OVERFITTING AND UNDERFITTING

Machine learning models should learn useful patterns from training data. When a model learns too little or too much, we get underfitting or overfitting.

- Underfitting means that the model is too simple and does not cover all real patterns in the data.
- Overfitting means that the model learns not just the underlying pattern, but also noise or random quirks in the training data. model memorizes training data
- A good model finds the right *spot, it is* complex enough to capture real patterns, but not so complex that it "memorizes" noise

What is Underfitting?

Underfitting happens when the model fails to learn important patterns. It performs poorly on both training and testing data. Underfitting happens due to:

- Model is too simple
- Very high regularization
- Features are weak or missing
- Not enough training
- High bias

Bias: *It is like assuming all birds can only be small and fly, so the model fails to recognize big birds like ostriches or penguins that can't fly and get biased with predictions.*

Bias–Variance Inside Underfitting

Underfitting mainly occurs due to high bias:

- High bias means model makes strong assumptions
- Ignores patterns
- Learns an overly simple representation
- Variance is low because the model gives similar outputs even if the data changes

Underfitting = High Bias + Low Variance

What is Overfitting?

Overfitting happens when the model learns too much from the training data, including noise and outliers. It performs very well on training data but poorly on test data. Overfitting happens due to:

- Model too complex
- Too many features
- Very little data
- No regularization
- High variance

Variance*: Error that happens when a machine learning model learns too much from the data, including random noise.*

Bias–Variance Inside Overfitting

Overfitting is mainly caused by high variance:

- High variance means model reacts too strongly to training data
- Learns noise as patterns
- Low bias because the model is extremely flexible

Overfitting = Low Bias + High Variance

- **Underfitting :** Straight line trying to fit a curved dataset but cannot capture the data's patterns, leading to poor performance on both training and test sets.
- **Overfitting**: A squiggly curve passing through all training points, failing to generalize performing well on training data but poorly on test data.