DATA SCIENCE INTERVIEW QUESTIONS WITH ANSWERS PDF

Ashwani Kumar

1.What is data Science?

Data science is the area of study which involves extracting insights from vast amount 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.What is Prior probability and likelihood?

Prior probability is the proportion of the dependent variable in the data set while the likelihood is the probability of classifying a given observant in the presence of some other variable.

5.Name three types of biases that can occur during sampling?

In the sampling process, there are three types of biases, which are:

* Selection bias

* Under coverage bias

* Survivorship bias

6.Explain Recommender system?

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

7.List out the libraries in python used for Data analysis and Scientific computation.

* SciPy

* Pandas

* Matplotlib

* NumPy

* Scikit

* Seaborn

8.What is bias?

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

9.Discuss ‘Naive’ in naïve bayes algorithm?

The Naïve 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.

10.What is 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 is the criterion variable.

11.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.

12.What is the aim of conducting A/B testing?

AB testing used to conduct random experiments with two variables, A and B. The goal of this testing method is to find out changes to a web page to maximize or increase the outcome of a strategy.

13. Explain Eigen value and Eigen vector.

Eigen vectors are for understanding linear transformations. Data scientist need to calculate the eigen vectors for a covariance matrix or correlation. Eigen values are the directions along using specific linear transformation acts by compressing, flapping, or stretching.

14.Define the term cross-validation.

Cross-validation is a validation technique for evaluating how the outcomes of statistical analysis will generalize for an independent dataset. This method is used in the backgrounds where the objective is forecast, and one needs to estimate how accurately a model will accomplish.

15.Explain the steps for a data analytics project.

The following are important steps involved in an analytics project:

- Understand the Business problem
- Explain the data and study it carefully.
- Prepare the data for modeling by finding missive 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.

16.What is 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.

17.What is 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.Expalin the difference between data scientist and Data analytics?

Data scientist 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 scientist has more technical knowledge then business analyst. Moreover, they don’t need understanding of the business required for data visualization.

19.Explain P – value?

When you conduct a hypothesis test in statistics, a p-value allows you to determine the strength of your results. It is a numerical number between 0 and 1. Based on the value it will help you to denote the strength of the specific result.

20.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).

21.What is Normal Distribution?

A normal distribution is a set of a continuous variable spread across a normal curve or in the shape of a bell curve. You can consider it as a continuous probability distribution which is useful in statistics. It is useful to analyze the variables and their relationships when we are using the normal distribution curve.

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

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

23.Name various types of Deep Learning Frameworks.

- PYTORCH
- Microsoft cognitive toolkit
- Tensor flow
- Caffe
- Chainer
- KERAS

24.Explain Auto -Encoder.

Autoencoders are learning networks. It helps you to transform into outputs with fewer numbers of errors. This means that you will get output to be as close to input as possible.

25.Discuss normal distribution.

Normal distribution equally distributed as such the mean, median and mode are equal.

26.What is recall?

A recall is a ratio of the true positive rate against the actual positive rate. It ranges from 0 to 1.

27.Explain the term Binomial Probability Formula?

The binomial distribution contains the probabilities of every possible success on N trials for independent events that have a probability of Pi of occurring.

28.State the difference between a Validation set and a Test set?

A Validation set mostly considered as a part of the training set as it is used for parameter selection which helps you to avoid overfitting of the model being built.

While a test set is used for testing or evaluating the performance of a trained machine learning model.

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

Following methods of variables 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.

30.Is it possible to capture the correlation between continuous and categorical variable?

Yes, we can use analysis of covariance technique to capture the association between continuous variables.

31.Treating a categorical variable as a continuous would result in a better predictive model?

Yes, the categorical value should be considered as a continuous variable only when the variable is ordinal in nature, so it is a better predictive model.