1.Variance and Bias (With Diagram, Overfitting & Underfitting Explanation)

Introduction to Bias and Variance

In the field of Machine Learning and Artificial Intelligence, building a model that makes accurate predictions is the primary goal. However, achieving high accuracy is not only about fitting the training data well — it is also about ensuring that the model performs well on unseen data. Two fundamental concepts that explain prediction errors in machine learning models are Bias and Variance.

Understanding bias and variance helps in diagnosing model performance, improving generalization, and selecting the right model complexity.

What is Bias?

Bias refers to the error that occurs when a model makes overly simplified assumptions about the data. It measures how far the model's predictions are from the actual values.

Key Characteristics of Bias:

- Occurs due to overly simple models.
- Leads to underfitting.
- Model fails to capture the underlying patterns in the data.
- High training error and high testing error.

Example:

If we try to fit a straight line (linear model) to a dataset that follows a complex curve, the model will not capture the real pattern. This situation results in high bias.

Causes of High Bias:

- Using a simple algorithm for complex data.
- Ignoring important features.
- Too much regularization

What is Variance?

Variance is the error due to the model being too sensitive to small fluctuations in the training data.

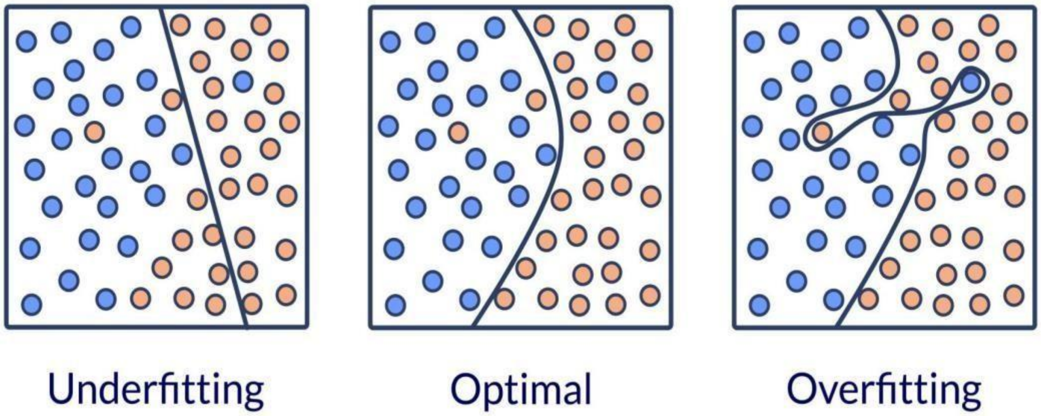
- High variance → Model is too complex
- It memorizes training data
- Leads to overfitting

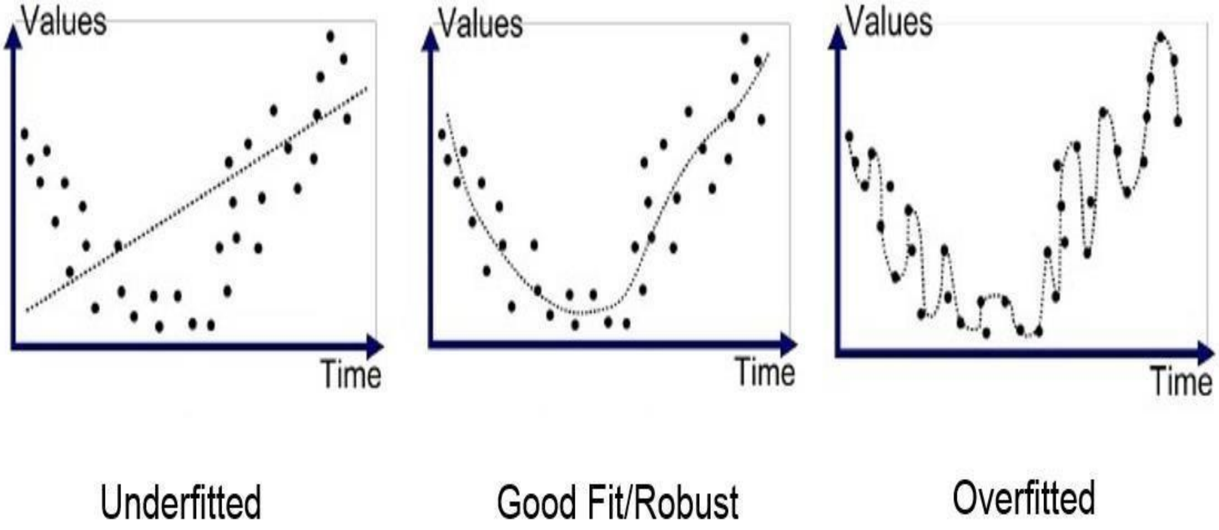
Characteristics of High Variance:

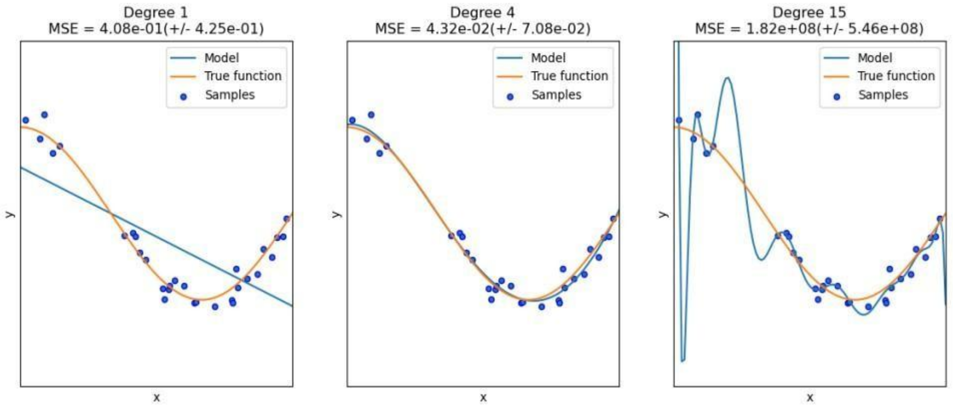
- Very low training error
- Very high testing error
- Model is too complex

Underfitting (High Bias)

s







What is Underfitting?

- Underfitting is a situation in machine learning where a model is unable to properly learn the relationships within the dataset. This usually occurs when the model lacks sufficient complexity to represent the actual patterns in the data.

Why Does It Happen?

Underfitting generally arises because the model has:

• Excessive bias

• Very low variance

Since the model is too basic, it makes strong assumptions and ignores important details in the dataset.

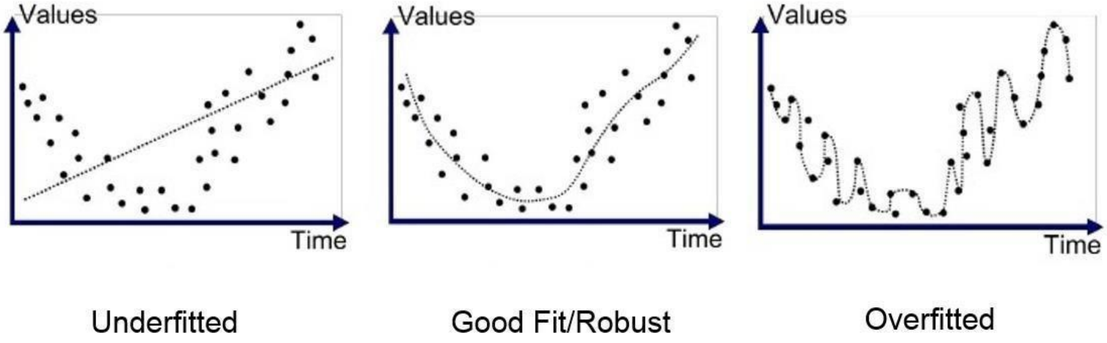
Illustrative Example

For instance, if a straight-line regression model is applied to data that follows a curved pattern, the model will fail to represent the true relationship between variables.

What Is the Impact?

As a result, the model performs poorly not only on new (test) data but also on the training data, indicating that it has not learned effectively

Overfitting (High Variance)



What is Overfitting?

Overfitting is a condition in machine learning where a model becomes excessively tailored to the training dataset. Instead of learning the general pattern, the model memorizes random fluctuations and noise present in the training data.

Why Does It Occur?

Overfitting typically happens when the model has:

- Very low bias
- Very high variance

The model becomes too complex and overly sensitive to small changes in the dataset.

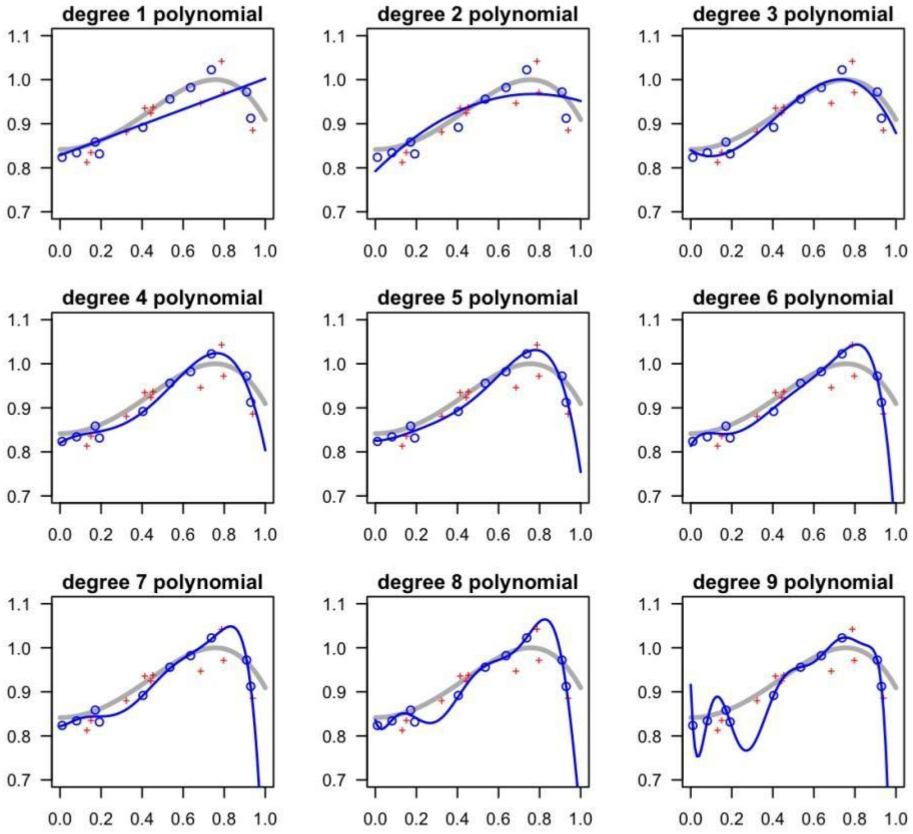
Example

For example, fitting a very high-degree polynomial to a dataset that could be explained with a simple curve may cause the model to perfectly match training points while failing to generalize.

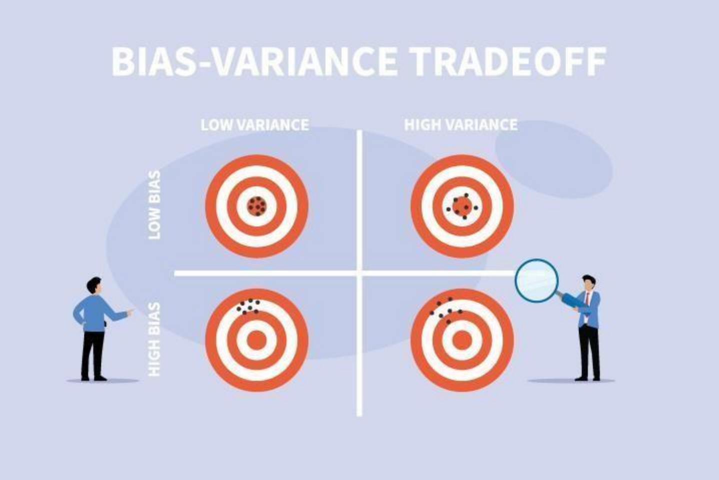
Consequences

- Extremely low error on training data
- Significantly higher error on unseen (test) data

This indicates that the model does not generalize well to new information



Bias–Variance Trade-off



There is a trade-off between bias and variance.

- As model complexity increases: ○ Bias decreases ○ Variance increases
- The goal is to find a balance.

What Characteristics Should an Ideal Model Have?

When selecting the most suitable machine learning model, we must determine the right combination of bias and variance. Consider the following possibilities:

- Low bias and high variance
- Low bias and low variance
- High bias and high variance
- High bias and low variance

Correct Choice

The ideal or best-fit model should have:

Low Bias and Low Variance

Why Is This the Best Combination?

- Low Bias ensures that the model accurately learns the actual relationship within the data.
- Low Variance ensures that the model performs consistently well on new, unseen data.

A model with these properties neither underfits nor overfits. Instead, it strikes a balance between simplicity and complexity.

Bias–Variance Trade-Off Perspective

On the bias–variance trade-off curve, this optimal model lies at the center point where total prediction error is minimized. It represents the perfect balance between learning the pattern correctly and maintaining strong generalization ability

Summary Table

Model Type	**Bias**	**Variance**	**Problem**
Underfitting	High	Low	Too simple
Overfitting	Low	High	Too complex
Best Fit Model	Low	Low	Balanced

Conclusion

In machine learning, prediction errors mainly arise due to two factors: bias and variance. These two elements directly influence how well a model performs on both training and unseen data.

- When bias is high, the model becomes too simple and results in underfitting.
- When variance is high, the model becomes too complex and results in overfitting.
- An effective model carefully balances both bias and variance.

Hence, an optimal or best-fit model is one that achieves low bias while also maintaining low variance, ensuring accurate learning and strong generalization ability.