Artificial Intelligence

Unit-6 Machine Learning –Linear Regression and Logistic Regression

Artificial Intelligence 01CE0702

Marwadi University

Department of Computer Engineering

Shilpa Singhal

Regression

- **Regression** is a statistical method that allows you to predict a dependent output variable based on the values of independent input variables.
- Regression, a type of **supervised learning**, finds the relationship between input and output values and, a given input data, to predict the output value. It does this by finding a mathematical, linear relationship between input and output values. It can have multiple inputs but has a single output.

Regression

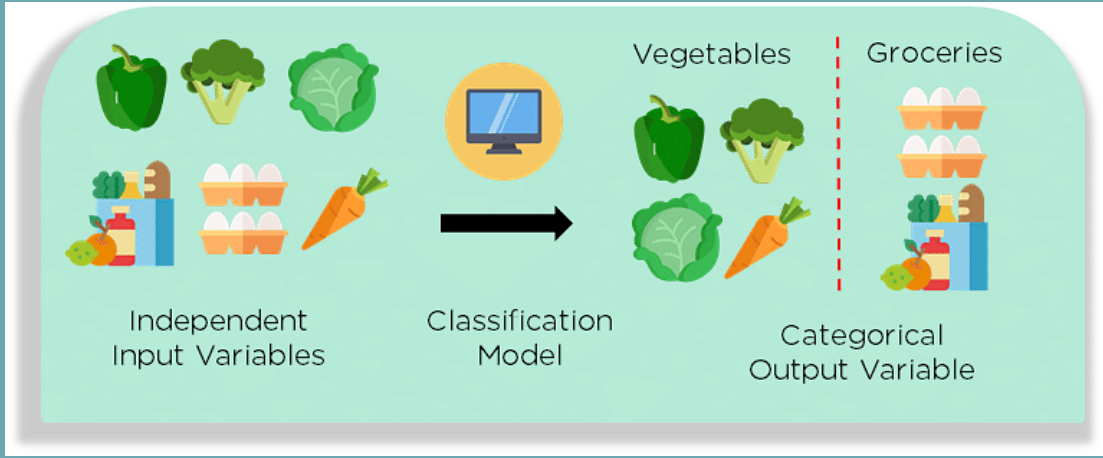
- You can understand regression better, using the diagram below. Using the given input variables or grocery ingredients, you can get a new output or dish. Here, Regression acts as a recipe used to find how these variables go together and the relationship between them.

Classification

- **Classification** allows you to divide a given input into some pre-defined categories. The output is a discrete value, i.e., distinct, like 0/1, True/False, or a pre-defined output label class.
- Simply put, classification is the process of segregating or classifying objects. It is a type of **supervised learning** method where input data is usually classified into output classes. It provides a mapping function to convert input values into known, discrete output classes. It can have multiple inputs and gives multiple outputs.

Classification

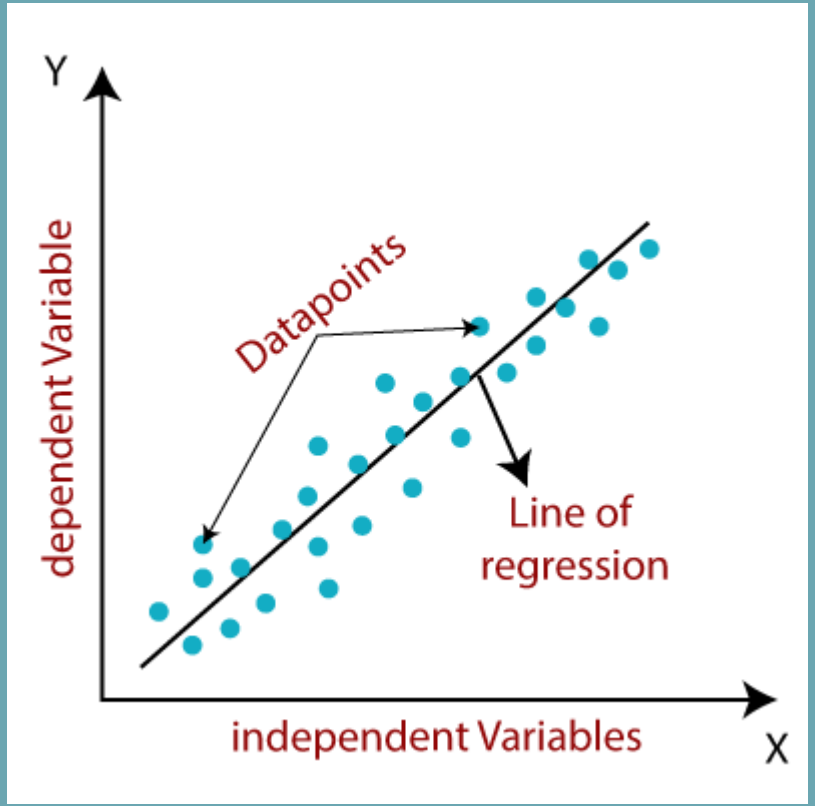
The diagram below clearly explains **classification**. Given a list of grocery items, you can separate them into different categories like vegetables, fruits, dairy products, groceries, etc., using classification.



Linear Regression

- Linear regression is one of the easiest and most popular Machine Learning algorithms.
- It is a statistical method that is used for predictive analysis.
- Linear regression makes predictions for continuous/real or numeric variables such as sales, salary, age, product price, etc.
- Linear regression algorithm shows a linear relationship between a dependent (y) and one or more independent (y) variables, hence called as linear regression.
- Since linear regression shows the linear relationship, which means it finds how the value of the dependent variable is changing according to the value of the independent variable.
- The linear regression model provides a sloped straight line representing the relationship between the variables.

Linear Regression



Linear Regression

- Mathematically, we can represent a linear regression as:

$$y = a0 + a1x + \varepsilon$$

Here,

Y= Dependent Variable (Target Variable)

X= Independent Variable (predictor Variable)

a0= intercept of the line (Gives an additional degree of freedom)

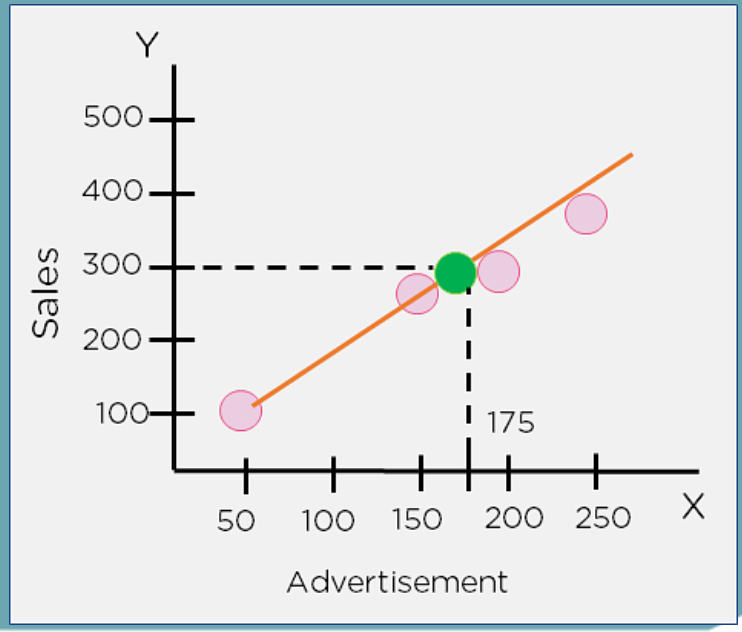
a1 = Linear regression coefficient (scale factor to each input value).

ε = random error

Linear Regression

- Consider the data that is displayed below, which tells you the sales corresponding to the amount spent on advertising.
- Using Linear Regression, you can plot the graph of Sales vs. Advertising, and find the line of best fit between them, and, using that, find the values of the missing variable.

Advertisement	Sales
50	100
150	275
200	300
250	400
175	??



Linear Regression

Using regression, given the advertisement amount, you can predict how many sales will take place.

Advertisement	Sales
50	100
150	275
200	300
250	400
175	300

Types of Linear Regression

- Linear regression can be further divided into two types of the algorithm:
- **Simple Linear Regression:**
  - If a single independent variable is used to predict the value of a numerical dependent variable, then such a Linear Regression algorithm is called Simple Linear Regression.
- **Multiple Linear regression:**
  - If more than one independent variable is used to predict the value of a numerical dependent variable, then such a Linear Regression algorithm is called Multiple Linear Regression.
- When working with linear regression, our main goal is to find the best fit line that means the error between predicted values and actual values should be minimized.
  - The best fit line will have the least error.

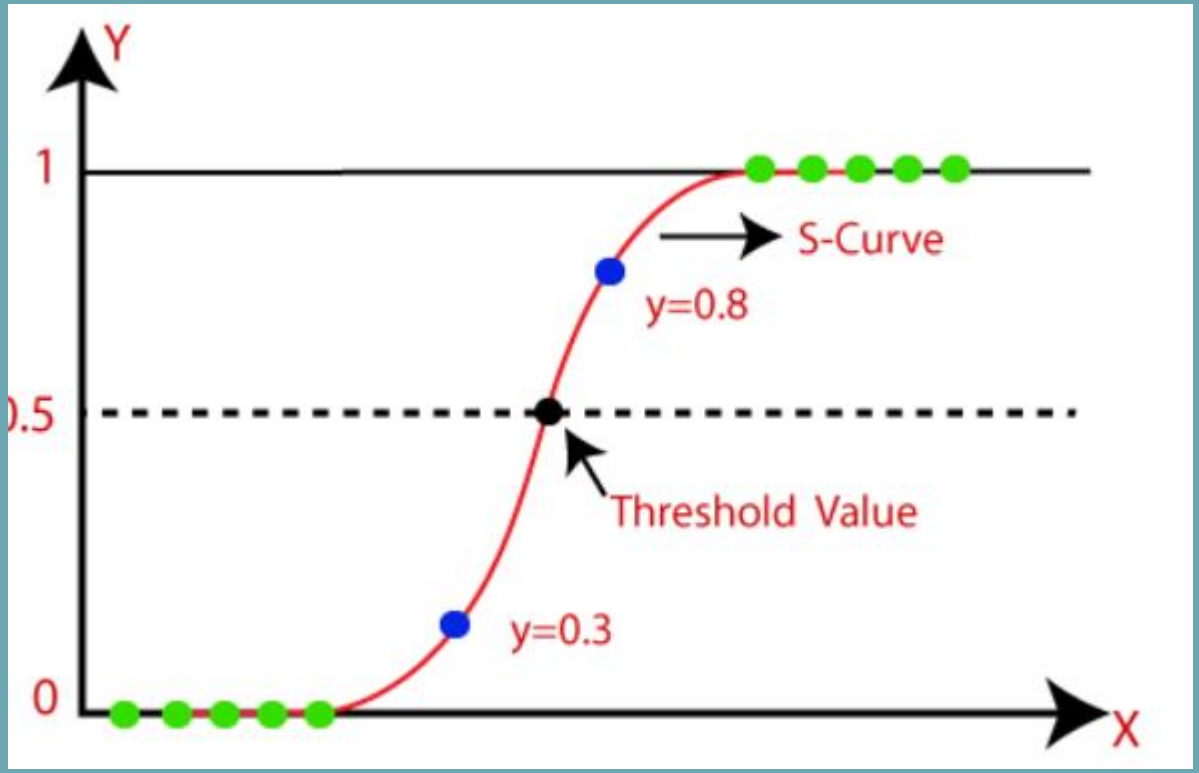
Logistic Regression

- Logistic regression is one of the most popular Machine Learning algorithms, which comes under the Supervised Learning technique.
- It is used for predicting the categorical dependent variable using a given set of independent variables.
- 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.
- Logistic Regression is much similar to the Linear Regression except that how they are used.
- Linear Regression is used for solving Regression problems, whereas Logistic regression is used for solving the classification problems.

Logistic Regression

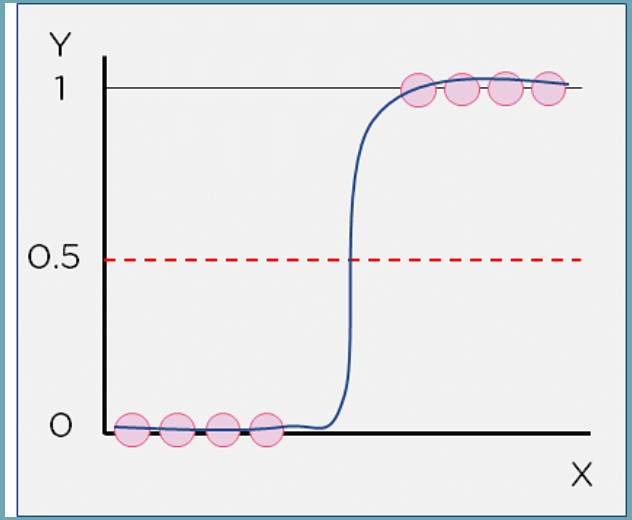
- In Logistic regression, instead of fitting a regression line, we fit an "S" shaped logistic function, which predicts two maximum values (0 or 1).
- The curve from the logistic function indicates the likelihood of something such as whether the cells are cancerous or not, a mouse is obese or not based on its weight, etc.
- Logistic Regression is a significant machine learning algorithm because it has the ability to provide probabilities and classify new data using continuous and discrete datasets.
- Logistic Regression can be used to classify the observations using different types of data and can easily determine the most effective variables used for the classification.

Logistic Regression



Logistic Regression

The data is plotted, and it draws a curve to represent the relationship between the points in our data, which joins the various classes in our output. To classify values into these two categories, you need to set a threshold value between them.



Logistic Regression

It maps the values of the input values onto a categorical variable depending on their position relative to the threshold value. Values of Y above this threshold will be classified as category 1, and it will take values below the threshold as category 0.

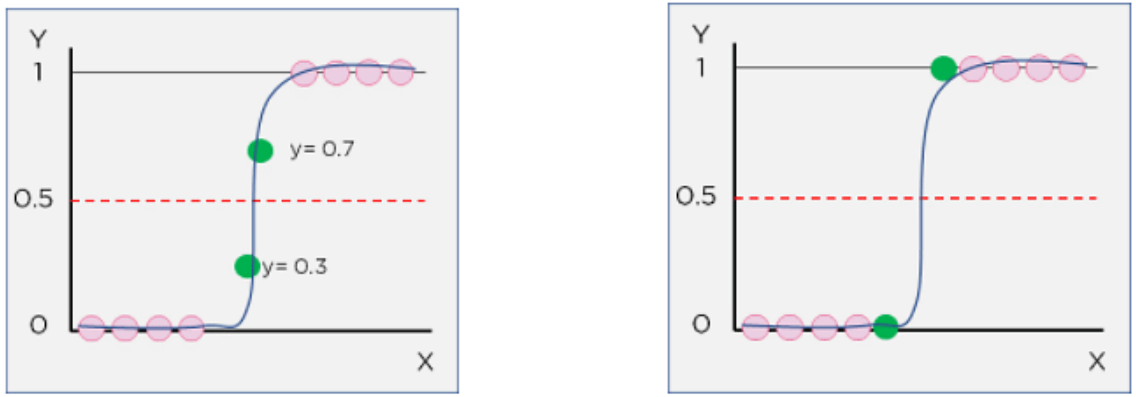


Figure 11: Dividing data into categories

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.

Logistic Regression

Logistic Regression finds the relationship between points by first plotting a curve between the output classes. This curve is called a sigmoid, and the given equation is used to represent a **sigmoid function**. Y is the probability of output, c is a constant, X is the various dependent variables, and b0, b1 gives you the intercept values.

$$\log \frac{Y}{1+Y} = b_0 + b_1 X1 + b_2 X2 + C$$

Logistic Regression Equation

- The Logistic regression equation can be obtained from the Linear Regression equation.
- The mathematical steps to get Logistic Regression equations are given below:
- We know the equation of the straight line can be written as:

$$y = b_0 + b_1x_1 + b_2x_2 + b_3x_3 + \cdots + b_nx_n$$

- In Logistic Regression y can be between 0 and 1 only, so for this let's divide the above equation by (1-y):

$$\frac{y}{1-y}$$; 0 for y= 0, and infinity for y=1

- But we need range between -[infinity] to +[infinity], then take logarithm of the equation it will become:

$$log\left[\frac{y}{1-y}\right] = b_0 + b_1x_1 + b_2x_2 + b_3x_3 + \cdots + b_nx_n$$

- The above equation is the final equation for Logistic Regression.

Linear Regression vs Logistic Regression

Linear Regression	Logistic Regression
Used to predict a dependent output variable based on independent input variable	Used to classify a dependent output variable based on independent input variable
Accuracy is measured using Least squares estimation	Accuracy is measured using Maximum Likelihood estimation
The best fit line is a straight line	The best fit is given by a curve
The output is a predicted integer value	The output is a binary value between 0 and 1 value
Used in business domain, forecasting stocks	Used for classification, image processing

- **MSE / Quadratic loss / L2 loss:**
  - Mean Squared Error, or MSE loss is the default loss to use for regression problems.

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(f_i - y_i)^2$$

where N is the number of data points, f_i the value returned by the model and y_i the actual value for data point i.

Say, $\mathbf{y_i}$ = [5,10,15,20] and f_i = [4.8,10.6,14.3,20.1]
MSE = 1/4 * (|5-4.8|^2+|10-10.6|^2+|15-14.3|^2+|20-20.1|^2) = 0.225

Performance Assessment of Regression Models

- Root Mean Squared Error (RMSE)

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{j=1}^{n}(y_j - \hat{y}_j)^2}$$

  - where y_i is the actual expected output and $\hat{y}_i$ is the model's prediction
  - RMSE = (0.225)^0.5 = 0.474
- Mean Absolute Error (MAE)

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

- Say, y_i = [5,10,15,20] and $\hat{y}_i$ = [4.8,10.6,14.3,20.1]
- Thus, MAE = 1/4 * (|5-4.8|+|10-10.6|+|15-14.3|+|20-20.1|) = 0.4

Performance Assessment of Regression Models

$$R^2 = 1 - \frac{SS_{RES}}{SS_{TOT}} = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \bar{y})^2}$$

$$Adjusted\ R^2 = 1 - \frac{(1 - R^2)(N - 1)}{N - p - 1}$$