Underfitting and Overfitting in Machine Learning

Bias

- **Definition:** Bias refers to the error introduced by approximating a real-world problem (complex or unknown function) with a simplified model. It is an assumption made by the model about the data.
- **Impact:** High bias usually results in **underfitting**, where the model fails to capture the complexity of the data.
- **Example:** A linear regression model trying to capture a quadratic relationship will have high bias because it cannot model the curve.
- **Characteristics:**
  - Models with high bias are too simple.
  - The predictions are consistently wrong in the same way across different datasets.

2. Variance

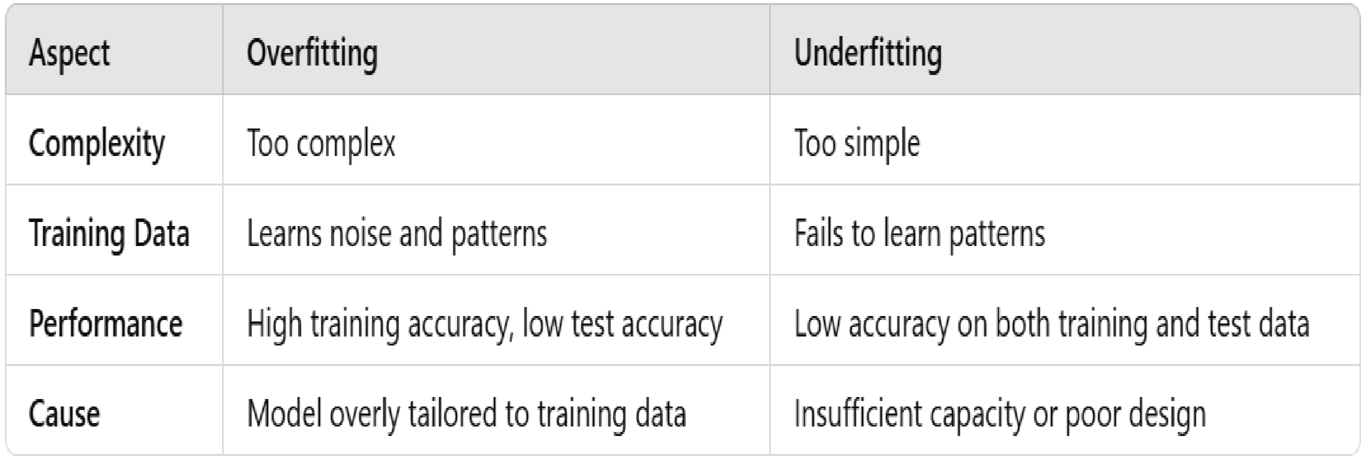
- **Definition:** Variance refers to the variability of model predictions for different training datasets. It captures the model's sensitivity to small changes in the training data.
- **Impact:** High variance usually results in **overfitting**, where the model captures noise in the training data and performs poorly on unseen data.
- **Example:** A very deep decision tree that learns specific patterns and noise in the training data will have high variance.
- **Characteristics:**
  - Models with high variance are overly complex.
  - Predictions change significantly with slight variations in the training data.

- **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.

3. Error

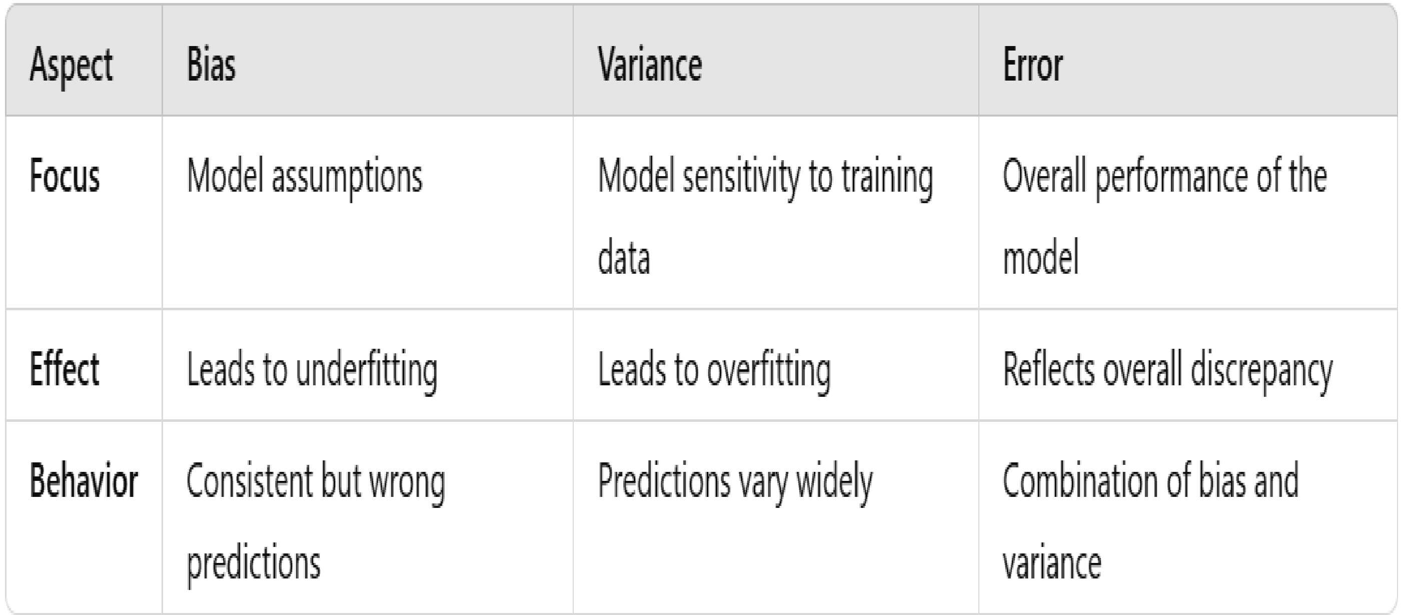
- **Definition:** Error is the overall discrepancy between the model's predictions and the actual values. It includes both bias and variance components.
- **Types:**
  - **Training Error:** The error the model makes on the training dataset.
  - **Testing Error:** The error the model makes on unseen data, which is more critical for evaluating generalization.



Aspect	Overfitting	Underfitting
Complexity	Too complex	Too simple
Training Data	Learns noise and patterns	Fails to learn patterns
Performance	High training accuracy, low test accuracy	Low accuracy on both training and test data
Cause	Model overly tailored to training data	Insufficient capacity or poor design

Bias-Variance Tradeoff

- In practice, there is a tradeoff between bias and variance:
  - Increasing model complexity reduces bias but increases variance.
  - Simplifying the model reduces variance but increases bias.
- The goal is to find a balance where the total error (bias + variance) is minimized, leading to good generalization.



Aspect	Bias	Variance	Error
Focus	Model assumptions	Model sensitivity to training data	Overall performance of the model
Effect	Leads to underfitting	Leads to overfitting	Reflects overall discrepancy
Behavior	Consistent but wrong predictions	Predictions vary widely	Combination of bias and 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.
- 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.

- **Definition:** Underfitting occurs when a model is too simple to capture the underlying structure of the data. It fails to learn adequately from the training data, resulting in poor performance on both the training and test data.
- **Characteristics:**
    - Low training accuracy and low test accuracy.
    - Model is too simple relative to the complexity of the data.
- **Example:** A linear regression model used for data with a nonlinear relationship.

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.

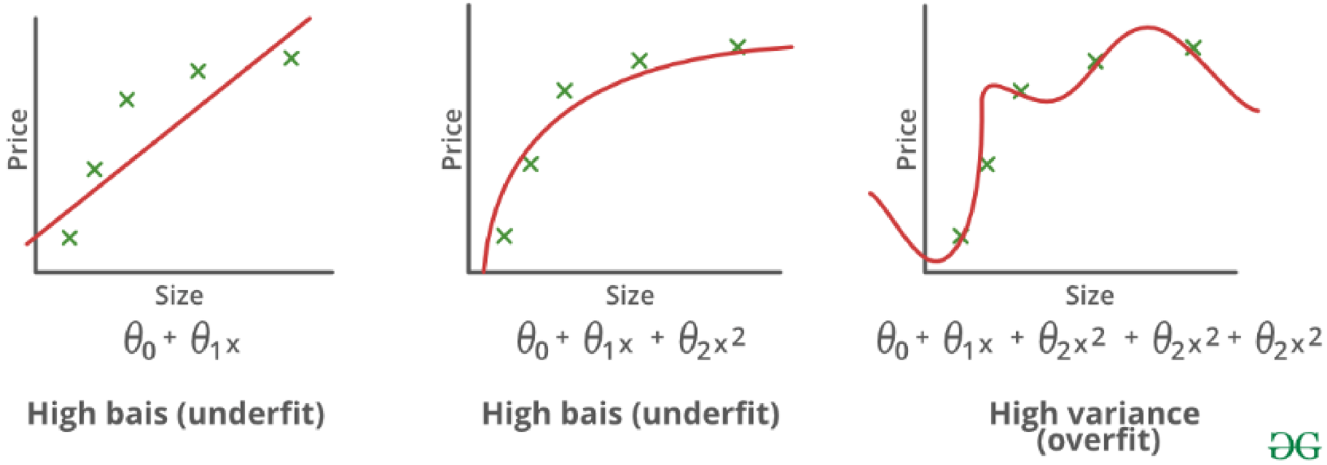
Overfitting:

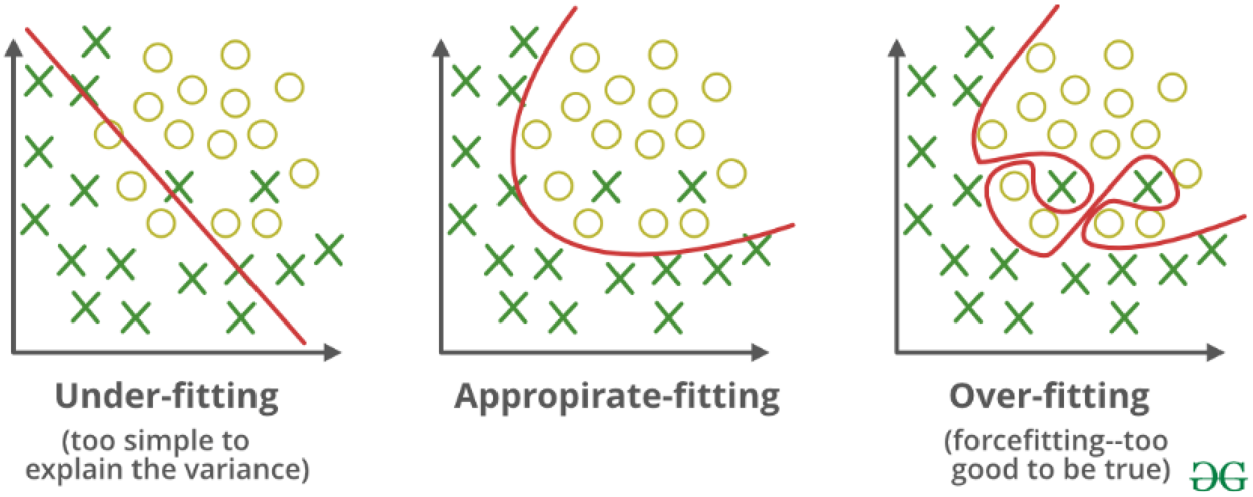
- A statistical model is said to be overfitted, when we train it with a lot of data
- When a model gets trained with so much of data, it starts learning from the noise and inaccurate data entries in our data set.
- Then the model does not categorize the data correctly, because of too many details and 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.

Contd.

- **Definition:** Overfitting occurs when a model learns not only the underlying patterns in the training data but also noise and random fluctuations. This makes the model highly specific to the training data, leading to poor generalization on new, unseen data.
- **Characteristics:**
- High training accuracy but low test accuracy.
- Model is too complex relative to the amount of data.
- **Example:** A deep decision tree that perfectly classifies training data but performs poorly on test data due to capturing noise.

Overfitting – High variance and low bias




Under-fitting
(too simple to explain the variance)
Appropirate-fitting
Over-fitting
(forcefitting--too good to be true)

Techniques to reduce overfitting :

1. Increase training data.
2. Reduce model complexity.
3. Early stopping during the training phase (have an eye over the loss over the training period as soon as loss begins to increase stop training).
4. Ridge Regularization and Lasso Regularization
5. Use dropout for neural networks to tackle overfitting.