International Journal of Scientific Research & Engineering Trends
Volume 8, Issue 1, Jan-Feb-2022, ISSN (Online): 2395-566X

Overfitting Vs Underfltting: Reat Illustration

Pradeep Bolleddu
Bachelor of Techno1ogy,
llT Ropar Ropar, lndia
20l9mcbl252@iitrpr.ac.in

Abstract- In this contemporary era,In the core field of Machine learning, We Have to determine the performance of data model in the case of underfitting (High bias and low variance), When the model has not learned enough from the training data which results erratic predictions and low gen-eralization, It can be avoided by using more data by reducing features using feature selection method. In the Over-fitting (High variance and low bias) J6] When the model gets trained with huge data, lt may cause noise give inaccurate data entries in data set, So to avoid overfitting, we can use linear algorithm and by increasing training data by reducing model complexity [Use Ridge regularization lasso regularization.

Keywords- Linear Regression, C1assification, Machine Learning, Data Models.

I. INTRODUCTION

Here,we want to explore and solving a fundamental data science problem in real wor1d, l took dataset from econometrics website describes sales and advertisement of their retail company so by using this data ,when we create useful machine learning system,the data model should make predictions [1] with 0 errors i.e good fit on the data, It is possible at a spot between overfitting, under-fitting,

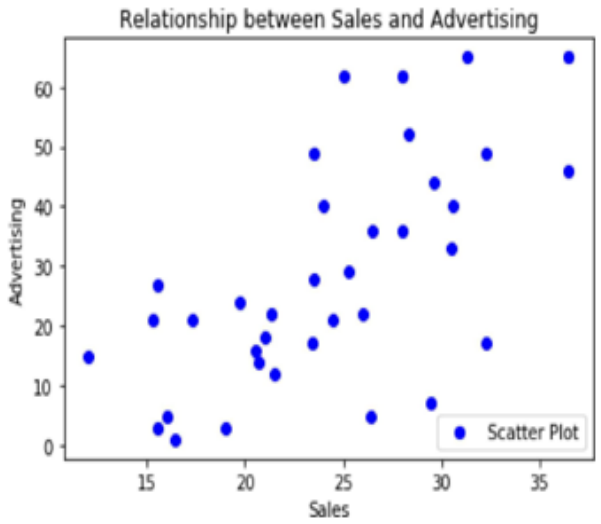

Fig 1. Visualize relationship & plotting a scatterplot between x and y.

To test performance of model with the passage of time, while our model is learning too long from training dataset,it will become more prone to Overfitting because of presence of noise.

In Order to get model with good fit skills on training dataset we have to stop at a point, [4] when error start increasing,so ln this paper i showed the main difference between over-fitting andunder-fitting in machine algorithms i.e Regression and classification.

II. STUDIES AND FINDINGS

In the context of regression analysis, overfitting leads to misleading of R Squared values, regression coefficients, and produce p-values,so when ever we have to decrease (or) avoid underfitting ,By gradually increasing the training time of our machine learning model and another alternative method is by increasing the number of features. Overfitting degrades the performance of machine learning. [4]

Here I mentioned some of the ways by which we can reduce the occurrence of overfitting in the ML model. Training with more data, removing features, early stopping the training data. Cross-Validation: it is a type of technique which involves reserving a specific sample of a dataset on which the model isn't trained and it actually used to decrease model from over-fitting. Especially if the amount of data available is limited.

It also had another name in the core field of ML, i.e rotation estimation Regulariza-tion : Regularization is another important method where we can decrease the errors by fitting function correctly on the given training set to reduce(or) to avoid overfitting. [6]

The most commonly used regularization techniques are: L1 regu1arization, L2 regularization and Dropout regularization Ensembling: it is one of the unique techniques where we can combine different individual predictive models to get an output with the final predic-tive model. [8]

There are 2 important techniques, by which we can get a good point for the model,i.e resampling method to

estimate model accuracy and another method is validation dataset.

III. USE OF DATA MODELS IN MACHINE LEARNING

1. Basics of Data Models:

(Data models) - simply it is a type of system for mapping inputs to outputs,From a training data set,All supervised models, [8]learns relationships between the inputs called [Features] and outputs called [labels] During training the model both features label's learns toMap internally ,later it is tested on a testing data set to give better prediction.so we will compare the predictions with known labels to calculate accuracy.

2. Training and Testing Data:

For making a model we should have training data set with x-values(features) and y-values(labels) that has underlying relationship,Later we can slice the single data set into training set and test set,[11] After validating trained model against test data so that we can find performance and calculate accuracy.

3. Model Building and Graphs of polynomial Expression:

The simplest model is linear regression,where we can get the outputs linearly [2] But to learn the relationship between between x and y ,I used polynomial regression,Here inputs will have different powers,The general equation is mentioned below Here,x=feature and y=label,All the B terms are model parameters and epsilon is the error present in any of the model,when model concentrates on the beta va1ues,we can take any value from x and get the corresponding y value for prediction polynomial degree of straight line is 1 polynomial degree of parabola is 2.

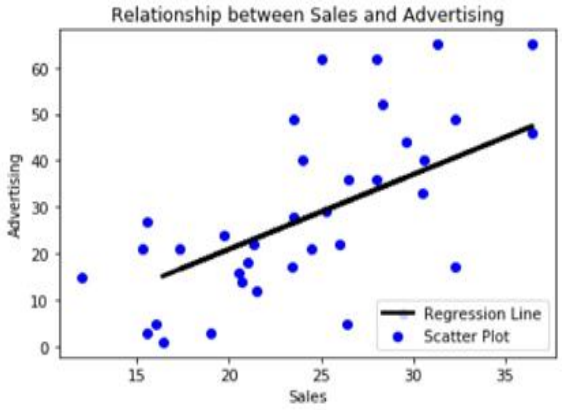

Fig 2. Plotting the regression line.

4. Flexibility in over-fitting Vs under-fitting:

If the model is caused by over-fitting,It is trying to memorize the training data instead of learning mapping from features to labe1s,It is one of the common problem actually occurs when we are training a model with so many parameters,that it can fit nearly any dataset, [9]The degree actually represents the flexibility in the data model,But in the case of under-fitting,it actually occurs when ever our machine learning model is not able to learn correctly from the training data ,so Automatically it reduces the accuracy and provides unreliable predictions.

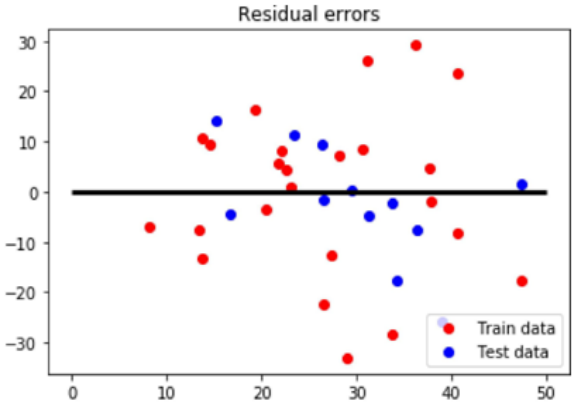

Fig 3. Plotting residual errors to check overfitting & underfitting.

IV. CONCLUSION

The techniques,scheme presented in this paper aims to help researchers and practitioners how to avoid under-fitting and overfitting This paper contributes to other studies in understanding the importance ,where the machine learning model is unpredictable or giving less accuracy on many Aspects.

In our study we mentioned many techniques to avoid overfitting, later found that 'early stopping method is best as compared to other methods, it is actually less sensitive to bias and variance, From the start we know that underfitting neural networks perform badly both on training and testing data set, but Income ML models, over-fitting networks may do very well on training sets but terribly on test data set.

V. ACKNOWLEDGEMENT

I would like to thank my Institute supervisor's for teaching machine learning techniques in a simple way and textbook authors for providing good knowledge results good basic foundations in this domain

REFERENCES

[1] A Theory of overfitting and under-fitting in Machine learning in Debug-ger Cafe by Sovit Ranjan Rath,March 18,2019
[2] Data modeling issues in simulating the dynamic processes in life cycle analysis of buildings By Nikaus Kohler In 2004 [Academic.edu—65116815]