07 Classification & Logistic Regression

Classification

- Classification is a type of supervised learning in machine learning where the goal is to predict the category or class of a given data point
- It involves training a model on a labeled dataset, where each data point has a known class label, and then using this model to classify new, unseen data

There are various types of classification problems

- **Binary Classification** : Involves two classes (e.g., spam or not spam)
- **Multiclass Classification** : Involves more than two classes (e.g., categorizing animals into cats, dogs, birds, etc.)
- **Multilabel Classification** : Each instance can belong to multiple classes simultaneously (e.g., tagging a movie with genres like action, comedy, drama)

Logistic Regression

- Logistic regression is a popular algorithm used for binary classification problems
- Despite its name, it's actually a regression model, but it is used for classification tasks
- The main idea is to model the probability that a given input belongs to a particular class
- **Interpretability**
  - The model is easy to understand and interpret, as it provides probabilities rather than just a binary decision
- **Efficiency**
  - Logistic regression is computationally efficient and works well for small to medium-sized datasets
- **Linearity**
  - It assumes a linear relationship between the input features and the log-odds of the output, which makes it effective for problems where this assumption holds true

Working of Logistic Regression

1. **Linear Combination**

   - Logistic regression calculates a weighted sum of the input features plus a bias (intercept). This is similar to linear regression

$$z = w_1 x_1 + w_2 x_2 + \cdots + w_n x_n + b$$

2. **Sigmoid Function**

- To convert the linear combination z into a probability, logistic regression applies the sigmoid function
- The sigmoid function maps any real-valued number to a value between 0 and 1, which can be interpreted as the probability of the instance belonging to the positive class

$$\sigma(z) = \frac{1}{1+e^{-z}}$$

3. **Decision Boundary**

- A threshold (commonly 0.5) is applied to the probability to classify the instance into one of the two classes
- If the probability is greater than the threshold, the instance is classified as the positive class; otherwise, it's classified as the negative class