Unit-1

Linear Methods for Regression and Classification:

- Classification is the process of predicting the class labels for the input data to categorize the data into classes.
- This can be done by building models. A classification model acts as a decision tree which separates the provided input data and gives each group of data (class) a label.
- These models are derived and can be built by providing labeled training data.

Hence, classification models are used to predict class labels for new, unlabeled data.

Ex: in email spam detection, a classification model can be trained on labeled emails (spam or not spam) to predict the class of incoming emails.

- Regression, on the other hand, is the process of predicting continuous numerical values based on the provided input features.
- It aims to find a mathematical relationship between the input variables and the output variable.

Ex: in house price prediction, regression can be used to build a model that predicts the price of a house based on factors like its area, number of rooms, location, etc.

Supervised Learning:

It is when the labeled data sets are provided as training data. The class labels for the new instances are predicted from that training labeled data.

Workflow in Supervised Learning

- **Data Collection and Preparation**: Gather and preprocess data, including labeling examples, cleaning, normalizing, and splitting into training, validation, and test sets.
- **Model Training**: Use the training set to help the model learn by iteratively adjusting parameters to minimize errors.
- **Model Evaluation**: Assess model performance on a validation set to fine-tune hyperparameters and on a test set to evaluate generalization.
- **Deployment**: Once validated, the model is deployed for real-world use, with regular monitoring for accuracy.

Common Algorithms

- **Linear Regression**: For regression tasks, fits a linear relationship between input variables and the output.
- **Logistic Regression**: For binary classification tasks, estimates probabilities to classify inputs into two categories.

- **Decision Trees and Random Forests**: Nonlinear models that are highly interpretable and suitable for both classification and regression.
- **Support Vector Machines (SVM)**: Powerful for classification by finding the optimal boundary between classes.
- **K-Nearest Neighbors (KNN)**: Classifies data points based on the majority class among the closest neighbors.
- **Neural Networks**: Highly flexible, able to approximate complex functions, suitable for tasks requiring high-dimensional data, such as image recognition.