Assignment - 1

Introduction to Bias and Variance

- In Machine Learning, especially in supervised learning, two important concepts that determine the performance of a model are bias and variance.
- These two factors directly influence whether a model performs well on unseen data or not.
- The ultimate goal of any predictive model is to achieve good generalization, meaning the model should perform well not only on training data but also on new, unseen data.
- Bias and variance are closely related to model complexity.
- If a model is too simple, it may not capture the underlying pattern of the data.
- If it is too complex, it may capture unnecessary noise along with the pattern.
- Therefore, understanding the balance between bias and variance is extremely important for building an effective model.

What is Bias?

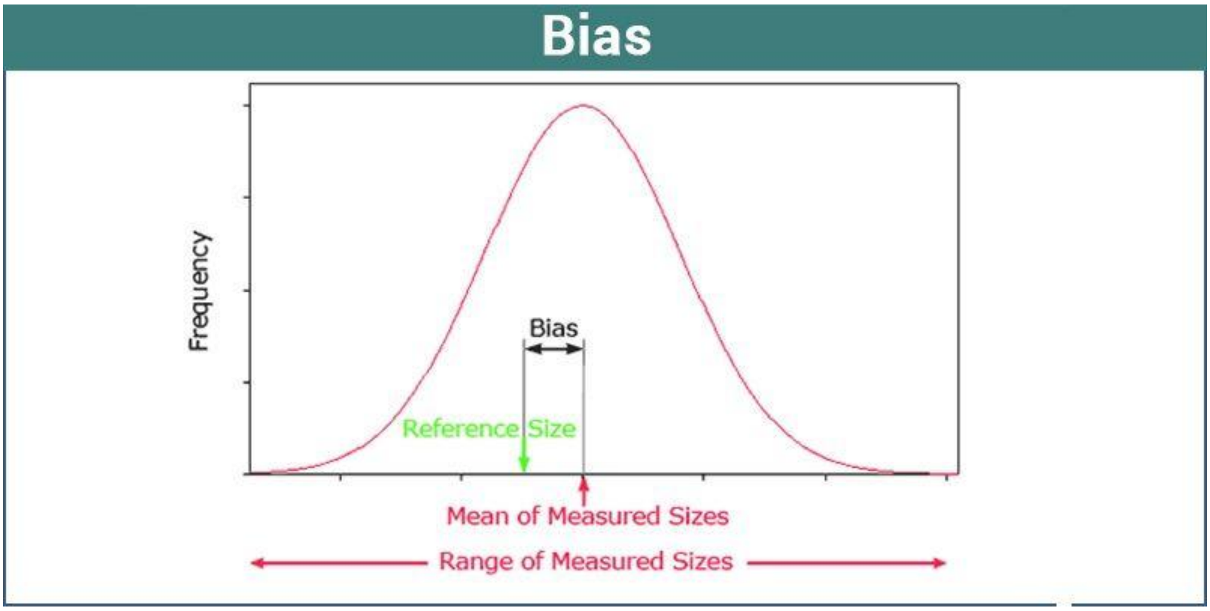
- Bias is a prejudice, inclination, or tendency for or against something, someone, or a group, usually in a way considered unfair.
- It represents a lack of objectivity, where preconceived notions or personal preferences influence judgment rather than facts or logic.
- Biases can be conscious or unconscious, learned, and often lead to systemic errors.
- Bias refers to the error caused by overly simplistic assumptions in the learning algorithm.
- In simple words, bias measures how far the model's predictions are from the actual values on average.
    - A high bias model pays very little attention to the training data.
    - It oversimplifies the problem.
    - It fails to capture important relationships between input and output variables.

When bias is high:

- The model makes strong assumptions.
- It does not fit the training data well.
- Both training error and testing error are high.

This situation is known as underfitting.

For example, if the actual relationship between variables is curved but we try to fit a straight line, the model will have high bias because it cannot capture the curve.



What is Variance?

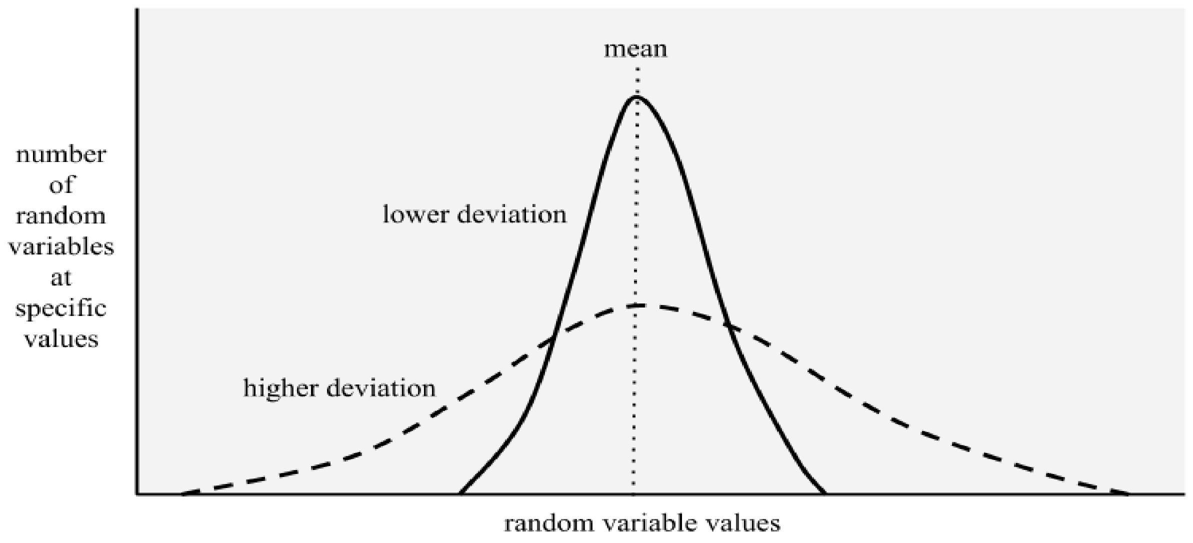
- Variance is a statistical measure of the dispersion or spread of data points around their mean (average) value.
- It quantifies how far numbers in a dataset are from the average; higher variance indicates greater volatility and wider spread, while lower variance indicates data points are closer to the mean.
- Variance refers to how much the model's predictions change when we use different training data.
- In simple terms, variance measures how sensitive the model is to small changes in the training dataset.
    - A high variance model pays too much attention to the training data.
    - It learns noise and random fluctuations in the data.
    - It performs very well on training data but poorly on testing data.

When variance is high:

- Training error is very low.
- Testing error is very high.

This situation is known as overfitting.

For example, if we use a very high-degree polynomial to fit a simple pattern, the model will pass through almost every training point, including noise, which leads to high variance.



Underfitting (High Bias, Low Variance)

- ❖ Underfitting occurs when the model is too simple to understand the underlying structure of the data.

Characteristics:

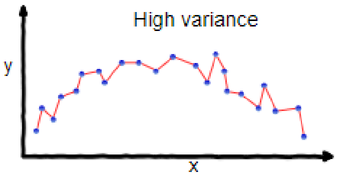
- o High bias
- o Low variance
- o High training error
- o High testing error

Overfitting (Low Bias, High Variance)

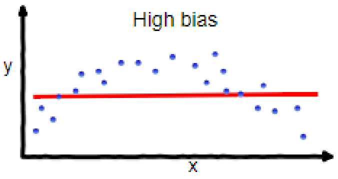
- ❖ Overfitting occurs when the model is too complex and captures noise along with actual patterns.

Characteristics:

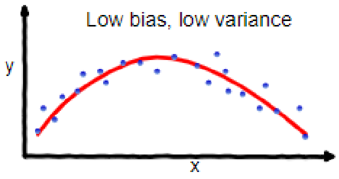
- o Low bias
- o High variance
- o Very low training error
- o High testing error



overfitting



underfitting



Good balance

Bias-Variance Tradeoff

- The bias-variance tradeoff describes the balance between bias and variance.

If we increase model complexity:

- Bias decreases
- Variance increases

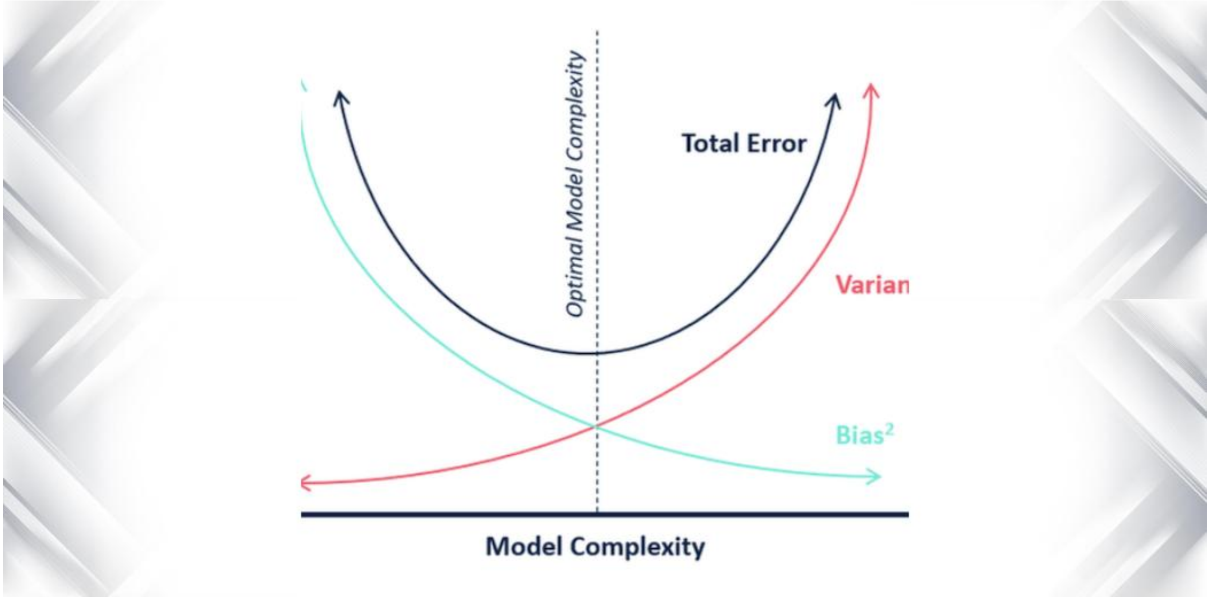
If we decrease model complexity:

- Bias increases
- Variance decreases

Total error = Bias² + Variance + Irreducible error

Irreducible error is noise in the data that we cannot remove.

Graphically, the relationship looks like:



Left side → High bias, underfitting

Right side → High variance, overfitting

Middle → Optimal balance

Practical Example

Consider predicting house prices:

- If we use a very simple linear model ignoring many features → high bias → underfitting.
- If we use a very complex neural network trained excessively → high variance → overfitting.
- If we choose a properly regularized model with appropriate features → low bias + low variance → best performance.

Techniques to reduce bias:

- Increase model complexity
- Add more features
- Reduce regularization

Techniques to reduce variance:

- Increase training data
- Use regularization
- Use cross-validation
- Apply pruning (in decision trees)
- Use ensemble methods like bagging

Conclusion:

- Bias and variance are fundamental concepts in machine learning that determine model performance and generalization ability.
- A model with high bias fails to learn the underlying pattern and results in underfitting, while a model with high variance learns too much from training data including noise and results in overfitting.
- The key objective in model building is to find a balance between these two extremes.
- The best fit model is achieved when both bias and variance are low, meaning the model is complex enough to capture the true pattern but simple enough to avoid learning noise.
- This balance is known as the bias-variance tradeoff, and achieving it is essential for building an accurate and reliable machine learning model.