Assignment 1:

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

Name : Akshay Kumar V Padmashali

USN : 01SU24CS015

Branch : CSE

Section : A

1. Introduction

In machine learning, building an accurate and reliable predictive model is not only about achieving high accuracy on training data. The real objective is to develop a model that generalizes well to unseen data. Two fundamental concepts that directly influence model performance are **bias** and **variance**. These two components determine whether a model underfits, overfits, or achieves the optimal balance known as the best fit.

The concepts of bias and variance are central to understanding why some models perform poorly even when trained on large datasets. A model may be too simple to capture patterns (high bias), or too complex and sensitive to noise (high variance). The delicate trade-off between these two factors is called the **bias-variance tradeoff**, and mastering it is essential for building effective machine learning systems.

This report provides a detailed explanation of bias and variance, supported by conceptual diagrams, real-world examples, mathematical interpretation, and discussion on overfitting and underfitting. Finally, it answers the key question: *For the best fit model, should we prefer low bias or high variance, low bias or low variance, high bias or high variance, or low bias or high variance?*

2. Understanding Bias

2.1 Definition of Bias

Bias refers to the error introduced by approximating a real-world problem with a simplified model. In other words, bias measures how far the average prediction of the model is from the true value.

A model with **high bias** makes strong assumptions about the data. Because of these assumptions, it may fail to capture the underlying relationship between input and output variables.

2.2 High Bias Characteristics

- Model is too simple.
- Cannot capture complex relationships.
- High training error.
- High testing error.
- Leads to **underfitting**.

2.3 Example of High Bias

Consider fitting a straight line (linear regression) to data that actually follows a curve. The straight line cannot capture the curvature of the data. This results in systematic error — this is high bias.

2.4 Mathematical View

If:

￼￼￼￼Error=Bias2+Variance+Irreducible Error

High bias increases the first term (Bias^2), leading to overall higher prediction error.

3. Understanding Variance

3.1 Definition of Variance

Variance measures how much the model's predictions change if different training data is used. A high variance model is very sensitive to fluctuations in the training data.

3.2 High Variance Characteristics

- Model is too complex.
- Captures noise as if it were a pattern.
- Very low training error.
- High testing error.
- Leads to **overfitting**.

3.3 Example of High Variance

If we fit a 10th-degree polynomial to a dataset with only 10 points, the model might pass exactly through all points. While training error becomes nearly zero, new unseen data points will likely produce large errors.

3.4 Mathematical View

In the same formula:

￼￼￼￼Error=Bias2+Variance+Irreducible Error

High variance increases the second term, again increasing total error.

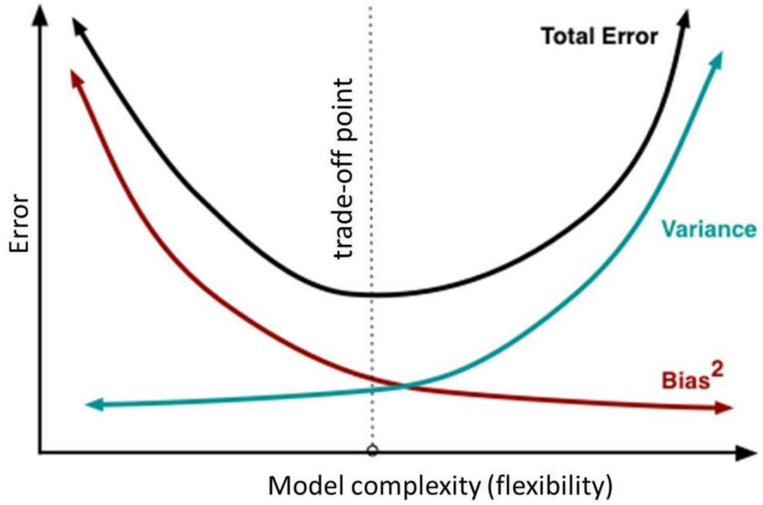
4. Bias-Variance Tradeoff

The bias-variance tradeoff describes the balance between two types of errors:

- Increasing model complexity → decreases bias but increases variance.
- Decreasing model complexity → increases bias but decreases variance.

The goal is to find a model complexity where total error is minimized.

Conceptual Diagram



- Left side → High Bias, Low Variance (Underfitting)
- Right side → Low Bias, High Variance (Overfitting)
- Middle → Balanced Bias and Variance (Best Fit)

5. Underfitting

5.1 Definition

Underfitting occurs when a model is too simple to capture the underlying structure of the data.

5.2 Causes

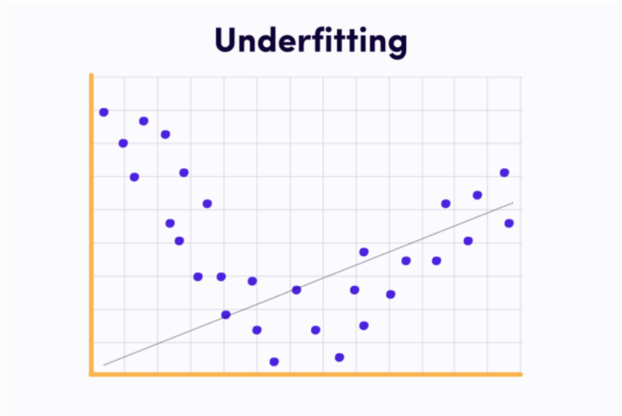
- Using a linear model for nonlinear data.

- Insufficient training time.
- Too much regularization.
- Too few features.

5.3 Characteristics

- High training error.
- High validation/testing error.
- Model performs poorly everywhere.

5.4 Diagram Representation



6. Overfitting

6.1 Definition

Overfitting occurs when a model learns the training data too well, including noise and random fluctuations.

6.2 Causes

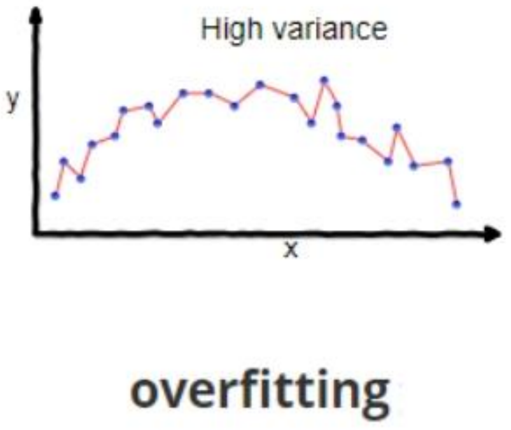
- Excessively complex model.
- Small dataset.
- Too many features.
- Lack of regularization.

6.3 Characteristics

- Very low training error.
- High testing error.

- Poor generalization.

6.4 Diagram Representation



7. Best Fit Model

The best fit model is achieved at the point where:

- Bias is low enough to capture true patterns.
- Variance is low enough to generalize well.

Correct Answer to the Question

For the best fit model, we should aim for:

Low Bias and Low Variance

Let us evaluate all given options:

1. **Low Bias or High Variance** → Wrong
   High variance leads to overfitting.
2. **Low Bias or Low Variance** → Correct
   The best model minimizes both bias and variance.
3. **High Bias or High Variance** → Wrong
   Both increase prediction error.
4. **Low Bias or High Variance** → Wrong
   High variance is undesirable.

Thus, the correct choice is:

Low Bias and Low Variance

However, in practice, achieving extremely low bias and extremely low variance simultaneously is difficult. Therefore, we aim for an optimal tradeoff where total error is minimized.

8. Real-World Example

Consider weather prediction.

- A very simple model that predicts "same temperature as yesterday" has high bias.
- A very complex model that memorizes every past weather condition has high variance.
- A balanced statistical model using historical patterns achieves the best performance.

Another example is stock price prediction:

- Linear model → High bias.
- Deep neural network with too many layers → High variance.
- Regularized model with optimal layers → Balanced.

9. Methods to Reduce Bias and Variance

9.1 Reducing High Bias

- Increase model complexity.
- Add more features.
- Reduce regularization.
- Use non-linear models.

9.2 Reducing High Variance

- Increase training data.
- Apply regularization (L1/L2).
- Prune decision trees.
- Use cross-validation.
- Apply ensemble techniques like bagging.

10. Graphical Explanation of Bias and Variance

Imagine a dartboard:

- **High Bias, Low Variance** → Darts tightly grouped but far from center.
- **Low Bias, High Variance** → Darts spread around center.
- **High Bias, High Variance** → Darts scattered far from center.
- **Low Bias, Low Variance** → Darts tightly clustered at center (Best Fit).

Best Model:
(• • •)
(center)

11. Mathematical Decomposition of Error

The expected squared error of a model can be decomposed as:

￼￼￼￼￼￼￼E[(y−f^ (x))2]=Bias2+Variance+Irreducible Error

- **Bias²**: Error due to wrong assumptions.
- **Variance**: Error due to sensitivity to training data.
- **Irreducible Error**: Noise inherent in data.

Our objective is to minimize:

￼Bias2+Variance

since irreducible error cannot be eliminated.

12. Practical Example Using Polynomial Regression

If degree = 1 → Underfitting (High Bias)
If degree = 15 → Overfitting (High Variance)
If degree = 3 or 4 → Balanced Model

This clearly demonstrates the tradeoff.

13. Importance in Modern AI Systems

Bias and variance are important in:

- Deep learning
- Decision trees
- Support vector machines
- Ensemble learning

In large neural networks:

- Too shallow → High bias.
- Too deep without regularization → High variance.

Techniques like dropout and batch normalization help control variance.

14. Conclusion

Bias and variance are two critical components that determine the predictive performance of a machine learning model. High bias leads to underfitting, where the model fails to capture important patterns. High variance leads to overfitting, where the model captures noise instead of meaningful structure.

The ultimate goal is to achieve a balance between bias and variance — minimizing total error while ensuring strong generalization ability. Among the given options:

- Low Bias & High Variance → Not ideal
- High Bias & High Variance → Worst
- High Bias & Low Variance → Underfitting
- **Low Bias & Low Variance → Best Fit Model**

Therefore, for the best fit model, we should aim for:

Low Bias and Low Variance

In practice, we rarely achieve zero bias and zero variance. Instead, we aim for the optimal balance where the combined error is minimized, ensuring strong predictive performance on unseen data.