Regression

What is Regression?

Regression is a statistical approach used to analyze the relationship between a dependent variable (target variable) and one or more independent variables (predictor variables). The objective is to determine the most suitable function that characterizes the connection between these variables.

Regression in Machine Learning

It is a supervised machine learning technique, used to predict the value of the dependent variable for new, unseen data. It models the relationship between the input features and the target variable, allowing for the estimation or prediction of numerical values.

Regression analysis problem works with if output variable is a real or continuous value, such as “salary” or “weight”. Many different models can be used, the simplest is the linear regression. It tries to fit data with the best hyper-plane which goes through the points.

Regression Algorithms

There are many different types of regression algorithms, but some of the most common include:

- **Linear Regression**
    - **Linear regression** is one of the simplest and most widely used statistical models. This assumes that there is a linear relationship between the independent and dependent variables. This means that the change in the dependent variable is proportional to the change in the independent variables.
- **Random Forest Regression**
    - **Random forest regression** is an ensemble method that combines multiple decision trees to predict the target value. Ensemble methods are a type of machine learning algorithm that combines multiple models to improve the performance of the overall model. Random forest

regression works by building a large number of decision trees, each of which is trained on a different subset of the training data. The final prediction is made by averaging the predictions of all of the trees.

Characteristics of Regression

Here are the characteristics of the regression:

- **Continuous Target Variable:** Regression deals with predicting continuous target variables that represent numerical values. Examples include predicting house prices, forecasting sales figures, or estimating patient recovery times.
- **Error Measurement:** Regression models are evaluated based on their ability to minimize the error between the predicted and actual values of the target variable. Common error metrics include mean absolute error (MAE), mean squared error (MSE), and root mean squared error (RMSE).
- **Model Complexity:** Regression models range from simple linear models to more complex nonlinear models. The choice of model complexity depends on the complexity of the relationship between the input features and the target variable.
- **Overfitting and Underfitting:** Regression models are susceptible to overfitting and underfitting.
- **Interpretability:** The interpretability of regression models varies depending on the algorithm used. Simple linear models are highly interpretable, while more complex models may be more difficult to interpret.

Regression Evaluation Metrics

Here are some most popular evaluation metrics for regression:

- **Mean Absolute Error (MAE):** The average absolute difference between the predicted and actual values of the target variable.
- **Mean Squared Error (MSE):** The average squared difference between the predicted and actual values of the target variable.
- **Root Mean Squared Error (RMSE):** The square root of the mean squared error.

- **Huber Loss:** A hybrid loss function that transitions from MAE to MSE for larger errors, providing balance between robustness and MSE's sensitivity to outliers.
- Root Mean Square Logarithmic Error
- **R2 – Score**: Higher values indicate better fit, ranging from 0 to 1.

Applications of Regression

- **Predicting prices:** For example, a regression model could be used to predict the price of a house based on its size, location, and other features.
- **Forecasting trends:** For example, a regression model could be used to forecast the sales of a product based on historical sales data and economic indicators.
- **Identifying risk factors:** For example, a regression model could be used to identify risk factors for heart disease based on patient data.
- **Making decisions:** For example, a regression model could be used to recommend which investment to buy based on market data.

Advantages of Regression

- Easy to understand and interpret
- Robust to outliers
- Can handle both linear and nonlinear relationships.

Disadvantages of Regression

- Assumes linearity
- Sensitive to multicollinearity
- May not be suitable for highly complex relationships

Linear Regression in Machine Learning

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

What is Linear Regression?

Linear regression is a type of **supervised machine learning** algorithm that computes the linear relationship between the dependent variable and one or more independent features by fitting a linear equation to observed data.

When there is only one independent feature, it is known as **Simple Linear Regression**, and when there are more than one feature, it is known as **Multiple Linear Regression**.

Similarly, when there is only one dependent variable, it is considered **Univariate Linear Regression**, while when there are more than one dependent variables, it is known as **Multivariate Regression**.

Why Linear Regression is Important?

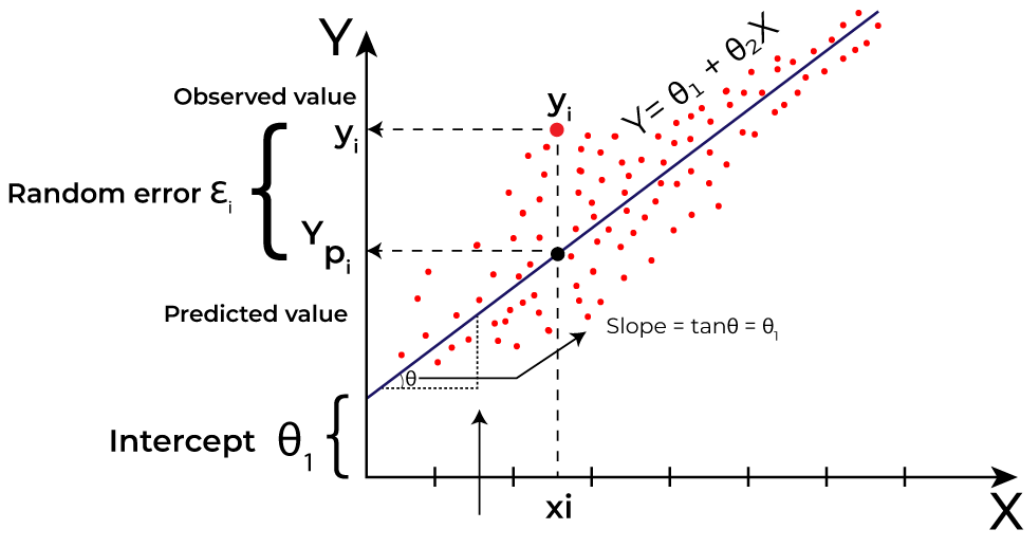
The interpretability of linear regression is a notable strength. The model's equation provides clear coefficients that elucidate the impact of each independent variable on the dependent variable, facilitating a deeper understanding of the underlying dynamics. Its simplicity is a virtue, as linear regression is transparent, easy to implement, and serves as a foundational concept for more complex algorithms.

Linear regression is not merely a predictive tool; it forms the basis for various advanced models. Techniques like regularization and support vector machines draw inspiration from linear regression, expanding its utility. Additionally, linear regression is a cornerstone in assumption testing, enabling researchers to validate key assumptions about the data.

What is the 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).



Here Y is called a dependent or target variable and X is called an independent variable also known as the predictor of Y. There are many types of functions or modules that can be used for regression. A linear function is the simplest type of function. Here, X may be a single feature or multiple features representing the problem.

Linear regression performs the task to predict a dependent variable value (y) based on a given independent variable (x)). Hence, the name is Linear Regression. In the figure above, X (input) is the work experience and Y (output) is the salary of a person. The regression line is the best-fit line for our model.

We utilize the cost function to compute the best values in order to get the best fit line since different values for weights or the coefficient of lines result in different regression lines.

Types of Linear Regression

There are two main types of linear regression:

Simple Linear Regression

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:

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

where:

- Y is the dependent variable
- X is the independent variable
- β0 is the intercept
- β1 is the slope

Multiple Linear Regression

This involves more than one independent variable and one dependent variable. The equation for multiple linear regression is:

$y = \beta 0 + \beta 1 X + \beta 2 X + \ldots\ldots\ldots \beta n X y = \beta 0 + \beta 1 X + \beta 2 X + \ldots\ldots\ldots \beta n X$

where:

- Y is the dependent variable
- X1, X2, ..., Xp are the independent variables
- β0 is the intercept
- β1, β2, ..., βn are the slopes

The goal of the algorithm is to find the best Fit Line equation that can predict the values based on the independent variables.

Gradient Descent

Gradient descent is an optimization technique used to tune the coefficient and bias of a linear equation.

A linear regression model can be trained using the optimization algorithm **gradient descent** by iteratively modifying the model's parameters to reduce the **mean squared error (MSE)** of the model on a training dataset. To update θ1 and θ2 values in order to reduce the Cost function (minimizing RMSE value) and achieve the best-fit line the model uses Gradient Descent. The idea is to start with random θ1 and θ2 values and then iteratively update the values, reaching minimum cost.

A gradient is nothing but a derivative that defines the effects on outputs of the function with a little bit of variation in inputs.

Finding the coefficients of a linear equation that best fits the training data is the objective of linear regression. By moving in the direction of the Mean Squared Error negative gradient with respect to the coefficients, the coefficients can be changed. And the respective intercept and coefficient of X will be if α α is the learning rate.