Polynomial Regression

Polynomial regression is a type of regression analysis used to model relationships between a dependent variable and one or more independent variables. Unlike simple linear regression, which fits a straight line to the data, polynomial regression involves fitting a curve to the data points.

The polynomial regression equation can be represented as:

$$y = a_0 + a_1 x + a_2 x^2 + a_3 x^3 + ... + a_n x^n + b$$

Here:

- y is the dependent variable.
- x is the independent variable.
- $a_0, a_1, a_2, ..., a_n$ are the coefficients of the polynomial terms.
- b represents the error term.

Why do we need Polynomial Regression?

Let us consider the Linear Regression model. We created our model and find out that it performs very badly, We observe that between the actual value and the best fit line,which we predicted and it seems that the actual value has some kind of curve in the graph and our line is nowhere near to the points.This is where polynomial Regression comes to the play,it predicts the best fit line that follows the pattern(curve) of the data. Polynomial Regression does not require the relationship between the independent and dependent variables to be linear in the

data set. Polynomial Regression is generally used when the points in the data are not captured by the Linear Regression Model.

As we increase the degree in the model,it tends to increase the performance of the model.However,increasing the degrees of the model also increases the risk of overfitting and underfitting the data.

To prevent this we can use,

Forward Selection:

This method increases the degree until it is significant enough to define the best possible model. It starts with an empty set of features and iteratively adds the most significant feature at each step. It begins by fitting the model with each individual feature separately and selecting the one that performs best based on a predetermined criterion. In next steps, it adds one feature at a time, considering all possible remaining features and selecting the one that improves the model the most until a stopping criterion is met.

Backward Selection:

This method decreases the degree until it is significant enough to define the best possible model. It starts with the full set of features and removes the least significant feature at each step. It begins by fitting the model with all features and eliminating the least significant feature based on a predetermined criterion. In next steps, it removes one feature at a time, re-fitting the model and removing the least significant feature until a stopping criterion is met.

K-Nearest Neighbors (KNN)

KNN is a simple yet powerful supervised machine learning algorithm used for both classification and regression tasks. It uses a distance metric (commonly Euclidean distance) to measure the similarity or distance between data points in the feature space. For each new data point needing a prediction, KNN calculates its distance to all other data points in the training set.

Once distances are calculated, the algorithm identifies the K-nearest neighbors (K data points with the shortest distances) to the input data point.

For classification tasks, KNN assigns the class that is most common among the K-nearest neighbors to the input data point. This can be determined by majority voting.

For regression tasks, KNN predicts the output value by taking the average (or weighted average) of the values of the K-nearest neighbors.

Why do we need a K-NN Algorithm?

Suppose there are two categories, i.e., Category A and Category B, and we have a new data point x1, so this data point will lie in which of these categories. To solve this type of problem, we need a K-NN algorithm. With the help of K-NN, we can easily identify the category or class of a particular dataset.

How to select the optimal K value?

There are no predefined statistical methods to find the most favorable value of K. We initialize a random K value and start computing.Choosing a small value of K leads to unstable decision boundaries. Then we can plot between error rate and K denoting values in a defined range. Then choose the K value as having a minimum error rate.