ML ASSIGNMENT PART – A

By Arpan Saha (B.SC. SEM-6)

1. Define the following:

a. Machine Learning :

Machine Learning (ML) is a subset of artificial intelligence (AI) that enables systems to learn patterns from data and make predictions or decisions without being explicitly programmed.

b. Supervised vs Unsupervised Learning

Supervised Learning: The model is trained on labeled data (i.e., input with corresponding output). Example: Spam detection.

Unsupervised Learning: The model is trained on unlabeled data and tries to find hidden patterns. Example: Customer segmentation.

c. Overfitting and Underfitting

Overfitting: The model learns the training data too well, including noise, and performs poorly on new data.

Underfitting: The model is too simple and fails to capture the patterns in the data.

d. Training, Validation, and Test Sets

Training Set: Used to train the model.

Validation Set: Used to tune hyperparameters and prevent overfitting.

Test Set: Used to evaluate the final model performance.

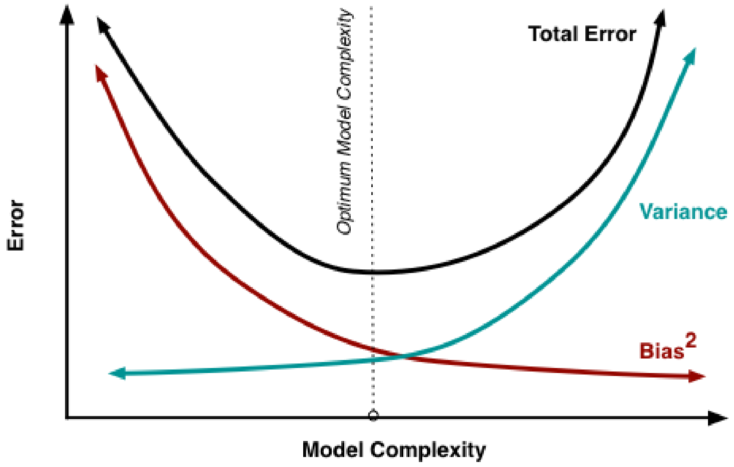
e. Cross-Validation

Cross-validation is a technique where the dataset is split into k subsets (folds). The model is trained on k-1 folds and validated on the remaining fold, repeating the process k times to ensure reliability.

2. Explain the bias-variance trade-off with a diagram and example.

ANS: Bias is error due to overly simplistic models (underfitting).

Variance is error due to overly complex models (overfitting).



Example:

- A linear model predicting house prices might **underfit** because it can't capture nonlinear trends (high bias).
- A deep neural network might **overfit** a small dataset of house prices (high variance).

3. Describe how the following models work, with advantages and limitations:

- Linear Regression:

  Linear Regression works by fitting a straight line through the data that minimizes the difference between predicted and actual values. It is simple and fast, and works well for linear relationships, but performs poorly if the relationship is non-linear or affected by outliers.

- Logistic Regression :

  Logistic Regression is used for binary classification tasks. It applies the sigmoid function to output probabilities between 0 and 1. It is effective for linearly separable data and is easy to interpret. However, it may not perform well on complex or non-linear problems without transformations.

- Decision Tree :

  Decision Tree models split the data based on feature values to form a tree-like structure where each internal node represents a decision, and each leaf node represents an outcome. They are intuitive and handle both numerical and categorical data but are prone to overfitting and can be unstable with small data changes.

- K-Nearest Neighbors (KNN) :

  K-Nearest Neighbors (KNN) is a non-parametric method where a data point is classified based on the majority class of its k closest neighbors. It is simple and requires no training phase, but is computationally expensive during prediction and sensitive to irrelevant features and feature scaling.

4. Discuss:

a. **One feature selection method (e.g., correlation, chi-square) :**

One common **feature selection method** is correlation analysis. It identifies features that have a strong relationship with the target variable and removes those that are redundant or irrelevant. This helps in simplifying the model and improving performance by reducing noise.

b. **One feature scaling technique (e.g., normalization, standardization) :**

A commonly used **feature scaling technique** is standardization. It transforms the features to have a mean of zero and a standard deviation of one. This is especially useful for algorithms like KNN and SVM that are sensitive to the scale of the input data.

5. What is regularization? How do L1 and L2 regularization reduce overfitting?

ANS: **Regularization** is a technique used to reduce overfitting by adding a penalty term to the model's loss function, discouraging it from fitting too closely to the training data.

L1 regularization, also known as Lasso, adds the absolute values of the coefficients as a penalty. It can shrink some coefficients to zero, effectively performing feature selection.

L2 regularization, or Ridge, adds the squared values of the coefficients as a penalty, which tends to shrink coefficients but not eliminate them. Both methods help in simplifying the model and improving its ability to generalize to new data.