Virtual Internship

Data Science

Article Writing

Name: Model Overfitting and Underfitting

Date: 30/07/2021

Internship Batch: LISUM01

Version: 1.0

Article by: William Ogweli Okomba

Data intake reviewer: Other Interns

Data storage location: https://github.com/williamokomba/DataGlacier_Internship-Data-science/tree/main/Article_Writting

Overfitting and underfitting of the models

Models are meant to predict the unseen data, therefore there is a need to ensure they perform this task well. However, this is not the case some time due to overfitting or underfitting.

I will dive into this topic to try to find out the reasons behind underfitting and overfitting, and how to rectify this problem.

Machine learning models can never make a perfect predictions: the test error is never exactly zero. This failure comes from a fundamental trade-off between modelling flexibility and limited size of the training dataset.

The only way to know whether the model is overfitting or underfitting is by comparing the train and test scores/errors. This means you cannot conclude that a model underfit or overfit because of the accuracy. You need to look at train /test score of the model to come to a conclusion.

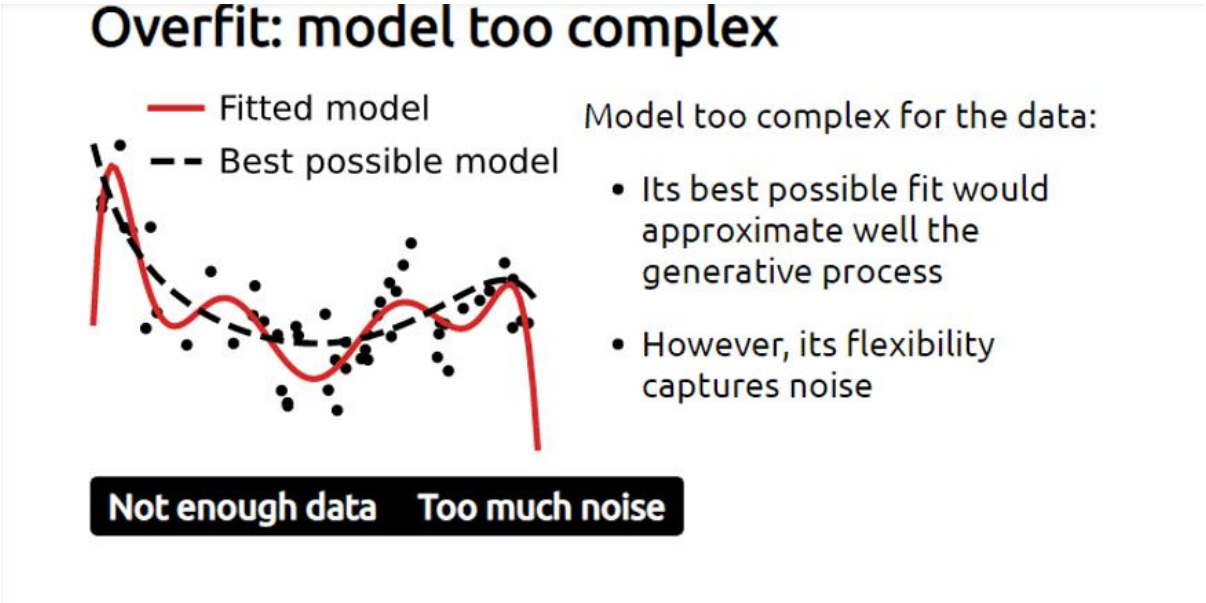
Let us first understand underfitting and overfitting.

1. Overfitting

Overfitting means the model is too complex for the data, it fits well the generative process (on train data); since its flexibility capture noise, this leads to the model performing poorly on test data.

The overfitting can be due to too much noise, and not having enough data.

This can be summarised as below



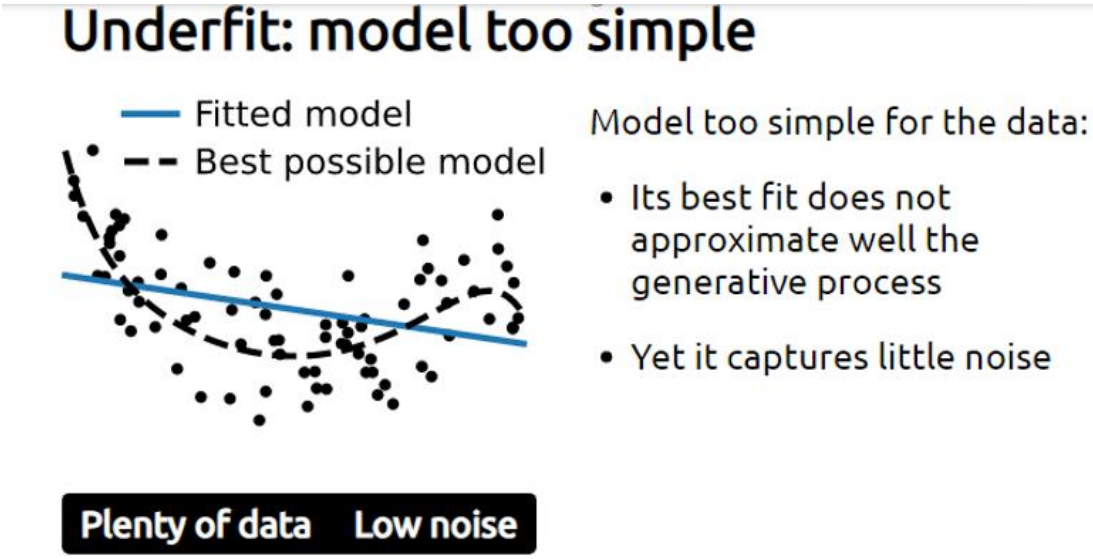
2. Underfitting

Underfitting means the model is too simple for data. Its best fit does not appropriate well the generative process yet it captures little noise.

Here the model does not suffer from the noise because they are too constrained to memorise the noise.

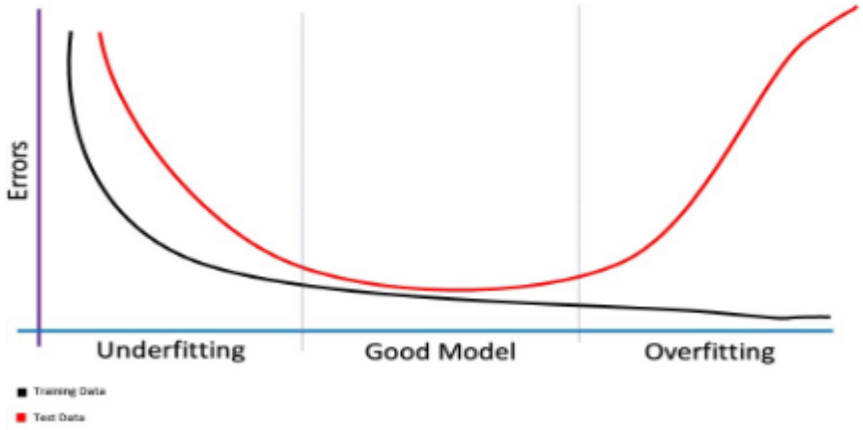
Therefore, unbefitting happens when there is plenty of data, and low noise, but you end up choosing the model that is too constrain for this dataset.

This can be summarised as below:



How to find the right trade-off

1. Underfitting – Validation and training error high
2. Good fit – Validation error low, slightly higher than the training error
3. Overfitting – Validation error is high, training error low



We can see underfitting occurs when both training and test error is high. Overfitting is where training error is low and test error is high.

A good model is when there is trade-off whereby both train and test error is low but test error is slightly higher.

Take Away

1. High bias leads to underfitting.
   i. This results to systematic prediction errors
   ii. The model prefers to ignore some aspect of data
   iii. Misspecified models; cannot adjust to the new model.

The bias can come from the choice of the model family for example using regression model on classification problem.

2. High variance leads t overfitting.

i. Results to prediction errors without obvious structure.

ii. Unstable models, model is too much flexible.

Thanks.

References: MOOC on Machine learning.