Introduction to machine Learning, Logistic Regression – Linear Regression

Introduction

Machine learning (ML) is a branch of artificial intelligence (AI) that enables computers to “self-learn” from training data and improve over time, without being explicitly programmed. Machine learning algorithms are able to detect patterns in data and learn from them, in order to make their own predictions. In short, machine learning algorithms and models learn through experience.

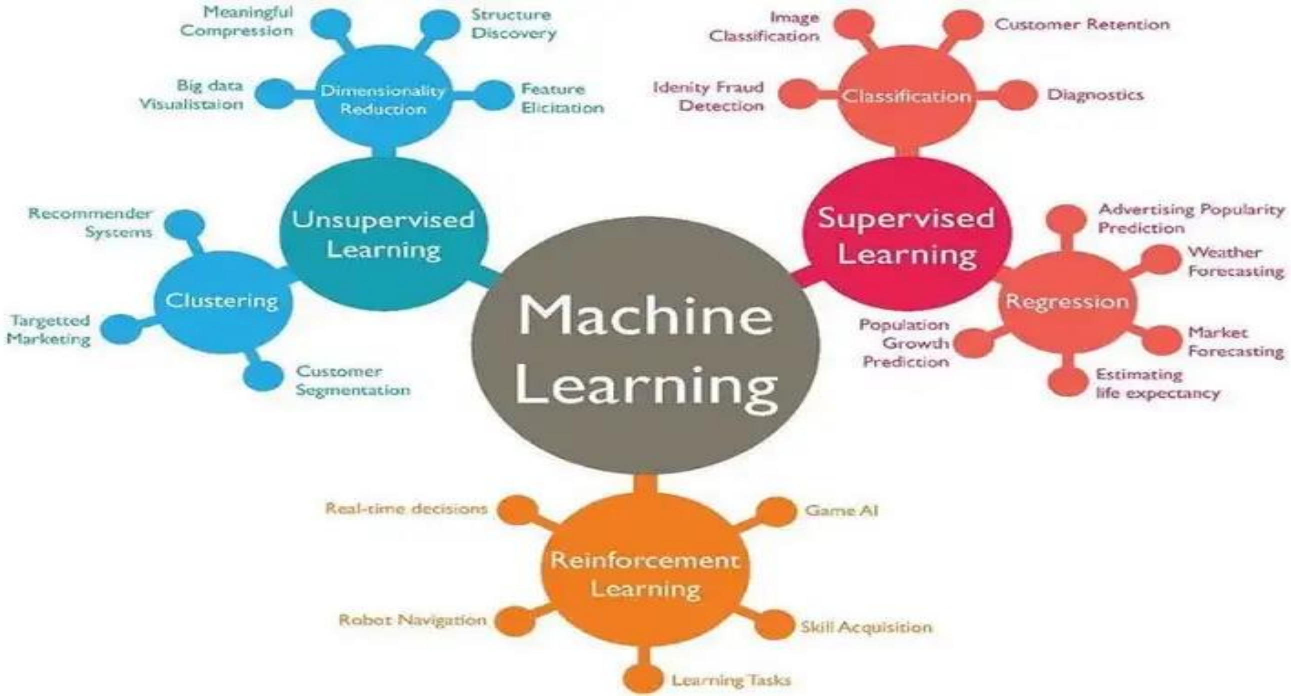
In traditional programming, a computer engineer writes a series of directions that instruct a computer how to transform input data into a desired output. Instructions are mostly based on an IF-THEN structure: when certain conditions are met, the program executes a specific action.

Machine learning, on the other hand, is an automated process that enables machines to solve problems with little or no human input, and take actions based on past observations.

Instead of programming machine learning algorithms to perform tasks, you can feed them examples of labeled data (known as training data), which helps them make calculations, process data, and identify patterns automatically.

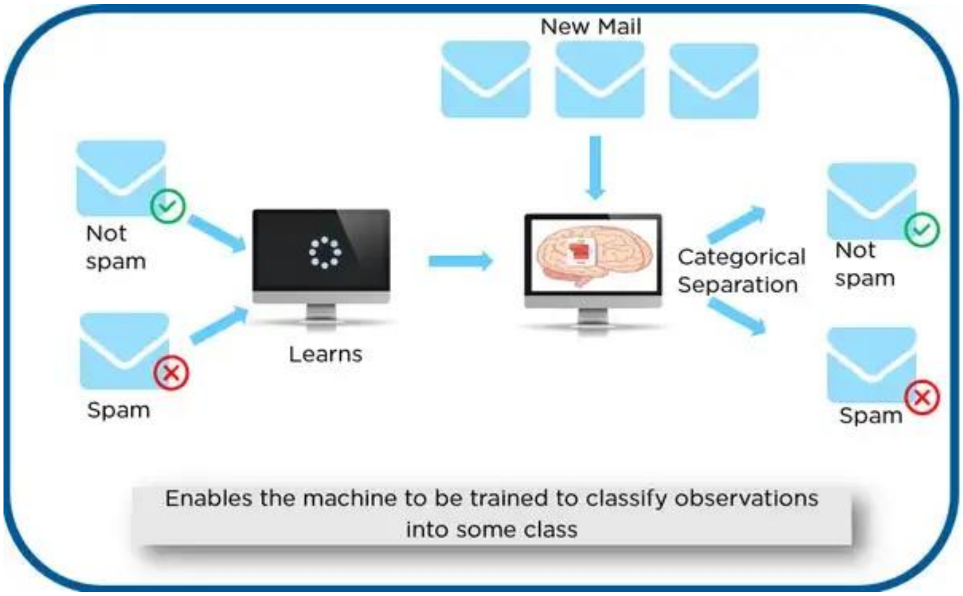
Machine learning can be put to work on massive amounts of data and can perform much more accurately than humans. It can help you save time and money on tasks and analyses, like solving customer pain points to improve customer satisfaction, support ticket automation, and data mining from internal sources and all over the internet.

Types of Machine Learning



Supervised Learning

Supervised learning algorithms and supervised learning models make predictions based on labeled training data. Each training sample includes an input and a desired output. A supervised learning algorithm analyzes this sample data and makes an inference – basically, an educated guess when determining the labels for unseen data.


New Mail
Not spam
Spam
Learns
Categorical Separation
Not spam
Spam
Enables the machine to be trained to classify observations into some class

For example, if you want to automatically detect spam, you would need to feed a machine learning algorithm examples of emails that you want classified as spam and others that are important, and should not be considered spam.

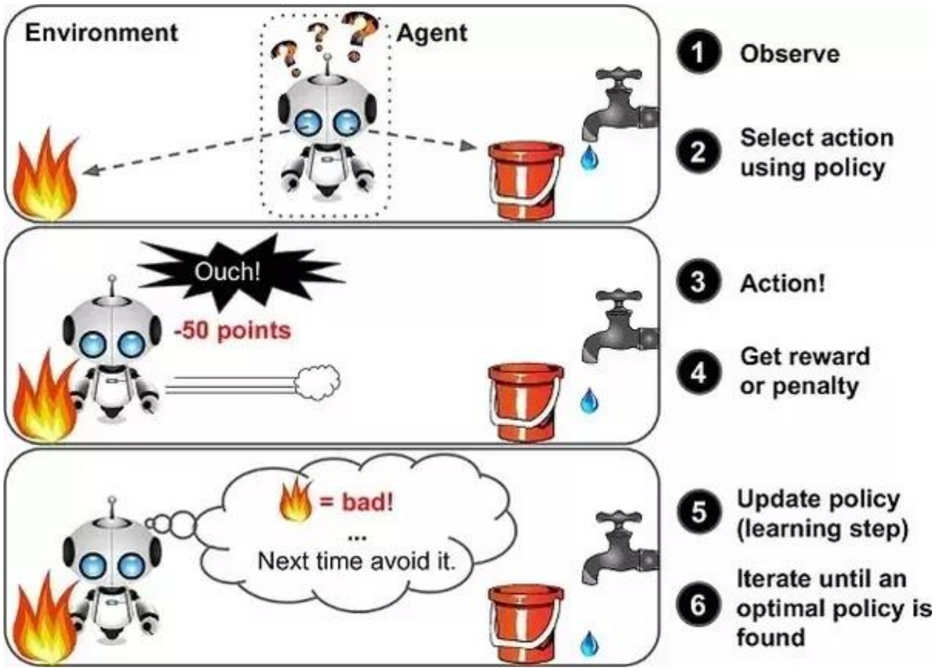
The two types of supervised learning tasks: classification and regression.

Unsupervised Learning

Unsupervised learning algorithms uncover insights and relationships in unlabeled data. In this case, models are fed input data but the desired outcomes are unknown, so they have to make inferences based on circumstantial evidence, without any guidance or training. The models are not trained with the "right answer," so they must find patterns on their own.

One of the most common types of unsupervised learning is clustering, which consists of grouping similar data. This method is mostly used for exploratory analysis and can help you detect hidden patterns or trends.

Reinforcement Learning


Environment
Agent
1 Observe
2 Select action using policy
Ouch!
-50 points
3 Action!
4 Get reward or penalty
= bad!
...
Next time avoid it.
5 Update policy (learning step)
6 Iterate until an optimal policy is found

Reinforcement learning (RL) is concerned with how a software agent (or computer program) ought to act in a situation to maximize the reward. In short, reinforced machine learning models attempt to determine the best possible path they should take in each situation. They do this through trial and error. Since there is no training data, machines learn from their own mistakes and choose the actions that lead to the best solution or maximum reward.

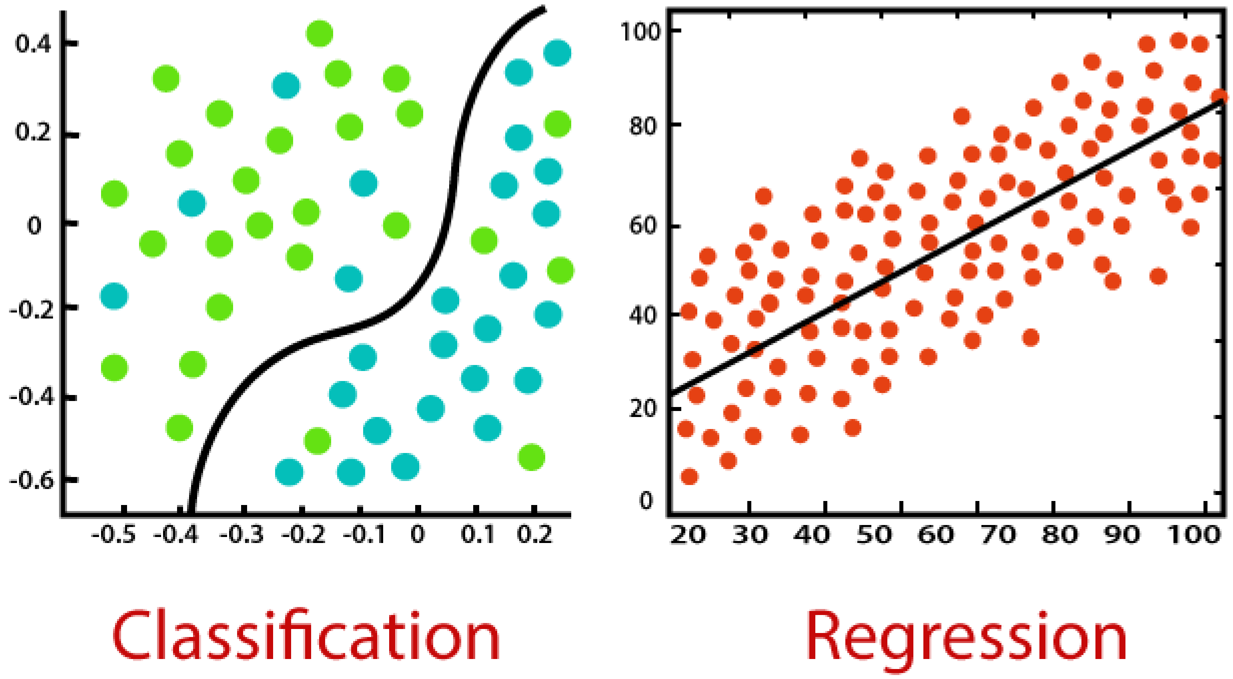
This machine learning method is mostly used in robotics and gaming. Video games demonstrate a clear relationship between actions and results and can measure success by keeping score. Therefore, they're a great way to improve reinforcement learning algorithms.

Linear vs Logistic Regression:

Linear Regression and Logistic Regression are two well-used Machine Learning Algorithms that both branch off from Supervised Learning. Linear Regression is used to solve Regression problems whereas Logistic Regression is used to solve Classification problems.

Classification is about predicting a label, by identifying which category an object belongs to based on different parameters.

Regression is about predicting a continuous output, by finding the correlations between dependent and independent variables.



Linear Regression is known as one of the simplest Machine learning algorithms that branch from Supervised Learning and is primarily used to solve regression problems.

The use of Linear Regression is to make predictions on continuous dependent variables with the assistance and knowledge from independent variables. The overall goal of Linear Regression is to find the line of best fit, which can accurately predict the output for continuous dependent variables. Examples of continuous values are house prices, age, and salary.

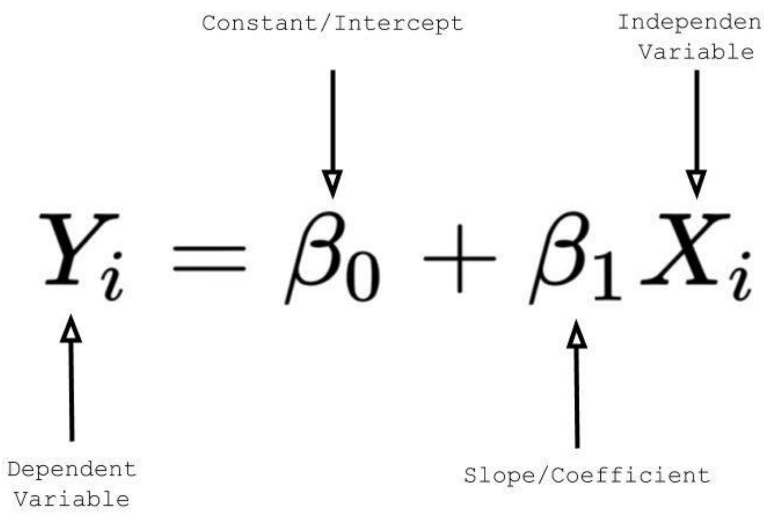
Simple Linear Regression is a regression model that estimates the relationship between one single independent variable and one dependent variable using a straight line. If there are more than two independent variables, we then call this Multiple Linear Regression.

Using the strategy of the line of best fits helps us to understand the relationship between the dependent and independent variable; which should be of linear nature.

The Formula for Linear Regression

If you remember high school Mathematics, you will remember the formula: y = mx + b and represents the slope-intercept of a straight line. ‘y’ and ‘x’ represent variables, ‘m’ describes the slope of the line and ‘b’ describe the y-intercept, where the line crosses the y-axis.

For Linear Regression, ‘y’ represents the dependent variable, ‘x’ represents the independent variable, $\boldsymbol{\beta}$0 represents the y-intercept and $\boldsymbol{\beta}$1 represents the slope, which describes the relationship between the independent variable and the dependent variable



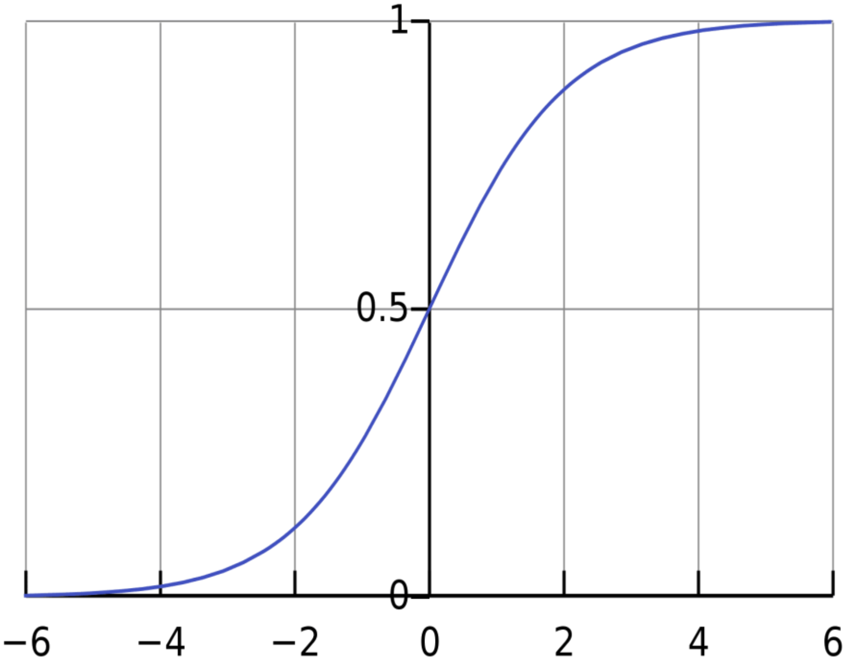
The Formula for Logistic Regression

Logistic Regression is also a very popular Machine Learning algorithm that branches off Supervised Learning. Logistic Regression can be used for both Regression and Classification tasks, however, it is mainly used for Classification.

An example of Logistic Regression predicting whether it will rain today or not, by using 0 or 1, yes or no, or true and false.

The use of Logistic Regression is to predict the categorical dependent variable with the assistance and knowledge of independent variables. The overall aim of Logistic Regression is to classify outputs, which can only be between 0 and 1.

In Logistic Regression the weighted sum of inputs is passed through an activation function called Sigmoid Function which maps values between 0 and 1.



$$f(x)=\frac{1}{1+\mathrm{e}^{-(x)}}$$

Cost Function

A Cost Function is a mathematical formula used to calculate the error, it is a difference between our predicted value and the actual value. It simply measures how wrong the model is in terms of its ability to estimate the relationship between x and y.

- **Linear Regression**

The Cost Function of a Linear Regression is root mean squared error or also known as mean squared error (MSE).

MSE measures the average squared difference between an observation's actual and predicted values. The cost will be outputted as a single number which is associated with our current set of weights. The reason we use Cost Function is to improve the accuracy of the model; minimising MSE does this.

The formula for MSE:

$$MSE = \frac{1}{N}\sum_{i=1}^{n}(y_i - (mx_i + b))^2$$

- **Logistic Regression**

The Cost Function of a Logistic Regression cannot use MSE because our prediction function is non-linear (due to sigmoid transform). Therefore, we use a cost function called Cross-Entropy, also known as Log Loss.

Cross-entropy measures the difference between two probability distributions for a given random variable or set of events.

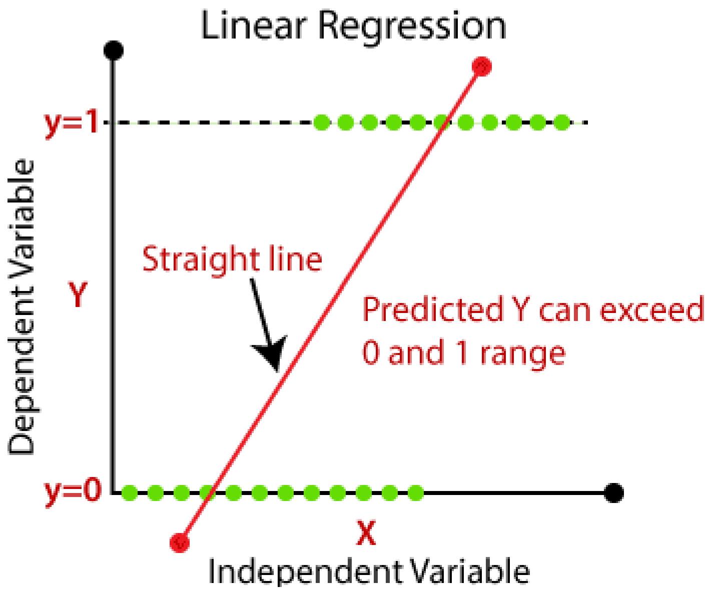
The formula for Cross Entropy:

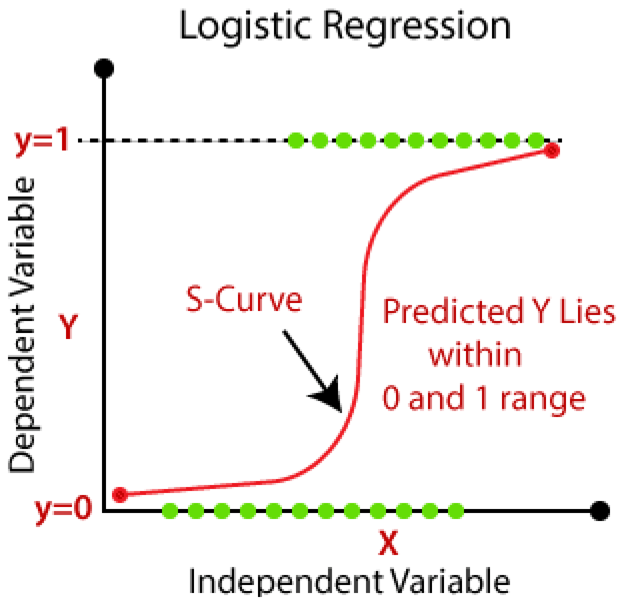
$$J(\theta) = \frac{1}{m}\sum_{i=1}^{m}\mathrm{Cost}(h_\theta(x^{(i)}), y^{(i)})$$

$$\begin{aligned}\mathrm{Cost}(h_\theta(x), y) &= -\log(h_\theta(x)) & \text{if } \mathrm{y} = 1\\ \mathrm{Cost}(h_\theta(x), y) &= -\log(1 - h_\theta(x)) & \text{if } \mathrm{y} = 0\end{aligned}$$

Linear vs Logistic Comparison

Linear Regression	Logistic Regression
Used to predict the continuous dependent variable using a given set of independent variables.	*Used to predict the categorical dependent variable using a given set of independent variables.*
The outputs produced must be a continuous value, such as price and age.	*The outputs produced must be Categorical values such as 0 or 1, Yes or No.*
The relationship between the dependent variable and independent variable must be linear.	*The relationship DOES NOT need to be linear between the dependent and independent variables.*
Used for solving Regression problems.	*Used for solving Classification problems.*
We are finding and using the line of best fit to help us easily predict outputs.	*We are using the S-curve (Sigmoid) to help us classify predicted outputs.*
Least square estimation method is used for the estimation of accuracy.	*Maximum likelihood estimation method is used for the estimation of accuracy.*
There is a possibility of collinearity between the independent variables.	*There should not be any collinearity between the independent variable.*


Linear Regression
y=1
Dependent Variable
Y
Straight line
Predicted Y can exceed
0 and 1 range
y=0
X
Independent Variable


Logistic Regression
y=1
Dependent Variable
Y
S-Curve
Predicted Y Lies
within
0 and 1 range
y=0
X
Independent Variable