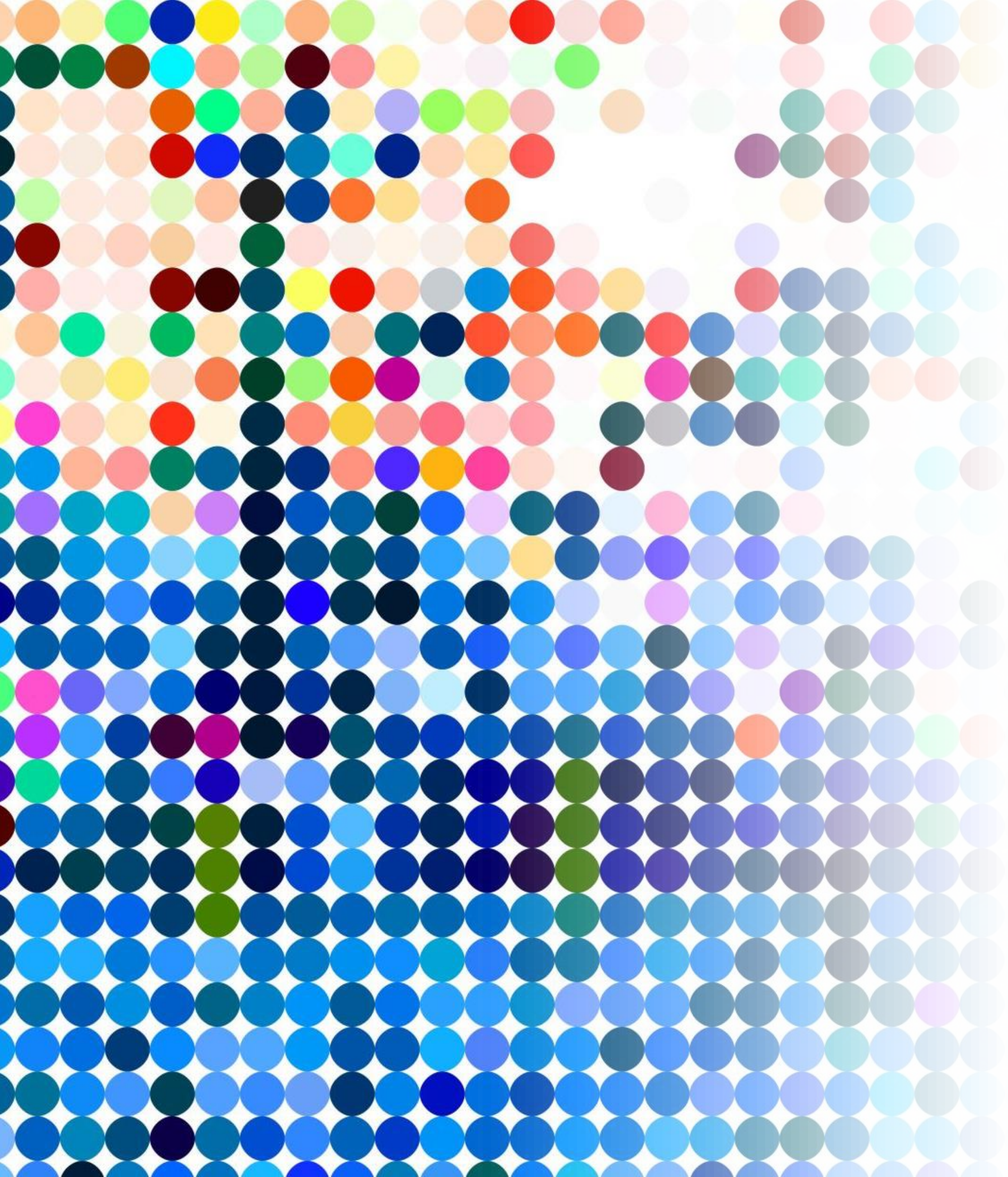


Machine Learning Concepts: Training, Testing, Overfitting, Underfitting, and Regularization

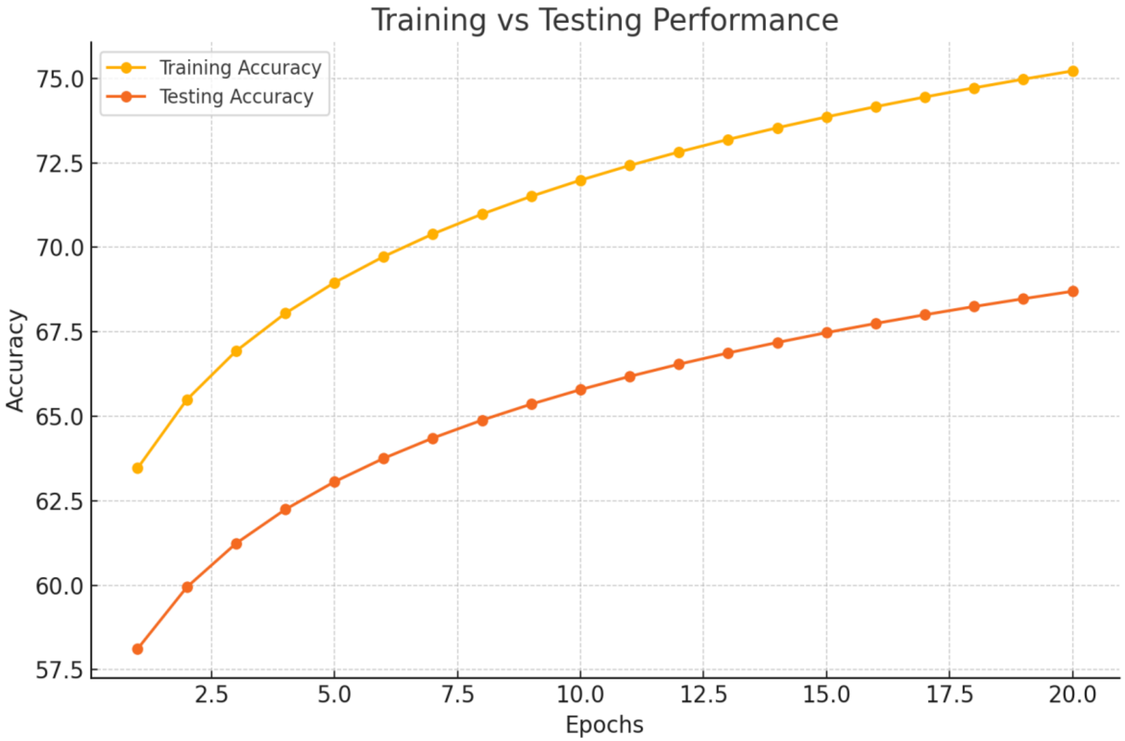
Understanding Core Principles

Training Data

- Training data is the dataset on which a machine learning model is trained.
- Purpose: Helps the model learn patterns and make predictions.
- Process:
  1. Collect data.
  2. Preprocess data (cleaning, normalization).
  3. Train the model using this data.
- Example: Image recognition - training on labeled images.

Testing Data

- Testing data is the dataset used to evaluate the performance of the trained model.
- Purpose: Measures the accuracy and generalization of the model.
- Process:
  1. Split dataset into training and testing sets (e.g., 80-20 split).
  2. Evaluate the model on the testing set.
- Example: Image recognition - testing on unseen labeled images.


Training vs Testing Performance
Training Accuracy
Testing Accuracy
75.0
72.5
70.0
67.5
65.0
62.5
60.0
57.5
Accuracy
2.5
5.0
7.5
10.0
12.5
15.0
17.5
20.0
Epochs

Overfitting

- Overfitting occurs when a model learns not only the underlying patterns but also the noise in training data.
- Symptoms: High accuracy on training data but poor performance on testing data.
- Causes:

  - Too complex model.

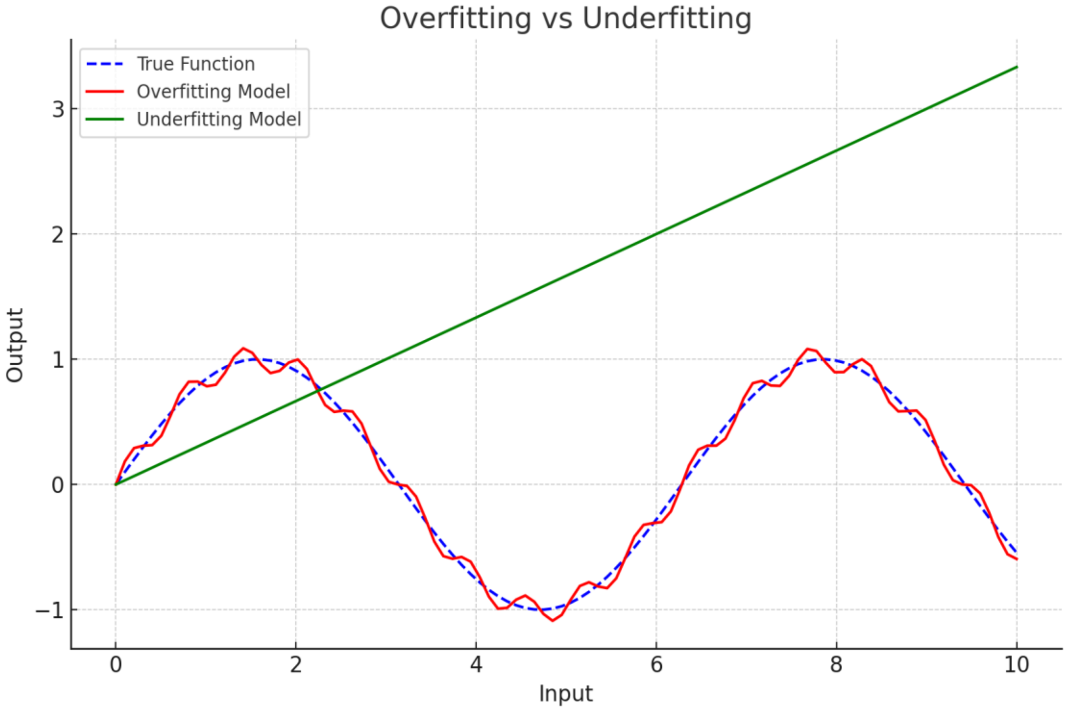
  - Insufficient training data.
- Example: A decision tree with too many branches fitting every single data point.

Underfitting

- Underfitting occurs when a model is too simple to capture the underlying patterns in the data.
- Symptoms: Poor performance on both training and testing data.
- Causes:

  - Model is too simple.

  - Insufficient training time.
- Example: A linear regression model applied to non-linear data.

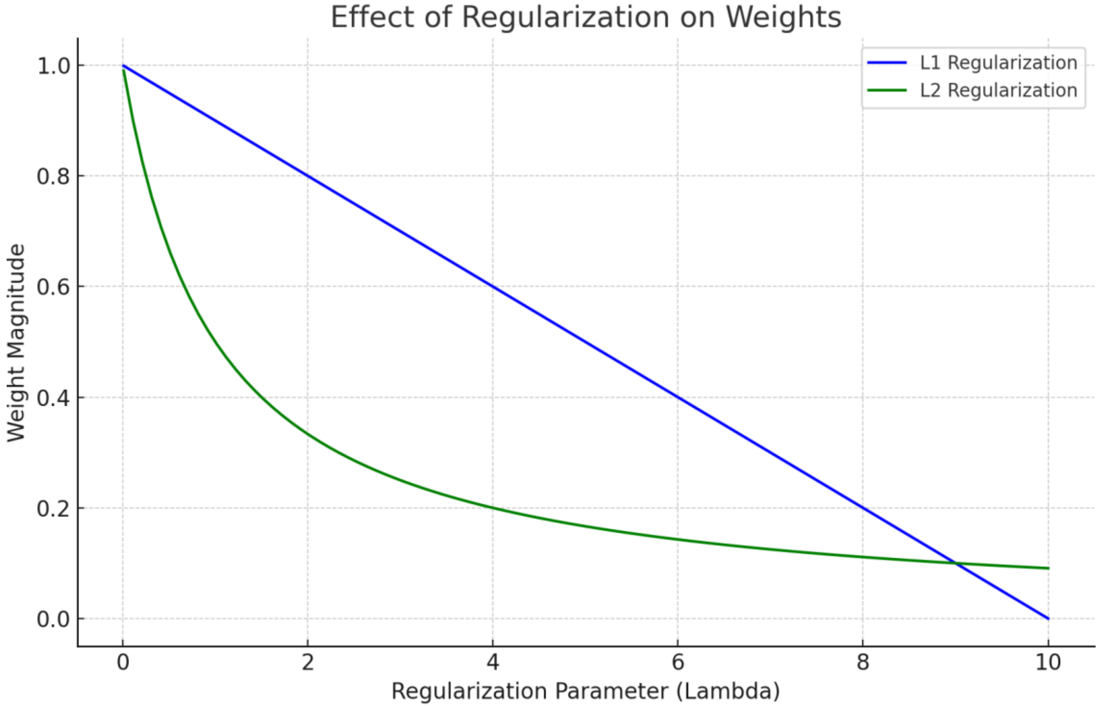

Overfitting vs Underfitting
True Function
Overfitting Model
Underfitting Model
Output
Input
3
2
1
0
−1
0
2
4
6
8
10

Regularization

- Utilize regularization to improve model performance.
- Regularization techniques are used to prevent overfitting by adding a penalty to the loss function.
- Types:
  - L1 Regularization (Lasso): Adds the absolute value of coefficients as penalty.
  - L2 Regularization (Ridge): Adds the squared value of coefficients as penalty.
- Process:
  1. Apply regularization during model training.
  2. Adjust regularization parameter to balance bias and variance.
- Example: Logistic regression with L2 regularization to prevent overfitting.

Regularization Techniques

- L1 Regularization: Encourages sparsity (many zero coefficients).
- L2 Regularization: Distributes error across all terms.
- Elastic Net: Combines L1 and L2 regularization.
- Dropout (in neural networks): Randomly drops neurons during training to prevent co-adaptation.
- Example: Regularization applied to neural network training.


Effect of Regularization on Weights
L1 Regularization
L2 Regularization
Weight Magnitude
1.0
0.8
0.6
0.4
0.2
0.0
0
2
4
6
8
10
Regularization Parameter (Lambda)

Comparing Overfitting and Underfitting

- Overfitting:
  - Complex model.
  - High variance.
  - Low bias.
  - Poor generalization.

- Underfitting:
  - Simple model.
  - Low variance.
  - High bias.
  - Poor generalization.