**AI/ML Interns – Bhavya**

**Learning Topics:**

* Bias-Variance Tradeoff.
* Difference between High-bias(Underfitting) and high variance(Overfitting).
* Visual Understanding of tradeoff.

1. **Bias-Variance Tradeoff:**

Understanding Bias

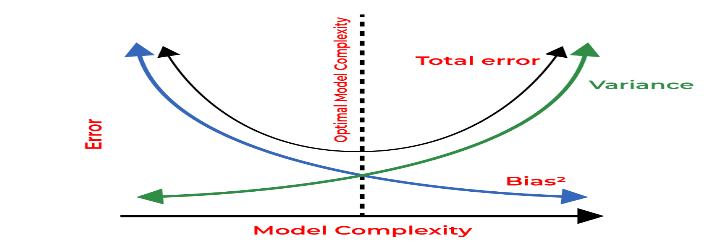
* **Definition:** Bias happens when a model is too simplistic and fails to capture important patterns in the data.
* **Example:** If you’re predicting Titanic survival based only on age, you’re ignoring critical factors like gender and passenger class. This results in a model that assumes, for instance, older people didn’t survive, missing the real complexity of the situation.
* **Characteristics:**
  + Over-generalizes data
  + Tends to underfit, performing poorly on both training and new data

Understanding Variance

* **Definition:** Variance refers to errors arising when a model is overly complex, capturing not just real patterns but also random noise in the data.
* **Example:** Consider a Random Forest model that uses every minute detail (even ticket numbers) to predict survival. It performs perfectly on training data but fails to generalize to new cases.
* **Characteristics:**
  + Fits noise rather than underlying patterns
  + Prone to overfitting (excellent on training data, poor on unseen data)

The Bias–Variance Tradeoff

* **Balance Required:** There is a constant tradeoff between bias and variance in model-building:
  + Increasing complexity (e.g., more trees in a Random Forest) reduces bias but increases variance.
  + Simplifying the model (e.g., using linear regression) reduces variance but increases bias.
* **Solution Strategies:**
  + Adjust model complexity to find a suitable balance.
  + Increase the amount of training data; for instance, more passenger records in the Titanic dataset help complex models generalize better.



1. **Differnce Between:**

| **Aspect** | **High Bias (Underfitting)** | **High Variance (Overfitting)** |
| --- | --- | --- |
| **Definition** | Model is too simple, misses key patterns in data. | Model is too complex, captures random noise as patterns. |
| **Analogy** | Predicting rain using only the day of the week (e.g., “Mondays are rainy”). Ignores temperature or humidity. | Predicting rain by memorizing exact weather details for each past day, including irrelevant factors like the neighbor’s dog barking. |
| **Training Performance** | Poor accuracy on past weather data (e.g., ~50%) because it oversimplifies the problem. | High accuracy on past weather data (e.g., ~95%) as it memorizes every detail. |
| **Test Performance** | Poor accuracy on new days (e.g., ~50%), similar to training, as it misses weather trends. | Lower accuracy on new days (e.g., ~60%) as it relies on noise not present in new data. |
| **Train-Test Gap** | Small gap (both poor), as model doesn’t learn enough from past data. | Large gap (high training, low test), as model learns irrelevant details from past data. |
| **Model Complexity** | Too simple (e.g., a rule-based model using only one factor like day of the week). | Too complex (e.g., a neural network tracking every minor weather fluctuation). |
| **Data Sensitivity** | Insensitive to changes in weather data, overly generalized (e.g., “It’s always cloudy”). | Overly sensitive, fits random variations (e.g., “Rain only when wind is exactly 12 mph”). |
| **Solution** | Add more factors (e.g., humidity, wind speed) or use a more complex model. | Simplify model (e.g., limit complexity) or collect more weather data to generalize better. |

1. **Visual Tradeoff:**

### Visual Representation

* **Graphical Illustration:**
  + **X-axis:** Model complexity (ranging from simple to complex).
  + **Y-axis:** Prediction error.
  + **High bias:** High error at low complexity.
  + **High variance:** Lower training error but higher test error at high complexity.
  + **Optimal zone:** The minimal total error (bias + variance) is located at a moderate level of complexity.

