

Q1. Applications of ML.

Answer:.

1. Product recommendations:

Machine learning is widely used by various e-commerce and entertainment companies such as Amazon, Netflix, etc., for product recommendation to the user.

2. Traffic prediction:

If we want to visit a new place, we take help of Google Maps, which shows us the correct path with the shortest route and predicts the traffic conditions.

It predicts the traffic conditions such as whether traffic is cleared, slow-moving, or heavily congested with the help of two ways:

Real Time location of the vehicle from Google Map app and sensors
Average time has taken on past days at the same time.

3. Speech Recognition

While using Google, we get an option of "Search by voice," it comes under speech recognition, and it's a popular application of machine learning.

4. Image Recognition:

Image recognition is one of the most common applications of machine learning. It is used to identify objects, persons, places, digital images, etc.

5. Social Media Features

Social media platforms use machine learning algorithms and approaches to create some attractive and excellent features.

Q2. Discuss Generalization Error.

Answer: The generalization error of a machine learning model is the difference between the empirical loss of the training set and the expected loss of a test set. In practice, it is measured by the difference between the error of the training data and the one of the test data.

Q3. What is Machine Learning?

Answer: Machine Learning is defined as the capability of a machine to imitate intelligent human behavior. Artificial intelligence systems are used to perform complex tasks in a way that is similar to how humans solve problems. It is a subfield of artificial intelligence. It basically focuses on analyzing and interpreting patterns and structures in data to enable learning, reasoning, and decision making outside of human interaction.

Q4. Write Types of ML.

Answer: Types of Systems of Machine Learning

There are different types of machine-learning systems.

1. Supervised
2. Unsupervised
3. Semi-supervised
4. Reinforcement Learning

Q5. What is Linear Regression?

Answer: Linear Regression is a machine learning algorithm based on supervised learning. It performs a regression task. Regression models a target prediction value based on independent variables. It is mostly used for finding out the relationship between variables and forecasting.

Linear regression can be further divided into two types of the algorithm:

1. Simple Linear Regression:

If a single independent variable is used to predict the value of a numerical dependent variable, then such a Linear Regression algorithm is called Simple Linear Regression.

2. Multiple Linear regression:

If more than one independent variable is used to predict the value of a numerical dependent variable, then such a Linear Regression algorithm is called Multiple Linear Regression.

Q6. What is Regression?

Answer: Regression is a technique for investigating the relationship between independent variables or features and a dependent variable or outcome. Predicting prices of a house given the features of house like size, price etc is one of the common examples of Regression. It is a supervised technique.

Q7. What is Multivariate Linear Regression?

Answer: Multivariate Regression is a supervised machine learning algorithm involving multiple data variables for analysis. Multivariate regression is an extension of multiple regression with one dependent variable and multiple independent variables.

Q8. What is Logistic Regression?

Answer: Logistic regression is one of the most popular Machine Learning algorithms, which comes under the Supervised Learning technique.

- It is used for predicting the categorical dependent variable using a given set of independent variables.
- Logistic regression predicts the output of a categorical dependent variable. Therefore the outcome must be a categorical or discrete value.

Q9. What are Decision Trees?

Answer: Decision trees classify instances by sorting them down the tree from the root to some leaf node, which provides the classification of the instance. Each node in the tree specifies a test of some attribute of the instance, and each branch descending

Q10. What is the fullform of CART?

Answer: CART stands for Classification And Regression Tree. CART algorithm is a classification algorithm for building a decision tree based on Gini's impurity index as a splitting criterion. CART is a binary tree built by splitting a node into two child nodes repeatedly.

Q11. What is the ROC curve?

Answer: ROC stands for receiver operating characteristic and it's a tool that is used with binary classifiers. An ROC curve is a graph showing the performance of a classification model at all classification thresholds. This curve plots two parameters: True Positive Rate, False Positive Rate.

Q12. Describe Kappa Statistics.

Answer: Kappa Statistics is a measure of how closely the instances classified by the machine learning classifier matched the data labeled as ground truth, controlling for the accuracy of a random classifier as measured by the expected accuracy.

Q13. Discuss F-measure.

Answer: The F-score is a measure of a model's accuracy on a dataset. It is used to evaluate binary classification systems, which classify examples into 'positive' or 'negative'.

Q14. What is K-fold cross validation?

Answer: Cross-validation is a resampling procedure used to evaluate machine learning models on a limited data sample. The procedure has a single parameter called k that refers to the number of groups that a given data sample is to be split into.

Q15. Write definition of XGBoost.

Answer: XgBoost stands for Extreme Gradient Boosting, which was proposed by the researchers at the University of Washington. It is a library written in C++ which optimizes the training for Gradient Boosting. XGBoost is an implementation of Gradient Boosted decision trees. XGBoost models majorly dominate in many Kaggle Competitions.

Q16. Discuss Bagging.

Answer: A Bagging classifier is an ensemble meta-estimator that fits base classifiers each on random subsets of the original dataset and then aggregate their individual predictions (either by voting or by averaging) to form a final prediction. Bagging reduces overfitting (variance) by averaging or voting, however, this leads to an increase in bias, which is compensated by the reduction in variance though.

Q17. What is Random Forest?

Answer: Every decision tree has high variance, but when we combine all of them together in parallel then the resultant variance is low as each decision tree gets perfectly trained on that particular sample data and hence the output doesn't depend on one decision tree but multiple decision trees. Random Forest has multiple decision trees as base learning models. We randomly perform row sampling and feature sampling from the dataset forming sample datasets for every model.

Q18. Describe Support Vector Machine SVM.

Answer: SVM is a supervised machine learning algorithm which can be used for classification or regression problems. They are used in applications like handwriting recognition, intrusion detection, face detection, email classification, gene classification, and in web pages. It can handle both classification and regression on linear and non-linear data.

Q19. Write about Margins and support vectors.

Answer: The objective of the support vector machine algorithm is to find a hyperplane in an N -dimensional space (N — the number of features) that distinctly classifies the data points.

Q20. What is Quadratic Programming?

Answer: Quadratic programming (QP) is the process of solving certain mathematical optimization problems involving quadratic functions.

Q21. Basics of Kernel Trick.

Answer: Kernel Trick is a simple method where a Non Linear data is projected onto a higher dimension space so as to make it easier to classify the data where it could be linearly divided by a plane.

Q22. Write about Support Vector Regression.

Answer: Support Vector Regression is a supervised learning algorithm that is used to predict discrete values. Support Vector Regression uses the same principle as the SVMs. The basic idea behind SVR is to find the best fit line.

Q23. What is Clustering ?

Answer: Clustering is the task of dividing the population or data points into a number of groups such that data points in the same groups are more similar to other data points in the same group and dissimilar to the data points in other groups. It is basically a collection of objects on the basis of similarity and dissimilarity between them.

Q24. What is Multiclass Classification?

Answer: multinomial classification is the problem of classifying instances into one of three or more classes

Q25. Overview of distance metrics.

Answer: Distance metrics are used in both supervised and unsupervised learning, generally to calculate the similarity between data points.

Q26. What is Graph Based Clustering?

Answer: Graph clustering is to group the vertices of a graph into clusters based on the graph structure and/or node attributes. Graph clustering aims at partitioning a set of graphs into different groups that share some form of similarity.

Q27. Describe DBSCAN.

Answer: The DBSCAN algorithm is based on this intuitive notion of “clusters” and “noise”. The key idea is that for each point of a cluster, the neighborhood of a given radius has to contain at least a minimum number of points.

DBSCAN algorithm requires two parameters:

1. **eps :**
 - It defines the neighborhood around a data point i.e. if the distance between two points is lower or equal to ‘eps’ then they are considered neighbors.
2. **MinPts:**

Minimum number of neighbors (data points) within eps radius.
Larger the dataset, the larger value of MinPts must be chosen.

Q28. What is Model based Clustering ?

Answer:Model-based clustering is a statistical approach to data clustering. The observed (multivariate) data is assumed to have been generated from a finite mixture of component models.

- Each component model is a probability distribution, typically a parametric multivariate distribution.

Q29. What is Dimensionality Reduction ?

Answer:Dimensionality reduction is a machine learning (ML) or statistical technique of reducing the amount of random variables in a problem by obtaining a set of principal variables. Principal Component Analysis (PCA), Backward Feature Elimination, Forward Feature Selection, Missing Value Ratio are some of the many dimension reduction techniques .

Q30. Discuss Principal Component Analysis.

Answer:Principal Component Analysis is a statistical process that converts the observations of correlated features into a set of linearly uncorrelated features with the help of orthogonal transformation.

- These new transformed features are called the Principal Components.
- It is one of the popular tools that is used for exploratory data analysis and predictive modeling.

Q31. What is Singular Value Decomposition ?

Answer:The Singular Value Decomposition (SVD) of a matrix is a factorization of that matrix into three matrices. It has some interesting algebraic properties and conveys important geometrical and theoretical insights about linear transformations.