4 Supervised Learning : Classification and Regression

4.1 Introduction to Supervised Learning

Supervised learning is a machine learning technique that involves training a model to make predictions based on labeled data. In supervised learning, the model learns from a dataset that includes both input features and corresponding output labels. The goal of supervised learning is to develop a model that can accurately predict the output for new input data that it has not seen before.

The labeled dataset used for supervised learning typically consists of a set of input features and corresponding output labels. The input features can be numerical or categorical, and the output labels can be binary (e.g. true or false) or multi-class (e.g. classification of different objects).

The process of supervised learning involves training a model using the labeled dataset, which involves finding the relationship between the input features and output labels. Once the model is trained, it can be used to predict the output for new input data that it has not seen before.

Supervised learning is commonly used in applications such as image classification, text classification, fraud detection, and recommendation systems. The key advantage of supervised learning is that it can produce highly accurate predictions when trained on a large and diverse dataset. However, it requires a labeled dataset, which can be costly and time-consuming to create.

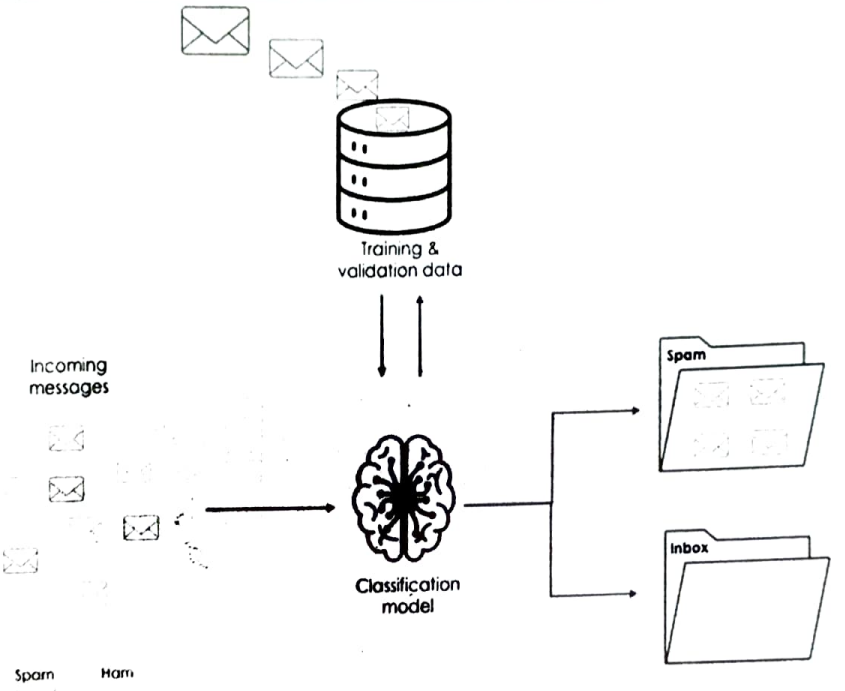
4.2 Classification Model

Classification is a type of supervised learning where the goal is to predict the categorical output variable for a given set of input features. In classification, the model learns to distinguish between different classes based on labeled data.

Classification is a type of supervised learning where the goal is to predict the categorical output variable for a given set of input features. In classification, the model learns to distinguish between different classes based on labeled data.

For example, in a binary classification problem, the goal is to predict one of two possible classes, such as "spam" or "not spam". In multi-class classification, the goal is to predict one of several possible classes, such as identifying different types of animals from an image.

To build a classification model, the first step is to collect labeled data. This involves creating a dataset where each data point includes a set of input features and a corresponding output label. The input features can be anything that might be relevant to the classification task, such as pixel values for image data or demographic data for predicting customer behavior.



Classification Model

Once the dataset is prepared, the next step is to choose an appropriate classification algorithm, such as logistic regression, decision trees, or neural networks. The algorithm is then trained on the labeled dataset to learn the relationship between the input features and the output labels.

During training, the algorithm adjusts its parameters to minimize the difference between the predicted output and the actual output in the labeled dataset. Once the model is trained, it can be used to predict the output label for new input data.

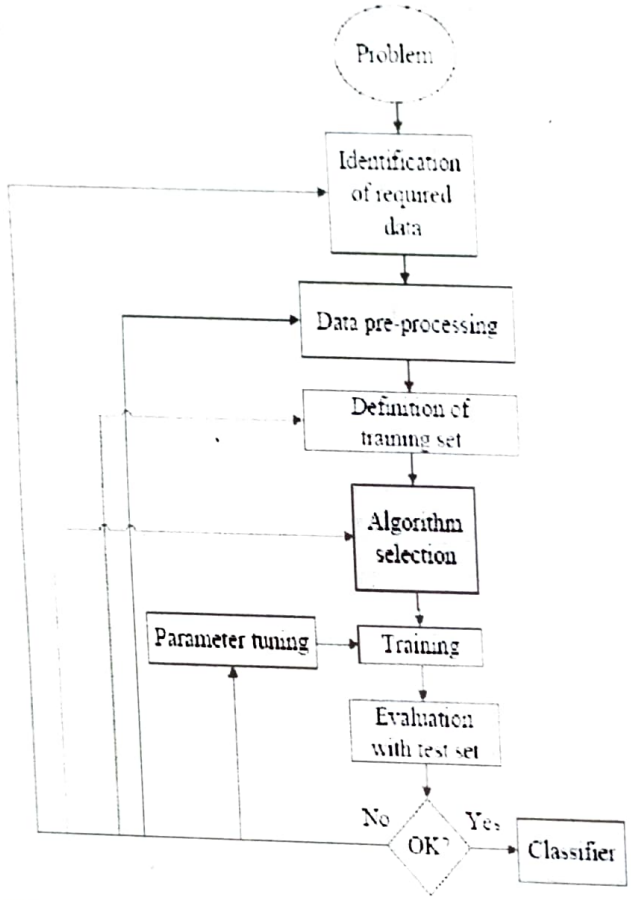
In summary, classification is a type of supervised learning where a target feature, which is of categorical type, is predicted for test data on the basis of the information imparted by the training data. The target categorical feature is known as class.

Evaluation of a classification model

involves measuring its accuracy on a held-out validation dataset or through cross-validation techniques. The accuracy of the model can be improved by using techniques such as feature engineering, regularization, and ensembling.

4.3 Learning Steps of Classification

Step 1: Problem Identification: Identifying the problem is the first step in the supervised learning model. The problem



must be well-posed problem.

Step 2: Identification of Required Data: Collecting and preparing a labeled dataset that includes input features and corresponding output labels.

Step 3: Data Preprocessing: Cleaning and preparing dataset, which may involve removing missing data, handling outliers and transferring the data into a suitable format.

Step 4: Definition of Training Data Set: Before starting the analysis, the user should decide type of data set is to be used as a training set. The training set needs to be actively representative of the real-world use of the given problem.

Step 5: Algorithm Selection: Choosing an appropriate classification algorithm based on the nature of the problem and the size and complexity of the dataset.

Step 6: Training: Using the labeled dataset to train the selected algorithm by adjusting its parameters to minimize the difference between the predicted output and the actual output. Also, adjusting the hyperparameters of the algorithm to optimize the model's performance on the validation dataset should be done, if it needs..

Step 7: Evaluation with the test dataset: Training data is run on the algorithm, and its performance is measured here. If a desired output is not obtained, again training of parameters may be required.

4.4 Classification Algorithms

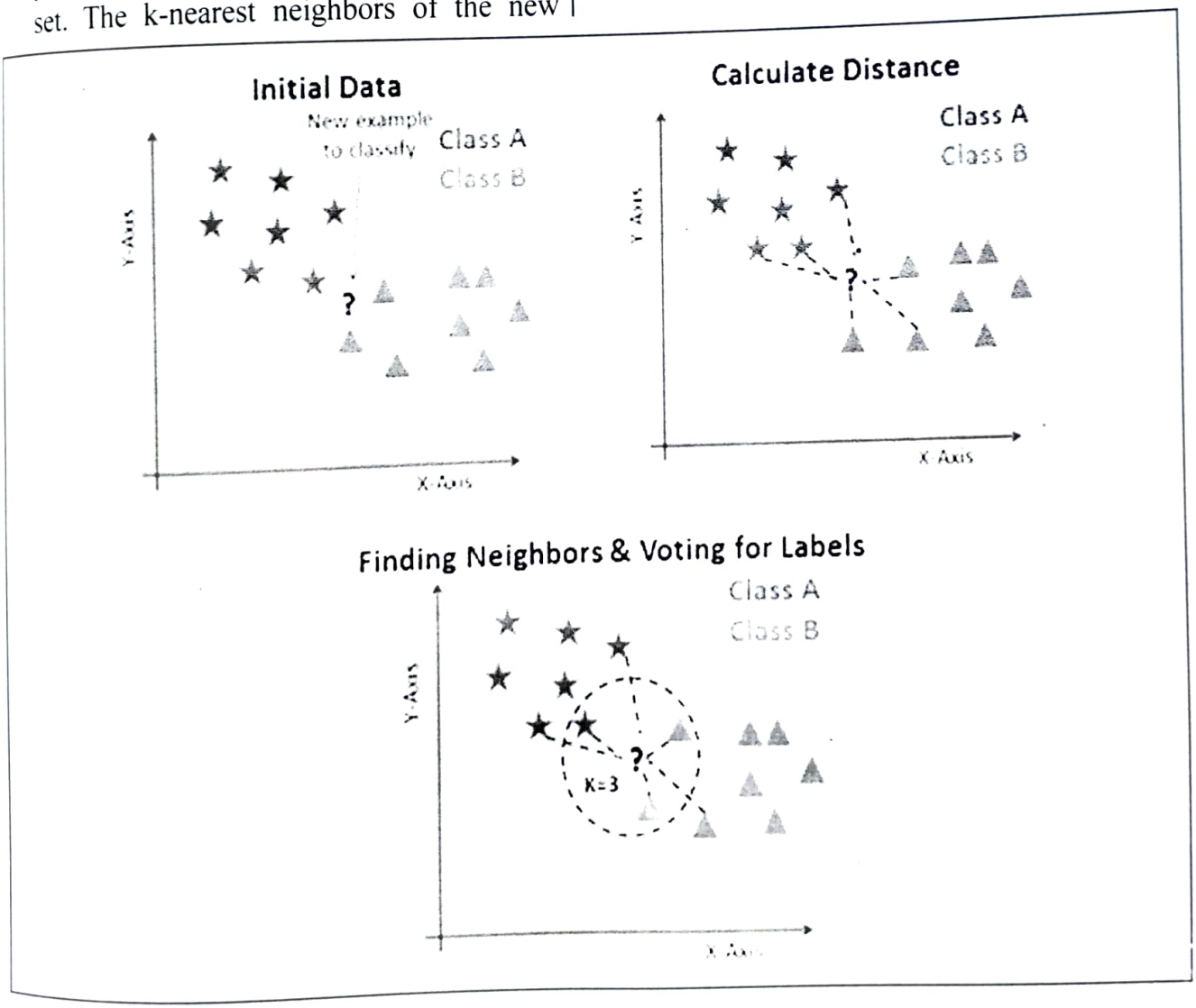
4.4.1 k-Nearest Neighbour Algorithm

The k-Nearest Neighbor (k-NN) algorithm is a simple, non-parametric algorithm used in supervised learning for classification and regression. In k-NN, the output is classified by a majority vote of its k nearest neighbors in the training dataset.

The algorithm works by first calculating the distances between the new input data point and all the data points in the training set. The k-nearest neighbors of the new data point are then determined based on the smallest distances.

In classification, the output label of the new data point is then determined by the majority class among its k-nearest neighbors. For regression, the output value is the average of the output values of its k-nearest neighbors.

The choice of k is an important parameter in the k-NN algorithm, and it can be determined by cross-validation techniques. If k is too small, the algorithm may be sensitive to noise and outliers, while if k is too large; the algorithm may over-generalize and not capture the underlying patterns in the data.



The k-NN algorithm is a simple and easy-to-understand algorithm, but it can be computationally expensive for large datasets. It also suffers from the curse of dimensionality, where the performance of the algorithm deteriorates as the number of input features increases.

Despite its limitations, the k-NN algorithm is still widely used for its simplicity and effectiveness in many applications, including image recognition, recommender systems, and anomaly detection.

For kNN, there is no learning happening in the real sense. Therefore, kNN falls under the category of lazy learner.

Here is the general k-Nearest Neighbor (k-NN) algorithm for classification:

Input :

A training dataset with labeled examples (X, y)

A new input data point x

The number of neighbours k

Output :

A predicted label for x

Algorithm:

1. Compute the distance between the new input data point x and all data points in the training set X.
2. Select the k-nearest neighbors of x based on the smallest distances.
3. Determine the majority class among the k-nearest neighbours.
4. Assign the predicted label to x based on the majority class.

Note: In step 1, generally Euclidean Distance is being used to calculate the distance. The formula for it is given below:

$$d(x, y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$$

Advantages of k-NN classification **algorithm :**

- Simple and easy to understand.
- Non-parametric method does not require any assumptions about the underlying data distribution.
- Can be used for both binary and multi-class classification problems.
- Can handle both continuous and categorical data types.
- Can work well with small datasets.

Disadvantages of k-NN classification algorithm:

- Computationally expensive for large datasets.
- Suffers from the curse of dimensionality where the performance deteriorates as the number of input features increases.
- Highly sensitive to the distance metric used, which can have a significant impact on the results.
- Requires careful selection of the value of k, which can affect the accuracy of the model.

Applications of k-NN classification algorithm:

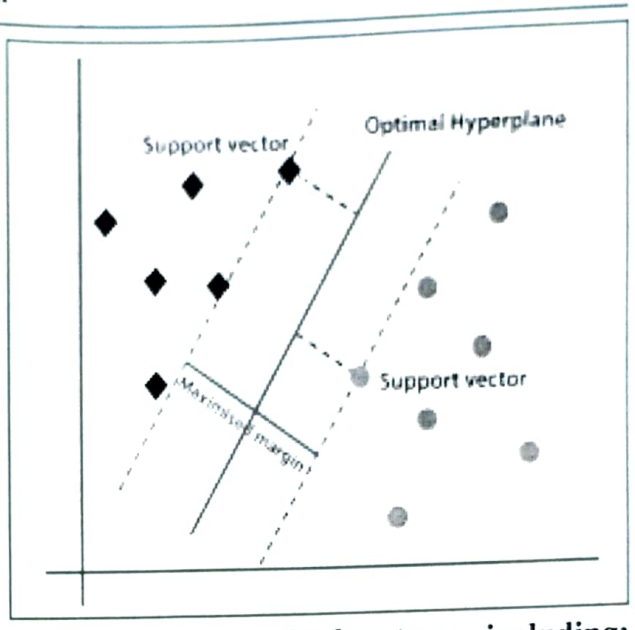
- Image recognition and classification.
- Recommendation systems for online shopping or entertainment.
- Credit risk assessment in finance.
- Medical diagnosis based on patient characteristics.
- Fraud detection in credit card transactions.

4.4.2 Support Vector Machines (SVM)

Support Vector Machine (SVM) is a powerful classification algorithm used in supervised learning. The algorithm works by finding the optimal hyperplane that separates the different classes in the input data.

Support vectors are the data points (representing classes), the critical component in a data set, which are near the identified set of lines (**hyperplane**). In the case of a binary classification problem, SVM aims to find the hyperplane that maximizes the margin between the two classes. The **margin** is defined as the distance between the hyperplane and the closest data points from each class. SVM finds the optimal hyperplane by solving a convex optimization problem that maximizes the margin while minimizing the classification error.

If the input data is not linearly separable, SVM can still be used by applying a kernel function that maps the data into a higher-dimensional feature space, where a linear hyperplane can be used to separate the classes. Some popular kernel functions used in SVM include linear, polynomial, radial basis function (RBF), and sigmoid.



SVM has several advantages, including:

- Effective for high-dimensional data.
- Can handle non-linearly separable data by applying kernel functions.
- Works well with small to medium-sized datasets.
- Robust to outliers.

However, SVM also has some limitations, including :

- SVM is applicable only for binary classification.
- Computationally expensive for large datasets.
- Sensitive to the choice of kernel function and parameters.
- Difficult to interpret the resulting model and extract insights from it.
- Overfitting may cause.

Applications of SVM include :

- Image classification and object recognition.
- Text classification and sentiment analysis.
- Bioinformatics and gene expression analysis.
- Financial forecasting and risk management.

- Anomaly detection in network intrusion detection systems.

4.5 Regression

Regression in supervised learning is a type of predictive modeling that involves estimating a continuous numerical output value based on input features. The goal of regression is to find a mathematical relationship between the input variables (also called predictors or independent variables) and the output variable (also called the response or dependent variable).

The output value in regression can be any numerical value, including integers, real numbers, or even negative values. Regression models can be used for both linear and nonlinear relationships between the input and output variables.

The main goal of regression is to create a function that can predict the output variable for new inputs with a certain degree of accuracy. In order to achieve this goal, regression algorithms use statistical techniques to find the optimal parameters of the model that minimize the difference between the predicted values and the actual values in the training dataset.

Applications of Regression :

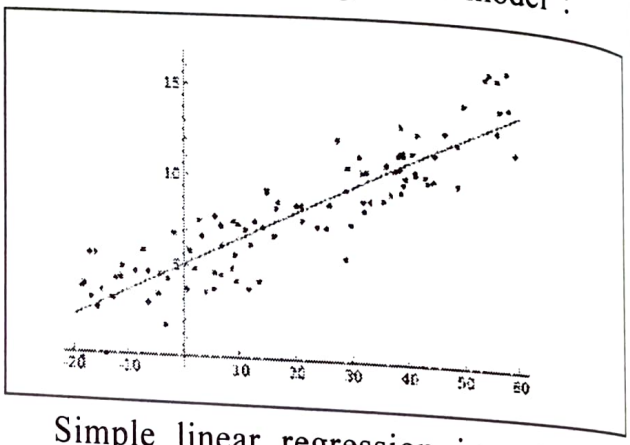
- Predicting sales or revenue based on marketing spend or other factors.
- Estimating the price of a house based on features such as location, size, and number of rooms.
- Forecasting demand for a product based on historical sales data.
- Predicting patient outcomes based on medical data.
- Analyzing trends in financial data such as stock prices or interest rates.

4.5.1 Simple Linear Regression

Linear regression is one of the most basic types of regression in machine learning. The linear regression model consists of a predictor variable and a dependent variable related linearly to each other. In case the data involves more than one independent variable, then linear regression is called multiple linear regression models.

$$y = mx + c + e$$

The below-given equation is used to denote the linear regression model :



Simple linear regression is a type of regression analysis that is used to model the relationship between two variables, where one variable is considered the predictor or independent variable, and the other is the response or dependent variable. The goal of simple linear regression is to find the best linear relationship between the two variables.

Here is an example of simple linear regression :

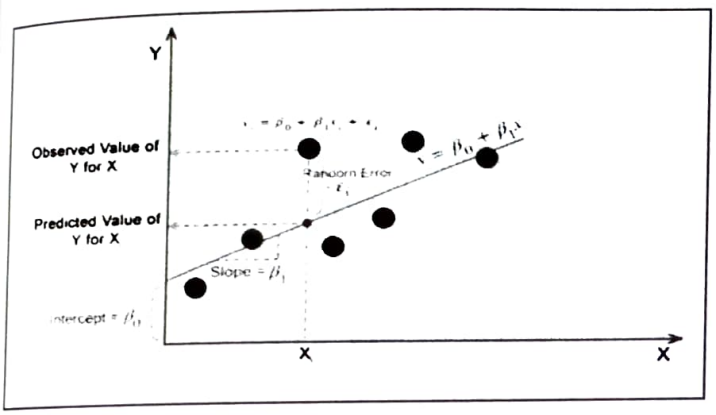
Suppose we have a dataset of 100 homes with two variables: square footage (independent variable) and sale price (dependent variable). We want to determine the relationship between these two variables, and specifically, how much the sale price of a home increases for each additional square foot of living space.

We can use simple linear regression to estimate the line of best fit for the data, which is a straight line that minimizes the sum of the squared differences between the predicted and actual values.

In simple linear regression, the model is expressed as:

$$Y = \beta 0 + \beta 1X + \varepsilon$$

where Y is the dependent variable, X is the independent variable, β0 is the intercept, β1 is the slope associated with the predictor variable X, and å is the random error term. The slope β1 indicates the change in the dependent variable associated with a one-unit change in the independent variable X.



The simple linear regression model can be fitted to the data using various techniques such as ordinary least squares (OLS) or maximum likelihood estimation (MLE). The fitted model can be used to make predictions for new values of the predictor variable or to test hypotheses about the relationship between the variables.

The validity of the simple linear regression model depends on certain assumptions such as linearity, normality, homoscedasticity, and independence of the error term. Violations of these assumptions can lead to biased and inefficient estimates and incorrect inference.

To estimate the values of β0 and β1, we can use the least squares method, which involves minimizing the sum of the squared errors between the predicted and actual values of the dependent variable. The equation for the least squares estimates of β0 and β1 are:

$$\beta 1 = (X_i - x\_mean)(Y_i - y\_mean) / (X_i - x\_mean)^2$$

$$\beta 0 = y\_mean - \beta 1 * x\_mean$$

Where xi and yi are the values of the independent and dependent variables for the ith observation, x_mean and y_mean are the mean values of the independent and dependent variables, and the sums are taken over all observations in the dataset.

Once we have estimated the values of β0 and β1, we can use the equation of the line to predict the sale price of a home based on its square footage.

For example, if the estimated equation of the line is:

$$y = 100000 + 100x$$

This means that for every additional square foot of living space, the sale price of the home is expected to increase by Rs. 100. So, a home with 1500 square feet of living space would be expected to sell for Rs. 250,000, (100000 + 100*1500).

Pros :

- Simple linear regression is easy to implement and interpret.
- It provides a simple way to quantify the strength and direction of the relationship between two variables.
- It can be used to make predictions and estimate the value of the response variable for a given value of the predictor variable.

Cons:

- Simple linear regression assumes that there is a linear relationship between the predictor and response variables, which may not be the case in reality.
- It assumes that the errors or residuals are normally distributed and have constant variance, which may not hold in practice.
- It can be sensitive to outliers, influential observations, and violations of the underlying assumptions.

Appliactions :

- Simple linear regression is commonly used in finance to model the relationship between two financial variables, such as stock prices and interest rates.
- It is used in marketing to analyze the impact of advertising on sales or the relationship between price and demand for a product.
- It is used in engineering to model the relationship between two physical variables, such as temperature and pressure.
- It is used in social sciences to study the relationship between two psychological variables, such as personality and job satisfaction.

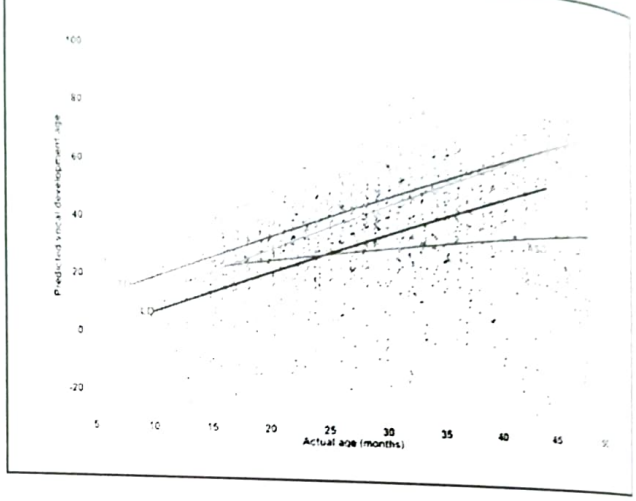
4.5.2 Multiple Linear Regression

Multiple linear regression is a statistical technique used to model the relationship between a response or dependent variable and two or more predictor or independent variables. It extends the simple linear regression model to multiple predictors, allowing for more complex relationships between the variables.

In multiple linear regression, the model is expressed as:

$$Y = \beta 0 + \beta 1X1 + \beta 2X2 + \ldots + \beta nXn + e$$

where Y is the dependent variable, $\beta 0$ is the intercept, $\beta 1$ to βn are the coefficients or slopes associated with the predictor variables X1 to Xn, and e is the random error term. The coefficients indicate the change in the dependent variable associated with a one-unit change in the corresponding predictor variable, holding other variables constant.



The multiple linear regression model can be fitted to the data using various techniques such as ordinary least squares (OLS) or maximum likelihood estimation (MLE). The fitted model can be used to make predictions for new values of the predictor variables or to test hypotheses about the relationships between the variables.

The validity of the multiple linear regression model depends on certain assumptions such as linearity, normality, homoscedasticity, and independence of the error terms. Violations of these assumptions can lead to biased and inefficient estimates and incorrect inference.

Pros :

- Multiple linear regression can capture the effects of multiple predictors on the response variable simultaneously, allowing for a more comprehensive analysis of the relationships between the variables.
- It can provide insight into the relative importance of different predictors in explaining the variability in the response variable.
- It can be used to make predictions and estimate the value of the response variable for a given set of predictor variable values.
- It can handle categorical predictor variables by encoding them as dummy variables.

Cons :

- Multiple linear regression assumes that there is a linear relationship between the predictors and the response variable, which may not be the case in reality.
- It assumes that the errors or residuals are normally distributed and have constant variance, which may not hold in practice.
- It can be sensitive to outliers, influential observations, and violations of the underlying assumptions.
- As the number of predictor variables increases, the model can become overfit and less interpretable.

Applications :

- Multiple linear regression is commonly used in economics to model the relationship between a dependent variable such as income, and multiple independent variables such as education, age, and experience.
- It is used in healthcare to study the relationship between health outcomes and multiple risk factors such as smoking, diet, and exercise.
- It is used in marketing to analyze the impact of multiple advertising channels on sales or the relationship between price, promotions, and demand for a product.

Comparison between multiple linear regression and simple linear regression:

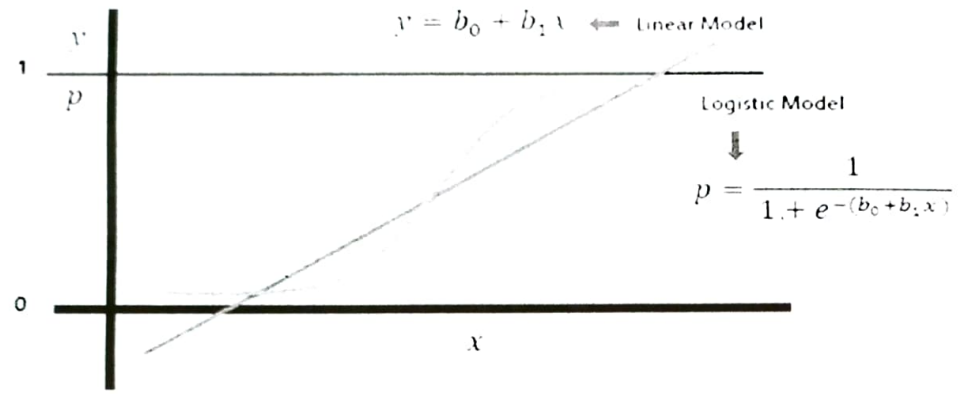
- Simple linear regression is a special case of multiple linear regression where there is only one predictor variable.
- Multiple linear regression can provide more insight into the relationships between the variables and can handle more complex scenarios where there are multiple predictors.
- However, simple linear regression can be easier to interpret and implement, and may be sufficient for simpler problems where there is only one predictor variable.

4.5.3 Logistic Regression

Logistic regression is a statistical technique used to model the relationship between a binary dependent or response variable and one or more independent or predictor variables. The goal of logistic regression is to estimate the probability of the dependent variable taking a particular value (e.g., 1 or 0) given the values of the predictor variables.

The logistic regression model assumes that the relationship between the predictor variables and the log-odds of the dependent variable is linear. The model

can be fitted to the data using various techniques such as maximum likelihood estimation (MLE) or Bayesian methods. The fitted model can be used to make predictions for new values of the predictor variables or to test hypotheses about the relationship between the variables.



In above formula of the logistic regression the constant (b0) moves the curve left and right and the slope (b1) defines the steepness of the curve.

Types of Logistic Regression:

1. Binary Logistic Regression

   The categorical response has only two 2 possible outcomes. Example: Spam or Not.

2. Multinomial Logistic Regression:

   Three or more categories without ordering. Example: Predicting which food is preferred more (Veg, Non-Veg, Vegan).

3. Ordinal Logistic Regression

   Three or more categories with ordering. Example: Movie rating from 1 to 5

Pros of logistic regression :

- Logistic regression is a powerful and widely used statistical technique for binary and ordinal classification problems, and can handle multiple predictor variables.
- The output of logistic regression is easy to interpret as it provides the predicted probabilities of the dependent variable taking a particular value.
- Logistic regression can handle both categorical and continuous predictor variables and can detect interactions between them.
- Logistic regression can be used to test hypotheses about the relationships between the variables and to estimate the impact of the predictor variables on the dependent variable.

Cons of logistic regression :

- Logistic regression assumes that the relationship between the predictor variables and the dependent variable is linear, which may not hold in all cases.
- Logistic regression assumes that the observations are independent, which may not hold in cases where the data are clustered or correlated.
- Logistic regression can be sensitive to outliers and influential observations.

which may affect the estimated coefficients and the goodness-of-fit of the model.

- Logistic regression may suffer from overfitting or underfitting if the model is too complex or too simple, respectively.

Applications of logistic regression:

- Logistic regression is widely used in various fields such as healthcare, marketing, finance, and social sciences for binary and ordinal classification problems such as disease diagnosis, customer churn, credit scoring, and voter prediction.
- Logistic regression is used in biomedical research to model the probability of a disease occurring based on risk factors such as age, sex, and lifestyle factors.
- Logistic regression is used in social sciences to model the likelihood of a particular behavior or attitude given demographic and psychographic characteristics of the individuals.
- Logistic regression is used in marketing to model the likelihood of a customer buying a product given their past behavior and demographic characteristics.

Classification vs. Regression :

Aspect	Classification	Regression
Goal	To predict the categorical outcome	To predict the numerical outcome
Output	Discrete values (classes/labels)	Continuous values (numbers)
Examples	Image classification, spam detection	Stock price prediction, housing prices
Evaluation	Accuracy, precision, recall	Mean squared error, root mean squared error
Algorithms	Logistic regression, decision trees	Linear regression, random forest
Problem types	Binary classification, multi-class	Linear regression, non-linear regression

Question Bank

■ Short-Answer Questions (3 or 4 marks) :

1. Give any three examples of supervised learning in Industry 4.0
2. Give any three examples of supervised learning in the field of healthcare.
3. Define following terms: Support Vectors, Hyperplane, Margin
4. Explain classification model in brief.
5. Give the difference between classification and regression.
6. Draw the flowchart which shows the classification learning process.
7. Explain different types of logistic regression.

8. Compare and contrast Single linear regression and multiple linear regression.
9. List applications of SVM algorithm.
10. State advantages and disadvantages of k-NN algorithm.

■ **Long-Answer Questions** (7 marks)

1. Define Classification. Explain classification learning steps in detail.
2. Write and discuss k-NN Algorithm.
3. Discuss the SVM model in detail with its pros and cons.
4. Explain logistic regression with advantage and disadvantage.
5. Write a short note on Single Linear Regression. Also, state applications of it.
6. Write a short note on Multiple Linear Regression. Also, state applications of it.
7. Explain any three applications of classification in detail.

■ **Check your knowledge :**

1. Which of the following is an example of a regression problem ?
   a) Predicting whether a customer will buy a product or not
   b) Estimating the age of a person based on their photograph
   c) Predicting the stock price of a company
   d) All of above
2. What is the main difference between classification and regression ?
   a) Classification predicts discrete outcomes while regression predicts continuous outcomes
   b) Classification predicts continuous outcomes while regression predicts discrete outcomes
   c) There is no difference between classification and regression
   d) Both are same.
3. Which of the following evaluation metrics is commonly used for regression problems?
   a) Accuracy
   b) Precision
   c) Mean squared error
   d) Specificity

   Ans.: c) Mean squared error
4. What is the goal of logistic regression?
   a) To estimate the mean of the dependent variable.
   b) To predict the dependent variable from the independent variables.
   c) To estimate the probability of the dependent variable taking a particular value given the independent variables.
   d) To test the significance of the independent variables.
5. What is the meaning of the parameter k in the kNN classification algorithm?
   a) The number of predictor variables in the model.
   b) The number of training observations used to fit the model.
   c) The number of nearest neighbors to consider when classifying a new observation.
   d) The level of significance for the hypothesis test of the model.

■