Data Science Questions

1. What is Data Science?

Data Science is the area of study which involves extracting insights from vast amounts of data using various scientific methods, algorithms, and processes. It helps you to discover hidden patterns from the raw data. The term Data Science has emerged because of the evolution of mathematical statistics, data analysis, and big data.

2. What is the Difference Between Data Science and Machine Learning?

Data Science is a combination of algorithms, tools, and machine learning technique which helps you to find common hidden patterns from the given raw data. Whereas Machine learning is a branch of computer science, that deals with system programming to automatically learn and improve with experience.

3. Discuss Decision Tree algorithm

A decision tree is a popular supervised machine learning algorithm. It is mainly used for Regression and Classification. It allows breaks down a dataset into smaller subsets. The decision tree can able to handle both categorical and numerical data.

4. Explain Recommender Systems?

It is a subclass of information filtering techniques. It helps you to predict the preferences or ratings which users likely to give to a product.

5. Name three disadvantages of using a linear model

Three disadvantages of the linear model are:

- The assumption of linearity of the errors.
- You can’t use this model for binary or count outcomes
- There are plenty of overfitting problems that it can’t solve

6. List out the libraries in Python used for Data Analysis and Scientific Computations.

- SciPy
- Pandas
- Matplotlib
- NumPy
- SciKit
- Seaborn

7. What is bias?

Bias is an error introduced in your model because of the oversimplification of a machine learning algorithm.” It can lead to underfitting.

8. Discuss ‘Naive’ in a Naive Bayes algorithm?

The Naive Bayes Algorithm model is based on the Bayes Theorem. It describes the probability of an event. It is based on prior knowledge of conditions which might be related to that specific event.

9. What is a Linear Regression?

Linear regression is a statistical programming method where the score of a variable ‘A’ is predicted from the score of a second variable ‘B’. B is referred to as the predictor variable and A as the criterion variable.

10. State the difference between the expected value and mean value

They are not many differences, but both of these terms are used in different contexts. Mean value is generally referred to when you are discussing a probability distribution whereas expected value is referred to in the context of a random variable.

11. What is Ensemble Learning?

The ensemble is a method of combining a diverse set of learners together to improvise on the stability and predictive power of the model. Two types of Ensemble learning methods are:

Bagging

Bagging method helps you to implement similar learners on small sample populations. It helps you to make nearer predictions.

Boosting

Boosting is an iterative method which allows you to adjust the weight of an observation depends upon the last classification. Boosting decreases the bias error and helps you to build strong predictive models.

12. Explain the steps for a Data analytics project

The following are important steps involved in an analytics project:

- Understand the Business problem
- Explore the data and study it carefully.
- Prepare the data for modeling by finding missing values and transforming variables.
- Start running the model and analyze the Big data result.
- Validate the model with new data set.
- Implement the model and track the result to analyze the performance of the model for a specific period.

13. Discuss Artificial Neural Networks

Artificial Neural networks (ANN) are a special set of algorithms that have revolutionized machine learning. It helps you to adapt according to changing input. So the network generates the best possible result without redesigning the output criteria.

14. What is Back Propagation?

Back-propagation is the essence of neural net training. It is the method of tuning the weights of a neural net depend upon the error rate obtained in the previous epoch. Proper tuning of the helps you to reduce error rates and to make the model reliable by increasing its generalization.

15. What is a Random Forest?

Random forest is a machine learning method which helps you to perform all types of regression and classification tasks. It is also used for treating missing values and outlier values.

16. What is the importance of having a selection bias?

Selection Bias occurs when there is no specific randomization achieved while picking individuals or groups or data to be analyzed. It suggests that the given sample does not exactly represent the population which was intended to be analyzed.

17. What is the K-means clustering method?

K-means clustering is an important unsupervised learning method. It is the technique of classifying data using a certain set of clusters which is called K clusters. It is deployed for grouping to find out the similarity in the data.

18. Explain the difference between Data Science and Data Analytics

Data Scientists need to slice data to extract valuable insights that a data analyst can apply to real-world business scenarios. The main difference between the two is that the data scientists have more technical knowledge then business analyst. Moreover, they don’t need an understanding of the business required for data visualization.

19. Define the term deep learning

Deep Learning is a subtype of machine learning. It is concerned with algorithms inspired by the structure called artificial neural networks (ANN).

20. Which language is best for text analytics? R or Python?

Python will more suitable for text analytics as it consists of a rich library known as pandas. It allows you to use high-level data analysis tools and data structures, while R doesn't offer this feature.

21. Name various types of Deep Learning Frameworks

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

21. Name commonly used algorithms.

Four most commonly used algorithm by Data scientist are:

- Linear regression
- Logistic regression
- Random Forest
- KNN

23. What is precision?

Precision is the most commonly used error metric is n classification mechanism. Its range is from 0 to 1, where 1 represents 100%

24. Explain cluster sampling technique in Data science

A cluster sampling method is used when it is challenging to study the target population spread across, and simple random sampling can't be applied.

25. While working on a data set, how can you select important variables? Explain

Following methods of variable selection you can use:

- Remove the correlated variables before selecting important variables
- Use linear regression and select variables which depend on that p values.
- Use Backward, Forward Selection, and Stepwise Selection
- Use Xgboost, Random Forest, and plot variable importance chart.
- Measure information gain for the given set of features and select top n features accordingly.