By OnlineInterviewQuestions.com

Machine Learning Interview Questions

Q1. What is machine learning?

Machine learning is the use of algorithms and data to perform specific tasks. **ML** is the process of giving the system the ability to learn from data using certain sophisticated algorithms. With the algorithms, you build a certain mathematical model on the data to explore patterns or predict.

ML today is used in a wide variety of applications like **Spotify**, **Netflix**, **Amazon**, **Google Search**, etc.

Q2. What supervised and unsupervised machine learning?

Supervised and **unsupervised** are the two types of **Machine learning algorithms** available.

In the **supervised type**, the algorithms are applied to the known labeled data to formulate a model. Labeled data means the data is tagged. With this data, the algorithm creates a model that is then applied to unknown data to predict its outcome or tag it. Linear regression is a good example of supervised learning.

In the **unsupervised type**, the algorithm is applied to the unlabeled data. The data neither is classified nor labeled, but the unsupervised algorithm is used to find the hidden structure with the unlabeled data. K-means clustering algorithm is a good example of this type.

Q3. What is ROC curve?

ROC curve is a graphical plot to illustrate the ability of a classifier system. Basically, this curve tells you how much a binary classifier system is capable of distinguishing between classes. This curve is plotted with **TPR (True Positive Rate)** on the **y-axis** and **FPR (False Positive Rate)** on the **x-axis**. TPR is also known as sensitivity recall or probability of detection and FPR is also known as the probability of false alarm.

Q4. What is the difference between classification and regression?

Regression is the process of estimating the mapping function (f) given the input value (x) to the continuous output value (y). It is used to predict a value given the data. Here, labeled data is used to create a model or function and this function is used to predict the value of unlabeled data. Linear or Logistic regression is a good example of this type.

Classification is the process of categorizing the data. The classification model is created from using the algorithm on the data so it is categorized mainly based on the similarity. Naive Bayes classifier is a good example of this type.

Q5. What is Ensemble learning?

Ensemble learning is the process of applying multiple learning algorithms on a dataset to get better predictive performance. By applying multiple algorithms the performance is improved while the likelihood of choosing a wrong algorithm is reduced. The ensemble is a supervised type of learning algorithm as it can be trained and used to make predictions.

Q6. Explain Navie Bayes?

Naive Bayes is a type of classification algorithm used to classify data based on the probabilistic classifiers. It is a collection of classification algorithms that uses Baye's theorem. This theorem finds the probability of an event occurring given the probability of an already occurred other event.

```
//Baye's Theorem mathematical equation
P(A/B) = P(B/A) * P(A) / P(B)
```

Q7. What is Reinforcement Learning?

Reinforced learning is a type of machine learning that employs a trial and error method to find a solution to the problem. It is used in many software and machines to find the best possible path to get to the solution The agent (i.e) the learning model takes the action to maximize the reward in a particular situation. There is no need for labeled data in this type as the reinforcement agent decides what to do with the data given the task.

Q8. What Is training Set and test Set in a Machine Learning Model?

The training data is used by the algorithm to create a model. It used this data to learn and fit the model. In a dataset, about 60 to 80 percent of data is allocated as training data. The testing data is used to test the accuracy of the model trained with the training data. The model from the training data predicts the testing data to see how well it works. Separating the dataset into training and testing data is important as you can minimize the effect of data discrepancies and better understand the characteristics of the model.

Q9. What is Confusion Matrix?

Confusion Matrix, also known as the error matrix, is a table to describe the performance of the classification model on the set of test data. The rows in this table represent the predicted class while the column presents the actual class. In this table, the number of correct and incorrect predictions are described with the count values so we can get insights into the errors and the type of errors made.

Q10. What are stages of building a model in Machine Learning?

The seven steps in building a machine learning model are,

Data Collection - In this step, we collect the data related to the problem.

Data Preparation - Here, we clean and organize the collected data based on the problem. We remove duplicate data, error data, fill missing data, etc in this process

Choosing an algorithm - As the name suggests, in this stage, you choose the appropriate algorithm for the problem.

Train the algorithm - We use the dataset to train the algorithm to create a model.

Evaluate the model - We use the test data from the dataset to find the accuracy of the model created.

Parameter Tuning - In this step, we tune the model parameters to improve its performance.

Make predictions - In this step, we apply the created model on a real dataset.

Q11. What Is Deep Learning?

Deep learning is a subfield of machine learning which uses an artificial neural network to learn from the dataset. It is now the most popular **ML technique** which is used in many areas such as **driverless cars**, **voice control**, **hand's free speakers**, and more. This technique uses data directly from the **image**, **sound**, or **video** to learn from it. The artificial neural network has multiple layers of nodes interconnected with each other. It is loosely inspired by biological neural networks. Deep Learning achieves good accuracy when compared to other models sometimes even exceeding human-level performance.

Q12. Explain KNN Algorithm?

KNN is a supervised **algorithm** used for both **classification** and **regression**. It uses labeled data to model a function to produce an output from the unlabeled data. It uses the Euclidean distance formula to calculate the distance between the data points for classification or prediction. It works on the principle that similar data points must be close to each other so it uses the distance to calculate the similar points that are close to each other.

Q13. What Is a Random Forest?

Random forest is a **supervised algorithm** that is mainly **used for classification problems**. It creates a decision tree from the data samples. Based on the decision tree, it predicts the result. Then, the voting process takes place in which voting is performed for every predicted result. Finally, the most voted prediction result is taken as the final prediction result. This technique is also used as regression as well.

Q14. What Is Decision Tree Classification?

The decision tree algorithm is a supervised learning algorithm that is used for **classification** as well as regression problems. In this type, we infer the simple decision rules from the training data and create a decision tree. We start from the root attribute of the decision tree with the record attribute and follow the branch of the root that corresponds to the match. In this way, we jump to the next branch until the final classification is reached.

Q15. What are collinearity and multicollinearity?

Collinearity is the association between two explanatory variables while the **multicollinearity** is the linear related association between two or more explanatory variables. Collinearity occurs when two predictor variables have a non-zero correlation in multiple regression. Non-collinearity occurs when two or more predictor variables are inter-correlated.