CLOUDYML

100+ DATA SCIENTIST

INTERVEIW QNA PDF COLLECTION

- Machine Learning
- Deep Learning
- PowerBi
- Tableau
- SQL & NoSQL
- Python/R
- Microsoft Excel
- And More Tools Are Covered

AKASH RAJ

Founder & CEO -CloudyML || Data Scientist

49K+ 48K+ 97K+

1. How does Stacking work?

The idea of stacking is to learn several different weak learners and combine them by training a meta-model to output predictions based on the multiple predictions returned by these weak models.
If a stacking ensemble is composed of L weak learners, then to fit the model the following steps are followed:
Split the training data into two folds.
Choose L weak learners and fit them to the data of the first fold.
For each of the L weak learners, make predictions for observations in the second fold.
Fit the meta-model on the second fold, using predictions made by the weak learners as inputs.

2. Can you provide me examples of when a scatter graph would be more appropriate than a line chart or vice versa?

A scatter graph would be more appropriate than a line chart when you are looking to show the relationship between two variables that are not linearly related. For example, if you were looking to show the relationship between a person's age and their weight, a scatter graph would be more appropriate than a line chart. A line chart would be more appropriate than a scatter graph when you are looking to show a trend over time. For example, if you were looking at the monthly sales of a company over the course of a year, a line chart would be more appropriate than a scatter graph.

3. Where is data stored in Power BI?

When data is ingested into Power BI, it is basically stored in Fact and Dimension tables.
Fact tables: The central table in a star schema of a data warehouse, a fact table stores quantitative information for analysis and is not normalized in most cases.
Dimension tables: It is just another table in the star schema that is used to store attributes and dimensions that describe objects stored in a fact table.

4. What is Cursor? How to use a Cursor?

After any variable declaration, DECLARE a cursor. A SELECT Statement must always be coupled with the cursor definition.
To start the result set, move the cursor over it. Before obtaining rows from the result set, the OPEN statement must be executed.
To retrieve and go to the next row in the result set, use the FETCH command.
To disable the cursor, use the CLOSE command.
Finally, use the DEALLOCATE command to remove the cursor definition and free up the resources connected with it.

5. What are decorators in Python?

Decorators are used to add some design patterns to a function without changing its structure. Decorators generally are defined before the function they are enhancing. To apply a decorator we first define the decorator function. Then we write the function it is applied to and simply add the decorator function above the function it has to be applied to. For this, we use the @ symbol before the decorator.

6. What is the ACID property in a database?

The full form of ACID is atomicity, consistency, isolation, and durability.
• Atomicity refers that if any aspect of a transaction fails, the whole transaction fails and the database state remains unchanged.
• Consistency means that the data meets all validity guidelines.
• Concurrency management is the primary objective of isolation.
• Durability ensures that once a transaction is committed, it will occur regardless of what happens in between such as a power outage, fire, or some other kind of disturbance.

7. What is the meaning of KPI in statistics?

KPI is an acronym for a key performance indicator. It can be defined as a quantifiable measure to understand whether the goal is being achieved or not. KPI is a reliable metric to measure the performance level of an organization or individual with respect to the objectives. An example of KPI in an organization is the expense ratio.

8. Explain One-hot encoding and Label Encoding. How do they affect the dimensionality of the given dataset?

One-hot encoding is the representation of categorical variables as binary vectors. Label Encoding is converting labels/words into numeric form. Using one-hot encoding increases the dimensionality of the data set. Label encoding doesn't affect the dimensionality of the data set. One-hot encoding creates a new variable for each level in the variable whereas, in Label encoding, the levels of a variable get encoded as 1 and 0.

9. What are autoencoders?

Autoencoders are artificial neural networks that learn without any supervision. Here, these networks have the ability to automatically learn by mapping the inputs to the corresponding outputs.
Autoencoders, as the name suggests, consist of two entities:
Encoder: Used to fit the input into an internal computation state
Decoder: Used to convert the computational state back into the output

10. Compare K-means and KNN Algorithms.

K-Means is unsupervised. K-Means is a clustering algorithm. The points in each cluster are similar to each other, and each cluster is different from its neighboring clusters. KNN is supervised in nature. KNN is a classification algorithm It classifies an unlabeled observation based on its K (can be any number) surrounding neighbors.

11. What is a Recursive Stored Procedure in SQL?

A stored procedure that calls itself until a boundary condition is reached, is called a recursive stored procedure. This recursive function helps the programmers to deploy the same set of code several times as and when required. Some SQL programming languages limit the recursion depth to prevent an infinite loop of procedure calls from causing a stack overflow, which slows down the system and may lead to system crashes.

12. SUM() vs SUMX(): What is the difference between the two DAX functions in Power BI?

The sum function (Sum()) takes the data columns and aggregates them totally but the SumX function (SumX()) lets you filter the data which you are adding.
SUMX(Table, Expression), where the table contains the rows for calculation. Expression is a calculation that will be evaluated on each row of the table.

13. How does a Decision Tree handle continuous(numerical) features?

Autoencoders are artificial neural networks that learn without any supervision. Here, these networks have the ability to automatically learn by mapping the inputs to the corresponding outputs.
Autoencoders, as the name suggests, consist of two entities:
Encoder: Used to fit the input into an internal computation state
Decoder: Used to convert the computational state back into the output

14. What are Loss Function and Cost Functions?

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.

15. What is the difference between Python Arrays and lists?

Arrays in python can only contain elements of same data types i.e., data type of array should be homogeneous. It is a thin wrapper around C language arrays and consumes far less memory than lists.
Lists in python can contain elements of different data types i.e., data type of lists can be heterogeneous. It has the disadvantage of consuming large memory.

16. What is root cause analysis? What is a causation vs. a correlation?

Root cause analysis: a method of problem-solving used for identifying the root cause(s) of a problem [5]
Correlation measures the relationship between two variables, range from -1 to 1. Causation is when a first event appears to have caused a second event. Causation essentially looks at direct relationships while correlation can look at both direct and indirect relationships.

17. Explain some cases where k-Means clustering fails to give good results

k-means has trouble clustering data where clusters are of various sizes and densities.Outliers will cause the centroids to be dragged, or the outliers might get their own cluster instead of being ignored. Outliers should be clipped or removed before clustering.If the number of dimensions increase, a distance-based similarity measure converges to a constant value between any given examples. Dimensions should be reduced before clustering them.

18. If your Time-Series Dataset is very long, what architecture would you use?

If the dataset for time-series is very long, LSTMs are ideal for it because it can not only process single data points, but also entire sequences of data. A time-series being a sequence of data makes LSTM ideal for it.For an even stronger representational capacity, making the LSTM's multi-layered is better.Another method for long time-series dataset is to use CNNs to extract information.

19. What are some common Data Preparation Operations you would use for Time Series Data?

Parsing time series information from various sources and formats. Generating sequences of fixed-frequency dates and time spans. Manipulating and converting date times with time zone information. Resampling or converting a time series to a particular frequency.

20. Describe the Difference Between Window Functions and Aggregate Functions in SQL.

The main difference between window functions and aggregate functions is that aggregate functions group multiple rows into a single result row; all the individual rows in the group are collapsed and their individual data is not shown. On the other hand, window functions produce a result for each individual row. This result is usually shown as a new column value in every row within the window.

21. What is Ribbon in Excel and where does it appear?

The Ribbon is basically your key interface with Excel and it appears at the top of the Excel window. It allows users to access many of the most important commands directly. It consists of many tabs such as File, Home, View, Insert, etc. You can also customize the ribbon to suit your preferences. To customize the Ribbon, right-click on it and select the “Customize the Ribbon” option.

22. Can you explain how the memory cell in an LSTM is implemented computationally?

The memory cell in an LSTM is implemented as a forget gate, an input gate, and an output gate. The forget gate controls how much information from the previous cell state is forgotten. The input gate controls how much new information from the current input is allowed into the cell state. The output gate controls how much information from the cell state is allowed to pass out to the next cell state.

23. What is CTE in SQL?

A CTE (Common Table Expression) is a one-time result set that only exists for the duration of the query. It allows us to refer to data within a single SELECT, INSERT, UPDATE, DELETE, CREATE VIEW, or MERGE statement's execution scope. It is temporary because its result cannot be stored anywhere and will be lost as soon as a query's execution is completed.

24. List the advantages NumPy Arrays have over Python lists?

Python's lists, even though hugely efficient containers capable of a number of functions, have several limitations when compared to NumPy arrays. It is not possible to perform vectorised operations which includes element-wise addition and multiplication. They also require that Python store the type information of every element since they support objects of different types. This means a type dispatching code must be executed each time an operation on an element is done.

25. What are Constraints in SQL?

Constraints are used to specify the rules concerning data in the table. It can be applied for single or multiple fields in an SQL table during the creation of the table or after creating using the ALTER TABLE command. The constraints are:
NOT NULL - Restricts NULL value from being inserted into a column.
CHECK - Verifies that all values in a field satisfy a condition.
DEFAULT - Automatically assigns a default value if no value has been specified for the field.
UNIQUE - Ensures unique values to be inserted into the field.
INDEX - Indexes a field providing faster retrieval of records.
PRIMARY KEY - Uniquely identifies each record in a table.
FOREIGN KEY - Ensures referential integrity for a record in another table.

26. What do you understand by sub-queries in SQL?

A subquery is a query inside another query where a query is defined to retrieve data or information back from the database. In a subquery, the outer query is called as the main query whereas the inner query is called subquery. Subqueries are always executed first and the result of the subquery is passed on to the main query. It can be nested inside a SELECT, UPDATE or any other query. A subquery can also use any comparison operators such as >,< or =.

27. What Would You Do If Some Countries/Provinces (Any Geographical Entity) are Missing and Displaying a Null When You Use Map View in Tableau?

When working with maps and geographical fields, unknown or ambiguous locations are identified by the indicator in the lower right corner of the view.
Click the indicator and choose from the following options:
Edit Locations - correct the locations by mapping your data to known locations
Filter Data - exclude the unknown locations from the view using a filter. The locations will not be included in calculations
Show Data at Default Position - show the values at the default position of (0, 0) on the map.

28. Explain the different layers in CNN

The different layers involved in the architecture of CNN are as follows:
1. Input Layer: The input layer in CNN should contain image data. Image data is represented by a three-dimensional matrix. We have to reshape the image into a single column.
For Example, Suppose we have an MNIST dataset and you have an image of dimension 28 x 28 =784, you need to convert it into 784 x 1 before feeding it into the input. If we have “k” training examples in the dataset, then the dimension of input will be (784, k).
2. Convolutional Layer: To perform the convolution operation, this layer is used which creates several smaller picture windows to go over the data.
3. ReLU Layer: This layer introduces the non-linearity to the network and converts all the negative pixels to zero. The final output is a rectified feature map.

29. What is the AdaBoost Algorithm?

AdaBoost also called Adaptive Boosting is a technique in Machine Learning used as an Ensemble Method. The most common algorithm used with AdaBoost is decision trees with one level that means with Decision trees with only 1 split. These trees are also called Decision Stumps. What this algorithm does is that it builds a model and gives equal weights to all the data points. It then assigns higher weights to points that are wrongly classified. Now all the points which have higher weights are given more importance in the next model. It will keep training models until and unless a lower error is received.

30. What is the Sliding Window method for Time Series Forecasting?

Time series can be phrased as supervised learning. Given a sequence of numbers for a time series dataset, we can restructure the data to look like a supervised learning problem.
In the sliding window method, the previous time steps can be used as input variables, and the next time steps can be used as the output variable.
In statistics and time series analysis, this is called a lag or lag method. The number of previous time steps is called the window width or size of the lag. This sliding window is the basis for how we can turn any time series dataset into a supervised learning problem.

31. Explain the Difference Between Tableau Worksheet, Dashboard, Story, and Workbook?

Tableau uses a workbook and sheet file structure, much like Microsoft Excel.
A workbook contains sheets, which can be a worksheet, dashboard, or a story.
A worksheet contains a single view along with shelves, legends, and the Data pane.
A dashboard is a collection of views from multiple worksheets.
A story contains a sequence of worksheets or dashboards that work together to convey information.

32. What are the steps involved in training a perceptron in Deep Learning?

There are five main steps that determine the learning of a perceptron:

- Initialize thresholds and weights
- Provide inputs
- Calculate outputs
- Update weights in each step
- Repeat steps 2 to 4

33. 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.
Hard-Margin SVMs are quite sensitive to outliers.
Soft-Margin SVMs try to find the best balance between keeping the margin as large as possible and limiting the margin violations.

34. What are the building blocks of Power BI?

The major building blocks of Power BI are:
Datasets: Dataset is a collection of data gathered from various sources like SQL Server, Azure, Text, Oracle, XML, JSON, and many more. With the GetData feature in Power BI, we can easily fetch data from any data source.
Visualizations: Visualization is the visual aesthetic representation of data in the form of maps, charts, or tables.
Reports: Reports are a structured representation of datasets that consists of multiple pages. Reports help to extract important information and insights from datasets to take major business decisions.
Dashboards: A dashboard is a single-page representation of reports made of various datasets. Each element is termed a tile.
Tiles: Tiles are single-block containing visualizations of a report. Tiles help to differentiate each report

35. What is the Right JOIN in SQL?

The Right join is used to retrieve all rows from the right-hand table and only those rows from the other table that fulfilled the join condition. It returns all the rows from the right-hand side table even though there are no matches in the left-hand side table. If it finds unmatched records from the left side table, it returns a Null value. This join is also known as Right Outer Join.

36. What are the uses of using RNN in NLP?

The RNN is a stateful neural network, which means that it not only retains information from the previous layer but also from the previous pass. Thus, this neuron is said to have connections between passes, and through time.
For the RNN the order of the input matters due to being stateful. The same words with different orders will yield different outputs.
RNN can be used for unsegmented, connected applications such as handwriting recognition or speech recognition.

37. How to remove values to a python array?

Array elements can be removed using pop() or remove() method. The difference between these two functions is that the former returns the deleted value whereas the latter does not.

38. What are the advantages and disadvantages of views in the database?

Advantages of Views:

- As there is no physical location where the data in the view is stored, it generates output without wasting resources.
- Data access is restricted as it does not allow commands like insertion, updation, and deletion.

Disadvantages of Views:

- The view becomes irrelevant if we drop a table related to that view.
- Much memory space is occupied when the view is created for large tables.

39. How to create a calculated field in Tableau?

Click the drop down to the right of Dimensions on the Data pane and select "Create > Calculated Field" to open the calculation editor.
Name the new field and create a formula.

40. How many types of points do we get after applying a DBSCAN Algorithm to a particular dataset?

We get three types of points upon applying a DBSCAN algorithm to a particular dataset – Core point, Border point, and noise point.
Core Point: A data point is considered to be a core point if it has a minimum number of neighboring data points (min_pts) at an epsilon distance from it. These min_pts include the original data points also.
Border Point: A data point that has less than the minimum number of data points needed but has at least one core point in the neighborhood.
Noise Point: A data point that is not a core point or a border point is considered noise or an outlier.

41. List the different types of relationships in SQL.

One-to-One - This can be defined as the relationship between two tables where each record in one table is associated with the maximum of one record in the other table.
One-to-Many & Many-to-One - This is the most commonly used relationship where a record in a table is associated with multiple records in the other table.
Many-to-Many - This is used in cases when multiple instances on both sides are needed for defining a relationship.
Self-Referencing Relationships - This is used when a table needs to define a relationship with itself.

42.What are the main difficulties when training RNNs? How can you handle them?

The two main difficulties when training RNNs are unstable gradients (exploding or vanishing) and a very limited short-term memory. These problems both get worse when dealing with long sequences.
To alleviate the unstable gradients problem, we can:
Use a smaller learning rate.
Use a saturating activation function such as the hyperbolic tangent (which is the default), and possibly use gradient clipping, Layer Normalization, or dropout at each time step.
To tackle the limited short-term memory problem, we can use a Long Short-Term Memory layer or a Gated recurrent unit layer.

43. How many types of points do we get after applying a DBSCAN Algorithm to a particular dataset?

We get three types of points upon applying a DBSCAN algorithm to a particular dataset – Core point, Border point, and noise point.
Core Point: A data point is considered to be a core point if it has a minimum number of neighboring data points (min_pts) at an epsilon distance from it. These min_pts include the original data points also.
Border Point: A data point that has less than the minimum number of data points needed but has at least one core point in the neighborhood.
Noise Point: A data point that is not a core point or a border point is considered noise or an outlier.

44. What is a dendrogram in Hierarchical Clustering Algorithm?

A dendrogram is defined as a tree-like structure that is mainly used to store each step as a memory that the Hierarchical clustering algorithm performs. In the dendrogram plot, the Y-axis represents the Euclidean distances between the observations, and the X-axis represents all the observations present in the given dataset.

45. What do you understand by the term silhouette coefficient?

The silhouette coefficient is a measure of how well clustered together a data point is with respect to the other points in its cluster. It is a measure of how similar a point is to the points in its own cluster, and how dissimilar it is to the points in other clusters. The silhouette coefficient ranges from -1 to 1, with 1 being the best possible score and -1 being the worst possible score.

46. What is the difference between trend and seasonality in time series?

Trends and seasonality are two characteristics of time series metrics that break many models. Trends are continuous increases or decreases in a metric's value. Seasonality, on the other hand, reflects periodic (cyclical) patterns that occur in a system, usually rising above a baseline and then decreasing again.

47.What is Bag of Words in NLP?

Bag of Words is a commonly used model that depends on word frequencies or occurrences to train a classifier. This model creates an occurrence matrix for documents or sentences irrespective of its grammatical structure or word order.

48. What is the difference between bagging and boosting?

Bagging is a homogeneous weak learners' model that learns from each other independently in parallel and combines them for determining the model average. Boosting is also a homogeneous weak learners' model but works differently from Bagging. In this model, learners learn sequentially and adaptively to improve model predictions of a learning algorithm