Linear Regression Model

What is linear regression?

Linear Regression tends to establish a relationship between a dependent variable(Y) and one or more independent variable(X) by finding the best fit of the straight line.

What Are the Basic Assumption?(favourite)

There are four assumptions associated with a linear regression model:

- Linearity: The relationship between X and the mean of Y is linear.
- Homoscedasticity: The variance of residual is the same for any value of X.
- Independence: Observations are independent of each other.
- Normality: For any fixed value of X, Y is normally distributed.

Advantages

- Linear regression performs exceptionally well for linearly separable data
- Easy to implement and train the model
- It can handle overfitting using dimensionlity reduction techniques and cross validation and regularization

Disadvantages

- Sometimes Lot of Feature Engineering Is required
- If the independent features are correlated it may affect performance
- It is often quite prone to noise and overfitting

How to find the best fit line in a linear regression model?

- To find the best fit line for our model we have to make the distance with respect to all the points minimum.
- We have to find that line which is closest to all the points.
- In statistics, this vertical distance is called residual.

Residual is equal to the difference between the observed value and the predicted value.

- - For data points above the line, the residual is positive,
- - and for data points below the line, the residual is negative.

Sum of Squared Residuals(SSE)

Why do we square the error instead of using modulus?

- the squared error is everywhere differentiable, while the absolute error is not (its derivative is undefined at 0).
- This makes the squared error more amenable to the techniques of mathematical optimization.
- To optimize the squared error, we can just set its derivative equal to 0 and solve.

_Explain Ordinary Least Squares Regression in brief.

- Ordinary least squares (OLS) regression is a statistical method of analysis that estimates the relationship between one or more independent variables and a dependent variable.
- The method estimates the relationship by minimizing the sum of the squares of the difference between the observed and predicted values of the dependent variable configured as a straight line.

What are the limitations of OLS?

OLS is computationally too expensive. It performs well with small data. For larger data Gradient Descent is preferred.

Can you briefly explain gradient descent?

- Gradient Descent is an optimization algorithm for finding a local minimum of a differentiable function.
- Gradient descent is simply used to find the values of a function's parameters (coefficients) that minimize a cost function as far as possible.

Explain the significance of learning rate.

Learning rate determines how big are the steps that gradient descent takes into the direction of the local minimum, which figures out how fast or slow we will move towards the optimal weights.

How to evaluate regression models?

There are five metrics used to evaluate regression models:

- Mean Absolute Error(MAE)
- Mean Squared Error(MSE)
- Root Mean Squared Error(RMSE)
- R-Squared(Coefficient of Determination)
- Adjusted R-Squared

R-squared tells us what percent of the prediction error in the y variable is eliminated when we use least-squares regression on the x variable.

R^2 is also called the coefficient of determination.

The value of R^2 varies from 0 to 1.

Can R^2 be negative? Negative R square is possible for the Linear Regression Models where fit is worse than the horizontal line.

What are the flaws in R-squared?

- Problem 1: R^2 increases with every predictor added to a model. As R^2 always increases and never decreases, it can appear to be a better fit with the more terms we add to the model. This can be completely misleading.
- Problem 2: Similarly, if our model has too many terms and too many high-order polynomials we can run into the problem of over-fitting the data. When we over-fit data, a misleadingly high R^2 value can lead to misleading predictions.

What is adjusted R^2?

- Adjusted R-squared is used to determine how reliable the correlation is between the independent variables and the dependent variable.
  - On addition of highly correlated variables the adjusted R-squared will increase whereas for variables with no correlation with dependent variable the adjusted R-squared will decrease.

Impact of outliers

- linear regression needs the relationship between the independent and dependent variables to be linear. It is also important to check for outliers since linear regression is sensitive to outlier effects.

Impact of Missing Values -- sensitive to missing values

Feature Scaling is required ---standardization ect..

Regularization adds the penalty as model complexity increases. The regularization parameter (lambda) penalizes all the parameters except intercept so that the model generalizes the data and won't overfit.

What is L1 Regularization (L1 = lasso)?

- Lasso Regression (Least Absolute Shrinkage and Selection Operator) adds "Absolute value of magnitude" of coefficient, as penalty term to the loss function.
- Lasso shrinks the less important feature's coefficient to zero; thus, removing some feature altogether. So, this works well for feature selection in case we have a huge number of features.
- Along with shrinking coefficients, the lasso performs feature selection, as well. Because some of the coefficients become exactly zero, which is equivalent to the particular feature being excluded from the model.

L2 Regularization(L2 = Ridge Regression)

- Ridge regression adds "squared magnitude" of coefficient as penalty term to the loss function.
- Ridge regularization forces the weights to be small but does not make them zero and does not give the sparse solution.
- Ridge is not robust to outliers as square terms blow up the error differences of the outliers, and the regularization term tries to fix it by penalizing the weights
- Ridge regression performs better when all the input features influence the output, and all with weights are of roughly equal size.
- L2 regularization can learn complex data patterns.

Lambda is a hyperparameter

Lambda directly proportional bias

Lambda inversely proportional variance

In [ ]: