VARIANCE AND BIAS IN MACHINE LEARNING

1. Introduction
   - Bias and variance are two major sources of error in machine learning models.
   - They determine whether a model performs well or poorly on unseen data.
   - The main objective of any model is to generalize well.
   - Improper balance between bias and variance leads to underfitting or overfitting.
   - Total prediction error consists of:
     - Bias² o Variance o
     - Irreducible error
2. Bias
   - 2.1 Definition
     - Bias is the error caused by overly simple assumptions in the learning algorithm.
     - It measures how far the model's predictions are from the true values on average.
   - 2.2 Characteristics of High Bias
     - Model is too simple.
     - Strong assumptions about data.
     - Cannot capture complex relationships.
     - High training error.
     - High testing error.
     - Leads to underfitting.
   - 2.3 Example
     - Using a straight-line model to fit curved data.
     - Linear regression applied to highly nonlinear data.
3. Variance

3.1 Definition

- Variance measures how much model predictions change with different training datasets.
- It indicates sensitivity to fluctuations in training data.

3.2 Characteristics of High

Variance □ Model is too

complex.

- Fits training data extremely well.
- Very low training error.
- High testing error.
- Learns noise along with patterns.
- Leads to overfitting.

3.3 Example

- Deep decision tree without pruning.
- Very high-degree polynomial regression.

4. Underfitting

4.1 Definition

- Underfitting occurs when the model is too simple to capture the underlying pattern.

4.2 Properties

- High Bias
- Low Variance
- Poor training performance
- Poor testing performance

4.3 Causes

- Model too simple

- Insufficient features
- Too much regularization
- Not enough training

4.4 Solution

- Increase model complexity
- Add more relevant features
- Reduce regularization
- Train longer

5. Overfitting

5.1 Definition

- Overfitting occurs when the model learns noise along with actual patterns.

5.2 Properties

- Low Bias
- High Variance
- Very low training error
- High testing error

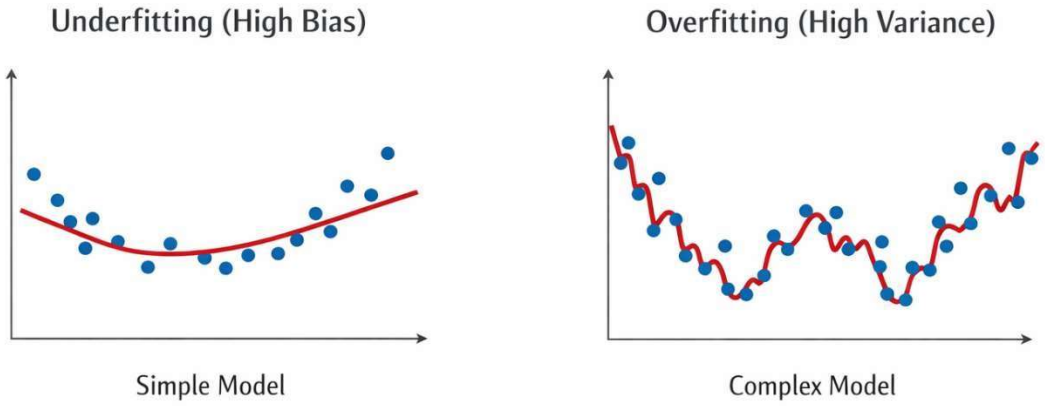
5.3 Causes

- Too complex model
- Too many parameters
- Small dataset
- No regularization

5.4 Solution

- Increase training data
- Apply regularization (L1, L2)
- Use cross-validation
- Prune decision trees

- Apply dropout in neural networks



6. Bias–Variance Tradeoff

- Increasing model complexity reduces bias.
- Increasing model complexity increases variance.
- Simple model → High Bias, Low Variance.
- Complex model → Low Bias, High Variance.
- Optimal model → Balanced Bias and Variance.

7. Error Decomposition

Total Error = Bias^2 + Variance + Irreducible Error

- Bias^2 → Error due to wrong assumptions.
- Variance → Error due to sensitivity to data.
- Irreducible error → Noise in data (cannot be removed).

8. Comparison Summary

Underfitting:

- Bias: High
- Variance: Low
- Training Error: High
- Testing Error: High

Overfitting:

- Bias: Low
- Variance: High
- Training Error: Low
- Testing Error: High

Best Fit Model:

- Bias: Low
- Variance: Low
- Training Error: Low
- Testing Error: Low

9. Conclusion

- Bias causes underfitting.
- Variance causes overfitting.
- The key objective in machine learning is to balance both.
- The best performing model has low bias and low variance.
- This balance ensures accurate and reliable predictions on unseen data.