INDIVIDUAL TASK 1

Variance and Bias (Diagram, overfit, underfit) For best fit model should high we have low bias or high variance, low bias or low variance, high bias or variance, low bias or high variance.

INTRODUCRION :

Bias and Variance – Short Summary

1. Bias :

Bias is error due to overly simple assumptions in the model.

High bias → model is too simple.

Leads to underfitting.

Both training and testing errors are high.

Example: Fitting a straight line to curved data.

2. Variance :

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

High variance → model is too complex.

Leads to overfitting.

Training error is very low, but testing error is high.

Example: A very complex curve that passes through every data point.

3. Underfitting

Caused by high bias and low variance.

Model cannot capture the true pattern.

Poor performance on both training and test data.

4. Overfitting

Caused by low bias and high variance.

Model memorizes noise.

Performs well on training data but poorly on new data.

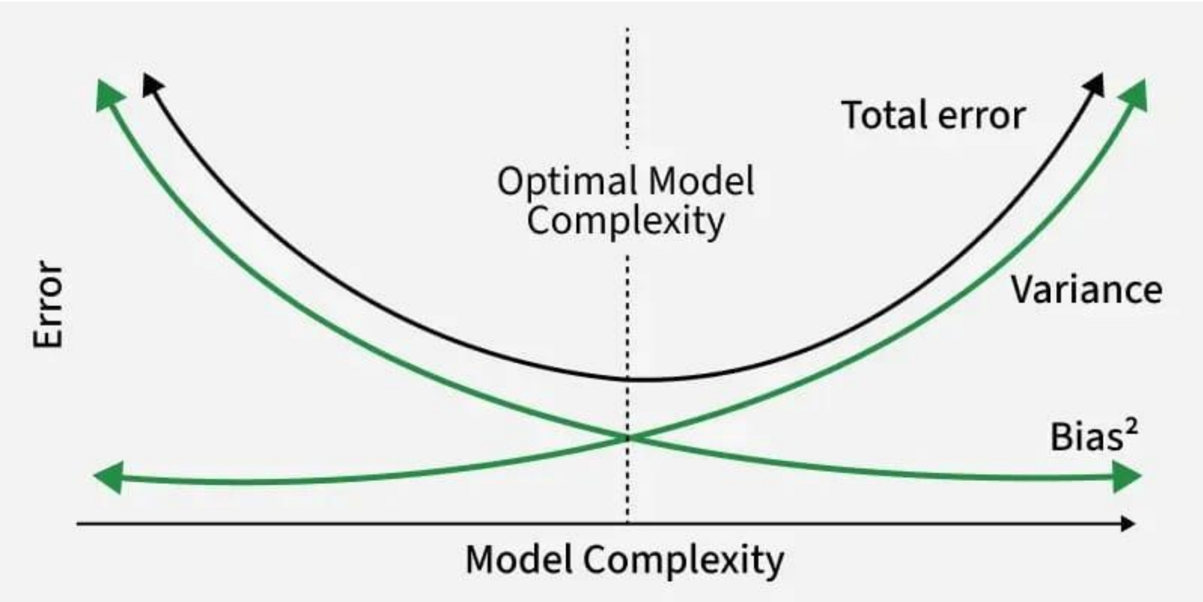
5. Best Fit Model

The best model should have:

✅ Low Bias

✅ Low Variance

In practice, we aim for a balance between bias and variance to minimize total error



What is Bias ?

Bias is the error that occurs when a machine learning model is too simple and makes strong assumptions about the data.

It measures how far the model's predictions are from the actual values because it cannot capture the true pattern.

In simple words:

Bias = Error due to oversimplification

Key Points:

High bias → Model is too simple

Misses important patterns

Causes underfitting

Leads to high training and testing error

Example:

If the real data follows a curve but the model uses a straight line, it cannot learn the curve properly. This error is called bias.

Short definition:

Bias is the error introduced when a model oversimplifies the problem.

Bias and variance are fundamental concepts in machine learning that affect model performance.

They help explain the errors made by models when predicting new data.

Understanding these concepts enables the development of better, more accurate models.

Causes of Bias :

Bias often arises from overly simplistic assumptions, such as linearity in a non-linear problem.

Using a model that is too simple to capture the underlying data complexity contributes to bias.

Choosing the wrong model or features can also increase bias in predictions

Using an Oversimplified Model

Choosing a simple model for complex data.

Example: Using linear regression for nonlinear data.

Leads to underfitting.

2. Wrong Assumptions About Data

Assuming data follows a certain pattern when it does not.

Example: Assuming a linear relationship when it is actually curved.

3. Too Few Features

Not including important input variables.

Missing important information increases bias.

4. High Regularization

Too much regularization (L1 or L2) forces the model to be very simple.

Can reduce model flexibility too much.

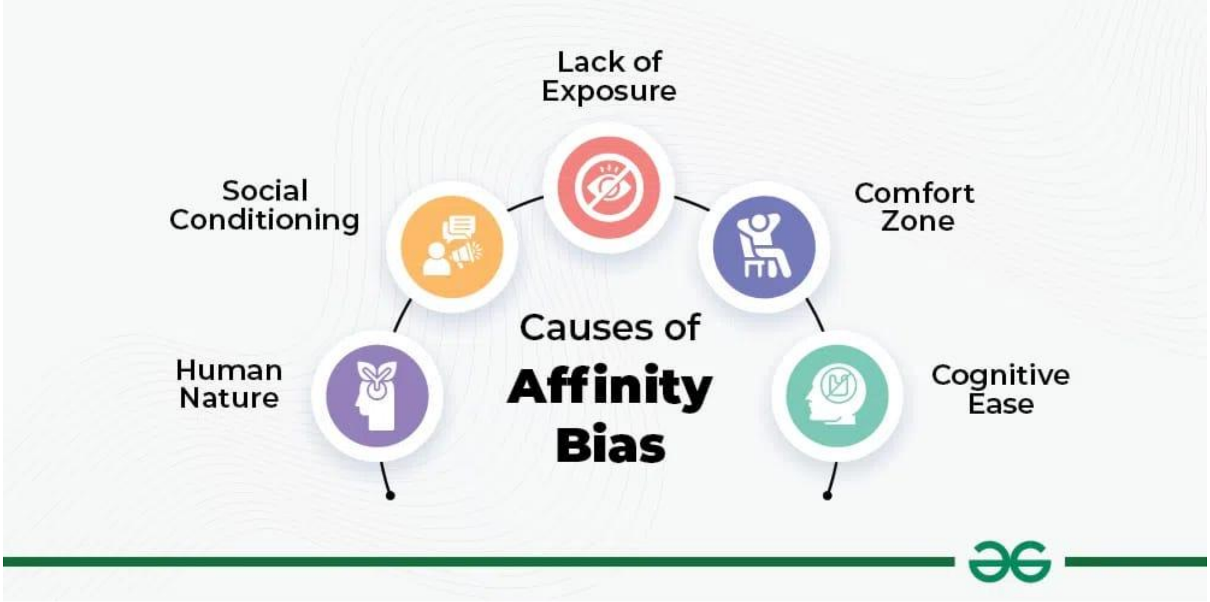
5. Insufficient Training Tim:

Stopping training too early (especially in neural networks).

Model does not learn enough from data.

6. Poor Data Representation :

Using inappropriate encoding or preprocessing.



Using an Oversimplified Model

Choosing a simple model for complex data.

Example: Using linear regression for nonlinear data.

Leads to underfitting.

2. Wrong Assumptions About Data :

Assuming data follows a certain pattern when it does not.

Example: Assuming a linear relationship when it is actually curved.

3. Too Few Features :

Not including important input variables.

Missing important information increases bias.

4. High Regularization :

Too much regularization (L1 or L2) forces the model to be very simple.

Can reduce model flexibility too much.

5. Insufficient Training Time :

Stopping training too early (especially in neural networks).

Model does not learn enough from data.

6. Poor Data Representation

Using inappropriate encoding or preprocessing.

Important patterns may not be captured properly.

Strategies to Reduce Bias:

Using more complex models can help decrease bias and better fit the data.

Incorporating additional features or using feature engineering can reduce bias.

Cross-validation helps identify the right model complexity to minimize bias.

Bias-Variance Trade off :

Bias and variance are often in opposition, creating a tradeoff in model tuning.

Reducing bias usually increases variance, and vice versa.

Achieving optimal model performance involves balancing these two factors.

Bias vs. Variance :

Bias is error due to overly simplistic assumptions, while variance is error due to model sensitivity.

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

Both need to be managed for optimal model performance

Bias is the error caused by overly simple assumptions in a model.

High bias makes the model too simple and leads to underfitting.

Understanding bias helps in selecting the appropriate model complexity.

A good machine learning model requires a proper balance between bias and variance to achieve better performance and generalization.

For the best-fit machine learning model, you must achieve low bias and low variance. This balance ensures the model is not too simple (avoiding underfit/high bias) nor too complex (avoiding overfit/high variance). The goal is to minimize both error types to ensure good performance on both training and new data.

Key Concepts :

Bias (Underfitting): Error from erroneous assumptions, making the model too simple (e.g., linear regression on non-linear data). High bias leads to missing relevant patterns.

Variance (Overfitting): Error from sensitivity to small fluctuations in the training set, making the model too complex (e.g., high-degree polynomial). High variance leads to modeling noise.

Best Fit: A model that accurately captures the underlying trend without fitting the noise, requiring low bias (accurate training) and low variance (generalizes well to test data).

Diagram Representation: Imagine a bullseye target. Low bias/Low variance means all shots hit the center. High bias/Low variance means shots are grouped together but far from the center. High variance/Low bias means shots are spread out around the center.

Trade-off: Increasing model complexity decreases bias but increases varian

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

High Variance → Model is too complex.

Causes Overfitting.

Training error is low, but testing error is high.

Underfitting (High Bias, Low Variance)

Model cannot capture the true pattern.

Too simple.

Poor performance on both training and test data.

Overfitting (Low Bias, High Variance)

Model memorizes noise.

Too complex.

Very good training accuracy but poor test accuracy

CONCLUSION :

Bias and variance are two fundamental concepts that determine how well a machine learning model performs.

High Bias occurs when the model is too simple. It cannot capture the underlying pattern of the data, leading to underfitting and high errors on both training and testing data.

High Variance occurs when the model is too complex. It learns not only the real patterns but also the noise in the training data, leading to overfitting and poor performance on new data.

The goal of model building is not to completely eliminate bias or variance, because that is usually impossible. Instead, we aim to find the right balance between them. This balance minimizes the total prediction error and improves the model's ability to generalize to unseen data.

In summary:

Too simple → High bias → Underfitting

Too complex → High variance → Overfitting

Balanced complexity → Low bias + Low variance → Best model performance

Therefore, the best fit model is one that maintains an optimal tradeoff between bias and variance, ensuring accuracy, stability, and good generalization.