforms 1σ, 2σ, 3σ, 4σ, and 5σ processes and is sufficiently reliable to deliver defect-free work.

31. What does KPI stand for in statistics?

A KPI is a quantifiable measure to evaluate whether the objectives are being met or not.
It is a reliable metric to measure the performance level of an organisation or individual.
An example of a KPI in an organisation such as the expense ratio.
In terms of performance, KPIs are an effective way of measuring whether an organisation or individual is meeting expectations.

32. What are exploding gradients?

Exploding Gradients is the problematic scenario where large error gradients accumulate to result in very large updates to the weights of neural network models in the training stage. In an extreme case, the value of weights can overflow and

result in NaN values. Hence the model becomes unstable and is unable to learn from the training data.

33.Explain the Law of Large Numbers.

The ‘Law of Large Numbers’ states that if an experiment is repeated independently a large number of times, the average of the individual results is close to the expected value.

There are two forms of the law of large numbers, but the differences are primarily theoretical.

The weak law of large numbers states that as n increases, the sample statistic of the sequence converges in probability to the population value.

The strong law of large numbers describes how a sample statistic converges on the population value as the sample size or the number of trials increases. For example, the sample mean will converge on the population mean as the sample size increases.

34. What is the importance of A/B testing.

The goal of A/B testing is to pick the best variant among two hypotheses, the use cases of this kind of testing could be a web page or application responsiveness, landing page redesign, banner testing, marketing campaign performance etc.

The first step is to confirm a conversion goal, and then statistical analysis is used to understand which alternative performs better for the given conversion goal.

35. Explain Eigenvectors and Eigenvalues.

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

The eigenvalue is the strength of the transformation in the direction of the eigenvector.

An eigenvector's direction remains unchanged when a linear transformation is applied to it.

36. Explain the data preprocessing steps in data analysis.

Data preprocessing transforms the data into a format that is more easily and effectively processed in data mining, machine learning and other data science tasks.

1. Data profiling.
2. Data cleansing.
3. Data reduction.
4. Data transformation.
5. Data enrichment.
6. Data validation.

37. What Are the Three Stages of Building a Model in Machine Learning?

The three stages of building a machine learning model are:

Model Building: Choosing a suitable algorithm for the model and train it according to the requirement

Model Testing: Checking the accuracy of the model through the test data

Applying the Model: Making the required changes after testing and use the final model for real-time projects

38. What are the subsets of SQL?

The following are the four significant subsets of the SQL:

Data definition language (DDL): It defines the data structure that consists of commands like CREATE, ALTER, DROP, etc.

Data manipulation language (DML): It is used to manipulate existing data in the database. The commands in this category are SELECT, UPDATE, INSERT, etc.

Data control language (DCL): It controls access to the data stored in the database. The commands in this category include GRANT and REVOKE.

Transaction Control Language (TCL): It is used to deal with the transaction operations in the database. The commands in this category are COMMIT, ROLLBACK, SET TRANSACTION, SAVEPOINT, etc.

39. What is a Parameter in Tableau? Give an Example.

A parameter is a dynamic value that a customer could select, and you can use it to replace constant values in calculations, filters, and reference lines.
For example, when creating a filter to show the top 10 products based on total profit instead of the fixed value, you can update the filter to show the top 10, 20, or 30 products using a parameter.

40. Explain how the filter function works in python?

The filter() method filters a series using a function that checks if each element in the sequence is true or not. The filter() function takes two arguments: function - a function and iterable - an iterable like sets, lists, tuples etc.

41. How to remove duplicate elements from a list?

First we have a List that contains duplicates. Create a dictionary, using the List items as keys. This will automatically remove any duplicates because dictionaries cannot have duplicate keys. Then, convert the dictionary back into a list.

42. Difference between a shallow and a deep copy?

It's faster to do shallow repetitions. It does, however, handle pointers and references in a "lazy" manner. It just copies over the pointer price rather than producing a current copy of the specific knowledge the pointer links to. As a result, each of the initial and subsequent copies can have pointers that relate to the same underlying knowledge. Deep repetition clones the underlying data completely. It is not shared by the first and, as a result, by the copy.

43. What is TF/IDF vectorization?

The TF-IDF statistic, which stands for term frequency–inverse document frequency, is a numerical measure of how essential a word is to a document in a collection or corpus. It's frequently used in information retrieval, text mining, and user modelling searches as a weighting factor. The tf–idf value rises in proportion to the number of times a word appears in a document and is offset by the number of documents in the corpus that

contain the term, which helps to compensate for the fact that some words appear more frequently than others.

44. Give some statistical methods that are useful for data

Two main statistical methods are used in data analysis: descriptive statistics, which summarizes data using indexes such as mean and median and another is inferential statistics, which draw conclusions from data using statistical tests such as student's t-test. The tests include ANOVA, Kruskal-Wallis H test, Friedman Test, etc.

45. Difference B/w Drop, Truncate and Delete?

Delete is used to delete one or more tuples of a table. With the help of the “DROP” command we can drop (delete) the whole structure in one go i.e. it removes the named elements of the schema. Truncate is used to delete all the rows of a table.

46. What is the importance of a dashboard in tableau?

Building dashboards with Tableau allows even non-technical users to create interactive, real-time visualizations in minutes. In just a few clicks, they can combine data sources, add filters, and drill down into specific information.

47. Explain the transformation phase of the data pipeline.

Transformation refers to operations that change data, which may include data standardization, sorting, deduplication, validation, and verification. The ultimate goal is to make it possible to analyze the data.

48. How is the first principal component axis selected in PCA?

In Principal Component Analysis (PCA) we look to summarize a large set of correlated variables (basically a high dimensional data) into a smaller number of representative variables, called the principal components, that explains most of the variability in the original set.

The first principal component axis is selected in a way such that it explains most of the variation in the data and is closest to all n observations.

49. What are Hard-Margin and Soft-Margin SVMs?

Hard-Margin SVMs have linearly separable training data. No data points are allowed in the margin areas. This type of linear classification is known as Hard margin classification.

Soft-Margin SVMs have training data that are not linearly separable. Margin violation means choosing a hyperplane, which can allow some data points to stay either in between the margin area or on the incorrect side of the hyperplane.

50. What is the empirical rule?

In statistics, the empirical rule states that every piece of data in a normal distribution lies within three standard deviations of the mean. It is also known as the 68–95–99.7 rule. According to the empirical rule, the percentage of values that lie in a normal distribution follow the 68%, 95%, and 99.7% rule. In other words, 68% of values will fall within one standard deviation of the mean, 95% will fall within two standard deviations, and 99.75 will fall within three standard deviations of the mean.

51. What is the left-skewed distribution and the right-skewed distribution?

In the left-skewed distribution, the left tail is longer than the right side.
Mean < median < mode
In the right-skewed distribution, the right tail is longer. It is also known as positive-skew distribution.
Mode < median < mean

52. What relationships exist between a logistic regression's coefficient and the Odds Ratio?

The coefficients and the odds ratios then represent the effect of each independent variable controlling for all of the other independent variables in the model and each coefficient can be tested for significance.

53. What's the relationship between Principal Component Analysis (PCA) and Linear & Quadratic Discriminant Analysis (LDA & QDA)

LDA focuses on finding a feature subspace that maximizes the separability between the groups. While Principal component analysis is an unsupervised Dimensionality reduction technique, it ignores the class label. PCA focuses on capturing the direction of maximum variation in the data set.The PC1 the first

principal component formed by PCA will account for maximum variation in the data.PC2 does the second-best job in capturing maximum variation and so on.

The LD1 the first new axes created by Linear Discriminant Analysis will account for capturing most variation between the groups or categories and then comes LD2 and so on.

54. What's the difference between logistic and linear regression? How do you avoid local minima?

Linear Regression is used to handle regression problems whereas Logistic regression is used to handle the classification problems.
Linear regression provides a continuous output but Logistic regression provides discreet output.
The purpose of Linear Regression is to find the best-fitted line while Logistic regression is one step ahead and fitting the line values to the sigmoid curve.
The method for calculating loss function in linear regression is the mean squared error whereas for logistic regression it is maximum likelihood estimation.
We can try to prevent our loss function from getting stuck in a local minima by providing a momentum value. So, it provides a basic impulse to the loss function in a specific direction and helps the function avoid narrow or small local minima by using stochastic gradient descent.

55. Explain the difference between type 1 and type 2 errors.

Type 1 error is a false positive error that 'claims' that an incident has occurred when, in fact, nothing has occurred. The best example of a false positive error is a false fire alarm – the alarm starts ringing when there's no fire. Contrary to this, a Type 2 error is a false negative error that 'claims' nothing has occurred when something has definitely happened. It would be a Type 2 error to tell a pregnant lady that she isn't carrying a baby.

CLOUDYML

SWITCH YOUR CAREER TO

DATA SCIENCE & ANALYTICS

"Make your career in this fastest growing Data Science industry without paying your lakh of rupees."

Features of this course :-

- ✅Get Hands-on Practical Learning Experience
- ✅Topic Wise Structured Tutorial Videos
- ✅Guided Practice Assignments
- ✅Capstone End-to-End Projects
- ✅1-1 Doubt Clearance Support Everyday
- ✅One Month Internship Opportunity
- ✅Interview QnA PDF Collection
- ✅Course Completion Certificate
- ✅Lifetime Course Content Access
- ✅No Prior Coding Experience Required to Join
- ✅Resume Review Feature
- ✅Daily Interview QnA Mail Everyday
- ✅Job Opening Mail & More.

Python
Macine Learning
Deep Learning
SQL
Maths & Stats.
Tablue
PowerBI
Excel
AWS Sagemaker
Google Data Studio
AWS QuickSight

Visit Our Website & Enroll Today

Chosen & Trusted by 8000+ Happy Learners

www.cloudyml.com