INDIAN INSTITUTE OF TECHNOLOGY
PATNA

SUBMITTED BY
BASKAR NATTARAJAN
2403RES19

CS501 DATABASE SYSTEMS & DATA MINING

Assignment - 4

Problem Statement:-

Write down the difference between linear regression and logical regression. provide mathematical formulation behind these two.

Linear Regression:-

Linear regression is a statistical method that uses a linear equation to model the relationship between two variables.

The variable being predicted is called the "dependent variable"

The variable used to predict is called the "Independent variable".

It is aimed to predict "continous" outcomes.

Linear regression predicts the relationship between two variables by assuming a linear connection between the independent and dependent variables.

It seeks the optimal line that minimizes the sum of squared differences between predicted & actual values.

Simple Linear Regression

There is a one independent variable and one dependent variable.

The model estimates the slope and intercept of the line of best fit, which represents the relationship between variables.

The slope represents the change in dependent variable for each unit change in independent variable

while the intercept represents the predicted value of dependent variable when the independent variable is zero.

Assume,

X - axis: in-dependent variable (predictor / output)

Y - axis : dependent variable (output)

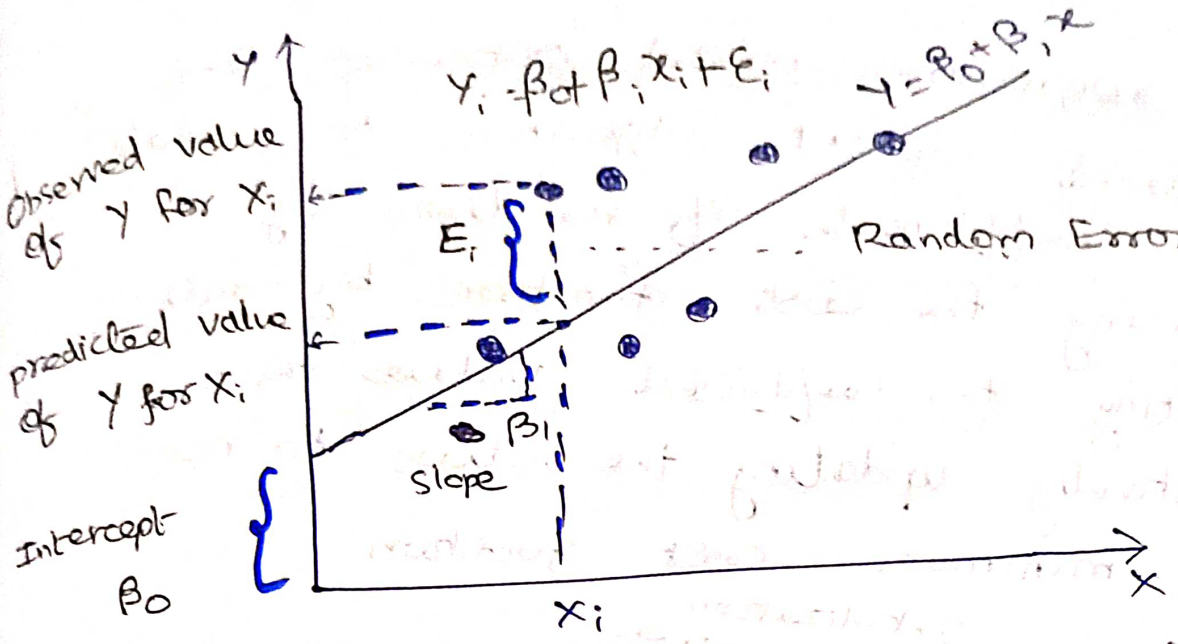
$$Y_i = \beta_0 + \beta_1 X_i$$

Y_i = dependent variable

β_0 = constant / intercept

β_1 = slope / intercept

X_i = Independent variable



In Linear Regression, generally "Mean Squared Error" (MSE) cost function is used. which is average of squared error that occured between the $Y_{predicted}$ and Y_i.

simple linear equation $Y = mx + b$.

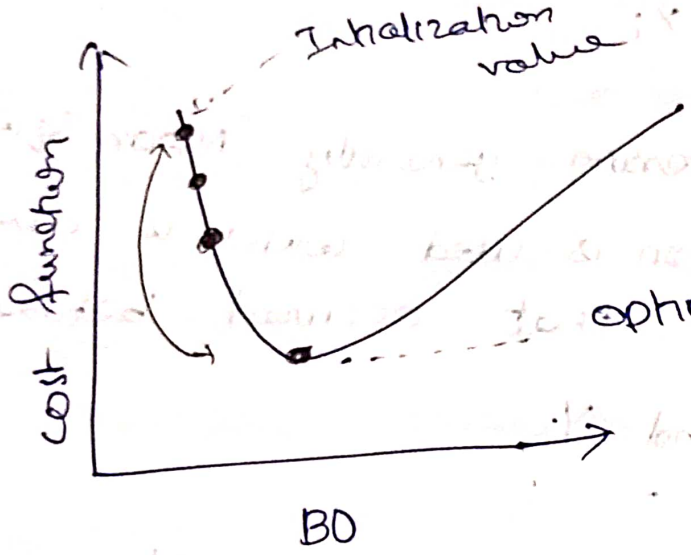
$$MSE = \frac{1}{N} \sum_{i=1}^{n} (y_i - (B1x_i + \beta 0))^2$$

Gradient Descent for Linear Regression.

It is an algorithm that optimize the cost function (objective function) to reach the optimal minimal solution.

To find the optimal solution we need to reduce the cost function (MSE) for all data points.

A regression model optimizes the gradient decent algorithm to update the coefficients of the line by reducing the cost function by randomly selecting the coefficient values then iteratively updating the values to reach the minimum cost function



Advantages of Linear Regression

* Relatively easy to understand and apply straight forward equation to show how one variable affects other or.
* It allows to predict future values based on existing data.
* Even complex algorithms often rely on linear regression as a starting point
* It is a versatile tool for uncovering relationships between variables.

Most used Metrics are.

1. Coefficient Determination or R-squared (R^2).
2. Root Mean Squared Error (RSME) & Residual Standard Error (RSE)

$$R^2 = 1 - (RSS / TSS)$$

Residual Sum of Errors (RSS)

Sum of squares of the residual for each data point in the plot/data.

Measures the difference between the expected and the actual observed output.

$$RSS = \sum_{i=1}^{n} (y_i - b_0 - b_1 x_i)^2$$

Total Sum of Squares (TSS)

Sum of errors of the data points from the mean of the response variable.

$$TSS = \sum (y_i - \bar{y}_i)^2$$

Root Mean Squared Error

- Root of the variance of the Residuals.
- It specifies the absolute fit of the model to the data
- how close the observed data points to the predicted values.

$$RMSE = \sqrt{\frac{RSS}{n}} = \sqrt{\sum_{i=1}^{n} (Y_i^{actual} - Y_i^{predicted})^2 / n}$$

To make the estimate unbiased, one has to devide the sum of squares residuals by the 'degree of freedom' rather than the total number of data points in the model. this term is called "Residual Standard Error (RSE)"

$$RSE = \sqrt{\frac{RSS}{df}} = \sqrt{\sum_{i=1}^{n} (Y_i^{Actual} - Y_i^{predicted})^2 / (n-2)}$$

Types of Linear Regression:

1. Simple linear Regression
   involves one independent variable
   $$y = \beta_0 + \beta_1 x + \epsilon$$
   Example:- predicting a Students test score based on the number of hours studied.

2. Multiple Linear Regression
   → two or more independent variables
   $$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n + \epsilon$$
   example: predicting house price based on area, no. of bedrooms, location et

3. Polynomial Regression:

→ Fits polynomial curve to data

$y = \beta_0 + \beta_1 x_1 + \beta_2 x^2 + \cdots \beta_n x^n + \epsilon$

example:- Modeling the relationship between temperature and gas consumption.

4. Ridge Regression (Regularized Linear Regression)

Adds a penalty terms to the coefficients to prevent overfitting.

Objective: Minimize $||y - X\beta||^2 + \alpha ||\beta||^2$.

Example:- When there are multicollinearities in the data set.

5. Lasso Regression (Least Absolute Shrinkage and Selection Operator)

→ Shrinks the coefficients and variable selection.

Objective:- Minimize $||y - X\beta||^2 + \alpha ||\beta||_1$.

example:- feature selection when dealing with high dimensional data.

6. Elastic Net Regression:

Combines the penalties of Ridge & Lasso

Objective: minimize $||y - X\beta||^2 + \alpha_1 ||\beta||_1 + \alpha_2 ||\beta||^2$.

Example: Balancing between feature Selection & regulation.

Disadvantages of Linear Regression:

→ assumes the relationship between Independent & dependent variables. If assumption not met, model may provide in-accurate predictions.

→ Sensitive to outlies, outliers can influence and reduce the model's predictive accuracy.

→ assuming the variance of residuals (Error) It may go wrong as well.

→ It is limited to linear data only.

→ multicollinearity can occur, which makes difficult to predict

→ overfitting :-
capturing the noise data.

underfitting :-
oversimplifying relationship.

→ Limited to continous outcome variables.

Logistic Regression:

It is a appropriate regression method when the dependent variable is binary

It is used to describe data and explain the relationship between one dependent binary variable and one or more nominal, ordinal, interval or ration-level independent variables.

Types of Logic Regression.

The logistic regression formula for predicting the probability of the dependent variable being (success) given the input x is

$$p(y=1/x) = \frac{1}{1+e^{-(\beta_0+\beta_1 x_1+\beta_2 x_2+\ldots\ldots\beta_n x_n)}}$$

where

- $p(y=1/x)$ is the probability that the outcome y is 1 given the input features x.
- $\beta_0, \beta_1 \ldots\ldots \beta_n$ are the coefficients to be estimated
- $x_1, x_2 \ldots\ldots x_n$ are the independent variables.
- the term $e^{-(\beta_0+\beta_1 x_1+\beta_2 x_2+\ldots\beta_n x_n)}$ is the exponential linear combination of the input features and coefficients.
- the denominator $(1+e^{-(\beta_0+\beta_1 x_1+\beta_2 x_2+\ldots\beta_n x_n)})$ is used to classify observations in to appropriate category.

Types of Logistic Regression:-

Binary logistic Regression:-

predict the probability of a binary outcome such as yes or no
true or false
0 or 1

For example:-

* predict whether a customer will buy or not.
* patient has disease or not.

Multinomial logical regression:-

* predict the probability of one of three or more possible outcomes.

example:-

* type of product customer will buy.
* The rating a customer will give for a product.

Ordinal logistic regression:-

* predicts the probability of outcomes that fall into predetermined order.

example:-

level of customer satisfication
severity of disease

Difference between Linear Regression and Logistic Regression

Purpose :-

Linear regression predicts Continous Outcomes.

Logical regression predicts Binary outcomes

fitting :-

Linear regression uses a "line" that best fits among data points

Logistic regression uses "S-Shaped" curve that predicts probability

Key difference between line & S shaped curve

→ Straight line predicts quantities.
how many, how much.
house price prediction
whether prediction etc.

→ S-shaped curve - predicts YES/NO, true/false, spam/not spam etc.

Mathematical Formula:

Linear Regression

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_n x_n + \epsilon$$

output :- continous numerical value, price or whether.

Logical Regression:-

$$P(y=1|x) = \frac{1}{1+e^{-(\beta_0+\beta_1 x_1+\beta_2 x_2+\dots+\beta_n x_n)}}$$

output :-

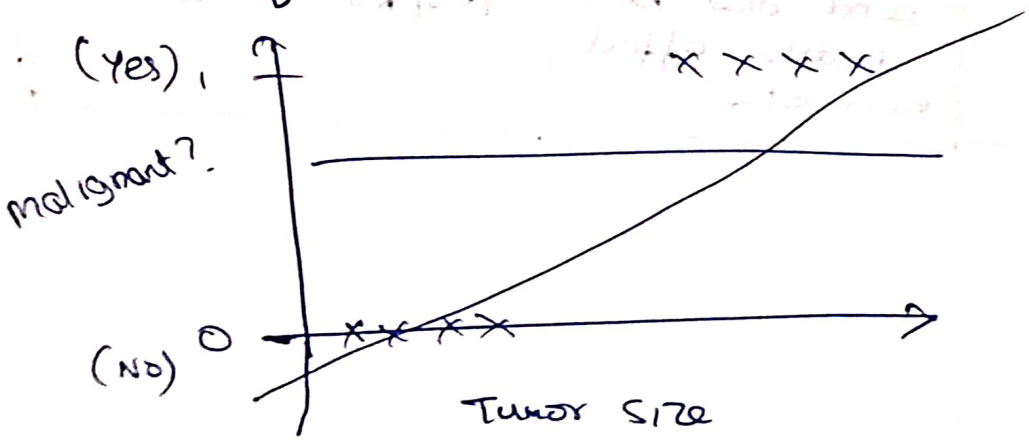
probability value between '0' and '1'

true. or false

Spam or not Spam.

Why do use logistic regression rather than Linear Regression?

→ When you have more outlier in the dataset best fit line in Linear regression may shift to that point.



Why do we use linear regression over logical regression.

→ If the dependent variable is continous.

→ If the value can not be binary

→ relationship trends continans to evolve

Feature	Linear Regression	Logistic Regression
outcome variable	continous (price/sales)	categorical (YES/NO) (Spam/not Spam)
Model Shape	Straight line	S-Shaped curve (Sigmoid function)
Datamining Task	predicting continous values	Identifying customers risk of chun
example	predicting house price	classify categories.
Output	1000 Rs per month rent	probability 1 or 0
Focus	Trends and how variables affect each other.	yes/no decisions.