

Supervised Learning - I

By:
Dr. Shikha Deep

Contents include:

- (1) Learning Curves**
- (2) Training-Validation Test**

Learning Curves:

- A learning model of a Machine Learning model shows how the error in the prediction of a Machine Learning model changes as the size of the training set increases or decreases.
- It helps us diagnose if a model is underfitting, overfitting, or well-generalized.

How is it drawn?

On the **X-axis** → the amount of training (e.g., size of training data or training iterations).

- On the **Y-axis** → the model's error or accuracy.
- **Two curves:**
 - **Training performance curve** (how well the model fits the training data).
 - **Validation performance curve** (how well the model performs on unseen validation data).

Patterns to Observe

Underfitting case:

- Training error is high.
- Validation error is also high.
- Both curves meet at a high error level → the model is too simple.

Overfitting case:

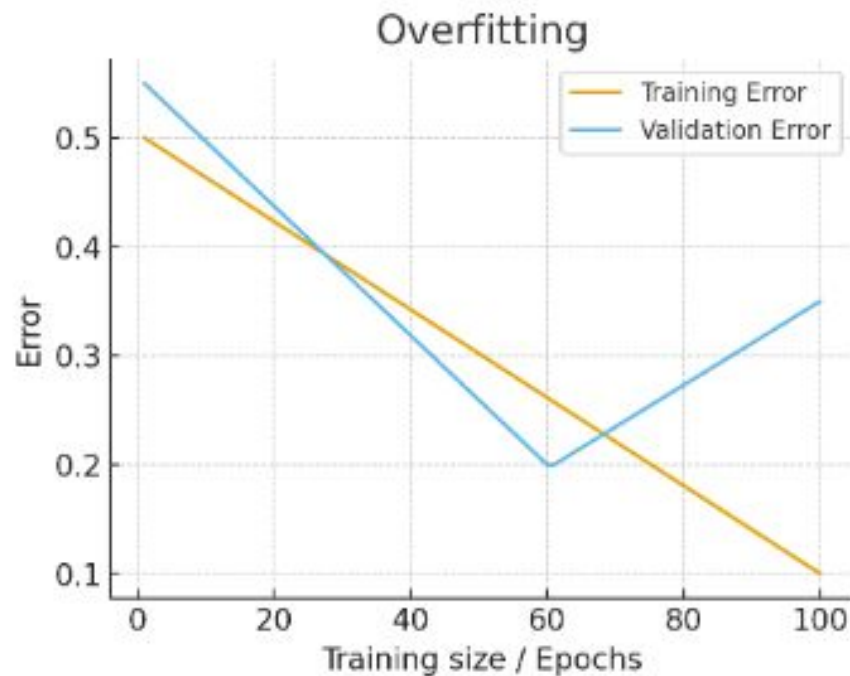
- Training error is very low.
- Validation error is high and does not decrease → the model memorizes instead of learning.

Good Fit case:

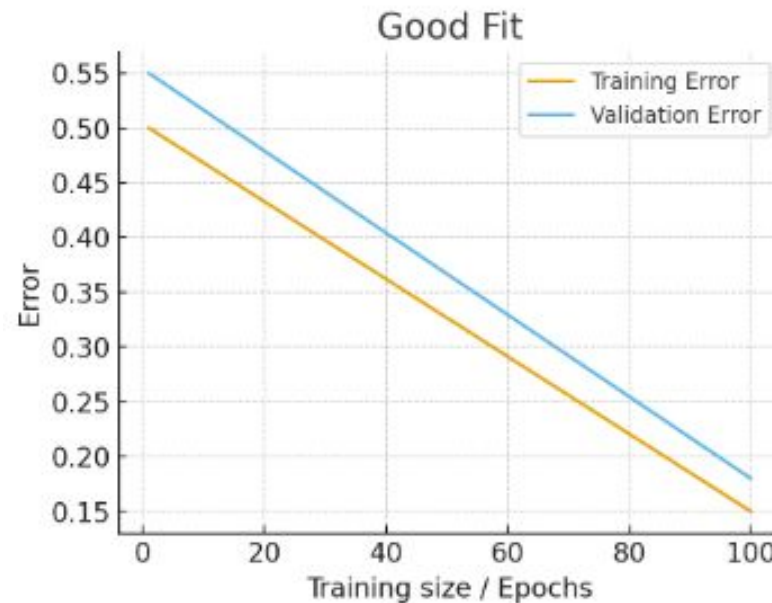
- Training error decreases steadily.
- Validation error also decreases and stays close to training error → balanced learning.



- The training error is high → the model cannot even perform well on the data it has already seen.
- The validation error is also high and close to the training error.
- Both curves meet at a high error level.



- The training error keeps decreasing to almost zero (the model memorizes training data).
- The validation error decreases initially but then increases again → meaning the model fails on unseen data.



- The **training error** decreases steadily but doesn't go to zero.
- The **validation error** also decreases and stays close to training error.
- Both curves stabilize at a low error.

Training–Validation–Test

Three-way split: Training, Validation, Test

- **Training set:**
 - Used to train the model.
 - Model learns patterns here.
- **Validation set:**
 - Used during model building to **choose the best version**.
 - Helps in hyperparameter tuning (e.g., how many trees in Random Forest).
 - Not used for actual training.
- **Test set:**
 - Used only at the **end** to evaluate final performance.



■ Training Set (Learn patterns) ■ Validation Set (Tune model) ■ Test Set (Final evaluation)

Training–Validation–Test split diagram:

- **Training (~70%)** → Learn patterns.
- **Validation (~15%)** → Tune the model (mock test).
- **Test (~15%)** → Final evaluation (board exam).

