Ayesha Ashraf

Task # 25

Overfitting and Underfitting in Machine Learning

Overfitting and Underfitting are the two main problems that occur in machine learning and degrade the performance of the machine learning models.

The main goal of each machine learning model is **to generalize well**. Here **generalization** defines the ability of an ML model to provide a suitable output by adapting the given set of unknown input. It means after providing training on the dataset, it can produce reliable and accurate output. Hence, the underfitting and overfitting are the two terms that need to be checked for the performance of the model and whether the model is generalizing well or not.

Before understanding the overfitting and underfitting, let's understand some basic term that will help to understand this topic well:

- **Signal:** It refers to the true underlying pattern of the data that helps the machine learning model to learn from the data.
- **Noise:** Noise is unnecessary and irrelevant data that reduces the performance of the model.
- **Bias:** Bias is a prediction error that is introduced in the model due to oversimplifying the machine learning algorithms. Or it is the difference between the predicted values and the actual values.
- **Variance:** If the machine learning model performs well with the training dataset, but does not perform well with the test dataset, then variance occurs.

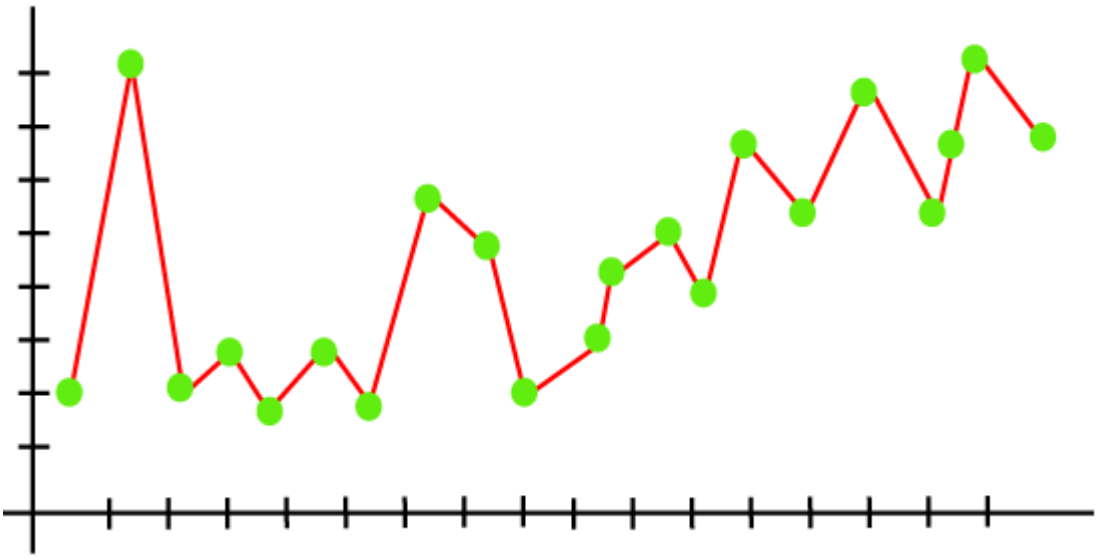
Overfitting

Overfitting occurs when our machine learning model tries to cover all the data points or more than the required data points present in the given dataset. Because of this, the model starts caching noise and inaccurate values present in the dataset, and all these factors reduce the efficiency and accuracy of the model. The overfitted model has **low bias** and **high variance.**

The chances of occurrence of overfitting increase as much we provide training to our model. It means the more we train our model, the more chances of occurring the overfitted model.

Overfitting is the main problem that occurs in supervised learning.

Example: The concept of the overfitting can be understood by the below graph of the linear regression output:



As we can see from the above graph, the model tries to cover all the data points present in the scatter plot. It may look efficient, but in reality, it is not so. Because the goal of the regression model to find the best fit line, but here we have not got any best fit, so, it will generate the prediction errors.

How to avoid the Overfitting in Model

Both overfitting and underfitting cause the degraded performance of the machine learning model. But the main cause is overfitting, so there are some ways by which we can reduce the occurrence of overfitting in our model.

- **Cross-Validation**
- **Training with more data**
- **Removing features**
- **Early stopping the training**

- **Regularization**
- **Ensembling**

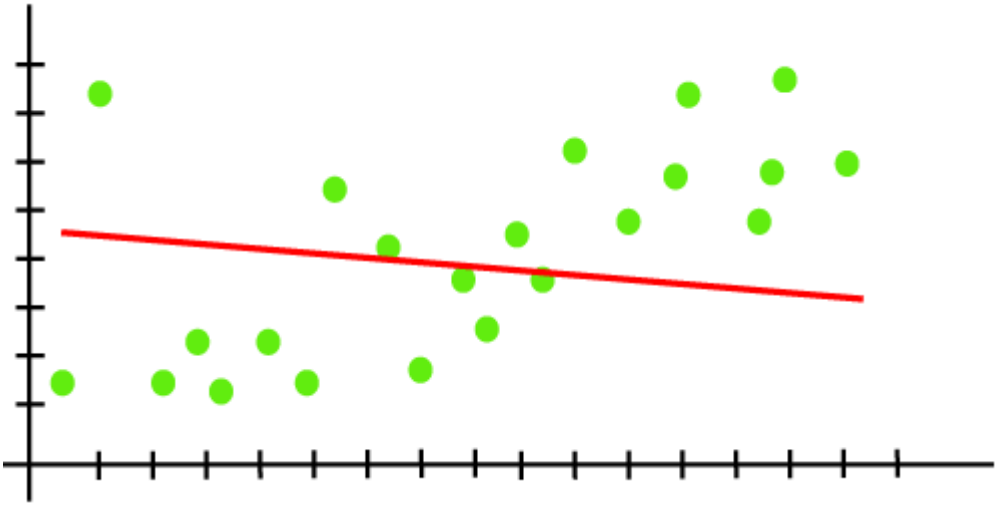
Underfitting

Underfitting occurs when our machine learning model is not able to capture the underlying trend of the data. To avoid the overfitting in the model, the fed of training data can be stopped at an early stage, due to which the model may not learn enough from the training data. As a result, it may fail to find the best fit of the dominant trend in the data.

In the case of underfitting, the model is not able to learn enough from the training data, and hence it reduces the accuracy and produces unreliable predictions.

An underfitted model has high bias and low variance.

Example: We can understand the underfitting using below output of the linear regression model:



As we can see from the above diagram, the model is unable to capture the data points present in the plot.

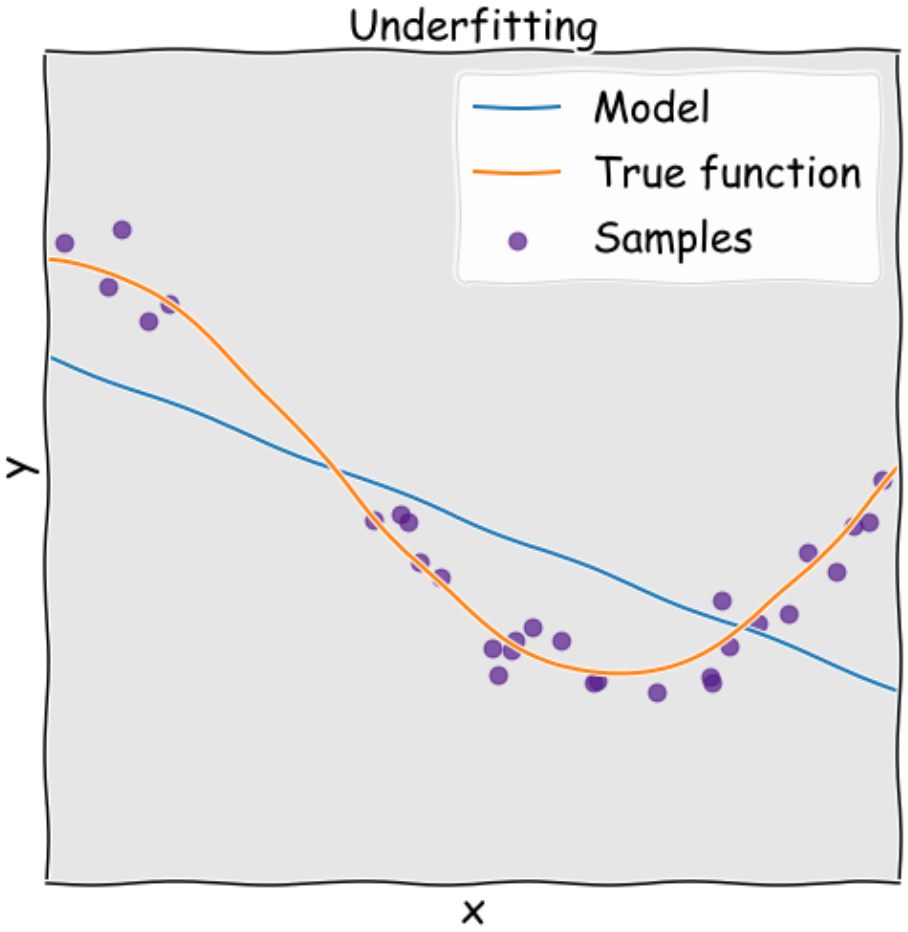
How to avoid underfitting:

- By increasing the training time of the model.
- By increasing the number of features.

Goodness of Fit

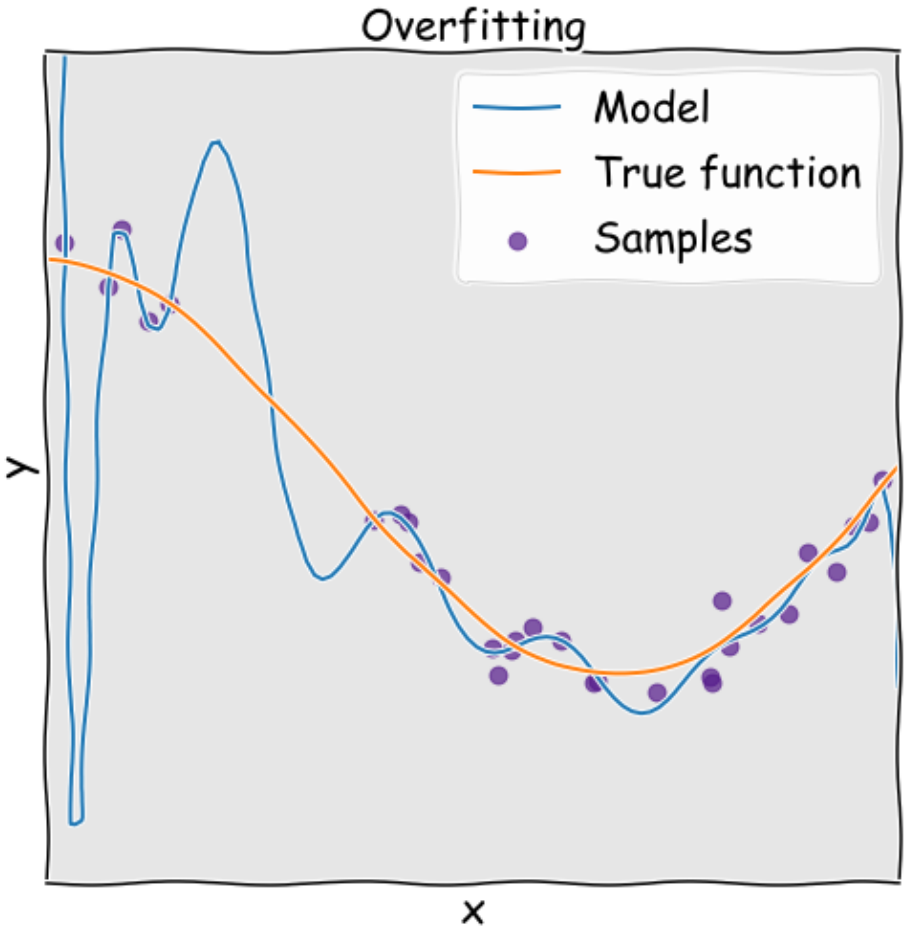
1. The "Goodness of fit" term is taken from the statistics, and the goal of the machine learning models to achieve the goodness of fit. In statistics modeling, *it defines how closely the result or predicted values match the true values of the dataset.*
2. The model with a good fit is between the underfitted and overfitted model, and ideally, it makes predictions with 0 errors, but in practice, it is difficult to achieve it.
3. As when we train our model for a time, the errors in the training data go down, and the same happens with test data. But if we train the model for a long duration, then the performance of the model may decrease due to the overfitting, as the model also learn the noise present in the dataset. The errors in the test dataset start increasing, *so the point, just before the raising of errors, is the good point, and we can stop here for achieving a good model.*
4. Bias and variance are two terms you need to get used to if constructing statistical models, such as those in machine learning. There is a tension between wanting to construct a model which is complex enough to capture the system that we are modelling, but not so complex that we start to fit to noise in the training data. This is related to underfitting and overfitting of a model to data, and back to the bias-variance tradeoff.
5. If we have an underfitted model, this means that we do not have enough parameters to capture the trends in the underlying system. Imagine for example that we have data that is parabolic in nature, but we try to fit this with a linear function, with just one parameter. Because the function does not have the required complexity to fit the data (two parameters), we end up with a poor predictor. In this case the model will have high **bias**. This means that we will get consistent answers, but consistently wrong answers.

An example of underfitting. The model function does not have enough complexity (parameters) to fit the true function correctly.

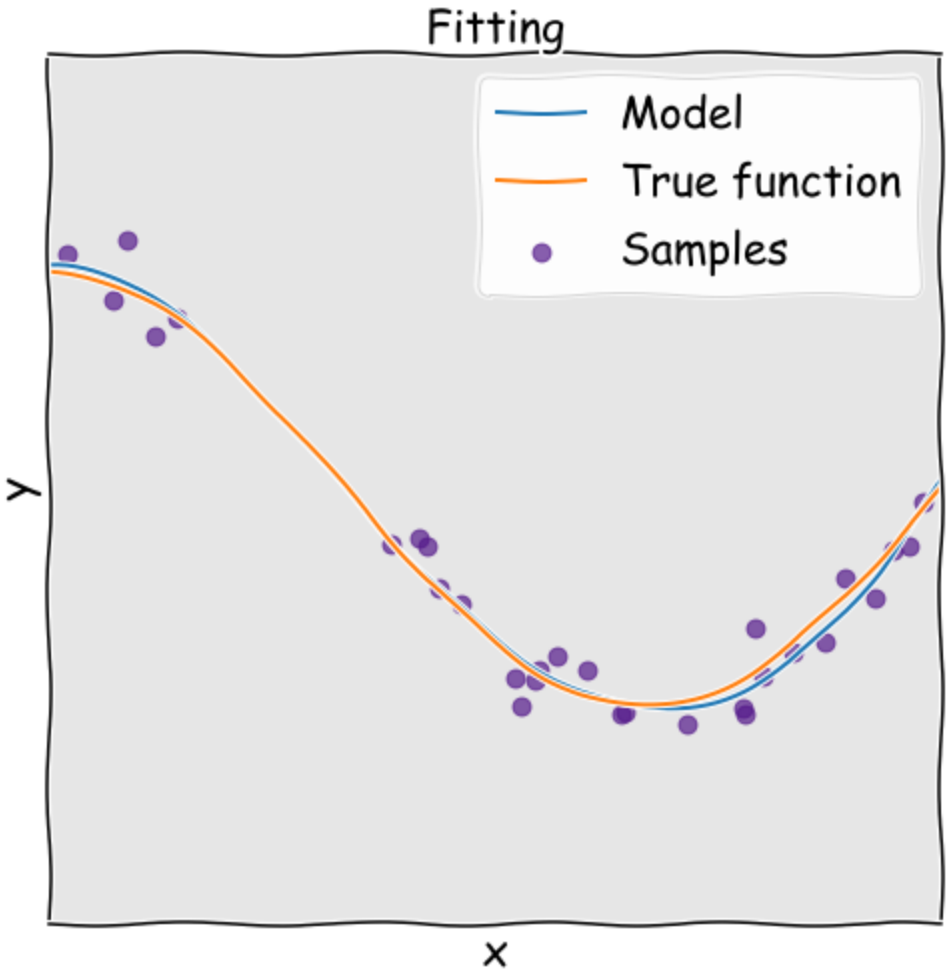


If we have overfitted, this means that we have too many parameters to be justified by the actual underlying data and therefore build an overly complex model. Again imagine that the true system is a parabola, but we used a higher order polynomial to fit to it. Because we have natural noise in the data used to fit (deviations from the perfect parabola), the overly complex model treats these fluctuations and noise as if they were intrinsic properties of the system and attempts to fit to them. The result is a model that has high **variance**. This means that we will not get consistent predictions of future results. For a striking and devastating example of the dangers of overfitting,

An example of overfitting. The model function has too much complexity (parameters) to fit the true function



In order to find the optimal complexity we need to carefully train the model and then **validate** it against data that was unseen in the training set. The performance of the model against the validation set will initially improve, but eventually suffer and dis-improve. The inflection point represents the optimal model. To illustrate this process below I have the Python code required to build a model


Fitting
Model
True function
Samples
y
x