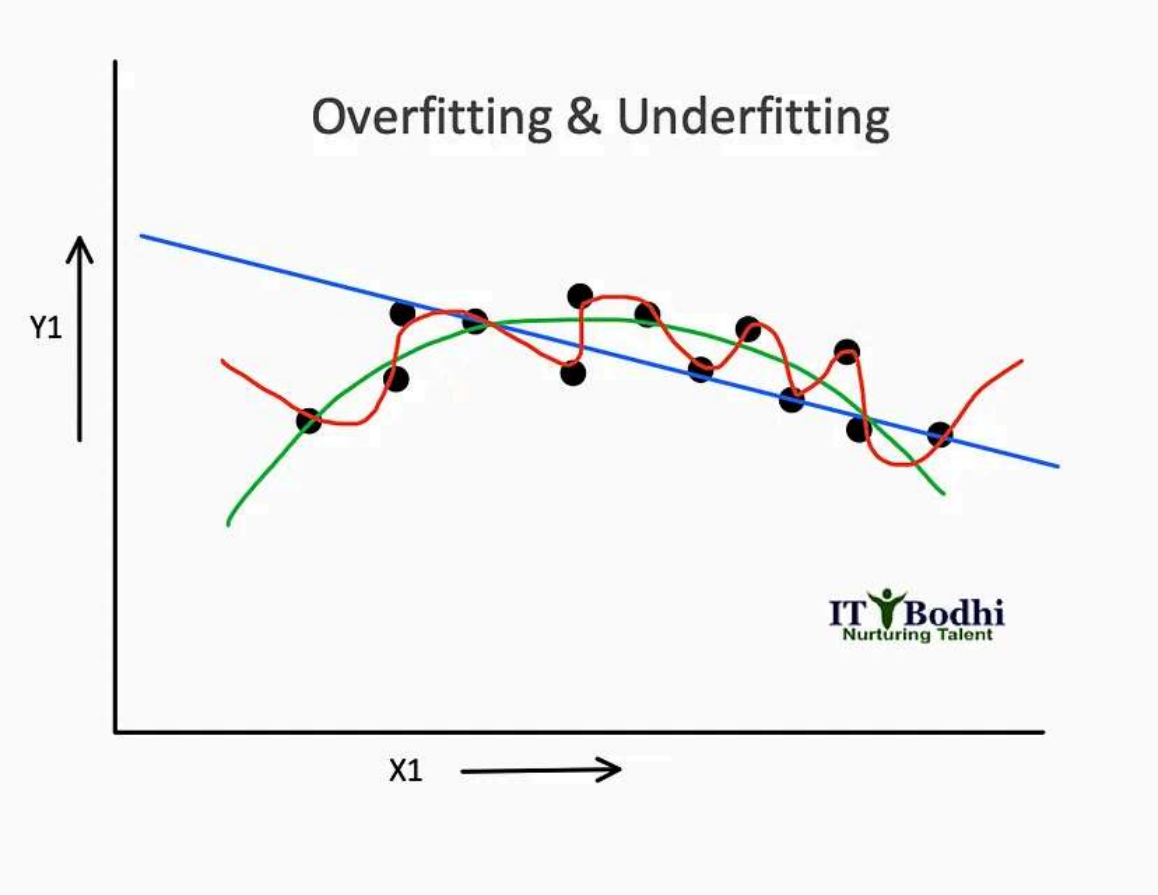
Overfitting and Underfitting

ITBodhi · Follow
6 min read · Jul 2, 2020



In Machine Leaning, model performance is evaluated on the basis of two important parameters. ***Accuracy and Generalisation.*** Accuracy means how well model predicts the right target value and Generalisation means how well model behaves on seen and unseen data.

Machine learning models are trained on given training data and performance is evaluated on the unseen test data. **Model is considered right when it behaves nearly same way on training and test data with highest accuracy.**

What is Underfitting?

> *Underfitting means model has low accuracy score on training data and test data both.*

Underfitting happens when algorithm used to build prediction model is very simple and not able to learn complex pattern from the training data. In that case accuracy will be on lower side on seen training data as well as unseen test data.

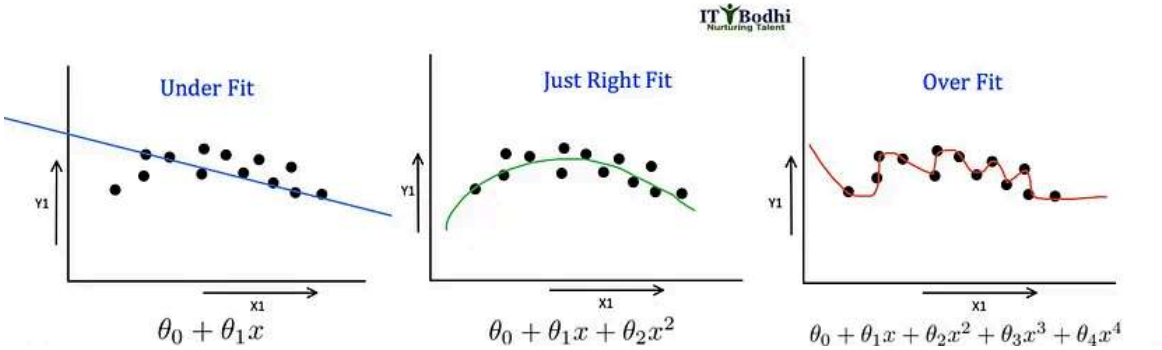
Underfitting also referred as High Bias

> *Check Bias and Variance Trade off*

Generally, It happens with Linear Algorithms. A underfit model makes incorrect assumptions about the dataset to make the target function easier to learn. If training data distribution is non linear and you apply linear algorithm to build the prediction model, in that case model would not be able to learn non linear relationship between target value and predictors(features) and accuracy will suffer.

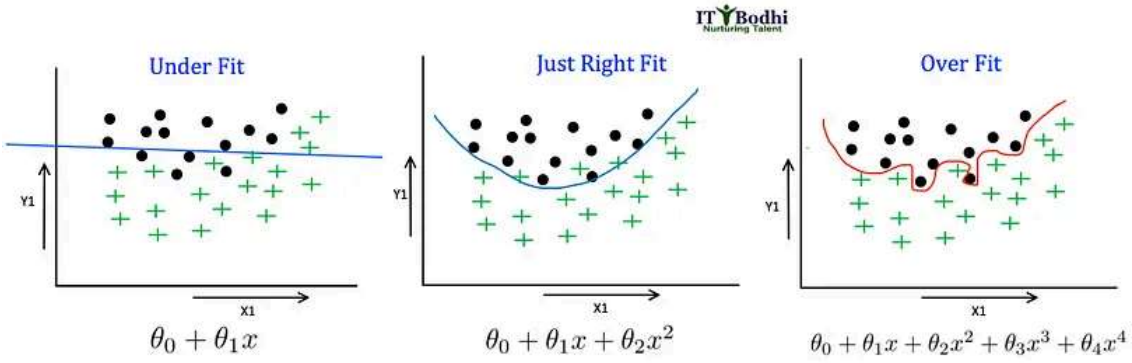
Underfit Model is very simple because of the assumptions made about the data(which may or may not be true) ***pays very little attention to the training data and oversimplifies the model. High Bias model always leads to high error on training as well as on test data.***

In a simple language, your model is very simple and not able to learn the complex pattern (or relationship in data) from the data to make right prediction.



Picture1 — Regression Example for Overfitting and Underfitting

Regression example ***Picture1*** — Regression Example for Overfitting and Underfitting, first Image represents model is Underfit. Data given for training is non linear and you are applying linear regression on the data for modelling. Model has high error because maximum of the data points are far from the line so not able to capture the existing pattern from the data.In this case model built is not right and will have low prediction accuracy.



Picture2 — Classification Example for Overfitting and Underfitting

Classification example *Picture2* — Regression Example for Overfitting and Underfitting, first Image represents model is Underfit. Data given for training is non linear and you are trying linear model for classification. When data distribution is non linear than linear model (in above example a Line) can not work as decision boundry to classify data in to different classes. Model has high error because many of the data points classified incorrectly so will have low prediction accuracy.

What is Overfitting?

> *Overfitting means model has High accuracy score on training data but low score on test data.*

Overfitting means your model is not Generalised.

Overfitting happens when algorithm used to build prediction model is very

data.Overfitting is error from sensitivity to small fluctuations in the training set. Overfitting can cause an algorithm to model the random noise in the training data, rather than the intended result.

Underfitting also referred as High Variance

> *Check Bias and Variance Trade off*

In supervised learning, **overfitting** happens when algorithms(Non Linear Algorithms) are strongly influenced by the specifics of the training data and try to learn patterns which are noisy and not generalized and only limited to training data set.

> In a simple language, your model is very complex and strongly influenced by the training data and will result in high score on training data but low score on unseen test data.

Regression example ***Picture-1***, third Image represents model is Overfit. Data given for training is non linear and you are using higher degree features for modelling. As you can observe liner model (line in above case) has been curved too much to go through each and every data points. In this case there will be minimum error on training data because model has over learned the patterns. But on test data when data distribution will differ from

the training data, prediction error will increase because line is curved to suit the training data only and not generalised to fit the unseen data.

Classification example ***Picture-2***, Third Image represents model is Overfit. Data given for training is non linear and you are trying linear model for classification. As you can observe, because of higher degree features, decision boundary has curved too much to classify each and every training example correctly. In this case there will be minimum error on training data because model has over learned the patterns and resulting in to zero classification error. But on test data when data distribution will differ from the training data, prediction error will increase because line is curved to suit the training data only and not generalised to classify the unseen data.

What is Right Fit?

Model is considered to be Right Fit when it is generalised and behaves more or less in the same way as on training data and on test data.

Model accuracy should be almost same on training and test data.

Right fit model is neither Underfit and nor Overfit, it is a generalised model that does not change for seen and unseen data. Middle image in Picture1–2 is right fit model.

> Practically, Overfitting and Underfitting are inverse to each other and you have to take trade of between

> Overfit and Underfit (Bias & variance) to get the right model.

Generally, You can see a general trend in the examples above:

- **Linear** machine learning algorithms often are **Underfit.** *Example:Linear Regression, Logistic Regression*
- **Nonlinear** machine learning algorithms often are **Overfit.** *Example: Decision Tree, SVM, Neural Networks*

How to find the Right Balance?

Lowering high Bias or Underfitting:

1. Use non Parameterised Algorithms

2. Make model more complex with more features

3. Use Non Linear Algorithms *Example(Polynomial Regression, Kernel Function in SVM*

Lowering high Variance or Overfitting:

1. Use More Data for training to make model learn maximum hidden pattern from the training data and model becomes generalised.

2. Use Regularization Techniques *Example: L1 , L2, Drop Out, Early Stopping(in case of Neural Networks)etc.*

3. Hyper Parameter Tuning to avoid Overfitting *Example: Higher value of K in KNN, Tuning of C and Gama for SVM, Depth of Tree in Decision Tree*

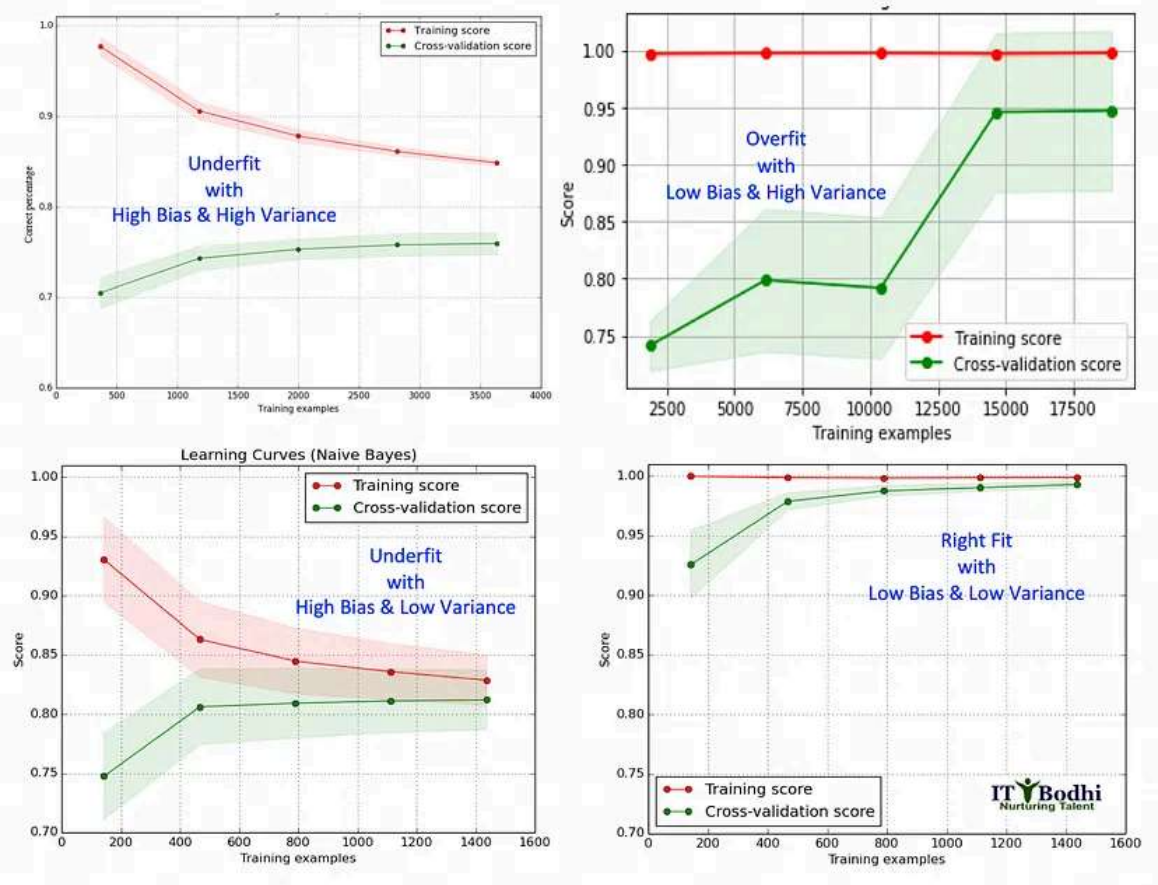
4. Use less number of features — *Manual or Feature Selection Algorithms or automated using L1, L2 Regularization*

5. Reduce complexity of Model — *Reduce polynomial degree in case of Polynomial regression and Logistic regression*

6. Use Advance techniques like Cross Validation, Stratified Cross Validation etc.

> Finding the right balance is an iterative process where model is trained with different combination of features, Hyperparameters, different set of data set for training and test to find the right combination. We stop when we reach the point where Low Bias and Low variance is achieved and model is neither underfit and nor overfit.i.e Prediction accuracy is same on train and test data.

Let's Visualise the things



Picture — 3 Accuracy Score on Training and Testing data during Cross Validation

In above picture we can have visualisation of Model Accuracy while Training/Validating on Train/Validation data.

First Image : Model has low accuracy on train and validation data + Model has difference in accuracy score in train and validation → Underfit with High Bias and High Variance

Second Image : Model has high accuracy on train and validation data + Model has difference in accuracy score in train and validation → Overfit with Low Bias and High Variance

Third Image : Model has high accuracy on train and validation data + Model has little difference in accuracy score in train and validation → RIGHT FIT with Low Bias and Low Variance

Fourth Image : Model has low accuracy on train and validation data + Model has little difference in accuracy score in train and validation → Underfit with High Bias and Low Variance

Thank you for reading this article and I hope it has helped you to understand the concept and will surely help you to make right model.

Happy Reading....

www.itbodhi.com

Overfitting | Bias And Variance | Bias Variance Tradeoff | Machine Learning | Machine Learning Course

Written by ITBodhi

44 Followers · 5 Following

Follow

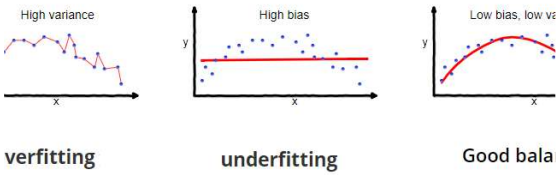
More from ITBodhi

ITBodhi

Handling Imbalanced data sets in Machine Learning

What are the Best Practices, Techniques and Tools to make Right Model with Imbalanced...

Jul 21, 2020 32 1



ITBodhi

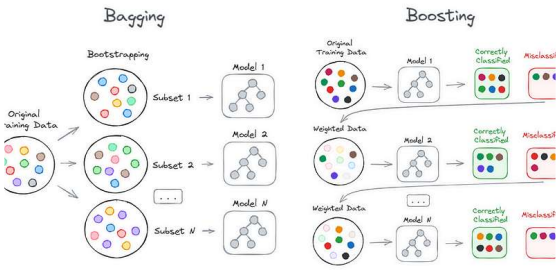
Bias and Variance Trade off

In Machine Learning, when we want to optimise model prediction, it is very...

Jul 1, 2020 266

See all from ITBodhi

Recommended from Medium



In Towards AI by Thomas A Dorfer

Bagging vs. Boosting: The Power of Ensemble Methods in Machine...

How to maximize predictive performance by creating a strong learner from multiple weak...

Jun 16, 2023 580 4

In Biased-Algorithms by Amit Yadav

What is Pruning in Machine Learning

You know, machine learning is all about building powerful models, but bigger isn't...

Sep 20

Lists

Predictive Modeling w/ Python

20 stories · 1700 saves

Practical Guides to Machine Learning

10 stories · 2068 saves

Natural Language Processing

1842 stories · 1466 saves

The New Chatbots: ChatGPT, Bard, and Beyond

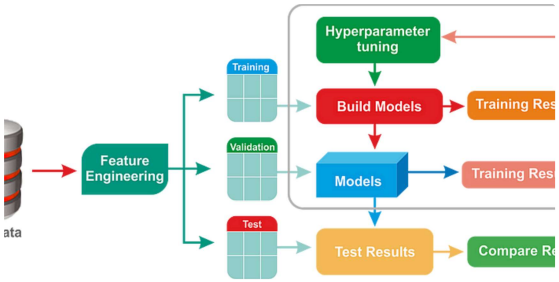
12 stories · 518 saves

Abhishek Jha

Cracking the Code: WOE Encoding and Binning for High Cardinality...

In machine learning, dealing with high-cardinality categorical variables can be trick...

Oct 10 10



In DevOps.dev by Shailendra Prajapati

Hyperparameter Tuning:

The Key to Unlocking Machine Learning Performance

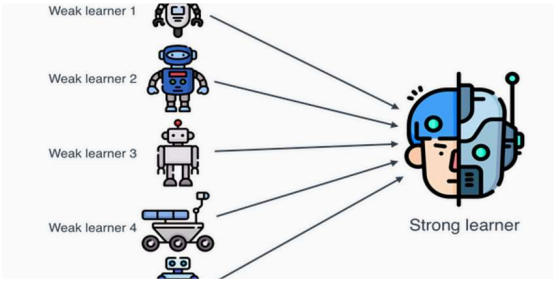
Aug 27

In Python in Plain Engli... by Mounica Kommajosy...

Mastering Model Selection: How to Choose the Right Model for Your...

When you are building a machine learning model, the process does not end after...

Nov 4 2



Avicsebooks

Part14:ML Ensemble models

Ensemble models are a powerful technique in machine learning where multiple models...

Jun 22 3

See more recommendations