MACHINE LEARNING

Generalization
Overfitting & Underfitting

Generalization

- Generalization refers to how well the concepts learned by a machine learning model apply to specific examples not seen by the model when it was learning.
- The goal of a good machine learning model is to generalize well from the training data to any data from the problem domain. This allows us to make predictions in the future on data the model has never seen.
- There is a terminology used in machine learning when we talk about how well a machine learning model learns and generalizes to new data, namely overfitting and underfitting.
- ***Overfitting and underfitting*** are the two biggest causes for poor performance of machine learning algorithms.

Generalization

- A model is said to be a good machine learning model if it generalizes any new input data from the problem domain in a proper way.
- This helps us to make predictions about future data, that the data model has never seen. Now, suppose we want to check how well our machine learning model learns and generalizes to the new data.
- For that, we have overfitting and underfitting, which are majorly responsible for the poor performances of the machine learning algorithms.

Overfitting and Underfitting

- 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 under fitting 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.

Bias & Variance in Machine Learning

- Bias: Bias refers to the error due to overly simplistic assumptions in the learning algorithm.
- These assumptions make the model easier to comprehend and learn but might not capture the underlying complexities of the data. It is the error due to the model's inability to represent the true relationship between input and output accurately.
- When a model has poor performance both on the training and testing data means high bias because of the simple model, indicating underfitting.
- Variance: Variance, on the other hand, is the error due to the model's sensitivity to fluctuations in the training data. It's the variability of the model's predictions for different instances of training data.
- High variance occurs when a model learns the training data's noise and random fluctuations rather than the underlying pattern.
- As a result, the model performs well on the training data but poorly on the testing data, indicating overfitting.

Model Underfitting in Machine Learning

- Underfitting occurs when your model is too simple for your data. overfitting occurs when your model is too complex for your data.
- Underfitting, In this case, train error is large and val/test error is large too.
- Overfitting means that your model makes not accurate predictions. In this case, train error is very small and val/test error is large

Under fitting in Machine Learning

- A machine learning algorithm is said to have underfitting when a model is too simple to capture data complexities.
- It represents the inability of the model to learn the training data effectively result in poor performance both on the training and testing data. In simple terms, an underfit model's are inaccurate, especially when applied to new, unseen examples.
- It mainly happens when we uses very simple model with overly simplified assumptions.
- To address underfitting problem of the model, we need to use more complex models, with enhanced feature representation, and less regularization.
- **Note: The under fitting model has High bias and low variance.**

Reasons for Under fitting

- The model is too simple, So it may be not capable to represent the complexities in the data.
- The input features which is used to train the model is not the adequate representations of underlying factors influencing the target variable.
- The size of the training dataset used is not enough.
- Excessive regularization are used to prevent the overfitting, which constraint the model to capture the data well.
- Features are not scaled.
- Data used for training is not cleaned and contains noise (garbage values) in it
- The model has a high bias
- The model is too simple

Techniques to Reduce Under fitting

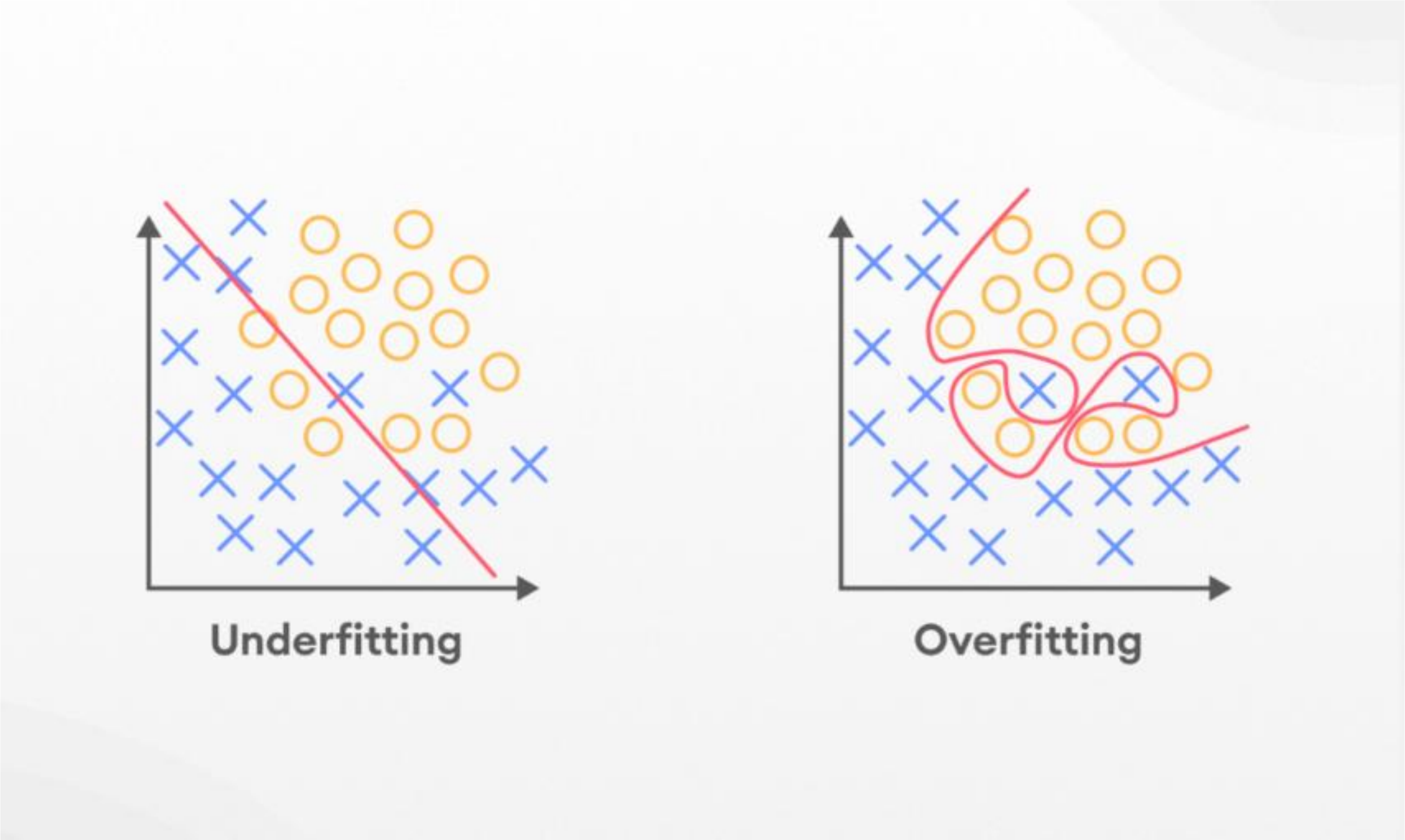
- Increase model complexity.
- Increase the number of features, performing feature engineering.
- Remove noise from the data.
- Increase the number of epochs or increase the duration of training to get better results.
- Increase the duration of training the data

Overfitting in Machine Learning

- A statistical model is said to be overfitted when the model does not make accurate predictions on testing data. When a model gets trained with so much data, it starts learning from the noise and inaccurate data entries in our data set.
- And when testing with test data results in High variance. Then the model does not categorize the data correctly, because of too many details and noise.
- The causes of overfitting are the non-parametric and non-linear methods because these types of machine learning algorithms have more freedom in building the model based on the dataset and therefore they can really build unrealistic models.
- In a nutshell, Overfitting is a problem where the evaluation of machine learning algorithms on training data is different from unseen data.

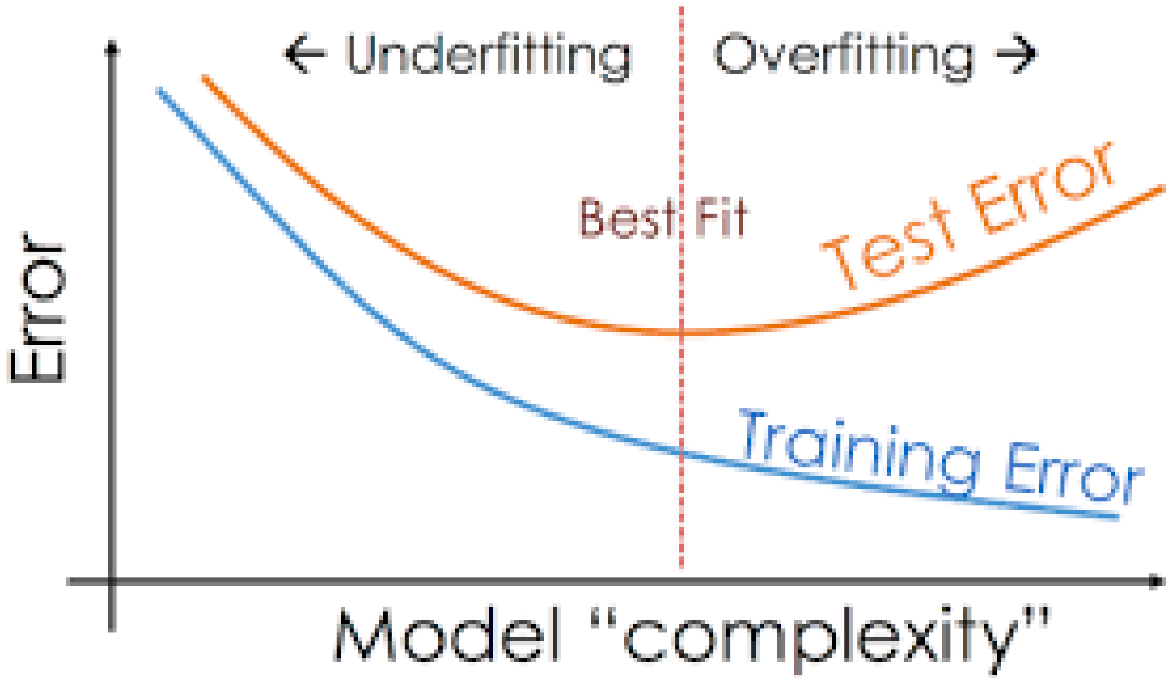
Reasons for Overfitting:

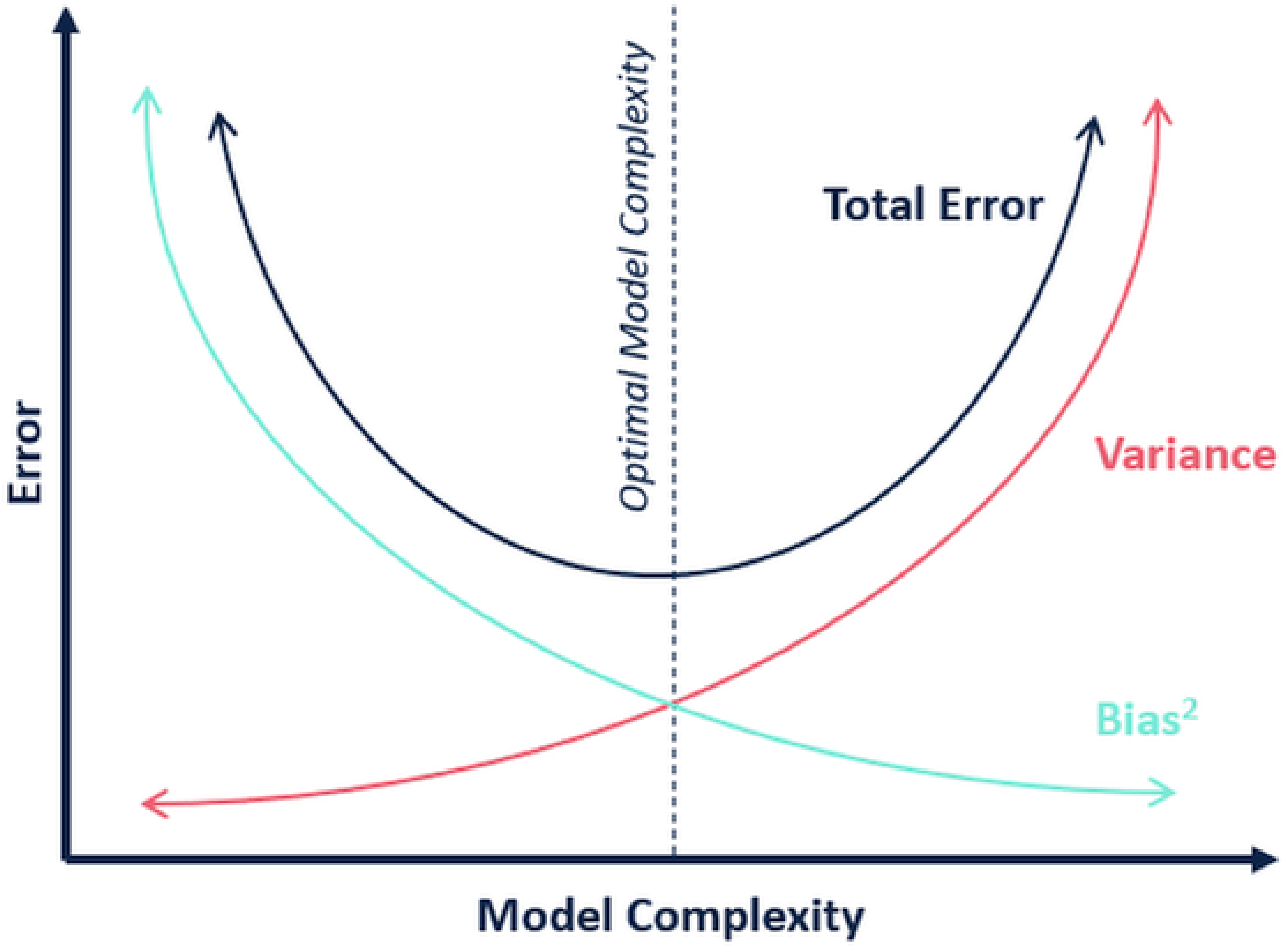
- The model is too complex.
- Data used for training is not cleaned and contains noise (garbage values) in it
- The model has a high variance
- The size of the training dataset used is not enough


Underfitting
Overfitting

Techniques to Reduce Overfitting

- Increase training data.
- Reduce model complexity.
- Early stopping during the training phase (have an eye over the loss over the training period as soon as loss begins to increase stop training).
- Ridge Regularization and Lasso Regularization.
- Use dropout for neural networks to tackle overfitting.


← Underfitting
Overfitting →
Best Fit
Test Error
Training Error
Error
Model "complexity"

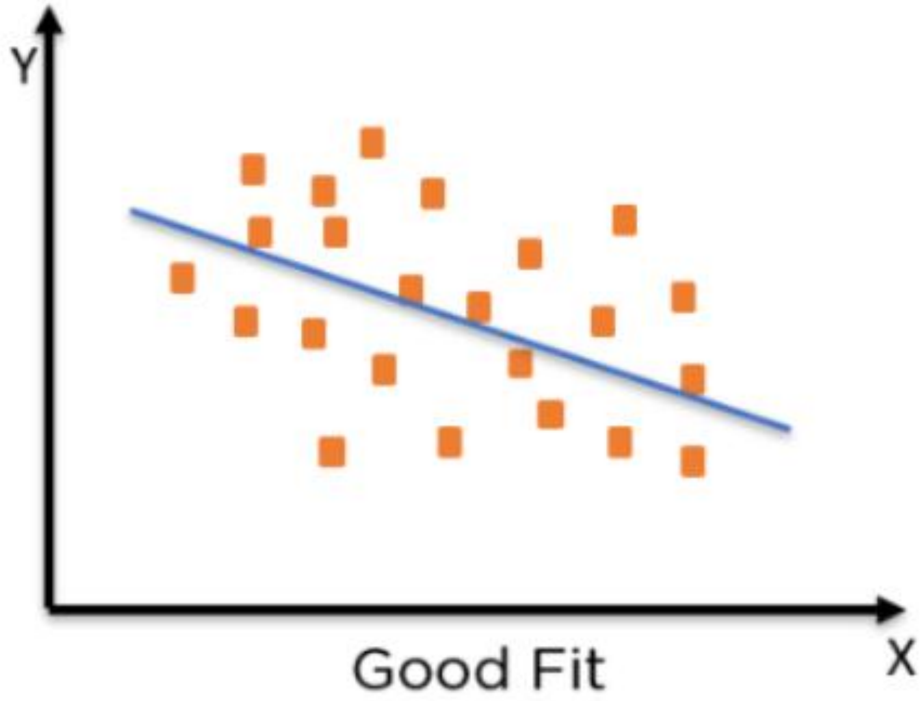

Error
Optimal Model Complexity
Total Error
Variance
Bias^2
Model Complexity

What Is a Good Fit In Machine Learning?

- To find the good fit model, you need to look at the performance of a machine learning model over time with the training data.
- As the algorithm learns over time, the error for the model on the training data reduces, as well as the error on the test dataset.
- If you train the model for too long, the model may learn the unnecessary details and the noise in the training set and hence lead to overfitting.
- In order to achieve a good fit, you need to stop training at a point where the error starts to increase.

Good Fit in a Statistical Model

- Ideally, the case when the model makes the predictions with 0 error, is said to have a good fit on the data.
- This situation is achievable at a spot between overfitting and underfitting. In order to understand it, we will have to look at the performance of our model with the passage of time, while it is learning from the training dataset.
- With the passage of time, our model will keep on learning, and thus the error for the model on the training and testing data will keep on decreasing.
- If it will learn for too long, the model will become more prone to overfitting due to the presence of noise and less useful details. Hence the performance of our model will decrease.
- In order to get a good fit, we will stop at a point just before where the error starts increasing. At this point, the model is said to have good skills in training datasets as well as our unseen testing dataset.


Y
X
Good Fit

Thank You!