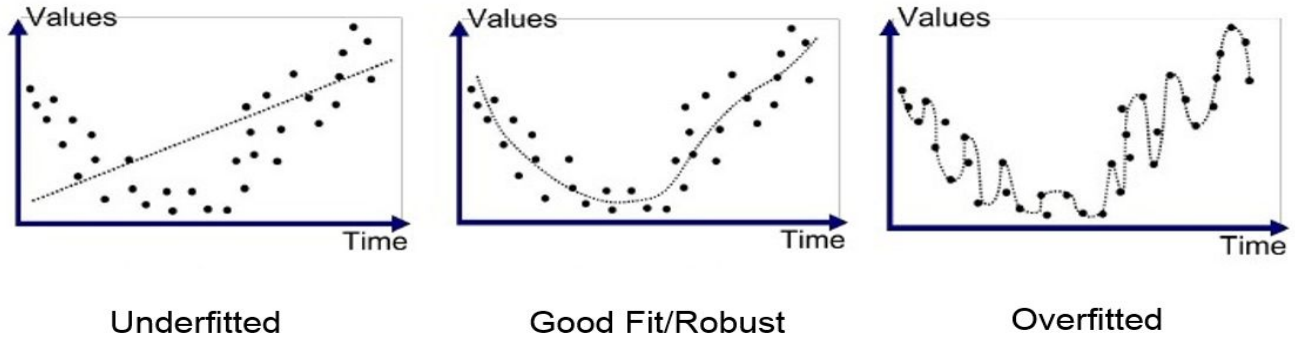
Overfitting And How To Avoid It

A statistical model is said to be overfitted, when we train it with a lot of data, making it relatable, imagine trying to fit into an oversized apparel.When a model fits more data than it actually needs, it starts catching the noisy data and thus cannot categorize data correctly. As a result, the model performs impressively in a training set, but performs poorly in a test set. In a nutshell, **Overfitting – High variance and low bias.**

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.



How to Avoid Overfitting In Machine Learning?

Cross-Validation

It is one of the most efficient techniques used in avoiding overfitting. The idea behind this is to use the initial training data to generate mini train-test-splits, and then use these splits to tune your model. In a standard k-fold validation, the data is partitioned into k-subsets also known as folds. After this, the algorithm is trained iteratively on k-1 folds while using the remaining folds as the test set, also known as holdout fold. The cross-validation helps us to tune the hyperparameters with only the original training set. It basically keeps the test set separately as a true unseen data set for selecting the final model. Hence, avoiding overfitting altogether.

Training With More Data

This technique might not work every time as the quantity of data is one of the causes of overfitting. Thus, in training with more data we have to make sure the data is clean and free from

randomness and inconsistencies. Because in some cases, the increased data can also mean feeding more noise to the model.

Removing Features

Although some algorithms have an automatic selection of features. For a significant number of those who do not have a built-in feature selection, we can manually remove a few irrelevant features from the input features to improve the generalization.

One way to do it is by deriving a conclusion as to how a feature fits into the model. It is quite similar to debugging the code line-by-line.

In case if a feature is unable to explain the relevancy in the model, we can simply identify those features. We can even use a few feature selection heuristics for a good starting point.

Early Stopping

When the model is training, you can actually measure how well the model performs based on each iteration. We can do this until a point when the iterations improve the model's performance. After this, the model overfits the training data as the generalization weakens after each iteration. So basically, early stopping means stopping the training process before the model passes the point where the model begins to overfit the training data. This technique is mostly used in deep learning.

Regularization

It basically means, artificially forcing your model to be simpler by using a broader range of techniques. It totally depends on the type of learner that we are using, for example, add a penalty parameter to the cost function in regression. Quite often, regularization is a hyperparameter as well. It means it can also be tuned through cross-validation.

Ensembling

This technique basically combines predictions from different Machine Learning models. Two of the most common methods for ensembling are listed below:

- Bagging attempts to reduce the chance of overfitting the models.
- Boosting attempts to improve the predictive flexibility of simpler models.

Even though they are both ensemble methods, the approach totally starts from opposite directions. Bagging uses complex base models and tries to smooth out their predictions while boosting uses simple base models and tries to boost its aggregate complexity.

What's the difference between bias and variance?

Bias and variance are used in supervised machine learning, in which an algorithm learns from training data or a sample data set of known quantities.

Bias	**Variance**
Bias deals with how far the predicted values are from the actual values.	Variance tells us how scattered the predicted values are from the actual value.
High bias causes algorithms to miss relevant relationships between input and output variable causing underfitting.	High variance causes overfitting as algorithms models absorb noise.
High bias is identified by High training error and Validation error or test error being the same as training error	High Variance can be identified when we observe Low training error, high validation error or high test error
Errors can be corrected by; ● Adding more input features ● Adding more complexity by introducing polynomial features ● Decreasing Regularization term As the error is due to the model being too simple.	Errors can be corrected by; ● Getting more training data ● Reducing input features ● Increasing Regularization term As the error is due to a model trying to fit most of the training dataset points and hence getting more complex.

When Will You Use Classification over Regression?

Regression is an algorithm in supervised machine learning that can be trained to predict real number outputs. Classification is an algorithm in supervised machine learning that is trained to identify categories and predict in which category they fall for new values.

The key difference between classification and regression is that in classification the dependent variables are categorical and unordered while in regression the dependent variables are continuous or ordered whole values. Thus classification is used over regression when dependent variables are categorical and unordered. An example is when we seek a yes-or-no prediction.