Regression

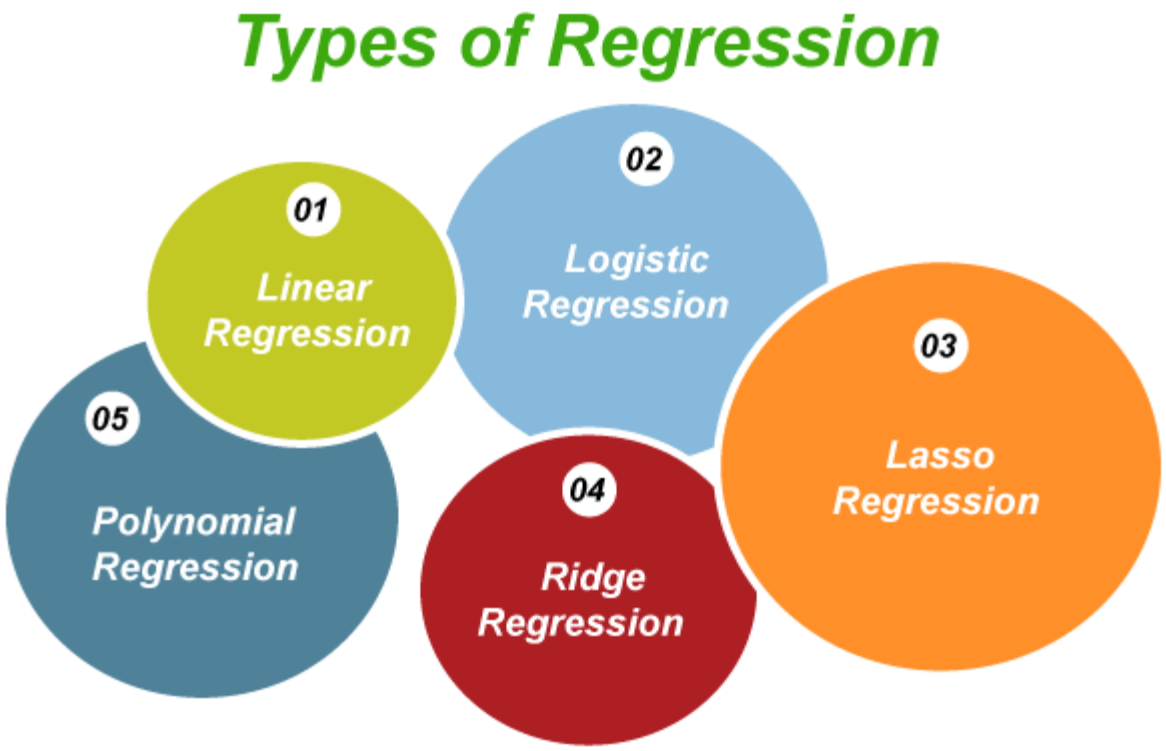
Regression refers to a type of supervised machine learning technique that is used to predict any continuous-valued attribute. Regression helps any business organization to analyze the target variable and predictor variable relationships. It is a most significant tool to analyze the data that can be used for financial forecasting and time series modeling.

Regression involves the technique of fitting a straight line or a curve on numerous data points. It happens in such a way that the distance between the data points and cure comes out to be the lowest.

The most popular types of regression are linear and logistic regressions. Other than that, many other types of regression can be performed depending on their performance on an individual data set.

Regression can predict all the dependent data sets, expressed in the expression of independent variables, and the trend is available for a finite period. Regression provides a good way to predict variables, but there are certain restrictions and assumptions like the independence of the variables, inherent normal distributions of the variables. For example, suppose one considers two variables, A and B, and their joint distribution is a bivariate distribution, then by that nature. In that case, these two variables might be independent, but they are also correlated. The marginal distributions of A and B need to be derived and used. Before applying Regression analysis, the data needs to be studied carefully and perform certain preliminary tests to ensure the Regression is applicable. There are non-Parametric tests that are available in such cases.

Types of Regression



Regression is divided into five different types

1. Linear Regression
2. Logistic Regression
3. Lasso Regression
4. Ridge Regression
5. Polynomial Regression

Linear Regression

Linear regression is the type of regression that forms a relationship between the target variable and one or more independent variables utilizing a straight line. The given equation represents the equation of linear regression

Y = a + b*X + e.

Where,

a represents the intercept

b represents the slope of the regression line

e represents the error

X and Y represent the predictor and target variables, respectively.

If X is made up of more than one variable, termed as multiple linear equations.

In linear regression, the best fit line is achieved utilizing the least squared method, and it minimizes the total sum of the squares of the deviations from each data point to the line of regression. Here, the positive and negative deviations do not get canceled as all the deviations are squared.

Polynomial Regression

If the power of the independent variable is more than 1 in the regression equation, it is termed a polynomial equation. With the help of the example given below, we will understand the concept of polynomial regression.

$Y = a + b * x^2$

In the particular regression, the best fit line is not considered a straight line like a linear equation; however, it represents a curve fitted to all the data points.

Applying linear regression techniques can lead to overfitting when you are tempted to minimize your errors by making the curve more complex. Therefore, always try to fit the curve by generalizing it o the issue.

Logistic Regression

When the dependent variable is binary in nature, i.e., 0 and 1, true or false, success or failure, the logistic regression technique comes into existence. Here, the target value (Y) ranges from 0 to 1, and it is primarily used for classification-based problems. Unlike linear regression, it does not need any independent and dependent variables to have a linear relationship.

Ridge Regression

Ride regression refers to a process that is used to analyze various regression data that have the issue of multicollinearity. Multicollinearity is the existence of a linear correlation between two independent variables.

Ridge regression exists when the least square estimates are the least biased with high variance, so they are quite different from the real value. However, by adding a degree of bias to the estimated regression value, the errors are reduced by applying ridge regression.

Lasso Regression

The term LASSO stands for Least Absolute Shrinkage and Selection Operator. Lasso regression is a linear type of regression that utilizes shrinkage. In Lasso regression, all the data points are shrunk towards a central point, also known as the mean. The lasso process is most fitted for simple and sparse models with fewer parameters than other regression. This type of regression is well fitted for models that suffer from multicollinearity.

Application of Regression

Regression is a very popular technique, and it has wide applications in businesses and industries. The regression procedure involves the predictor variable and response variable. The major application of regression is given below.

- Environmental modeling
- Analyzing Business and marketing behavior
- Financial predictors or forecasting
- Analyzing the new trends and patterns.

Difference between Regression and Classification in data mining

Regression and classification are quite similar to each other. Classification and Regression are two significant prediction issues that are used in data mining. If you have given a training set of inputs and outputs and learn a function that relates the two, that hopefully enables you to predict outputs given inputs on new data. The only difference is that in classification, the outputs are discrete, whereas, in regression, the outputs are not. But the concepts are blurred, as in "logistic regression", which can be interpreted as either a classification or a regression method. So, it becomes difficult for the user to understand when to use classification and regression.

Difference between Regression and Classification in data mining

Regression	Classification
Regression refers to a type of supervised machine learning technique that is used to predict any continuous-valued attribute.	Classification refers to a process of assigning predefined class labels to instances based on their attributes.

In regression, the nature of the predicted data is ordered.	In classification, the nature of the predicated data is unordered.
The regression can be further divided into linear regression and non-linear regression.	Classification is divided into two categories: binary classifier and multi-class classifier.
In the regression process, the calculations are basically done by utilizing the root mean square error.	In the classification process, the calculations are basically done by measuring the efficiency.
Examples of regressions are regression tree, linear regression, etc.	The examples of classifications are the decision tree.

The regression analysis usually enables us to compare the effects of various kinds of feature variables measured on numerous scales. Such as prediction of the land prices based on the locality, total area, surroundings, etc. These results help market researchers or data analysts to remove the useless feature and evaluate the best features to calculate efficient models.