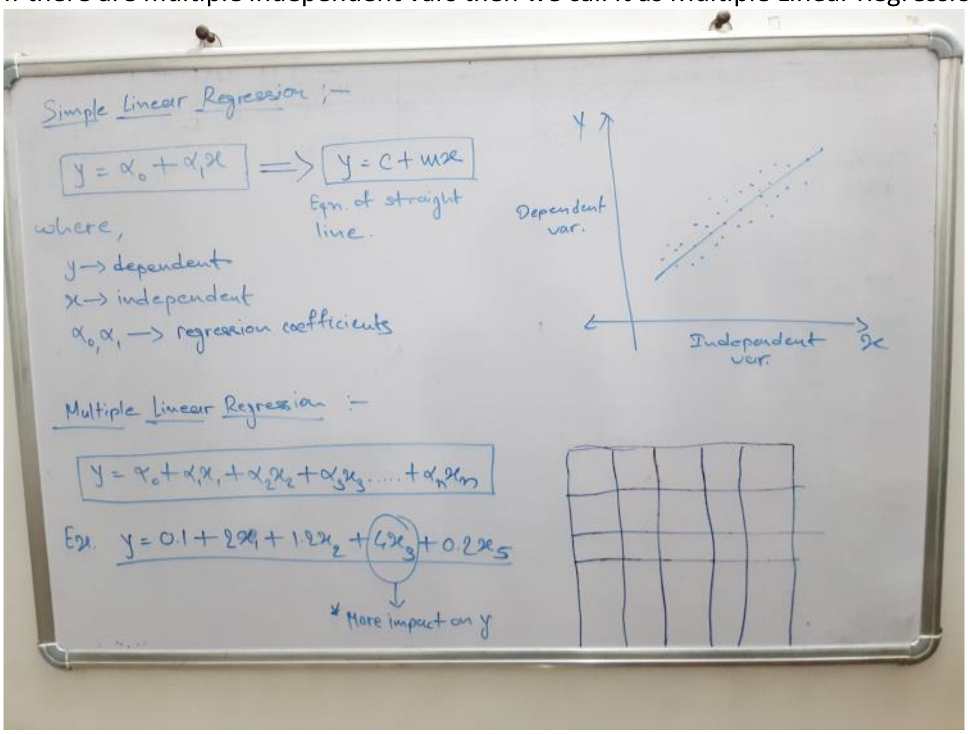
Supervised Learning - Regression

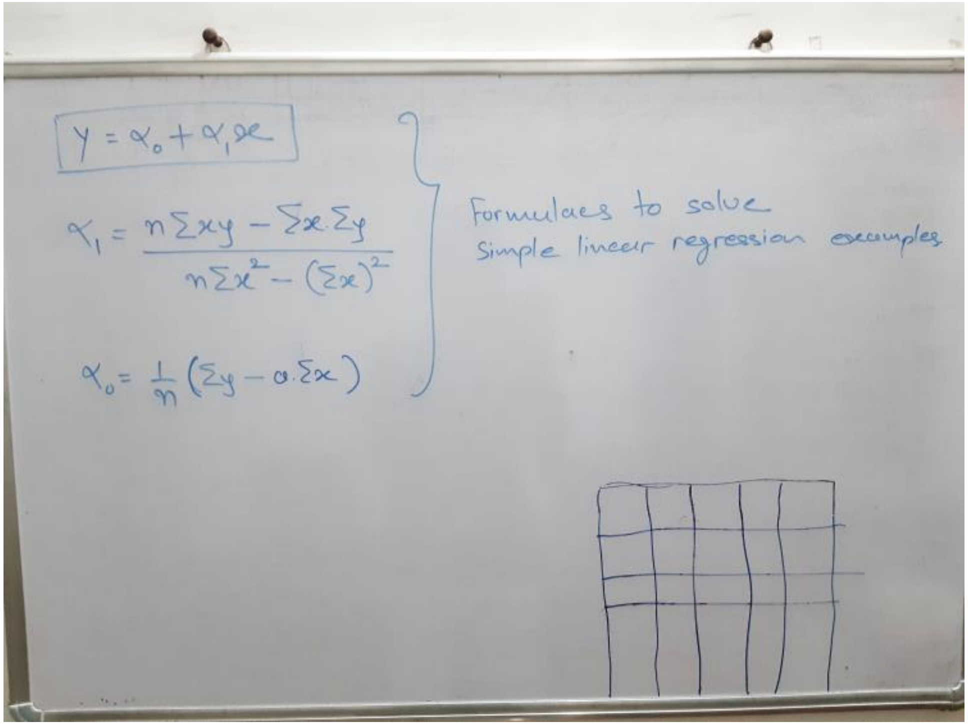
Thursday, December 15, 2022 5:24 PM

Linear Regression -

- Regression is a supervised learning algo and used to establish relationship between dependent and independent vars
- the dependent variable must be continuos or numeric in nature
- Linear regression is one of the type of the regression
- It is used to establish linear relation between dependent and independent vars
- y = θx + b # Linear Equation
- If there the dependent var only depends upon one var then we call it as Simple linear regression
- If there are multiple independent vars then we call it as Multiple Linear Regression

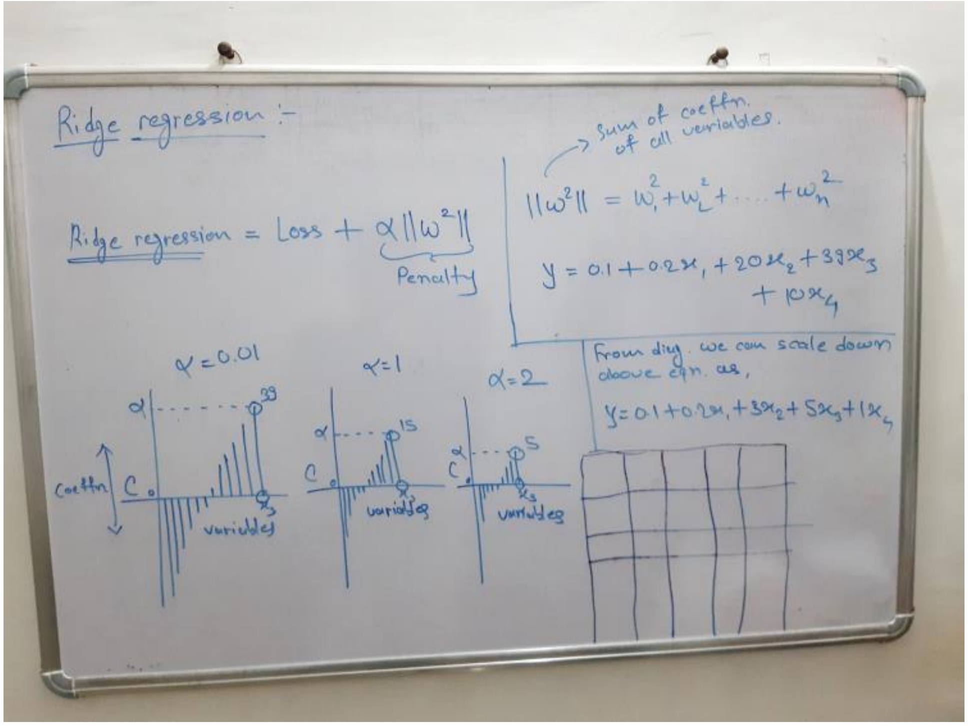


- the regression coefficient represents the amount of impact a variable is having on dependent var



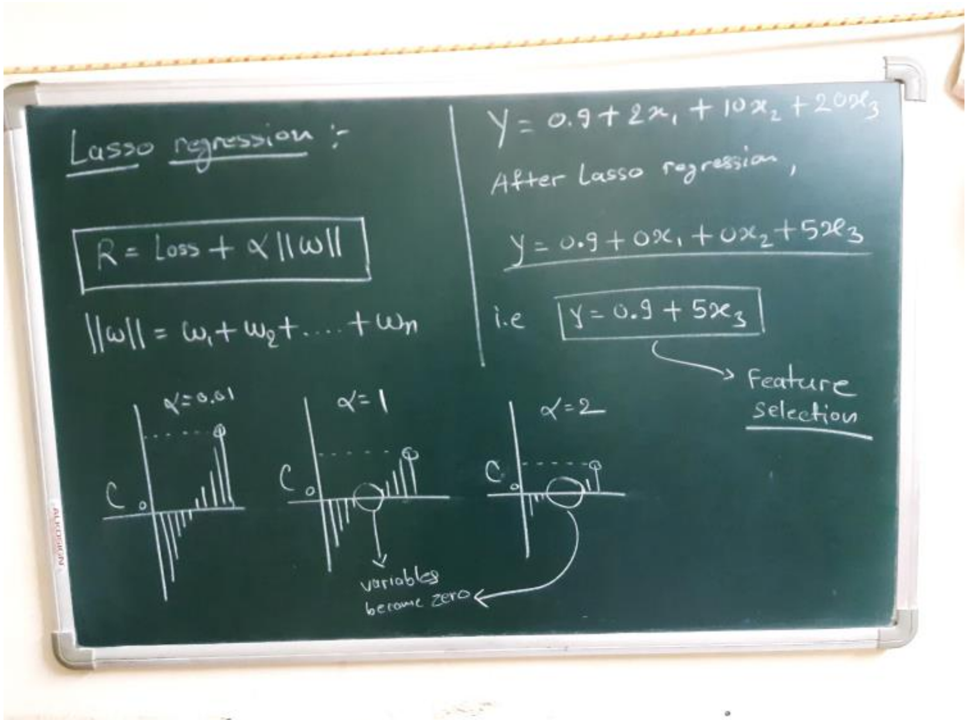
Ridge regression -

- In regression analysis the computation cost may affect the performance of the model
- Larger values of coefficients may cause complex computation
- To scale down the magnitude of the coefficient and to reduce the cost of computation we use Regularization
- Regularization is of two type
  1.Ridge regression
  2.Lasso regression
- Loss = difference between predicted value and actual value
- Penalty = compensation provided to reduce the penalty
- Ridge regression scale down some variables to almost zero



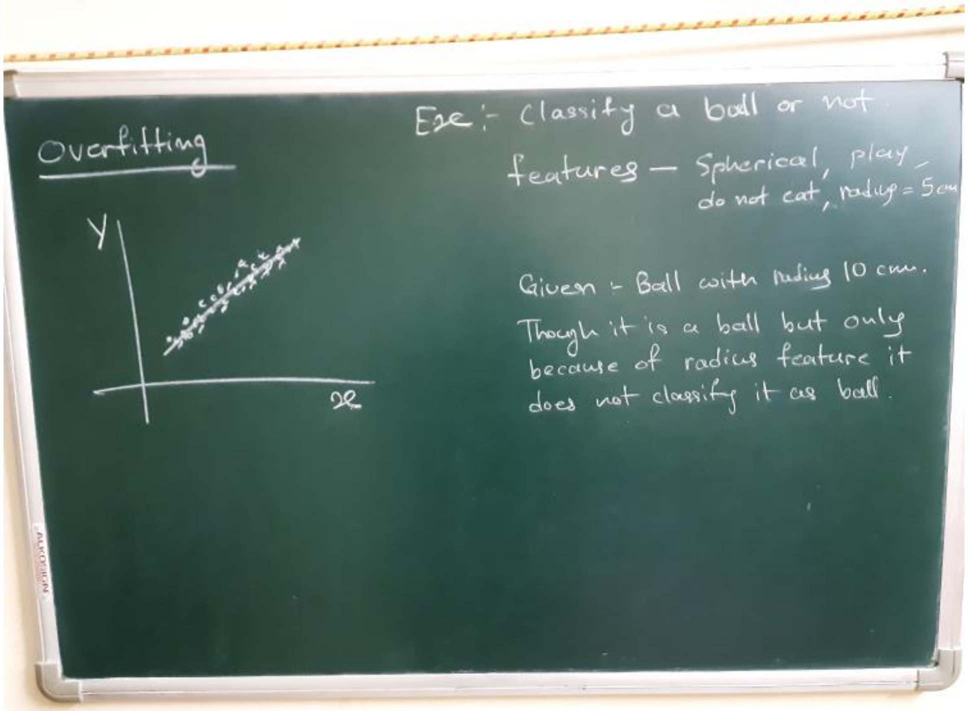
Lasso regression -

- Similar to Ridge regression
- In this some variables scale down to zero
- This is also used for feature selection
- Used to avoid overfitting problem
- Removes non impacting variables

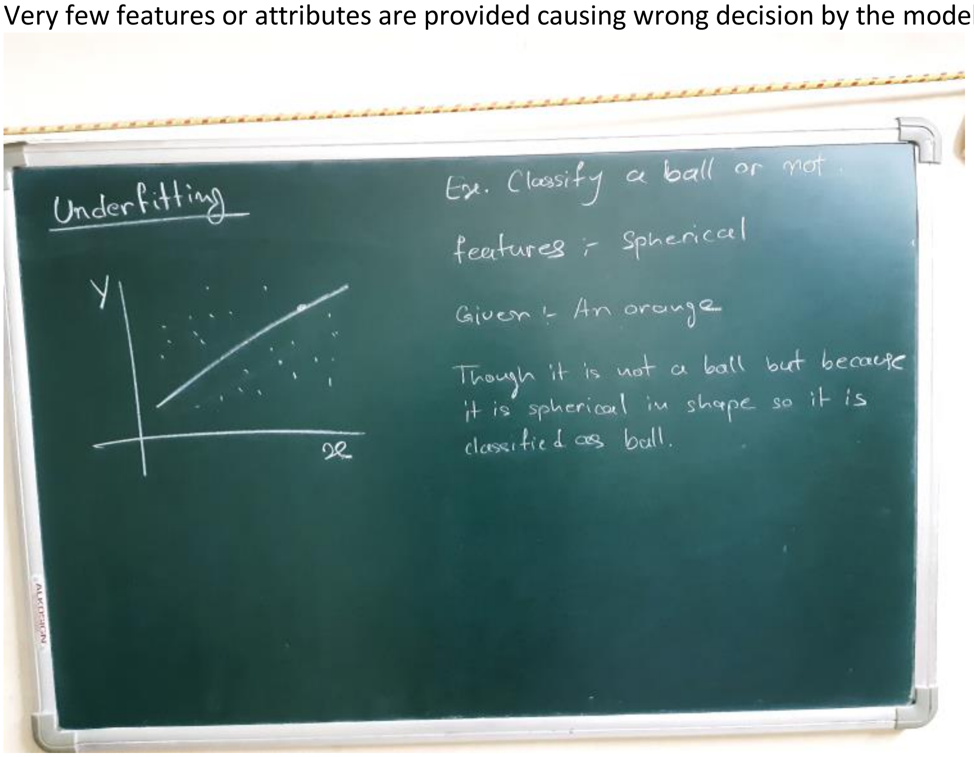


Overfitting and Underfitting -

- When a model tries to cover all the data points on the x-y plane then it is called as overfitting
- In layman terms overfitting is providing too much information to the model
- During training phase the model is provided with too much info
- This too much info then creates confusion while testing the model in real time and gives wrong classification
- Too much features or attributes are provided causing confusion while taking decision by a model

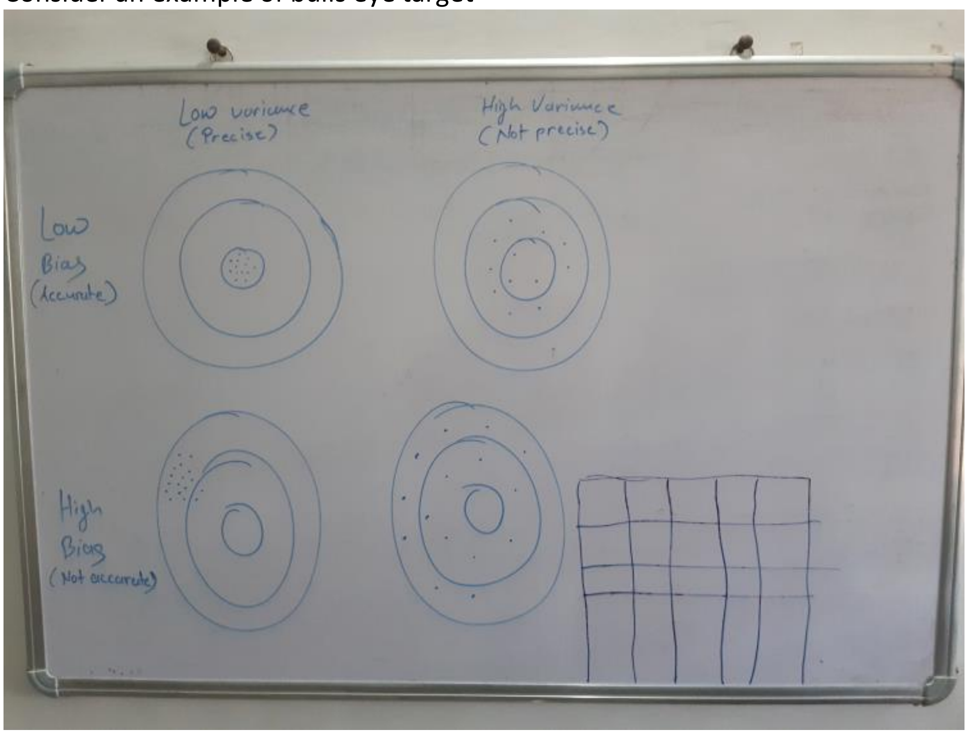


- When a model does not does not relate to data points in the x-y plane and operates on its own is known as underfitting
- In layman term underfitting is providing very less info to the model
- During training phase the model is trained with very less data
- This less info causes model to classify in wrong manner because of lack of info.
- Very few features or attributes are provided causing wrong decision by the model



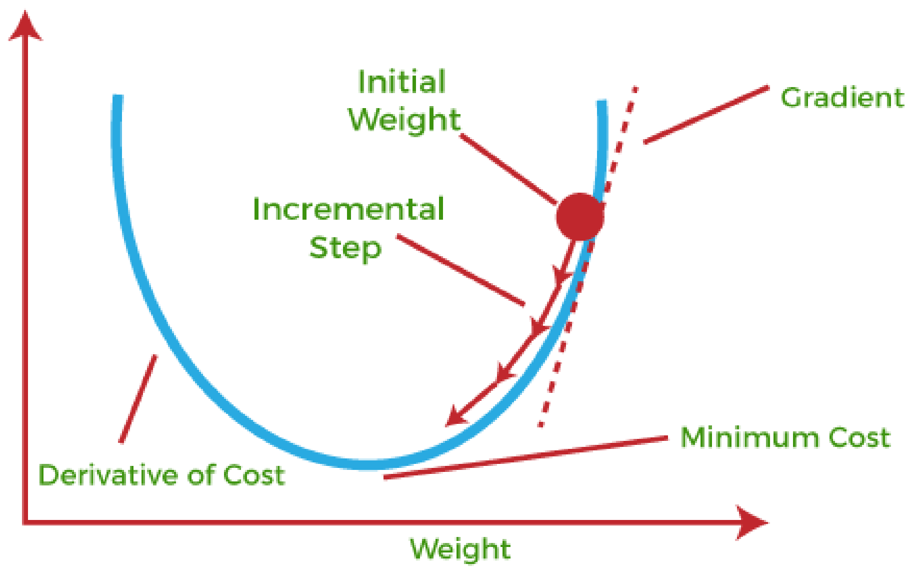
Bias and Variance -

- Bias is the gap between the predicted value and the actual value by a model.
- High bias means the gap between actual and predicted value is high
- Low bias is reverse of the high bias
- Bias is used to check accuracy of the model
- Variance is measurement of scatteredness of predicted values
- High variance means the predicted values are very scattered or far from each other.
- Low variance is reverse of high variance.
- Variance is used to check precision of the model
- Consider an example of bulls eye target

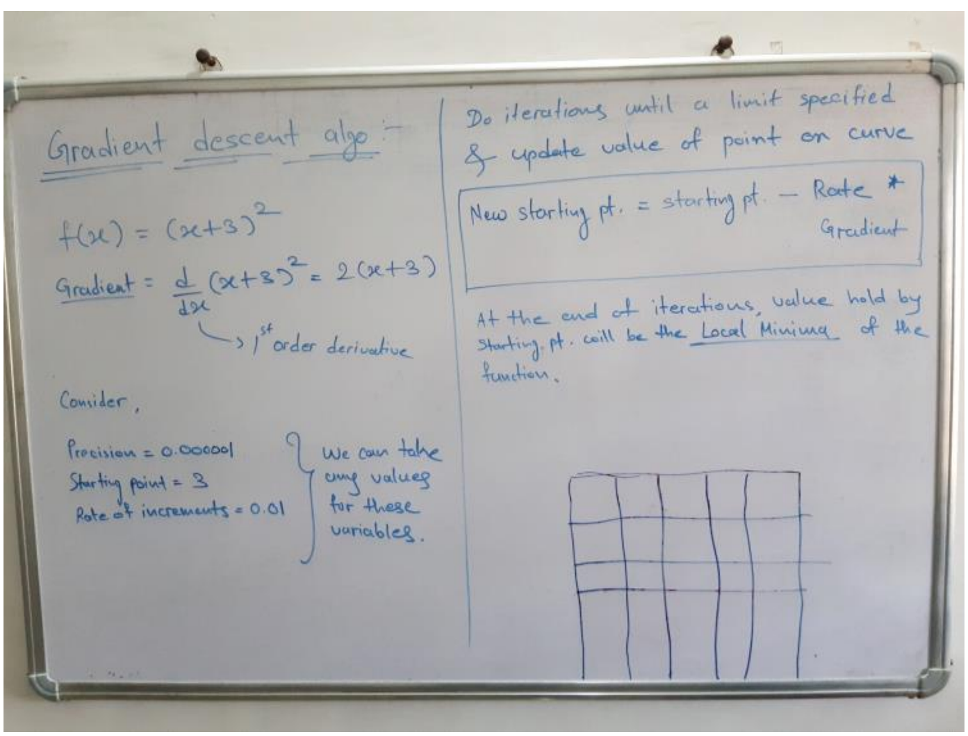


Gradient descent -

- Gradient descent is an optimization and iterative algo
- It is used to train ML models by reducing errors between actual and expected values
- Optimization includes minimization or maximization of the outcome
- Gradient descent algo means finding local minima of a given function
- Gradient is the slope of the given function.
- Cost function is the measurement of difference between actual and expected values



- Steps -
  1.Calculate gradient of the function by first order derivation
  2.Select a starting point on the graph of the function
  3.Move incrementally away from the gradient
  4.The value we got after iteration is the local minima of the function



Evaluation metrics -

These are used to evaluate performance of the ML model developed.
We build a model and it generates some output.
This output then evaluated and feedback is given by evaluation metrics.
Then using this feedback we improve or make changes in our model.

MAE (Mean Absolute Error) -

It is the mean of absolute difference between actual and predicted values

$$\text{MAE} = \frac{1}{n}\sum_{j=1}^{n}|y_j - \hat{y}_j|$$

Where,
N = no. of observations
Y = actual value
Y^ = predicted value

RMSE (Root Mean Squared Error) -

It is simply square root of Mean Squared Error

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{j=1}^{n}(y_j - \hat{y}_j)^2}$$

R2 (R Squared) -

It is used to calculate Goodness of Fit

$$R^2 = 1 - \frac{SS_{RES}}{SS_{TOT}} = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \overline{y})^2}$$

SSRes = Squared Sum error of regression line
SStot = Squared Sum error of total

If R*R = 0, then Regression Line and Mean line overlaps and this means that model performance is worst
If R*R = 1, then model is perfect (ideally)
The value of R*R should lie between 0 and 0.8