Machine Learning: Important Terminologies and Their Purpose

Understanding key terminologies in machine learning is crucial for grasping its concepts and implementation. Here are the important terms along with their purposes:

1. Dataset

- **Definition**: A collection of data used to train and evaluate models.
- **Purpose**: Provides input features and target labels to the model for learning patterns.

2. Features (Independent Variables)

- **Definition**: The input variables (X) used for training a model.
- **Purpose**: They represent the characteristics or attributes that influence the output.

3. Label (Target, Dependent Variable)

- **Definition**: The actual outcome or value (Y) that the model predicts.
- **Purpose**: Serves as the ground truth for model training and evaluation.

4. Model

- **Definition**: A mathematical function trained on data to make predictions.
- **Purpose**: Learns patterns from data and generalizes to make predictions on new data.

5. Training Set

- **Definition**: A subset of data used to train the model.
- **Purpose**: Enables the model to learn patterns and relationships.

6. Validation Set

- **Definition**: A subset of data used to tune hyperparameters and avoid overfitting.

- **Purpose**: Helps in model selection and fine-tuning.

7. Test Set

- **Definition**: A separate dataset used to evaluate the model's performance.
- **Purpose**: Measures how well the model generalizes to unseen data.

8. Overfitting

- **Definition**: When a model learns noise along with patterns, performing well on training data but poorly on new data.
- **Purpose**: Identifies models that have memorized rather than generalized from training data.

9. Underfitting

- **Definition**: When a model is too simple and fails to learn patterns.
- **Purpose**: Indicates the need for more complex models or additional training.

10. Bias

- **Definition**: Error due to overly simplistic assumptions in the learning algorithm.
- **Purpose**: Helps understand whether the model is underfitting the data.

11. Variance

- **Definition**: Error due to the model's sensitivity to small fluctuations in training data.
- **Purpose**: Helps identify overfitting in models.

12. Hyperparameters

- **Definition**: Configuration settings that control the learning process (e.g., learning rate, tree depth).
- **Purpose**: Tuneable parameters that optimize model performance.

13. Parameters

- **Definition**: Internal values learned by the model during training (e.g., weights in neural networks).
- **Purpose**: Define the model's structure and performance.

14. Loss Function

- **Definition**: A function that measures the error between actual and predicted values.
- **Purpose**: Helps in adjusting the model to reduce prediction errors.

15. Optimization Algorithm

- **Definition**: A method (e.g., Gradient Descent) used to minimize the loss function.
- **Purpose**: Adjusts parameters to improve model accuracy.

16. Supervised Learning

- **Definition**: A learning method where the model is trained using labeled data.
- **Purpose**: Used for classification and regression tasks.

17. Unsupervised Learning

- **Definition**: A learning method where the model is trained using unlabeled data.
- **Purpose**: Used for clustering and anomaly detection.

18. Reinforcement Learning

- **Definition**: A learning approach where an agent learns by interacting with the environment.
- **Purpose**: Used in robotics, gaming, and AI-driven decision-making.

19. Feature Scaling

- **Definition**: A technique to normalize or standardize features (e.g., Min-Max Scaling, Standardization).

- **Purpose**: Improves model performance and convergence speed.

20. Feature Engineering

- **Definition**: The process of transforming raw data into meaningful features.

- **Purpose**: Enhances the predictive power of machine learning models.

Real-Time Examples for Machine Learning Terminologies

1. Dataset

- **Example:** A bank collects customer transaction data, including account balance, credit score, and transaction history, to detect fraudulent activities.

2. Features (Independent Variables)

- **Example:** In a loan approval system, features include **age, income, credit score, and existing loans**.

3. Label (Target, Dependent Variable)

- **Example:** The target variable in a **loan approval model** is **"Loan Approved" (Yes/No).**

4. Model

- **Example:** A **spam email detection model** classifies emails as **spam or not spam** based on past email patterns.

5. Training Set

- **Example:** To build a handwriting recognition system, 80% of the collected handwritten digits dataset is used for training the model.

6. Validation Set

- **Example:** In facial recognition for an **attendance system**, a validation set helps adjust hyperparameters like the number of facial landmarks used.

7. Test Set

- **Example:** A **medical diagnosis AI model** is tested on unseen patient reports to verify its accuracy before deployment.

8. Overfitting

- **Example:** A stock price prediction model **memorizes** past trends but fails when **new economic events** occur.

9. Underfitting

- **Example:** A basic **house price prediction model** only considers area size but ignores **location, number of bedrooms, and market trends**, making poor predictions.

10. Bias

- **Example:** A facial recognition system trained **only on light-skinned faces** may fail to recognize **dark-skinned individuals**, showing bias in training data.

11. Variance

- **Example:** A medical AI trained on limited hospitals performs **very well on those hospitals** but poorly in different regions.

12. Hyperparameters

- **Example:** In a **self-driving car**, tuning the **learning rate** decides how quickly the model adjusts based on errors.

13. Parameters

- **Example:** In a **house price prediction model**, the learned coefficients (weights) for **size, number of rooms, and location** are model parameters.

14. Loss Function

- **Example:** In an **AI-powered handwriting recognizer**, the loss function calculates how different the model's predicted text is from the actual handwritten text.

15. Optimization Algorithm

- **Example:** In a **speech-to-text AI system**, **gradient descent** adjusts model parameters to improve recognition accuracy.

16. Supervised Learning

- **Example: Email spam detection** learns from past labeled emails (Spam/Not Spam) and applies that knowledge to classify new emails.

17. Unsupervised Learning

- **Example: Customer segmentation in e-commerce** groups customers into different categories (e.g., frequent buyers, occasional buyers) without predefined labels.

18. Reinforcement Learning

- **Example: Autonomous drones** learn optimal flying paths by getting rewards for avoiding obstacles and penalties for collisions.

19. Feature Scaling

- **Example:** In **credit card fraud detection**, transaction amounts may range from **₹1 to ₹10,00,000**, requiring feature scaling for better model performance.

20. Feature Engineering

- **Example:** In **weather forecasting**, raw temperature and humidity values are converted into **"feels-like temperature"**, which is more meaningful for predictions.

(PTO)

Model Evaluation performance