Points covered for Linear Regression.

1. Definition of algorithm
2. Aim/Goal
3. Objective
4. Advantages
5. Disadvantages
6. Performance Matrix
7. Regularization/Optimization

1. Definition:

It is the process of establishing relationships between dependent & independent variables. When the relationship is linear in nature we call it Linear Regression. It is based on the **Least Squared method. It is a linear Model.**

2. Aim/Goal:

Here aim is to create the best fit line by using equation

y=mx+c where, y⇒ dependent variable

m⇒ Slope/Gradient/Weight/Coeff.of .Regression

x⇒ independent variable

c⇒ intercept (Point where regression line touches y axis)

3. Objective:

- Establishing the relationship between x & y.
- Forecast the new observation.

4. Advantage:

- Easy to interpret.

5. Disadvantage:

- Highly affected by outliers,missing values and skewness.

6. Performance Matrix:

- **Mean Absolute Error(MAE)**

  Mean absolute error (MAE) is **a loss function used for regression**. Use MAE when you are doing regression and don't want outliers to play a big role. The loss is the mean over the absolute differences between true and predicted values, deviations in either direction

from the true value are treated the same way. **The lower the value the better** and 0 means the model is perfect.

- **Mean Squared Error(MSE)** Speaks about variance

  It gives a clear idea about the scattered ness of the data. It is used in real world case studies. It has one drawback, variance has squared quantity on the y axis but x is singular so it is **not interpretable.**

  The Mean Squared Error measures how close a regression line is to a set of data points.

- **Root Mean Squared Error(RMSE)** Speaks about standard deviation

  It shows how far the predicted data fall from measured true values using the Euclidean distance.**one of the most commonly used measures for evaluating the quality of predictions**.

  The **lower the RMSE**, the **better a given model** is able to "fit" a dataset.

- **R2 Score/R squared (used to measure accuracy)**

  It tells how close the data is from the fitted regression line .i.e. It measures the accuracy of the model. It is used for Model Evaluation.

  It is also known as the **Coefficient of Multiple Determination.**

  The **R squared** value lies between **0 and 1** where 0 indicates that this model doesn't fit the given data and 1 indicates that the model fits perfectly to the dataset provided.

7. Regularization/Optimization

It is the technique which is used for optimizing the model performance by adding some penalty terms in your error function.

The error added is in the form of slope. This is also called penalizing.

When we create the ML model by using linear regression, we use the training data for the regression line. The values of M & C are calculated by using these training data. So logically the regression line that we will get will be best fit for training data, but it is not able to predict the value for testing data.

But in ML our aim is to create the best fit line which is generalized (Best for training and testing data) .

Error in Training set of MSE: Low Bais

Error in Testing set of MSE: High Variance

} **leads to over fitting**

Error in Training set of MSE: High Bais
Error in Testing set of MSE: Low Variance
} **leads to under fitting**

For this we have **<u>two types of Regularization</u>**:

1. Lasso Regularization (L1) Will minimize the coefficients of unwanted features to zero. i.e. Columns having low strength slope will be reduced to zero

2. Ridge Regularization (L2) Will minimize the coefficients of unwanted features but not till zero.

Alpha is the hyper parameter Tuning parameter, which is adjusted for getting good accuracy.

Note: **L2 will add more error than L1**

Again it is a parametric type of model.

- Creates simplified assumptions.
- Requires less data, hence trains model fast.
- Gives low performance/Accuracy
- Gives Under Fitted Results.

Assumptions of Linear Regression:

- There should be a **linear relationship between features and target.**
  Use scatter plot,corr,heatmap to check
- Relationship between features and target should be **Homoskedastic** (Variance should be constant)
- Residuals should be **normally distributed**.
- There should **not be any multicollinearity** between columns/features.(There should not be any correlation between the columns of features)