# Machine Learning

# Linear Regression

## Linear Regression

Simple linear regression is useful for finding relationship between two continuous variables. One is predictor or independent variable and other is response or dependent variable. It looks for statistical relationship but not deterministic relationship. The core idea is to obtain a line that best fits the data. The best fit line is the one for which total prediction error (all data points) are as small as possible. Error is the distance between the point to the regression line.

## Ridge Regression

In a multiple LR, there are many variables at play. This sometimes poses a problem of choosing the wrong variables for the Machine Learning, which gives undesirable output as a result. Ridge regression is used in order to overcome this. This method is a regularisation technique in which an extra variable (tuning parameter) is added and optimised to offset the effect of multiple variables in Linear Regression.

## Lasso Regression

Least absolute shrinkage and selection operator, abbreviated as LASSO or lasso, is an LR technique which also performs regularisation on variables in consideration. In fact, it almost shares a similar statistical analysis evident in ridge regression, except it differs in the regularisation values. This means, it considers the absolute values of the sum of the regression coefficients (hence the term was coined on this ‘shrinkage’ feature). It even sets the coefficients to zero thus reducing the errors

completely.

## RANSACK

RANSAC stands for Random sample consensus, which is an iterative method to estimate parameters of a mathematical model from a set of observed data that contains outliers, when outliers are to be accorded no influence on the values of the estimates. Therefore, it also can be interpreted as an outlier detection method. It is a non-deterministic algorithm in the sense that it produces a reasonable result only with a certain probability, with this probability increasing as more iterations are allowed. A basic assumption is that the data consists of inliers which is the data whose distribution can be explained by some set of model parameters, though may be subject to noise, and outliers which are data that do not fit the model. The outliers can come, for example, from extreme values of the noise or from erroneous measurements or incorrect hypotheses about the interpretation of data. RANSAC also assumes that, given a set of inliers, there exists a procedure which can estimate the parameters of a model that optimally explains or fits this data.

## Inference on Big Mart sales Dataset

Linear Regression requires 2 attributes with high correlation. Using correlation Matrix, the 2 attributes were found to be ‘Item_MRP’ and ‘Item_Outlet_Sales’.

While fitting the linear regression model with learning rate of 0.001, the model didn’t converge. The model started converging near the rate of 0.000001 and converged steeply at the rate of 0.00005.

On applying different techniques on the above selected data, we get the following results with the slope, intercept and the accuracy:

| TECHNIQUE | SLOPE | INTERCEPT | ACCURACY |
|---|---|---|---|
| LINEAR REGRESSION | 0.575 | 0,010 | 0.31 |
| RIDGE REGRESSION | 0.575 | 0.010 | 0.31 |
| LASSO REGRESSION | 0.475 | 0.009 | 0.31 |
| RANSACK | 0.592 | -0.025 | 0.321 |

It is clear from the table that Ridge Regression has the same output as that of Linear Regression. That means that adding a penalty equal to the sum of squares of coefficients to the cost function has no effect on the dataset.

Whereas Lasso Regression's output has different slope and intercept but has same accuracy.

In Ransack, there is a change in the slope and intercept of the model with an overall increase in the accuracy. This means that more no of points that fit with the inliers of the model are better points that train the model better than all the points together.