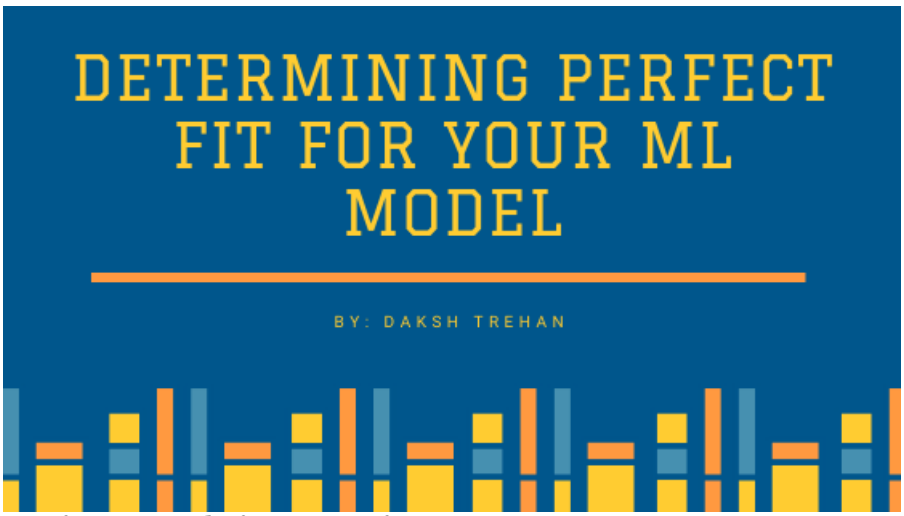
Determining perfect fit for your ML model.

Teaching Overfitting vs Underfitting vs Perfect fit in easiest way.



Overfitting vs Underfitting vs Perfect Fit

Before getting into detailed explanation about each algorithm through medium of series known as "<algorithm_name> Explained ". *It would be better if we could understand the way we want to tune our model.*

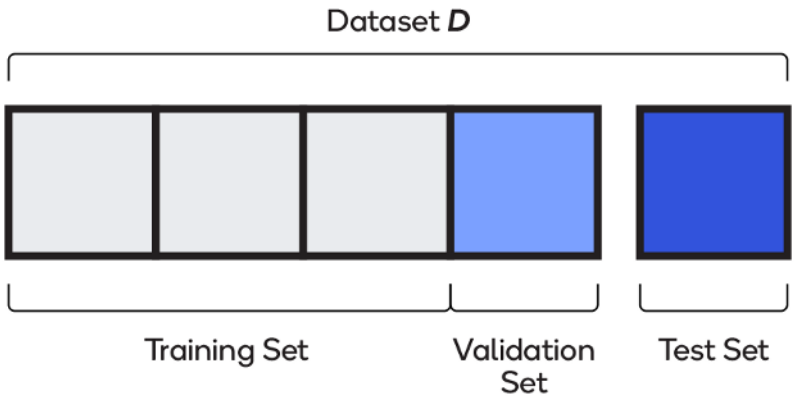
But before everything else, let's recall definition of machine learning.

Machine learning is an ability we provide to computer to learn without being programmed explicitly; too formal? so in laymen words, consider yourself as guardian and your machine as your child; now, as usual, you must teach him how to do different chores? Now either you'll teach your infant through your experience or let him survive based on his ups and downs.

> For more detailed explanation follow "What exactly Machine Learning is?" ,to know more about Data Science follow "Serving Data Science to a Rookie." and to learn how we can relate machine learning to our daily life follow "Relating Machine Learning techniques to Real life."

Accuracy is percentage of how many problems we solved right and loss is the percentage how many problems were wrong from our part.

So, when we mine our data we split it into three parts that is training set, testing set and validation set.



Training set is used to teach our model, using it it tries evaluates the complex function between input and output. Depending upon different algorithms it works in different manner.

For example, when using *Linear Regression*, the points in the training set are used to draw the line of best fit. In *K-Nearest Neighbors*, the points in the training set are the points that could be the neighbors.

Again relating it to real life example; suppose we want our child to learn any concept of math; what we usually do is teach him with help of textbook, but the scope of questions in textbook is limited that's what exactly training data is, it is the limited data on which we can train our model.

Validation set: After training is completed, validation data is used to compute accuracy or error of the classifier. The key insight here is that we know the true labels of every point in the validation set, but we're temporarily going to pretend like we don't. We can use every point in the validation set as input to our classifier. We'll then receive a classification for that point. We can now compare it with the true label of the validation point and see whether we got it right or not. If we do this for every point in the validation set, we can compute the validation error!

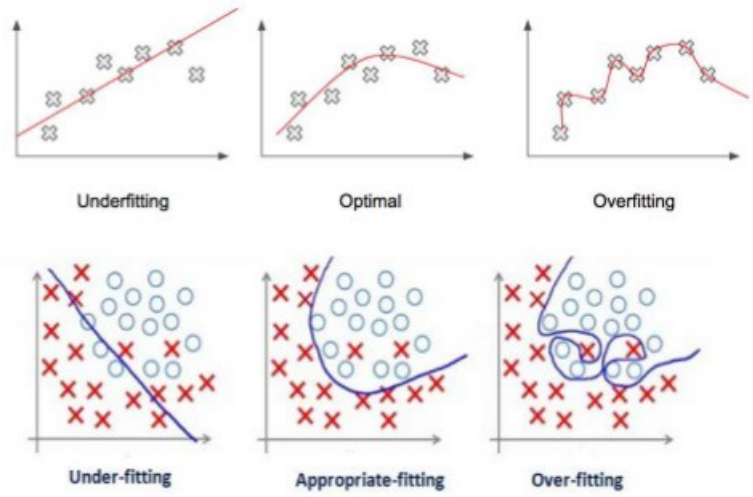
Or it is the the test before the actual exam where we're going to get unseen questions the only difference is here we try to test our child on already taught problems.

Testing set is unseen data, that we use to test our model accuracy. The data is completely unseen by model, but that follows the same probability distribution as the training dataset .Therefore we expect it to predict/classify as correct as possible.

Or it is the questions that might appear in the exam and we expect our child to solve them correctly. Similarly the data in testing data is unseen and we expect our model to generate as more accuracy as possible.

When we develop a machine learning model, we want it to produce greater accuracy on testing data, because that defines how well our model has understood the trained data. The better understanding the better accuracy and a better fit. But what's a better fit?

There are three types of fit that we expect after training our model:



1. **Overfitting**—This is modelling error that occurs when the training done is for particular data or the function is too closely fit to a limited set of data points i.e. it can easily predict, classify or cluster upon the validation data but on unseen data it faces difficulty in producing correct outputs. It is simply when we cramp problems upon which we are taught and then we fail in our exams where we get some twisted or different questions.
2. **Underfitting**—This occurs when the model that can neither model the training data nor generalize to new data. An underfit machine learning model is not a suitable model and will be obvious as it will have poor performance on the training data.
3. **Just right fit**—This is the sweet spot, this is highly desirable and occurs when our model can produce good accuracies on both testing and validation data. This can happen with correct set of all hyperparameters. Relating it to the above example this happens when our child is able to learn as well as score better marks in exams.

Now the question comes, How to differentiate between overfitting and underfitting?

Solving the issue of bias and variance is really about dealing with over-

fitting and under-fitting. Bias is reduced and variance is increased in relation to model complexity. As more and more parameters are added to a model, the complexity of the model rises and variance becomes our primary concern while bias steadily falls.

Lets first learn what is bias, variance and their importance in predicting model.

Bias: It happens to be the tendency of an ML model to consistently learn the wrong relations by not taking in account all the features given for the training.

A ML model with high bias won't be able to learn relations between features effectively and hence would Underfit on the dataset leading to low accuracy while predicting.

It gives us how closeness is our predictive model's to training data after averaging predict value. Generally algorithm has high bias which help them to learn fast and easy to understand but are less flexible. That looses it ability to predict complex problem, so it fails to explain the algorithm bias. This results in underfitting of our model.

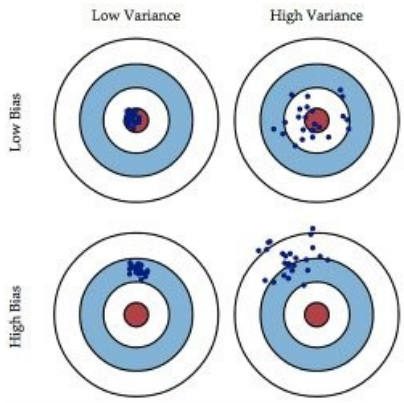
Variance: It's explained as the amount by which the target function changes while it's being trained on data. Alternatively, it's the flexibility of the Model to tune itself with the data points in the given training dataset

An ML model with High variance causes the it to become highly flexible with respect to the data points of the dataset. Such a condition causes a model to Overfit on the training data leading to low accuracy while predicting.

It define as deviation of predictions, in simple it is the amount which tell us when its point data value change or a different data is use how much the predicted value will be affected for same model or for different model respectively. Ideally, the predicted value which we predict from model should remain same even changing from one training data-sets to another, but if the model has high variance then model predict value are affect by value of data-sets.

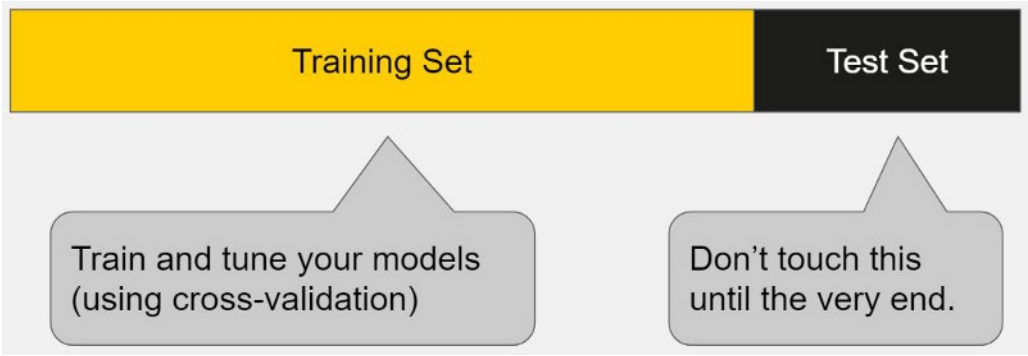
Balancing Bias and Variance also called bias variance Trade-off proves out to be the best way to ensure that model is sufficiently fit on the data

and performs well on new data.



Bias-variance trade off in Machine Learning:

1. Divide data into training, testing and validation split.
2. Start with some configuration.
3. Make sure you're using right hyperparameters.
4. Model training and validation error(don't touch test data)

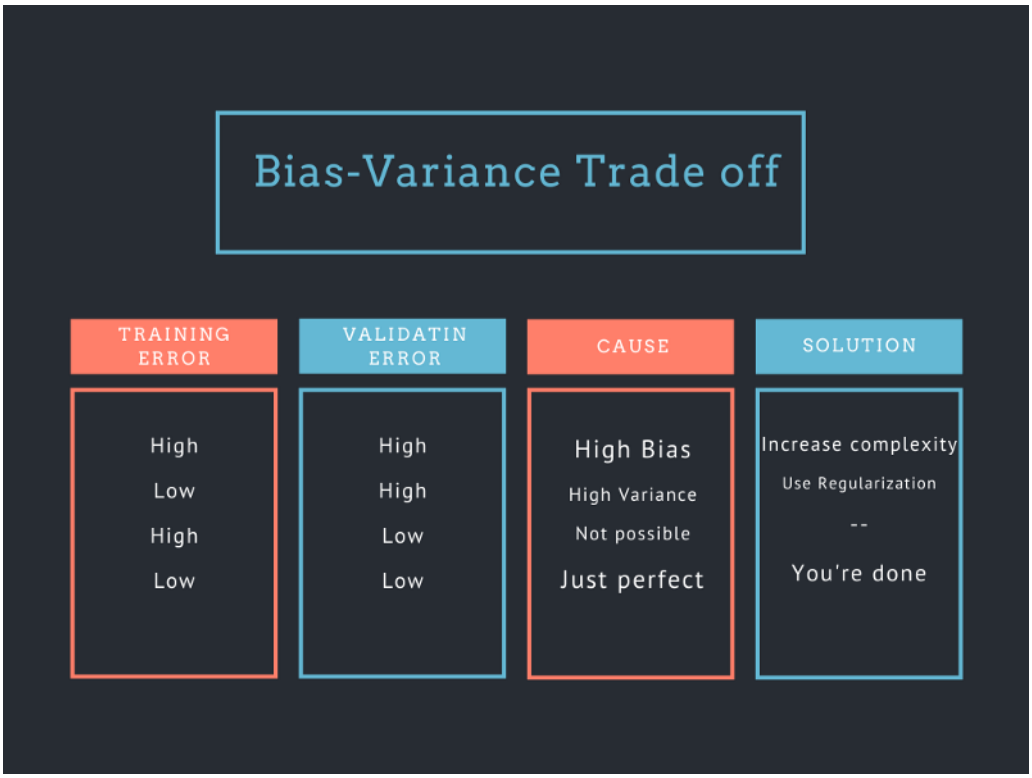


High bias leads to underfitting and high variance leads to overfitting.

Simple model — High bias, Low Variance

Complex model — Low bias, High Variance

Ideal model — Low bias, Low variance

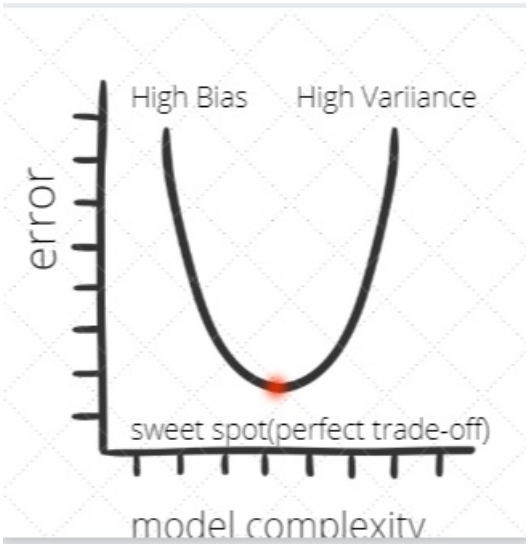


If training error and validation error are high that means we are hit by high bias and it can be corrected by either increasing complexity or training for more epochs.

If training error is low but validation error is high that means we are hit by high variance and it can be corrected by either using Regularization techniques or by early stopping.

If both training error and validation error is low then we got what we wanted, it's a perfect model.

It is not possible to have high training error but low validation error.



Conclusion

Hopefully, this article will help you to tune your model in best way and also assisted you to know more machine learning concepts in better way and even made you realize machine learning is not difficult and is already happening in your daily life.

As always, thanks so much for reading, and please share this article if you found it useful!

Feel free to connect:

LinkedIN ~ https://www.linkedin.com/in/dakshtrehan/

Instagram ~ https://www.instagram.com/_daksh_trehan_/

Github ~ https://github.com/dakshtrehan

Follow for further Machine Learning/ Deep Learning blogs.

Medium ~ https://medium.com/@dakshtrehan

Cheers.

By Daksh Trehan on April 25, 2020.

Canonical link

Exported from Medium on July 15, 2020.