**Underfitting and Overfitting in Machine Learning**

Let us consider that we are designing a machine learning model. 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 in the future data, that 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.

Before diving further let’s understand two important terms:

Bias – Assumptions made by a model to make a function easier to learn.  
Variance – If you train your data on training data and obtain a very low error, upon changing the data and then training the same previous model you experience high error, this is variance.

**Underfitting:**  
A statistical model or a machine learning algorithm is said to have underfitting when it cannot capture the underlying trend of the data. *(It’s just like trying to fit undersized pants!)* Underfitting destroys the accuracy of our machine learning model. Its occurrence simply means that our model or the algorithm does not fit the data well enough. It usually happens when we have less data to build an accurate model and also when we try to build a linear model with a non-linear data. In such cases the rules of the machine learning model are too easy and flexible to be applied on such minimal data and therefore the model will probably make a lot of wrong predictions. Underfitting can be avoided by using more data and also reducing the features by feature selection.

In a nutshell, **Underfitting – High bias and low variance**

Techniques to reduce underfitting :  
1. Increase model complexity  
2. Increase number of features, performing feature engineering  
3. Remove noise from the data.  
4. Increase the number of epochs or increase the duration of training to get better results.

**Overfitting:**  
A statistical model is said to be overfitted, when we train it with a lot of data *(just like fitting ourselves in oversized pants!)*. When a model gets trained with so much of data, it starts learning from the noise and inaccurate data entries in our data set. 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. A solution to avoid overfitting is using a linear algorithm if we have linear data or using the parameters like the maximal depth if we are using decision trees.

In a nutshell, **Overfitting – High variance and low bias**