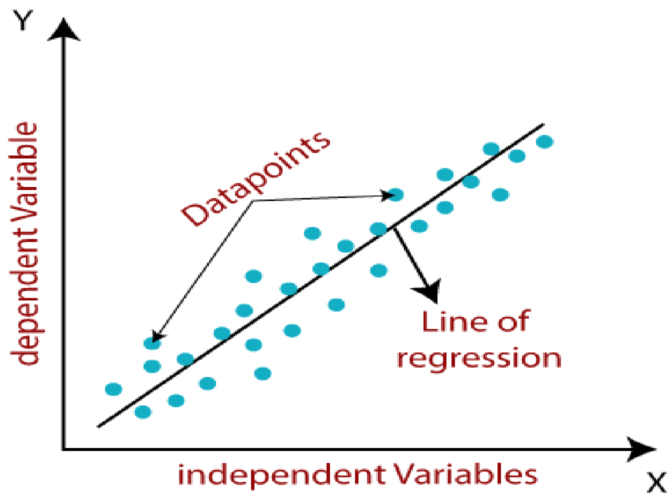
Linear Regression Algorithm

- Rahul Kendre

Linear regression is a supervised machine learning algorithm used to find a linear relationship between a dependent (y) and one or more independent (y) variables, hence called linear regression



❖ Types of Linear Regression

1. 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.

y = mx + c

2. 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.

y = m1x1 + m2x2 + ... + mNxN + c

y = Dependent Variable (Target Variable)

x = Independent Variable (predictor Variable)

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

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

❖ Loss Function(Cost Function):

- The loss is the error in our predicted value of m and c. Our goal is to minimize this error to obtain the most accurate value of m and c.
- We will use the Mean Squared Error function to calculate the loss

$$E = \frac{1}{n} \sum_{i=0}^{n} (y_i - \bar{y}_i)^2$$

$$E = \frac{1}{n} \sum_{i=0}^{n} (y_i - (mx_i + c))^2$$

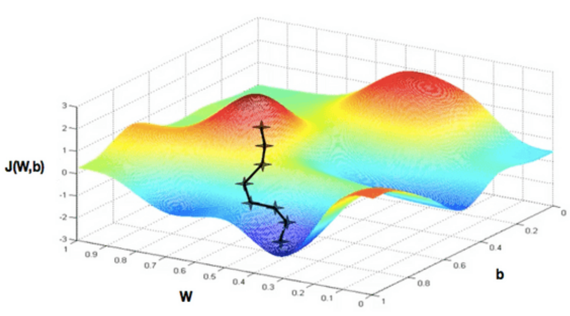
❖ Gradient Descent Algorithm

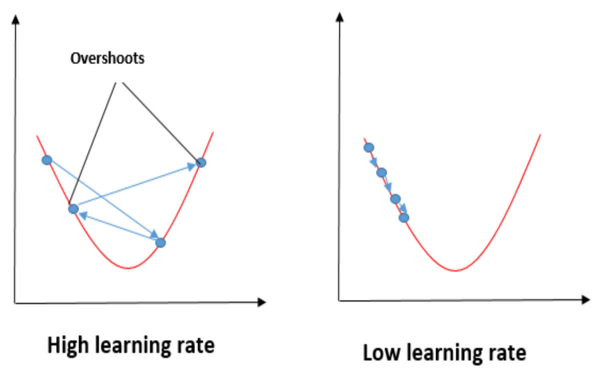
- Gradient descent is an iterative optimization algorithm to find the minimize the Loss Function.
- Gradient descent is a method of updating m and c values to minimize the cost function (MSE) and to get the best fit line(regression line)
- **Best Fit Line**: When working with linear regression, our main goal is to find the best fit line which means the error between predicted values and actual values should be minimized. The best fit line will always have the least Mean squares error.
- A regression model uses gradient descent to update the coefficients of the line (**m and c**) by reducing the cost function by a random selection of coefficient values and then iteratively updating the values to reach the minimum cost function.
- To update m and c, we take gradients from the cost function. To find these gradients, we take partial derivatives for m and c.

$$D_m = \frac{1}{n} \sum_{i=0}^{n} 2(y_i - (mx_i + c))(-x_i)$$

$$D_m = \frac{-2}{n} \sum_{i=0}^{n} x_i(y_i - \bar{y}_i)$$

- Calculate the partial derivative of the loss function with respect to m, and plug in the current values of x, y, m, and c in it to obtain the derivative value D.
    - $$D_c = \frac{-2}{n} \sum_{i=0}^{n} (y_i - \bar{y}_i)$$
      $$m = m - L \times D_m$$
    - $$c = c - L \times D_c$$
- **Global minima:** It is a point that obtains the absolute lowest value of our function
- **Learning Rate**: It determines the size of the steps that are taken by the gradient descent algorithm
- If α (Learning Rate) is very small, it would take a long time to converge and become computationally expensive.
- If α is large, it may fail to converge and overshoot the minimum.
- The most commonly used rates are : 0.001, 0.003, 0.01, 0.03, 0.1, 0.3.





- ## Gradient Descent variants:

  There are three types of gradient descent methods based on the amount of data used to calculate the gradient:

  1. **Batch gradient descent**
  2. **Stochastic gradient descent**
  3. **Mini-batch gradient descent**

❖ Assumptions of Linear Regression:

1. **Linearity**: Relationship between the independent and dependent variables to be linear.
2. **No Multicollinearity** (Independence): Observations are independent of each other.
3. **Normality of Residual**
4. **Homoscedasticity**: The variance of residual is the same for any value of X.

1. Linearity:

Relationship between the independent and dependent variables to be linear

1.1 How to check Linearity:

1. Coefficient of correlation
2. Scatter Plot
3. Correlation matrix

1.2 How to Handle Linearity if get violated:

Apply a nonlinear transformation to the independent and/or dependent variable

1. Log transformation
2. Square root transformation
3. Reciprocal transformation

1.1.1 Coefficient of correlation(R): Correlation coefficients are used to measure how strong a relationship is between two variables.
There are several types of correlation coefficients, but the most popular is Pearson's.

$$\text{Correlation Coefficient(r)} = \frac{Covariance(x,y)}{Std\ dev\ (x) * Std\ dev\ (y)}$$

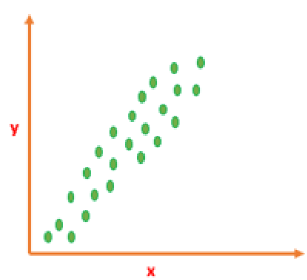
$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$$

Where,

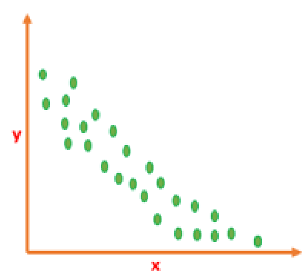
r = Pearson Correlation Coefficient

x_i = x variable samples $\quad$ y_i = y variable sample

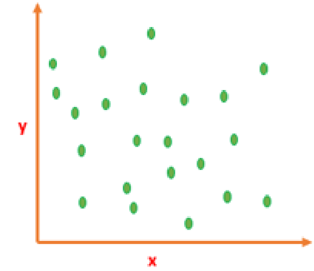
$\bar{x}$ = mean of values in x variable $\quad$ $\bar{y}$ =mean of values in y variable



Positive Correlation
r ≈ 1



Negative Correlation
r ≈ -1



No Correlation
r ≈ 0

The range of **R-Value** is between **-1 to +1**

R = 1 >> indicates a strong positive relationship. R = 1 >> indicates a strong positive relationship.

R = -1 >> indicates a strong negative relationship. if one variable increases, other variable decreas

R = 0 >> It means there is no linear relationship. It doesn't mean that there is no relationship

2. No Multicollinearity:

- Multicollinearity (or collinearity) occurs when one independent variable in a regression model is linearly correlated with another independent variable.
- This means that an independent variable can be predicted from another independent variable in a regression model
- Multicollinearity can be a problem in a regression model because we would not be able to distinguish between the individual effects of the independent variables on the dependent

- Y = M1X1 + M2X2 + C

- Coefficient M1 is the increase in Y for a unit increase in X1 while keeping X2 constant. But since X1 and X2 are highly correlated, changes in X1 would also cause changes in X2 and we would not be able to see their individual effect on Y.
- Multicollinearity may not affect the accuracy of the model as much. But we might lose reliability in determining the effects of individual features in your model – and that can be a problem when it comes to interpretability.

2.1 How to detect Multicollinearity:

1. VIF (Variable Inflation Factors):
   The VIF score of an independent variable represents how well the variable is explained by other independent variables.
   VIF = 1 → No correlation
   VIF = 1 to 5 → Moderate correlation
   VIF >10 → High correlation
2. Correlation matrix / Correlation plot
3. Scatter plots

2.1 How to handle Multicollinearity

1. Dropping variables
2. Combining variables

3. Normality of the residuals

Residuals: The difference between the actual y value and the estimated y value

Residuals = (Ya-Yp)

A normal distribution has some important properties:

1. The mean, median, and mode all represent the center of the distribution.
2. the distribution is a bell shape
3. ≈68% of the data falls within 1 standard deviation of the mean, ≈95% of the data falls within 2 STD of the mean and ≈99.7% of the data falls within 3 STD of the mean

3.1 How to check normality:

1. Graphs for Normality test:

1. Distribution curve, Histogram (sns. distplot, sns.kdeplot)
2. Q-Q or Quantile-Quantile Plot

2. Statistical Tests for Normality(Hypothesis Testing):

1. Shapiro-Wilk test
2. Kolmogorov-Smirnov test
3. D'Agostino's K-squared test

3.2 How to handle Normality:

1. Check and remove outliers
2. Apply a nonlinear transformation to the independent and/or dependent variable
    a. Log transformation
    b. Square root transformation
    c. Reciprocal transformation

4. Homoscedasticity:

- Residuals have constant variance at every level of x. This is known as **homoscedasticity**. When this is not the case, the residuals are said to suffer from **heteroscedasticity**.
- When heteroscedasticity is present in a regression analysis, the results of the analysis become hard to trust

4.1 How to Check Homoscedasticity:

1. Scatter plot between fitted value and residual plot.

4.2 How to handle :

1. **Transform the dependent variable(Y)**: log transformation of the dependent variable
2. **Redefine the dependent variable**
3. **Use weighted regression:** This type of regression assigns a weight to each data point based on the variance of its fitted value.

❖ Advantages:

1. Simple to implement and easier to interpret the output coefficients.
2. When you know the dependent and independent variables have a linear relationship, this algorithm is the best to use because it’s less complex as compared to other algorithms.
3. Linear Regression is prone to over-fitting but it can be avoided using some dimensionality reduction techniques, regularization (L1 and L2) techniques, and cross-validation.

❖ Disadvantages:

1. If the independent features are correlated it may affect performance.
2. it is only efficient for linear data(High Corr between x and Y)
3. Sometimes a lot of feature engineering is required.
4. Scaling is Required: predictors have a mean of 0.
5. It is often quite prone to noise and overfitting.
6. It is sensitive to missing values.
7. It is sensitive to Outliers

❖ Applications:

1. Forecasting the data
2. Analyzing the time series
3. Price Prediction
4. Salary Prediction

❖ Evaluation Metrics for Linear Regression:

1. Mean Absolute Error(MAE): It is most Robust to outliers.
2. Mean Squared Error(MSE)
3. Root Mean Squared Error(RMSE)
4. R-squared or Coefficient of Determination:
    a. SSE(Sum of Squared Error
    b. SSR(Sum of Squares due to Regression)
    c. SST(Sum of Squares Total or Total Error)
5. Adjusted R Squared

Variance = $\sum_{i=1}^{n}(y_i - \bar{y})^2$ → SST (Sum of squares of total)

SSE = $\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$ → Sum of squares of Errors – Unexplained Variance

SSR= $\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$ → Sum of Squares of Regression –Explained Variance

SST= SSR+SSE

y_i - Actual value of y
$\bar{y}$ - Mean value of y
$\hat{y}_i$ - Predicted value of y

$$R^2 = \frac{SSR}{SST} \quad \text{Or} \quad R^2 = 1 - \frac{SSE}{SST}$$