INTERVIEW QUESTIONS OF DATA SCIENCE

Q1. What is Data Science?

Answer:

Data Science deals with the processes of data mining, cleansing, analysis, visualization, and actionable insight generation. Data Science is the mining and analysis of relevant information from data to solve analytically complicated problems. It is most widely used technique between Artificial Intelligence and Machine Learning. For example, when you logged on any e-commerce website and browsed some categories and products before purchase, you are generating data, which will be helpful for analysists to know your behavior about purchase.

Q2. What are the differences between supervised and unsupervised learning?

Answer:

In supervised learning, all the data is labeled and the algorithms study to forecast the output form the input data, whereas, in unsupervised learning, all data is unlabeled and algorithms study to inherent structure from the input data.

Supervised machine learning can be categorized into the following:-

i. **Classification** – where the output variable is a category like black or white, plus or minus. Naïve Bayes, Support Vector Machine, Decision Tree are the most popular supervised machine learning algorithms.

ii. **Regression** – where the output variable is a real value like quantity.

Un-supervised machine learning can be categorized into the following: -

i. **Clustering** – where you findout the inherent groupings like grouping clients by procuring behavior. K-means clustering, hierarchical clustering and density based spatial clustering are more popular clustering algorithms.

ii. **Association** – where you find out rules that label large slices of your data.

Q3. What are Recommender Systems?

Answer:

A subclass of data sifting frameworks that are intended to anticipate the inclinations or evaluations that a client would provide for an item. Recommender systems are generally utilized in music, pictures, research, news, articles, social labels, and so on.

Q4. Can you utilize machine learning for time series analysis?

Answer:

Yes, machine learning can be utilized for time series analysis but it depends on the applications.

Q5. How will you assess the statistical importance of an insight? whether it is a real insight or just by chance?

Answer:

By utilizing Hypothesis Testing, we can assess the statistical significance of an insight.

Q6. For text analytics, Python or R which one would you give the preference?

Answer:

Python is the best choice for text analytics as it has Pandas library that provides easy to use data structures and better performance data analysis gadgets.

Q7. Which method is utilized to forecast categorical responses?

Answer:

Supervised machine learning i.e. Classification technique is widely utilized in mining for classifying data sets.

Q8. What are the basic expectations to be made for linear regression?

Answer:

Statistical independence of errors, normality of error distribution, linearity and additivity.

Q9. What is the difference between Data Science and Machine Learning?

Answer:

Data Science deals with the processes of data mining, cleansing, analysis, visualization, and actionable insight generation, whereas, machine Learning is the part of Data Science which enables the system to process datasets autonomously without any human interference by utilizing various algorithms to work on massive volume of data generated and extracted from numerous sources.

Q10. What is the formula to calculate R-square?

Answer:

R-Square can be calculated as:-

1 - (Residual Sum of Squares/ Total Sum of Squares)

Q11. What basic knowledge required for Data Scientist?

Answer:

Data Scientist must have the basic knowledge of mathematics, computer programming and statistics to solve the complex data problems in an efficient way to boost the business revenue.

Q12. Names of basic models of Machine Learning?

Answer:

There are two basic models of Machine learning are:-

i. Supervised Machine Learning

ii. Unsupervised Machine Learning

Q13. Do you know about Interpolation and Extrapolation?

Answer:

Interpolation is assessing a value from two known values from a list of values, whereas, extrapolation is assessing a value by extending a known set of values or evidences.

Q14. What are the basic benefits of Data Science?

Answer:

Data Science helps in finding and refining of target viewers. It ensure better communication between service providers and service utilizers. Also improved business value and better risk analysis

Q15. Do you know power analysis?

Answer:

Power Analysis is an experimental design method for determining the effect of a given sample size.

Q16. What are the basic expertise required for Data Science?

Answer:

- Mathematics
- Statistics
- Programming Skills
- Data warehousing
- Machine Learning
- Software Engineering
- Data visualization & communication

Q17. What is Collaborative filtering?

Answer:

It is used by the recommender systems to find patterns or information by collaborating viewpoints, several data sources and various agents.

Q18. What are the top tools utilized in Data Science?

Answer:

- R (a language for statistical computing and graphics)
- Python
- Tableau
- Keras
- Jupyter Notebook

Q19. Are expected value and mean value different or otherwise?

Answer:

No difference, but the terms are used in different situations. Generally, mean is referred when we talking about a probability distribution or sample population, while, expected value is referred in a random variable situation. For sampling data, mean value is the only value that comes from the sampling data, whereas, expected value is the mean of all the means (the value that is built from several samples). For distributions, mean value and expected value are same regardless of the distribution, under the condition that the distribution is in the similar population.

Q20. What are the main process of Data Science?

Answer:

1. Data Exploration:
2. Modeling:
3. Model Testing:
4. Model deployment:

Q21. Is data cleaning plays an important role in analysis?

Answer:

Yes, data cleaning is played an important role in analysis as the number of data sources increases, so, the time consume to clean this data also increases due to the number of sources and the volume of data generated in these sources. About 80% of the time increased for just cleaning data, so, it is an important part of analysis.

Q22. Name any industry players of Data Science?

Answer:

Google – Google hire best data scientists from all over the world and offers the absolute best data science pay rates.

Amazon – Amazon is a worldwide online business and distributed computing mammoth that is contracting data scientists on a major scale. They hire data scientist to get some answers concerning the client mentality, upgrade the geographical contact of both the web based business area and cloud space among different business-driven objectives.

Visa – It is online money related portal for the majority of the organizations and Visa does exchanges in the scope of several millions throughout a day. Because of this, the necessity for data scientists is colossal at Visa to create more income, check false exchanges, and alter the items and administrations according to the client prerequisites.

Q23. What is the difference between univariate, bivariate and multivariate analysis.

Answer:

Univariate, Bivariate and Multivariate analysis are descriptive statistical analysis techniques that can be distinguished on the number of variables involved at a given point of time. For instance, the pie charts of sales based on area involve only one variable, so, it is known as univariate analysis. If the analysis goes to understand the difference between two variables at a time as in a scatter plot, then it is known as bivariate analysis. For instance, analyzing the volume of sale and spending can be measured as an instance of bivariate analysis. Multivariate analysis deals with more than two variables.

Q.24 What is the difference between Cluster and Systematic Sampling?

Answer:

Cluster sampling – It is a technique which can be utilized used when it becomes hard to study the target population spread across an extensive area and simple random sampling cannot be functional.

Systematic sampling – It is a statistical technique which can be utilized where elements are nominated from an ordered selection frame. Equal probability is a best example for systematic sampling.

Q25. Do gradient descent methods always converge to the same point?

Answer:

Gradient descent methods don't always converge to the same point as in few cases it reaches a local minima or a local optima point but we don't reach the global optima point as it based on the data and starting situations.

Q26. What is the basic purpose of A/B Testing?

Answer:

Basically, A/B Testing is a statistical hypothesis testing for randomized research with two variables A and B. The basic purpose of A/B Testing is to recognize any changes to the web page in order to increase or maximize the result of an interest. For instance, recognizing the click through rate for a banner advertisement.

Q27. What is Machine Learning?

Answer:

Machine Learning is the part of Data Science which enables the system to process datasets autonomously without any human interference by utilizing various algorithms to work on massive volume of data generated and extracted from numerous sources. A social media platform i.e. Facebook is a decent example of machine learning implementation where fast and furious algorithms are used to gather the behavioral information of every user on social media and recommend them appropriate articles, multimedia files and much more according to their choice.

Q28. What are the applications of Data Science?

Answer:

- Internet Search Engines
- Speech Recognition
- Recommender Systems
- Self-driving Cars
- Image Recognition
- Comparative analysis of Price
- Fraud and risk detection
- Gaming
- Robotics
- Airline route planning

Q29. What is the difference between a Test Set and a Validation Set?

Answer:

Validation set is used for parameter selection and to avoid overfitting of the model being made, so, it can be considered as a part of the training set, whereas, the test set is used for testing or assessing the performance of a trained machine leaning model. Furthermore, training set is to fit the parameters while validation set is to tune the parameters.

Q30. How can you assess a good logistic model?

Answer:

Various techniques are being used to assess the outcome of a logistic regression analysis-

i. By utilizing Classification Matrix to see the true negatives and false positives.

ii. Harmony which helps identify the ability of the logistic model to distinguish between the event happening or not.

Q31. What is the basic objective of clustering?

Answer:

The basic aim of clustering is to group the related entities in a way that the entities within a group are alike to each other but the groups are dissimilar from each other. In K-Means clustering, “K” defines the number of clusters.

Q.32 What is the difference between Eigen Value and Eigen Vector?

Answer:

Eigen Vectors are used for understanding linear transformation and we usually calculate the eigenvector for correlation or covariance matrix, whereas, Eigen Value can be referred to as the strength of the transformation in the direction of Eigen Vector.

Q33. What steps are involved in making a Decision Tree.

Answer:

i. Take the whole data set as input.

ii. Look for a split that maximize the division of the classes. A split is any test that divides the data into two sets.

iii. Apply the split to the input data (divide step).

iv. Re-apply steps I to II to the separated data.

v. Stop when you meet some stopping criteria.

vi. This step called pruning. Clean up the tree if you went too far doing splits.

Q34. Do you know about selective bias.

Answer:

Selection bias is a problematic situation in which error is launch due to a non-random population section.

Q35. What types of biases can occur during sampling?

Answer:

i. Selection bias

ii. Survivorship bias

iii. Under coverage bias

Q36. What are the different kernels functions in Support Vector Machine?

Answer:

Four types of kernels in Support Vector Machine.

i. Linear Kernel

ii. Polynomial kernel

iii. Sigmoid kernel

iv. Radial basis kernel

Q37. What is pruning in Decision Tree?

Answer:

The process of removing sub-nodes of a decision node is called pruning or reverse process of splitting.

Q38. What is deep learning?

Answer:

It is a subfield of machine learning inspired by structure and role of brain called Artificial Neural Network (ANN). Deep learning is an extension of Neural Network while there are a lot of algorithms under machine learning like Linear Regression, Support Vector Machine (SVM), Neural Network, etc.

Q39. What statistical methods are useful for data-scientist?

Answer:

i. Bayesian method

ii. Markov process

iii. Simplex algorithm

iv. Mathematical optimization

v. Spatial and cluster processes

vi. Rank statistics, percentile, outlier's detection

vii. Imputation techniques

Q40. What tools are utilized for data analysis?

Answer:

- RapidMiner
- Tableau
- KNIME
- Google Fusion Tables
- Google Search Operators
- Solver
- io
- NodeXL
- OpenRefine

Q41. What are the properties of clustering algorithms?

Answer:

i. Hard and soft

ii. Hierarchical or flat

iii. Iterative

iv. Disjunctive

Q42. In many domain Time Series Analysis performed?

Time Series analysis can be performed in following two domains:-

i. Time domain

ii. Frequency domain.