Machine Learning Interview Questions

Q.1: What are the Basic assumption in Naïve Bayes?

Ans: Features are independent

Q.2 Advantages of Naïve Bayes?

Ans: work very well with many number of features and works well with large training dataset. It converges faster when we are training the model and also performs well with categorical features.

Q:3 Disadvantage .

Ans: Correlated features affects performance.

Q:4 Whether features Scaling is required?

Ans: No.

Q:5 What is Overfitting ? And how can you avoid it?

Ans: Overfitting occurs when the model learns the training set too well. It takes up random fluctuations in the training data as concepts. These impact the model's ability to generalize and don't apply to new data.

Q:6 What is 'Training set' and 'Test set ' in a Machine Learning model? How much data will you allocate for your training validation and test set?

Ans: Training set- Training set are examples given to the model to analyze and learn. This is labelled data used to train the model.

Test set- Test set is used to test the accuracy of the hypothesis generated by the model. We test without labeled data and then verify results with labels.

Q:7 How do you handle missing or corrupted data in a dataset?

Ans: The ways to handle missing /corrupted data is to drop those rows/columns or replace them completely with some other value.

There are two useful methods in panda:

a.) Isnull() and dropna() will help finding the columns/ rows with missing data and drop them.
b.) Fillna() will replace the wrong values with a place holder values(0).

Q:8 How can you choose a classifier based on training set size?

Ans: When the training set is small, a model that has a high bias and low variance seems to work better because they area less likely to overfit. For e.g. Naïve Bayes works best.When the training set is large, models with low bias and high variance tend to perform better as they work fine with complex relationships. E.g. Decision Tree.

Q:9 Explain confusion matrix with respect to ML algorithms.

Ans: Confusion Matrix is a specific table that is usewd to measure the performance of an algorithm. It's mostly used in supervised machine learning. Confusion matrix has two dimensions:

a.) Actual
b.) Predicted

It also has identical sets of features in both these dimensions.

Q:10 What are the three steps to build a model in machine learning ?

Ans: Model Building:- a.) Choose the suitable algorithm

b.)Train the model using training dataset.

Model Testing:- a.)Test the model with new data

b.)Check the accuracy of the mode

Applying the model:- a.) Make required changes after testing.

b.)Apply for real time projects.

Q:11 What is deep learning?

Ans: Deep Learning involves systems that think and learn like humans using artificial neural networks.

Some points:-

a.) Best features are selected by the system
b.) It's subset of machine learning
c.) Lesser testing fame
d.) Better scalability
e.) Problems solved in an end-to-end method
f.) Performance improves with more data

Q12. How do you handle corrupted or missing data in a dataset?

Ans:- You could find corrupted or missing data in a dataset by either dropping those rows or columns or replacing them with another value. There are two methods - isnull() and dropna(). These methods will help you find columns of missing/corrupted data. If you want to fill invalid values with a placeholder value, you could use the fillna() method.

Q13. Do you know what Spark is?

Ans:- This is one of the commonly asked Machine Learning interview questions that test your knowledge and experience in Spark. A spark is a great tool used to handle massive datasets with speed. It is presently in demand.

Q14. What is a hash table?
Ans:- A hash table is a data structure that produces an associative array. A hash table is used for database indexing. In a hash table, a key is mapped to certain values through a hash function.

Q15. How are foreign and primary keys related to SQL?
Ans:- If you are attempting Machine Learning interview questions at top-ranked technical or FAANG companies, you must have a profound knowledge of various data formats. SQL is one of those. You should be familiar with how to manipulate SQL databases. The key differences between a primary and foreign key are:

Q16. What are data types supported by JSON?
Ans:- You must be adept in JSON to answer these types of Machine Learning interview questions. You can manipulate six data types in JSON - numbers, strings, objects, null values, arrays, and booleans.

Q17. What is the difference between bias and variance?
Ans:- Bias is an error that occurs due to overly simplistic assumptions or erroneous assumptions in the learning algorithm. If you use bias, it can lead to the model underfitting your data with low predictive accuracy.

On the other hand, variance is an error due to complexity in the learning algorithm. In variance, your data gets highly sensitive to high degrees of variation, leading your model to overfit the data. You'll end up carrying noise from your training data for your model to be useful for your test data.

Q18. How will you differentiate between supervised and unsupervised machine learning?
Ans:-Supervised learning required training labeled data. For instance, to classify a supervised learning task, you must first label the data you'll use to train the model. Contrastingly, unsupervised learning does not require labeling data explicitly.

Q19. How does a ROC curve work?
Ans:- ROC is a graphical representation of the contrast between true positive rates and the false positive rate at various thresholds. You should know that it's often used as a proxy for the trade-off between the true positives (sensitivity of the model) vs. the false positives (fall-out or probability to trigger a false alarm).

Q20. What is Bayes' theorem? How is it useful in machine learning?
Ans:- Using Bayes' theorem, you can get the posterior probability of an event given that is known as prior knowledge. Bayes' theorem notably includes the Naive Bayes classifier.

Q21. What is 'Naive' Bayes naive?
Ans:-Naive Bayes is considered naive because it makes assumptions impossible to see in real-life data. Despite its practical applications, especially in text mining, the resulting

probability implies the absolute independence of features, which is a condition that can never be met in real life.

Q22. What is the difference between L1 and L2 regularization?
Ans:- L1 is binary/sparse, with many variables assigned a 1 or 0 in weighting. It corresponds to setting a Laplacean before the terms. In contrast, L2 regularization tends to spread error among all the terms. L2 corresponds to a Gaussian prior.

Q23.How to use backpropagation and certain principles
Ans:- It represents an unsupervised learning algorithm that learns data representations through the use of neural networks.

Q1. Do you have research experience in machine learning?
Q2. What are your favorite use cases of machine learning models?

Q3. Where do you usually source datasets?
Q4. What are your favorite APIs to explore?
Q5. How do you think quantum computing will affect machine learning?

What are the different types of machine learning?

What is overfitting, and how to avoid it?

What are the "training Set" and "test Set" in a Machine Learning Model?

How much data will you allocate for your training, validation, and test sets?

Explain the confusion matrix with respect to machine learning algorithms.

- **Write a pseudo code for a parallel implementation.**
- **What data visualization tools do you use?**
- **What are the three stages of building a model in machine learning?**
- **Given two strings, A and B, of the same length n, find whether it is possible to cut both strings at a common point such that the first part of A and the second part of B form a palindrome.**
- **How do XML and CSVs compare in terms of size?**
- **What is the difference between inductive machine learning and deductive machine learning?**
- **How would you build a data pipeline?**
- **How is KNN different from k-means clustering?**
- **What is PCA? When do you use it?**
- **Will you explain your favorite algorithm in less than a minute?**
- **What are type I and II errors?**
- **What is a Fourier transform?**
- **What is the difference between probability and likelihood?**
- **What is an F1 score? How will you use it?**
- **Explain the SVM algorithm in detail.**
- **Which one is more important - model performance or model accuracy?**
- **How is a decision tree pruned?**
- **What is the difference between generative and discriminative models?**
- **How will you handle an imbalanced dataset?**

- **How do you ensure you're not overfitting with a model?**
- **What is a kernel trick? How is it useful?**
- **What do you think of the current data process?**
- **How can you use machine learning skills to generate revenue?**
- **How would you implement a recommendation system for our company's users?**
- **What do you think is the most valuable data in our business?**
- **How will you approach the Netflix Prize competition?**
- **How do you think Google is training data for self-driving cars?**
- **How would you simulate the approach AlphaGo took to beat Lee Sedol at Go?**
- **What are your thoughts on GPT-3 and OpenAI's model?**
- **How do you design an email spam filter?**
- **What models do you train for fun, and what GPU/hardware do you use?**