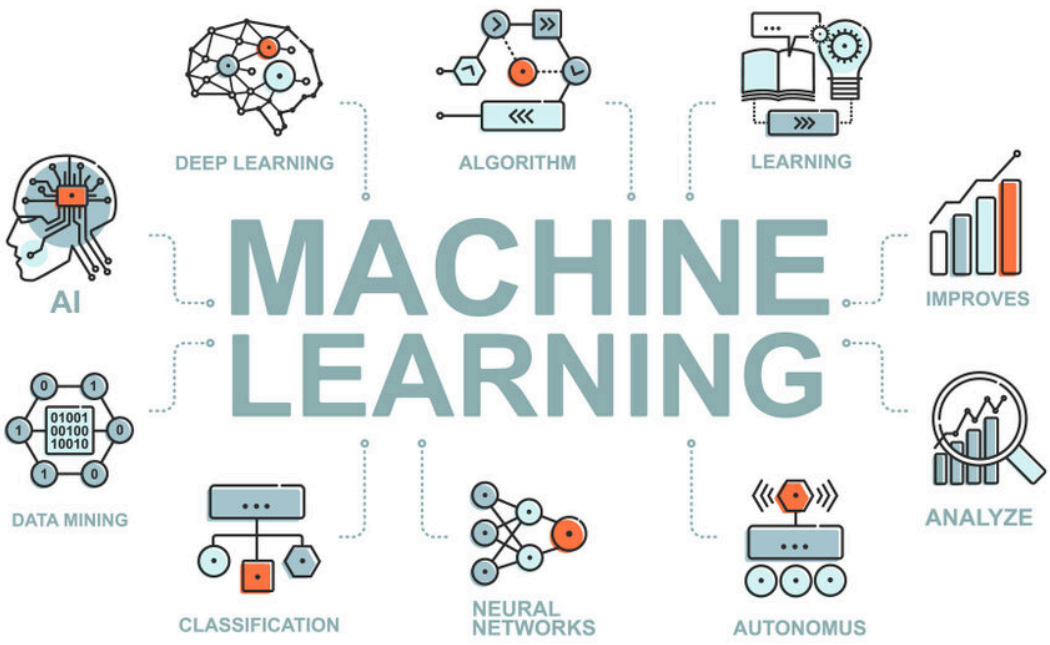


Project Report

—

Group Members

Dheeraj Kumar

Luv Valecha

Dhruv Sharma

Adithya Subhash

Overview

In this project, we explore various machine learning algorithms, including linear regression, logistic regression, and KNN. Additionally, we discuss the importance of data scaling and its impact on model performance.

Goals

1. Implement and compare linear regression, logistic regression, and KNN models.
2. Evaluate model performance using appropriate metrics.
3. Understand the significance of data scaling in machine learning

Specifications

1. Dataset

Dataset is taken from the csv file. It includes both training and test files. File names are taken from the user as input.

2. Algorithms

a. Linear Regression

Linear regression is also a type of machine-learning algorithm, more specifically a supervised machine-learning algorithm that learns from the labeled datasets and maps the data points to the most optimized linear functions. which can be used for prediction on new datasets.

Simple Linear Regression is This is the simplest form of linear regression, and it involves only one independent variable and one dependent variable. The equation for simple linear regression is:

$y = \beta 0 + \beta 1 X$

where:

Y is the dependent variable

X is the independent variable

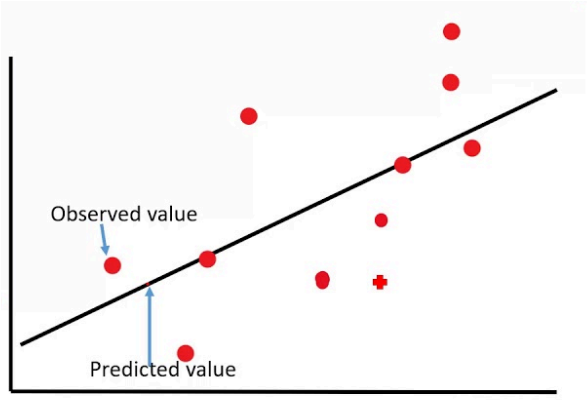
β0 is the intercept

β1 is the slope

Best Fit line:

Our primary objective while using linear regression is to locate the best-fit line, which implies that the error between the predicted and actual values should be kept to a minimum. There will be the least error in the best-fit line.

The best Fit Line equation provides a straight line that represents the relationship between the dependent and independent variables. The slope of the line indicates how much the dependent variable changes for a unit change in the independent variable(s).

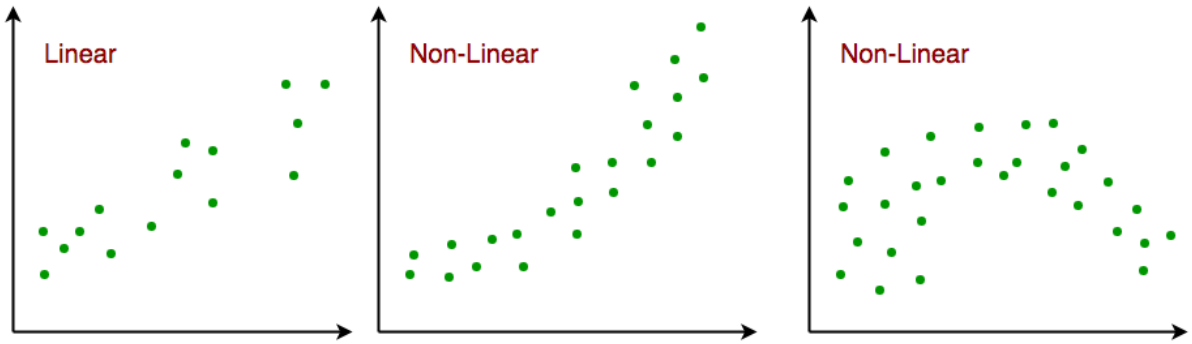


Assumptions:

Linear regression is a powerful tool for understanding and predicting the behavior of a variable, however, it needs to meet a few conditions in order to be accurate and dependable solutions.

Linearity: The independent and dependent variables have a linear relationship with one another. This implies that changes in the dependent variable follow those in the independent variable(s) in a linear fashion. This means that there should be a straight line that can be drawn through the data points. If the relationship is not linear, then linear regression will not be an accurate model.

Independence: The observations in the dataset are independent of each other. This means that the value of the dependent variable for one observation does not depend on the value of the dependent variable for another observation. If the observations are not independent, then linear regression will not be an accurate model.



Cost Function:

The cost function or the loss function is nothing but the error or difference between the predicted value and the true value.

Evaluation Metrics:

1. Mean Squared Error:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2$$

2. Mean Absolute Error:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|\hat{y}_i - y_i|$$

3. Root Mean Squared Error:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}$$

b. Logistic Regression

Logistic regression is a supervised machine learning algorithm used for classification tasks where the goal is to predict the probability that an instance belongs to a given class or not. Logistic regression is a statistical algorithm which analyzes the relationship between two data factors. Logistic regression is used for binary classification where we use sigmoid function, that takes input as independent variables and produces a probability value between 0 and 1

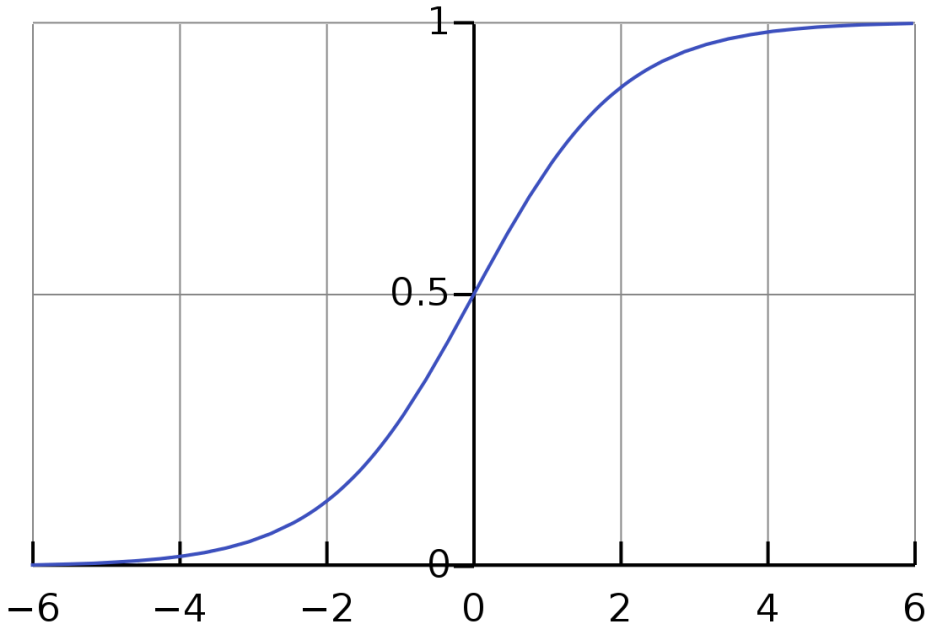
Sigmoid Function:

The sigmoid function is a mathematical function used to map the predicted values to probabilities.

It maps any real value into another value within a range of 0 and 1. The value of the logistic regression must be between 0 and 1, which cannot go beyond this limit, so it forms a curve like the “S” form.

The S-form curve is called the Sigmoid function or the logistic function.

In logistic regression, we use the concept of the threshold value, which defines the probability of either 0 or 1. Such as values above the threshold value tends to 1, and a value below the threshold values tends to 0.



Key Points:

Logistic regression predicts the output of a categorical dependent variable. Therefore, the outcome must be a categorical or discrete value.

It can be either Yes or No, 0 or 1, True or False, etc. but instead of giving the exact value as 0 and 1, it gives the probabilistic values which lie between 0 and 1.

In Logistic regression, instead of fitting a regression line, we fit an "S" shaped logistic function, which predicts two maximum values (0 or 1)

Assumptions:

Independent observations: Each observation is independent of the other. meaning there is no correlation between any input variables.

Binary dependent variables: It takes the assumption that the dependent variable must be binary or dichotomous, meaning it can take only two values. For more than two categories SoftMax functions are used.

Linearity relationship between independent variables and log odds: The relationship between the independent variables and the log odds of the dependent variable should be linear.

No outliers: There should be no outliers in the dataset.

Large sample size: The sample size is sufficiently large

Cost Function:

$$J = -\sum_{i=1}^{m} y_i \log\left(h_\theta\left(x_i\right)\right) + \left(1 - y_i\right) \log\left(1 - h_\theta\left(x_i\right)\right)$$

where,

m is the number of training examples

y_i is the true class label for the i-th example (either 0 or 1).

$h_\theta(x_i)$ is the predicted probability for the i-th example, as calculated by the logistic regression model.

θ is the model parameters

c. K-Nearest Neighbors (KNN)

(K-NN) algorithm is a versatile and widely used machine learning algorithm that is primarily used for its simplicity and ease of implementation. It does not require any assumptions about the underlying data distribution. It is a non-parametric method that makes predictions based on the similarity of data points in a given dataset. K-NN is less sensitive to outliers compared to other algorithms.

The K-NN algorithm works by finding the K nearest neighbors to a given data point based on a distance metric, such as Euclidean distance. The class or value of the data point is then determined by the majority vote or average of the K neighbors. This approach allows the algorithm to adapt to different patterns and make predictions based on the local structure of the data.

Euclidean Distance

This is nothing but the cartesian distance between the two points which are in the plane/hyperplane. Euclidean distance can also be visualized as the length of the straight line that joins the two points which are into consideration. This metric helps us calculate the net displacement done between the two states of an object.

$$\text{distance}(x, X_i) = \sqrt{\sum_{j=1}^{d} (x_j - X_{i_j})^2]}$$

How to choose the value of k for KNN Algorithm?

The value of k is very crucial in the KNN algorithm to define the number of neighbors in the algorithm. The value of k in the k-nearest neighbors (k-NN) algorithm should be chosen based on the input data. If the input data has more outliers or noise, a higher value of k would be better. It is recommended to choose an odd value for k to avoid ties in classification. Cross-validation methods can help in selecting the best k value for the given dataset.

Working Steps:

Step 1: Selecting the optimal value of K

K represents the number of nearest neighbors that needs to be considered while making a prediction.

Step 2: Calculating distance

To measure the similarity between target and training data points, Euclidean distance is used. Distance is calculated between each of the data points in the dataset and target point.

Step 3: Finding Nearest Neighbors

The k data points with the smallest distances to the target point are the nearest neighbors.

Step 4: Voting for Classification or Taking Average for Regression

The class labels are determined by performing majority voting. The class with the most occurrences among the neighbors becomes the predicted class for the target data point.

Advantages of the KNN Algorithm:

Easy to implement as the complexity of the algorithm is not that high.

Adapts Easily – As per the working of the KNN algorithm it stores all the data in memory storage and hence whenever a new example or data point is added then the algorithm adjusts itself as per that new example and has its contribution to the future predictions as well.

Few Hyperparameters – The only parameters which are required in the training of a KNN algorithm are the value of k and the choice of the distance metric which we would like to choose from our evaluation metric.

Disadvantages of the KNN Algorithm:

Does not scale – As we have heard about this, the KNN algorithm is also considered a Lazy Algorithm. The main significance of this term is that this takes lots of computing power as well as data storage. This makes this algorithm both time-consuming and resource exhausting.

Curse of Dimensionality – There is a term known as the peaking phenomenon according to this the KNN algorithm is affected by the curse of dimensionality which implies the algorithm faces a hard time classifying the data points properly when the dimensionality is too high.

Prone to Overfitting – As the algorithm is affected due to the curse of dimensionality it is prone to the problem of overfitting as well. Hence generally feature selection as well as dimensionality reduction techniques are applied to deal with this problem.

3. Data Scaling

In the machine learning process, data scaling falls under data preprocessing, or feature engineering. Scaling your data before using it for model building can accomplish the following:

Scaling ensures that features have values in the same range

Scaling ensures that the features used in model building are dimensionless

Scaling can be used for detecting outliers

Standard Scaler:

StandardScaler follows Standard Normal Distribution (SND). Therefore, it makes mean = 0 and scales the data to unit variance.

$$X_{std}^{(i)} = \frac{X^{(i)} - \bar{X}}{\sigma X}$$

Min Max Scaler:

MinMaxScaler scales all the data features in the range [0, 1] or else in the range [-1, 1] if there are negative values in the dataset.

$$X_{norm}^{(i)} = \frac{X^{(i)} - X_{min}}{X_{max} - X_{min}}$$

GNU Plot:

It is a tool used for plotting and printing the graph which is used in the linear regression function.

4. Results:

After training the model, it asks if you want to print the predicted values, also calculating the evaluation metrics. Additionally, in the linear regression, an option for printing the graph is also available.

Different Files in our project:

b23cs1016_b23cm1022_b23ci1016_b23mt1004_machine_learning.h : Common header file for all the functions, to be imported in the main file.

b23cs1016_b23cm1022_b23ci1016_b23mt1004_main.c : Main file where the main menu for the user is available and calls for all the functions as per user input.

b23cs1016_b23cm1022_b23ci1016_b23mt1004_models.c : This c file contains the code for all the machine learning models.

b23cs1016_b23cm1022_b23ci1016_b23mt1004_model_functions.c : This c file contains some functions that are common in different machine learning models.