Question Bank

1. **Define Bias and Variance.**

   Ans: 1.Bias: Refers to the error introduced by approximating a real-world problem, assuming it is simpler than it actually is.
   -High bias can lead to underfitting, where the model is too simplistic and fails to capture the underlying patterns in the data.
   2.Variance: Represents the model's sensitivity to fluctuations in the training data. High variance can lead to overfitting, where the model performs well on the training data but fails to generalize to new, unseen data.

2. **What is the structure of Learning?**

   Ans: i)Data Collection: Gathering relevant data that the model will learn from.
   ii)Data Preprocessing: Cleaning, organizing, and preparing the data for analysis.
   iii)Feature Selection/Engineering: Choosing relevant features (input variables) or creating new ones to improve model performance.
   iv)Model Selection: Choosing a suitable machine learning algorithm or model architecture for the task at hand.
   v)Training: Using the prepared data to teach the model to make predictions or decisions.
   vi)Evaluation: Assessing the model's performance on new, unseen data to ensure it generalizes well.
   vii)Hyperparameter Tuning: Adjusting the model's settings to optimize its performance.

3. **How to make predictions in Linear regression?**

   Ans: 1.In linear regression, making predictions involves using the equation of the fitted line. The general form of a simple linear regression equation is:
   $y\hat{} = mx + b$
   Where:

- y^ is the predicted output.
- x is the input feature.
- m is the slope of the line.
- b is the y-intercept.

2.For multiple linear regression with multiple features:

y^ = b0 + b1x1 + b2x2 +.....+bnxn

Where:

- b0 is the intercept.
- b1,b2,.....,bn are the coefficients for each feature x1,x2,....,xn

3.To make predictions:

1. Plug in the values of the input features into the regression equation.
2. Calculate the predicted output (y^)

For example, if you have a simple linear regression equation y^ = 2x + 1, y^ for x = 3:

y^ = 2(3) + 1 = 7

So, when x = 3, the predicted y^ is 7.

4. What is NO-Free Lunch theorem?

Ans: 1.The "No Free Lunch" (NFL) theorem in machine learning essentially states that there is no one-size-fits-all algorithm that performs best for every type of problem.

2.In other words, no single machine learning algorithm is universally superior across all possible datasets or tasks.

3.The theorem was introduced by David Wolpert in the late 1990s.

4.It emphasizes that the performance of a machine learning algorithm is contingent on the specific characteristics and nature of the data it is applied to.

5.As a consequence of the NFL theorem, it underscores the importance of selecting and designing algorithms based on the specific properties of the data at hand.

6.This highlights the need for experimentation, careful consideration of the problem domain, and sometimes the development of specialized algorithms tailored to the unique

features of a given task. Essentially, there is no free lunch in the sense that there is no algorithm that universally outperforms all others in every scenario.

5. What do you mean by stopping criterion in CART?

Ans: 1.The context of Classification and Regression Trees (CART) in machine learning, a stopping criterion is a condition that determines when the construction of the decision tree should halt or stop growing. The purpose of a stopping criterion is to prevent the tree from becoming overly complex, which could lead to overfitting.

2.Common stopping criteria in CART include:

i)Minimum Samples Per Leaf: Specify a threshold for the minimum number of samples required to create a terminal (leaf) node. If a node has fewer samples than this threshold, it won't be split further.

ii)Maximum Depth: Set a limit on the maximum depth or levels of the tree. Once the tree reaches this depth, it stops growing.

iii)Minimum Impurity Decrease: Define a threshold for the minimum amount of impurity decrease required for a split to be considered. If a potential split doesn't improve the overall impurity by at least this amount, the split is not performed.

iv)Maximum Number of Leaves: Limit the total number of terminal nodes (leaves) in the tree.

6.What are the libraries used to do visualization?

Ans: 1.Matplotlib: A versatile 2D plotting library for creating static, animated, and interactive visualizations in Python. It is often used for basic plotting and visualization.

-Website: Matplotlib

2.Seaborn: Built on top of Matplotlib, Seaborn provides a high-

level interface for drawing attractive and informative statistical graphics. It simplifies the process of creating complex visualizations.

-Website: Seaborn

3.Plotly: A powerful and interactive graphing library that can be used to create a variety of plots. It supports both Python and JavaScript.

-Website: Plotly

4.Bokeh: A Python interactive visualization library that targets modern web browsers for presentation. It allows for the creation of interactive plots and dashboards.

-Website: Bokeh

5.Altair: A declarative statistical visualization library for Python that is simple, concise, and produces interactive visualizations.

-Website: Altair

7.Explain the Framework of inductive learning.

Ans: 1.The inductive learning process involves iteratively refining the hypothesis based on the observed training examples, aiming to generalize well to new, unseen instances.

2.The goal is to find a hypothesis that approximates the target function with high accuracy on both the training and testing data.

3.Inductive learning is commonly associated with supervised learning, where the algorithm is provided with labeled examples to learn from. 4.The learned hypothesis can then be applied to make predictions on new, unlabeled instances.

5.This framework forms the basis for various machine learning algorithms, including decision trees, support vector machines, neural networks, and more.

8.Define Training set and Test set.

Ans: 1.Training Set:

1.The training set is a subset of the dataset that is used to train a machine learning model.

2.It consists of input-output pairs (instances and their corresponding labels) that the model uses to learn the underlying patterns and relationships in the data.

2.Test Set:

1.The test set is another subset of the dataset that is distinct from the training set.

2.It is used to evaluate the performance and generalization ability of the trained machine learning model.

9.Define Overfitting and Underfitting.

Ans: 1.Overfitting: Overfitting occurs when a machine learning model learns the training data too well, capturing noise and random fluctuations instead of the underlying patterns. As a result, the model may perform exceptionally well on the training set but fails to generalize to new, unseen data.

2.Underfitting: Underfitting happens when a machine learning model is too simple to capture the underlying patterns in the training data. The model fails to learn important relationships and performs poorly both on the training set and on new, unseen data.

10.What is Aggregation?

Ans: 1.In machine learning, aggregation refers to the process of combining the predictions of multiple models to produce a single, more robust prediction. 2.Ensemble learning techniques, such as bagging and boosting, commonly use aggregation to improve the overall performance and generalization of a model. 3.Common aggregation methods include averaging (for regression problems) and voting (for classification problems).

4.Popular ensemble algorithms, such as Random Forest

(bagging) and AdaBoost (boosting), employ aggregation to enhance predictive accuracy.

11.Give any 3 applications of ML.

Ans: Three applications of machine learning include:

1.Image Recognition: Machine learning is extensively used in image recognition applications, such as facial recognition, object detection, and autonomous vehicles. Convolutional Neural Networks (CNNs) are commonly employed to analyze and classify visual data.

2.Natural Language Processing (NLP): ML is applied in NLP for tasks like sentiment analysis, language translation, chatbots, and speech recognition. Recurrent Neural Networks (RNNs) and Transformer models are often used in NLP applications.

3.Healthcare Predictive Analytics: Machine learning is used to analyze medical data for disease prediction, diagnosis, and treatment recommendations. ML models can predict patient outcomes, identify potential health risks, and optimize personalized treatment plans.

12.Define: Gini cost.

Ans: In the context of decision trees, the Gini cost is a measure of impurity or the degree of disorder in a dataset. The Gini impurity is used to evaluate how well a particular split separates the data into classes. It is commonly employed in the construction of decision trees, specifically in algorithms like CART (Classification and Regression Trees).

13.What do you mean Gradient Descent?

Ans: 1.Gradient Descent is an iterative optimization algorithm used to minimize the cost function in machine learning models, particularly in training processes.

2.The goal is to find the optimal parameters (weights and biases) for a model that minimize the difference between the predicted and actual outcomes.

3.The basic idea is to iteratively adjust the model parameters in the direction of steepest descent (negative gradient) of the cost function.
4.The algorithm computes the gradient of the cost function with respect to each parameter and updates the parameters in proportion to the negative of the gradient.
5..There are different variants of gradient descent, such as batch gradient descent (updating parameters using the entire training dataset), stochastic gradient descent (updating parameters using a single randomly chosen training sample), and mini-batch gradient descent (updating parameters using a small batch of training samples).

14.Explain limit overfitting. How to handle the sequence?

Ans: Limiting Overfitting:Overfitting occurs when a model learns the training data too well and fails to generalize to new, unseen data. To limit overfitting, several techniques can be employed:

1.Cross-validation: Split the dataset into training and validation sets to evaluate the model's performance on unseen data during training.

2.Regularization: Add penalties to the model's parameters to discourage overly complex models. Common regularization techniques include L1 regularization (Lasso) and L2 regularization (Ridge).

3.Feature Selection: Choose relevant features and discard irrelevant ones to reduce model complexity.

4.Early Stopping: Monitor the model's performance on a validation set during training and stop training when performance starts to degrade.

5.Ensemble Methods: Use ensemble methods like Random Forests, which combine multiple models to improve generalization and reduce overfitting.

15.How Random forest using ML can influence the efficiency of the algorithm?

Ans:

1.Random Forest is an ensemble learning method that combines the predictions of multiple decision trees to improve predictive accuracy and control overfitting. Here's how Random Forest influences the efficiency of the algorithm:

i.Diversity: Random Forest builds multiple decision trees by considering random subsets of features and data points. This diversity helps reduce overfitting and increases the model's ability to generalize to new data.

ii.Feature Importance: Random Forest provides a measure of feature importance, indicating the contribution of each feature to the model's predictive performance. This information can guide feature selection and improve model interpretability.

iii.Robustness: Random Forest is less sensitive to outliers and noisy data compared to individual decision trees. The aggregation of multiple trees tends to smooth out individual errors.

iv.Parallelization: Random Forest training can be parallelized, making it computationally efficient and suitable for large datasets.

16.How do you check the accuracy in ML model.

Ans: 1.Accuracy in machine learning is commonly measured using the accuracy score, which is the ratio of correctly predicted instances to the total instances in the dataset. The formula for accuracy is:

Accuracy=Number of Correct Predictions/Total Number of Predictions

2.It is a straightforward metric for classification problems. However, in some cases, accuracy alone may not provide a complete picture, especially when dealing with imbalanced datasets. Additional metrics like precision, recall, and F1 score may be used for a more comprehensive evaluation.

17.Can ML increase the predictions efficiency? Justify your answer.

Ans: Yes, machine learning can increase prediction efficiency. Machine learning models, when properly trained on relevant data, can learn intricate patterns and relationships in the data, allowing them to make accurate predictions on new, unseen instances. ML algorithms can handle complex tasks, automate decision-making processes, and adapt to changing patterns in the data, contributing to increased prediction efficiency.

18.Give the formula for finding the prediction values?

Ans: In the context of a simple linear regression model, the formula for finding prediction values (y^) is:

y^ = mx + b

Where:

- y^ is the predicted output.
- x is the input feature.
- m is the slope of the line.
- b is the y-intercept.

2.For multiple linear regression with multiple features:

y^ = b0 + b1x1 + b2x2 +.....+bnxn

Where:

- b0 is the intercept.
- b1,b2,.....,bn are the coefficients for each feature x1,x2,....,xn

19.How Precision and Recall calculated in ML?

Ans: Precision and Recall Calculation in ML:

- Precision:TP/TP+FP

-Recall (Sensitivity): TP/TP+FN

20.How CART model can be created from DATA?

Ans: A CART (Classification and Regression Trees) model is created by recursively splitting the dataset based on features that minimize impurity (e.g., Gini impurity for classification, mean squared error for regression). The process continues until a stopping criterion is met, such as a predefined tree depth or the achievement of a certain level of purity in the leaf nodes.

21.How errors are calculated in Linear Regression?

Ans: - Errors in linear regression are typically measured using metrics like Mean Squared Error (MSE) or Mean Absolute Error (MAE).(write also the formula of MSE and MAE)

22.Explain Tree Pruning.

Ans: Tree pruning is the process of removing branches (subtrees) from a decision tree to prevent overfitting. It involves simplifying the tree by removing nodes that contribute less to overall performance. Pruning helps improve the model's ability to generalize to new data.

23.Explain in detail the Learning methods of ML.

Ans: 1.Supervised Learning: Models learn from labeled data, making predictions or decisions.

2.Unsupervised Learning: Models find patterns in unlabeled data, often used for clustering or dimensionality reduction.

3.Reinforcement Learning: Agents learn by interacting with an environment, receiving feedback in the form of rewards or penalties.

24.How Non-linear algorithm different from Linear

algorithm?

Ans: 1.Linear Algorithms: Follow linear relationships. Examples include linear regression and linear support vector machines.

2.Non-linear Algorithms: Capture complex patterns and relationships. Examples include decision trees, random forests, and neural networks.

25.Explain Generalization in ML.

Ans: 1.Generalization: The ability of a model to perform well on new, unseen data.

2.Evaluation: Techniques like cross-validation assess a model's generalization by testing its performance on data not used during training.

3.Goal: ML models aim for good generalization to make reliable predictions in real-world scenarios.

27.Explain Linear Regression Model.

Ans: 1.definition: Predictive modeling method for linear relationships.

2.Equation: y^ =mx+b.

3.Objective: Minimize the difference between predicted (y^) and actual (y) values.

4.Steps in Linear Regression:

-Data Collection: Gather a dataset with predictor and outcome variables.

-Data Preprocessing: Clean and prepare the data for analysis.

-Feature Selection/Engineering: Choose relevant features or create new ones.

-Model Selection: Choose linear regression as the algorithm.

-Training: Adjust coefficients to minimize prediction errors.

-Evaluation: Assess model performance on a test dataset.

-Prediction: Apply the model to make predictions on new data.

38.How ML and AI are coordinated?

Ans: 1.Machine Learning (ML) is a part of Artificial Intelligence (AI).

2.ML focuses on creating systems that learn from data, allowing them to improve over time without explicit programming.

3.AI, on the other hand, is a broader concept encompassing various technologies, including ML.

4.ML helps AI systems recognize patterns, make predictions, and adapt to new information, contributing to their intelligence.

5.It's like ML is the learning engine within the larger framework of AI, working together to create smart and adaptable systems.

NOTE: (Q.26,28,29,30,31,32,33,34,35,36 and 37 are repeated question…)