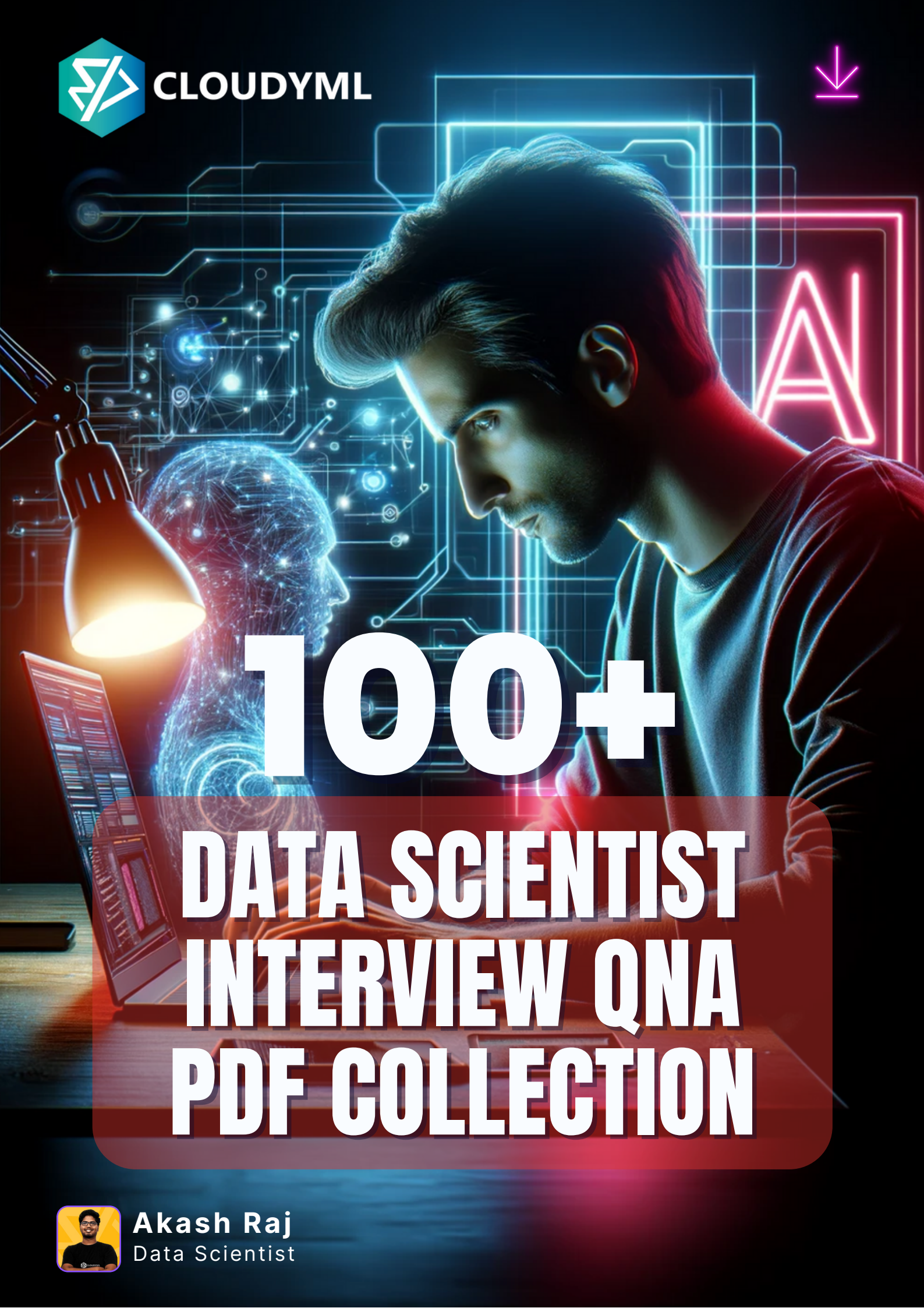
CLOUDYML



100+

DATA SCIENTIST INTERVIEW QNA PDF COLLECTION

Akash Raj
Data Scientist

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.

India #1 Industry-Oriented Program

Our Unique Ways Of Teaching Data Science

1. Assignments, Quizzes & Projects Driven Course

For each topic, there are assignments, quizzes and Real World End-to-End Projects.

Along with Theoretical Knowledge through Recorded Tutorial Videos, You will be getting **Complete Hands-on Practical Learning Experience from Day 1**

2. 1-1 Doubt Clearance Support Daily

Get **1-1 Personal Chat Support for Doubt Clearance everyday** between 6PM to 9AM (including weekends also) over our Mobile App and Web Portal.

And between 8PM to 9PM, you will get Live Zoom Session for Doubt Clearance daily.

This way you will get a smooth and clear learning experience and complete guidance.

3. Do Industrial Internship

While doing the course, you can do internship in our Tie-Up Companies and get real world experience.

At the end of your course, you will not just get the course certificate but Internship Experience Letter as well.

4. Job Hunting Techniques

Simultaneously with the course, you will be learning **Linkedin Growth Hacks, Pro Resume Building** and Proven techniques to find jobs through various online platforms.

Also you will get **Interview QnA PDF collection, General Aptitude Course** and **recordings of Mock Interviews.**

5. One Course For All Job Roles

After completing our course, you will be able to crack interviews of all the Data Role like **Data Scientist, Data Analyst, Data Engineer, Business Analyst etc.**

Python SQL Statistics Machine Learning Deep Learning DSA

Tableau PowerBI Scala Sqoop Airflow Kafka HBASE

Excel Big Data Hadoop MapReduce Hive Spark Cloud

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

49. What is a Self-Join?

A self-join is a type of join that can be used to connect two tables. As a result, it is a unary relationship. Each row of the table is attached to itself and all other rows of the same table in a self-join. As a result, a self-join is mostly used to combine and compare rows from the same database table.

50. What are the types of views in SQL?

In SQL, the views are classified into four types. They are:
Simple View: A view that is based on a single table and does not have a GROUP BY clause or other features.
Complex View: A view that is built from several tables and includes a GROUP BY clause as well as functions.
Inline View: A view that is built on a subquery in the FROM clause, which provides a temporary table and simplifies a complicated query.
Materialized View: A view that saves both the definition and the details. It builds data replicas by physically preserving them.

51. What is the difference between data mining and data profiling.

Data mining Process: It generally involves analyzing data to find relations that were not previously discovered. In this case, the emphasis is on finding unusual records, detecting dependencies, and analyzing clusters.

Data Profiling Process: It generally involves analyzing that data's individual attributes. In this case, the emphasis is on providing useful information on data attributes such as data type, frequency, etc.

52. Explain the KNN imputation method.

A KNN (K-nearest neighbor) model is usually considered one of the most common techniques for imputation. It allows a point in multidimensional space to be matched with its closest k neighbors. By using the distance function, two attribute values are compared. Using this approach, the closest attribute values to the missing values are used to impute these missing values.

53. Explain Hierarchical clustering.

This algorithm group objects into clusters based on similarities, and it is also called hierarchical cluster analysis. When hierarchical clustering is performed, we obtain a set of clusters that differ from each other.

This clustering technique can be divided into two types:

Agglomerative Clustering (which uses bottom-up strategy to decompose clusters)

Divisive Clustering (which uses a top-down strategy to decompose clusters)

54. Explain N-gram

N-gram, known as the probabilistic language model, is defined as a connected sequence of n items in a given text or speech. It is basically composed of adjacent words or letters of length n that were present in the source text. In simple words, it is a way to predict the next item in a sequence, as in (n-1).

55.What is RDBMS? How is it different from DBMS?

RDBMS stands for Relational Database Management System that stores data in the form of a collection of tables, and relations can be defined between the common fields of these tables.

56. What is ETL in SQL?

ETL stands for Extract, Transform and Load. It is a three-step process, where we would have to start off by extracting the data from sources. Once we collate the data from different sources, what we have is raw data. This raw data has to be transformed into the tidy format, which will come in the second phase.Finally, we would have to load this tidy data into tools which would help us to find insights.

57. Explain Hierarchical clustering.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.

58. What do you understand by the F1 score?

The F1 score represents the measurement of a model's performance. It is referred to as a weighted average of the precision and recall of a model. The results tending to 1 are considered as the best, and those tending to 0 are the worst. It could be used in classification tests, where true negatives don't matter much.

59.What do Tableau's sets and groups mean?

Data is grouped using sets and groups according to predefined criteria. The primary distinction between the two is that although a set can have only two options—either in or out—a group can divide the dataset into several groups. A user should decide which group or sets to apply based on the conditions.

60. What do you mean by a Bag of Words (BOW)?

It is used for word frequency or occurrences to train a classifier.
It contains a text representation that describes the frequency with which words appear in a document.
It has two steps:
-A list of terms that are well-known.
-A metric for determining the existence of well-known terms.

61. What are Nested Triggers?

Triggers may implement DML by using INSERT, UPDATE, and DELETE statements. These triggers that contain DML and find other triggers for data modification are called Nested Triggers.

62. What is a True positive rate and a false positive rate?

True positive rate or Recall: It gives us the percentage of the true positives captured by the model out of all the Actual Positive class.

TPR = TP/ (TP+FN)

False Positive rate: It gives us the percentage of all the false positives by my model prediction from the all Actual Negative class.

FPR = FP/(FP+TN)

63. What is the meaning of dropout in Deep Learning?

Dropout is a technique that is used to avoid overfitting a model in Deep Learning. If the dropout value is too low, then it will have minimal effect on learning. If it is too high, then the model can under-learn, thereby, causing lower efficiency.

64. What is the meaning of dropout in Deep Learning? What are sets in Tableau?

Sets are custom fields that define a subset of data based on some conditions. A set can be based on a computed condition, for example, a set may contain customers with sales over a certain threshold. Computed sets update as your data changes. Alternatively, a set can be based on specific data point in your view.

65. What is the difference between DROP and TRUNCATE commands?

Triggers may implement DML by using INSERT, UPDATE, and DELETE statements. These triggers that contain DML and find other triggers for data modification are called Nested Triggers.

66. What is slicing in Python?

Slicing is used to access parts of sequences like lists, tuples, and strings. The syntax of slicing is-[start:end:step]. The step can be omitted as well. When we write [start:end] this returns all the elements of the sequence from the start (inclusive) till the end-1 element. If the start or end element is negative i, it means the ith element from the end.

67. What are the different types of Pooling? Explain their characteristics.

DMax pooling: Once we obtain the feature map of the input, we will apply a filter of determined shapes across the feature map to get the maximum value from that portion of the feature map. It is also known as subsampling because from the entire portion of the feature map covered by filter or kernel we are sampling one single maximum value.
Average pooling: Computes the average value of the feature map covered by kernel or filter, and takes the floor value of the result.
Sum pooling: Computes the sum of all elements in that window.

68. What is a Moving Average Process in Time series?

In time-series analysis, moving-average process, is a common approach for modeling univariate time series. The moving-average model specifies that the output variable depends linearly on the current and various past values of a stochastic term.

69. What is the difference between SQL having vs where?

The WHERE clause specifies the criteria which individual records must meet to be selected by a query. It can be used without the GROUP by clause. The HAVING clause cannot be used without the GROUP BY clause . The WHERE clause selects rows before grouping. The HAVING clause selects rows after grouping. The WHERE clause cannot contain aggregate functions. The HAVING clause can contain aggregate functions

70. What is Relative cell referencing in excel?

n Relative referencing, there is a change when copying a formula from one cell to another cell with respect to the destination. cells' address Meanwhile, there is no change in Absolute cell referencing when a formula is copied, irrespective of the cell's destination. This type of referencing is there by default. Relative cell referencing doesn't require a dollar sign in the formula.

71. Define Entity, Entity type, and Entity set.

Entity can be anything, be it a place, class or object which has an independent existence in the real world.

Entity Type represents a set of entities that have similar attributes.

Entity Set in the database represents a collection of entities having a particular entity type.

72. Why is the KNN Algorithm Considered a "Lazy Learner?"

The KNN method simply saves the training data when it receives it; it does not learn and create a model. Instead of using the training data to discover any discriminative functions, instance-based learning is used, and the training data is also used for making predictions on datasets that have not yet been viewed.
As a result, KNN is referred to as a "Lazy Learner" since it delays learning a model instead of doing so right away.

73. What is the meaning of term weight initialization in neural networks?

In neural networking, weight initialization is one of the essential factors. A bad weight initialization prevents a network from learning. On the other side, a good weight initialization helps in giving a quicker convergence and a better overall error. Biases can be initialized to zero. The standard rule for setting the weights is to be close to zero without being too small.

74. 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

75. What are dimensionality reduction and its benefits?

The Dimensionality reduction refers to the process of converting a data set with vast dimensions into data with fewer dimensions (fields) to convey similar information concisely.
This reduction helps in compressing data and reducing storage space. It also reduces computation time as fewer dimensions lead to less computing. It removes redundant features; for example, there's no point in storing a value in two different units (meters and inches).

76. How can outlier values be treated?

Try a different model. Data detected as outliers by linear models can be fit by nonlinear models. Therefore, be sure you are choosing the correct model.

Try normalizing the data. This way, the extreme data points are pulled to a similar range.

You can use algorithms that are less affected by outliers; an example would be random forests.

77. What do you understand about the true-positive rate and false-positive rate?

TRUE-POSITIVE RATE: The true-positive rate gives the proportion of correct predictions of the positive class. It is also used to measure the percentage of actual positives that are accurately verified.

FALSE-POSITIVE RATE: The false-positive rate gives the proportion of incorrect predictions of the positive class. A false positive determines something is true when that is initially false.

78. What information is gained in a decision tree algorithm?

The Dimensionality reduction refers to the process of converting a data set with vast dimensions into data with fewer dimensions (fields) to convey similar information concisely.
This reduction helps in compressing data and reducing storage space. It also reduces computation time as fewer dimensions lead to less computing. It removes redundant features; for example, there's no point in storing a value in two different units (meters and inches).

79. How can outlier values be treated?

Try a different model. Data detected as outliers by linear models can be fit by nonlinear models. Therefore, be sure you are choosing the correct model.

Try normalizing the data. This way, the extreme data points are pulled to a similar range.

You can use algorithms that are less affected by outliers; an example would be random forests.

80. What do you understand about the true-positive rate and false-positive rate?

TRUE-POSITIVE RATE: The true-positive rate gives the proportion of correct predictions of the positive class. It is also used to measure the percentage of actual positives that are accurately verified.

FALSE-POSITIVE RATE: The false-positive rate gives the proportion of incorrect predictions of the positive class. A false positive determines something is true when that is initially false.

Become Job-Ready In 5 Stepts With CloudyML

01 You will be learning all the technical topics from basics to advanced level and solving assignments to get complete hands-on practical learning experience.

STEP 1

STEP 2

02 After that, you will be doing Industrial Projects that will boost up your confidence and enhance your resume

STEP 3

03 simultaneously with the course, you will be learning LinkedIn Growth Hacks, Pro Resume Building and Proven techniques to find jobs through various online platforms.

STEP 4

04 After learning all the core topics and doing projects, you can do Industrial Internship with Us and get Real-World exposure that will super enhance your skills

STEP 5

05 With our Placement Guarantee Program, which comes with Job Referrals, Unlimited Mock Interviews, Company Tie-up Interviews etc, you will be able to get your Dream Job and you can start your career in Data Science, Analytics & Engineering Domain.

Visit Our Website Now

81. How can you select k for k-means?

We use the elbow method to select k for k-means clustering. The idea of the elbow method is to run k-means clustering on the data set where 'k' is the number of clusters.
Within the sum of squares (WSS), it is defined as the sum of the squared distance between each member of the cluster and its centroid.

82. 'People who bought this also bought...' recommendations seen on Amazon are a result of which algorithm?

The recommendation engine is accomplished with collaborative filtering. Collaborative filtering explains the behavior of other users and their purchase history in terms of ratings, selection, etc.
The engine makes predictions on what might interest a person based on the preferences of other users. In this algorithm, item features are unknown.

83. What is a Confusion Matrix?

The Confusion Matrix is the summary of prediction results of a particular problem. It is a table that is used to describe the performance of the model. The Confusion Matrix is an n*n matrix that evaluates the performance of the classification model.

84. What is the difference between a box plot and a histogram?

The frequency of a certain feature's values is denoted visually by both box plots and histograms.
Boxplots are more often used in comparing several datasets and compared to histograms, take less space and contain fewer details. Histograms are used to know and understand the probability distribution underlying a dataset.

85. What are the common problems that data analysts encounter during analysis?

The common problems steps involved in any analytics project are:

Handling duplicate data
Collecting the meaningful right data at the right time
Handling data purging and storage problems
Making data secure and dealing with compliance issues

86. Explain the Type I and Type II errors in Statistics?

In Hypothesis testing, a Type I error occurs when the null hypothesis is rejected even if it is true. It is also known as a false positive.

A Type II error occurs when the null hypothesis is not rejected, even if it is false. It is also known as a false negative.

87. Name an example where ensemble techniques might be useful?

Ensemble techniques use a combination of learning algorithms to optimize better predictive performance. They typically reduce overfitting in models and make the model more robust (unlikely to be influenced by small changes in the training data). You could list some examples of ensemble methods (bagging, boosting, the "bucket of models" method) and demonstrate how they could increase predictive power.

CLOUDYML

DATA SUPERSTAR

The Highest Paying Job Role Of 21st Century

- Get Practical Learning Experience
- 50+ Industry Oriented Projects
- 1-1 Doubt Clearance Support
- Industrial Internship Opportunity
- 100% Placement Guarantee

Learn From Scratch
@ Most Affordable Price

VISIT OUR WEBSITE

WWW.CLOUDYML.COM

Trusted By 18,000+ Students From India, USA, UK, Canada, Germany, Australia & 20+ Other Countries

88. What's the difference between a generative and discriminative model?

A generative model will learn categories of data while a discriminative model will simply learn the distinction between different categories of data. Discriminative models will generally outperform generative models on classification tasks.

89. What cross-validation technique would you use on a time series dataset?

Instead of using standard k-folds cross-validation, you have to pay attention to the fact that a time series is not randomly distributed data—it is inherently ordered by chronological order. If a pattern emerges in later time periods, for example, your model may still pick up on it even if that effect doesn't hold in earlier years!

You'll want to do something like forward chaining where you'll be able to model on past data then look at forward-facing data.

- Fold 1 : training [1], test [2]
- Fold 2 : training [1 2], test [3]
- Fold 3 : training [1 2 3], test [4]
- Fold 4 : training [1 2 3 4], test [5]
- Fold 5 : training [1 2 3 4 5], test [6]

90. What's the "kernel trick" and how is it useful?

The Kernel trick involves kernel functions that can enable in higher-dimension spaces without explicitly calculating the coordinates of points within that dimension: instead, kernel functions compute the inner products between the images of all pairs of data in a feature space. This allows them the very useful attribute of calculating the coordinates of higher dimensions while being computationally cheaper than the explicit calculation of said coordinates.

91. 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.

92. What is the bin in tableau?

Bins in tableau are containers of equal size used to store data values fitting in bin size. In other words, bins group the data into groups of equal size or data which can be used in systematic viewing of data. All the discrete fields in tableau can also be considered as set of bins

93. What's a Fourier transform?

A Fourier transform is a generic method to decompose generic functions into a superposition of symmetric functions. Or as this more intuitive tutorial puts it, given a smoothie, it's how we find the recipe. The Fourier transform finds the set of cycle speeds, amplitudes, and phases to match any time signal. A Fourier transform converts a signal from time to frequency domain—it's a very common way to extract features from audio signals or other time series such as sensor data.

94. What are Superkey and candidate key in SQL?

A super key may be a single or a combination of keys that help to identify a record in a table. Know that Super keys can have one or more attributes, even though all the attributes are not necessary to identify the records.

A candidate key is the subset of Superkey, which can have one or more than one attribute to identify records in a table. Unlike Superkey, all the attributes of the candidate key must be helpful to identify the records

95. Explain split(), sub(), subn() methods of "re" module in Python.

To modify the strings, Python's "re" module is providing 3 methods. They are:

split() – uses a regex pattern to "split" a given string into a list.

sub() – finds all substrings where the regex pattern matches and then replace them with a different string

subn() – it is similar to sub() and also returns the new string along with the no. of replacements.

96. When should you use a t-test vs a z-test?

-> z tests are a statistical way of testing a hypothesis when either:

We know the population variance, or
We do not know the population variance but our sample size is large n ≥ 30
If we have a sample size of less than 30 and do not know the population variance, then we must use a t-test.

- > t-tests are a statistical way of testing a hypothesis when:
We do not know the population variance
Our sample size is small, n < 30

97. What is Dropout?

Dropout is a regularization technique to avoid overfitting thus increasing the generalizing power. Generally, we should use a small dropout value of 20%-50% of neurons with 20% providing a good starting point.

98. What is a stored procedure?

Stored Procedure is a function consists of many SQL statements to access the database system. Several SQL statements are consolidated into a stored procedure and execute them whenever and wherever required.

99. 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.

100. Explain how a ROC curve works.

The ROC curve is a graphical representation of the contrast between true positive rates and the false positive rate at various thresholds. It's often used as a proxy for the trade-off between the sensitivity of the model (true positives) vs the fall-out or the probability it will trigger a false alarm (false positives).

101. What is P-value?

P-values are used to make a decision about a hypothesis test. P-value is the minimum significant level at which you can reject the null hypothesis. The lower the p-value, the more likely you reject the null hypothesis

102. What is the data analysis process?

Data analysis generally refers to the process of assembling, cleaning, interpreting, transforming, and modeling data to gain insights or conclusions and generate reports to help businesses become more profitable.

- **Collect Data:** The data is collected from a variety of sources and is then stored to be cleaned and prepared. This step involves removing all missing values and outliers.
- **Analyze Data:** As soon as the data is prepared, the next step is to analyze it. Improvements are made by running a model repeatedly. Following that, the model is validated to ensure that it is meeting the requirements.
- **Create Reports:** In the end, the model is implemented, and reports are generated as well as distributed to stakeholders.

103. Explain two different ways to detect outliers.

- **Box Plot Method:** According to this method, the value is considered an outlier if it exceeds or falls below 1.5*IQR (interquartile range), that is, if it lies above the top quartile (Q3) or below the bottom quartile (Q1).
- **Standard Deviation Method:** According to this method, an outlier is defined as a value that is greater or lower than the mean ± (3*standard deviation)

104. How can we relate standard deviation and variance?

Standard deviation refers to the spread of your data from the mean. Variance is the average degree to which each point differs from the mean i.e. the average of all data points. We can relate Standard deviation and Variance because it is the square root of Variance.

105. What is Perceptron? And how does it Work?

If we focus on the structure of a biological neuron, it has dendrites which are used to receive inputs. These inputs are summed in the cell body and using the Axon it is passed on to the next biological neuron. Similarly, a perceptron receives multiple inputs, applies various transformations and functions and provides an output. A Perceptron is a linear model used for binary classification. It models a neuron which has a set of inputs, each of which is given a specific weight. The neuron computes some function on these weighted inputs and gives the output.

106. What are the activation functions?

Activation function translates the inputs into outputs. Activation function decides whether a neuron should be activated or not by calculating the weighted sum and further adding bias with it. The purpose of the activation function is to introduce non-linearity into the output of a neuron.
There can be many Activation functions like:

Linear or Identity

Unit or Binary Step

Sigmoid or Logistic

Tanh

ReLU

Softmax.

107. 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.

108. What do you mean by Associative Rule Mining (ARM)?

Associative Rule Mining is one of the techniques to discover patterns in data like features (dimensions) which occur together and features (dimensions) which are correlated. It is mostly used in Market-based Analysis to find how frequently an itemset occurs in a transaction. Association rules have to satisfy minimum support and minimum confidence at the very same time. Association rule generation generally comprised of two different steps:

"A min support threshold is given to obtain all frequent item-sets in a database."
"A min confidence constraint is given to these frequent item-sets in order to form the association rules."
Support is a measure of how often the "item set" appears in the data set and Confidence is a measure of how often a particular rule has been found to be true.

109. What is the lambda function?

Lambda functions are an anonymous or nameless function.
These functions are called anonymous because they are not declared in the standard manner by using the def keyword. It doesn't require the return keyword as well. These are implicit in the function.
The function can have any number of parameters but can have just one statement and return just one value in the form of an expression. They cannot contain commands or multiple expressions.
An anonymous function cannot be a direct call to print because lambda requires an expression.
Lambda functions have their own local namespace and cannot access variables other than those in their parameter list and those in the global namespace.
Example:

```
x = lambda i,j: i+j

print(x(7,8))
```

Output: 15

110. 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