Regularization

Regularization is an important technique in machine learning (ML) used to prevent overfitting and improve the generalization of models. It helps to reduce the complexity of a model without substantially increasing its error rate.

Role Of Regularization

1. **Preventing Overfitting**: One way to prevent overfitting is to use regularization, which penalizes large coefficients and constrains their magnitudes,
2. **Balancing Bias and Variance**: Regularization can help balance the trade-off between model bias (underfitting) and model variance (overfitting) in machine learning, which leads to improved performance.
3. **Feature Selection**: Some regularization methods, such as L1 regularization (Lasso), promote sparse solutions that drive some feature coefficients to zero. This automatically selects important features while excluding less important ones.
4. **Generalization**: Regularized models learn underlying patterns of data for better generalization to new data, instead of memorizing specific examples.

Overfitting:

Overfitting occurs when a model learns the training data too well, including noise and random fluctuations. An overfit model performs exceptionally well on the training data but poorly on new, unseen data.

Characteristics of overfitting:

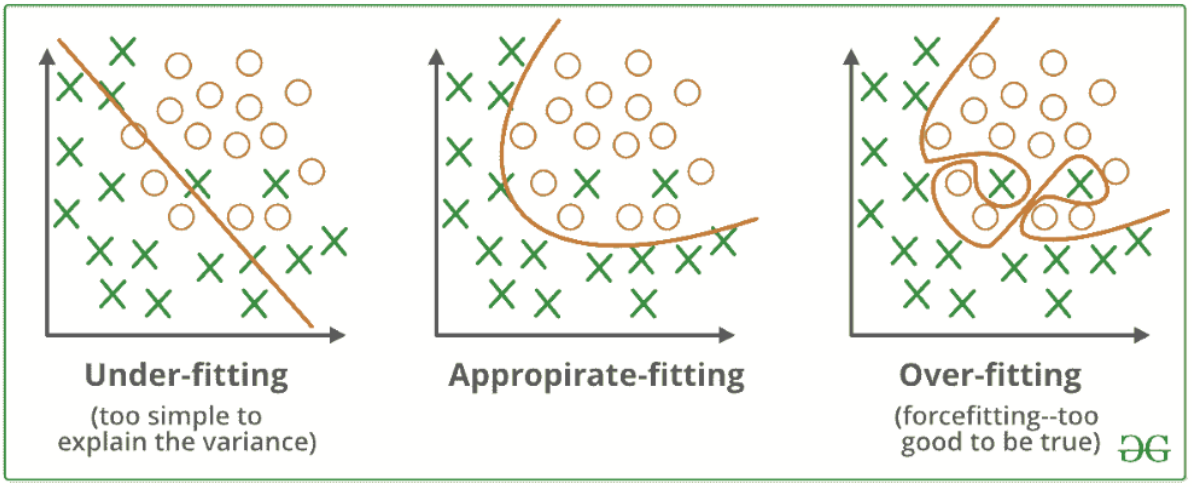
- Low bias, high variance
- Complex model that captures noise in the data

Underfitting:

Underfitting happens when a model is too simple to capture the underlying pattern in the data. An underfit model performs poorly on both training and test data.

Characteristics of underfitting:

- High bias, low variance
- Too simple to capture the data's complexity
- Poor performance on both training and test data



Bias:

Bias is the error introduced by approximating a real-world problem with a simplified model. It's the difference between the expected predictions of the model and the true values.

High bias:

- Model is too simple (underfit)
- Makes strong assumptions about the data

Low bias:

- Model is flexible enough to capture the underlying pattern
- Makes fewer assumptions about the data structure

Variance:

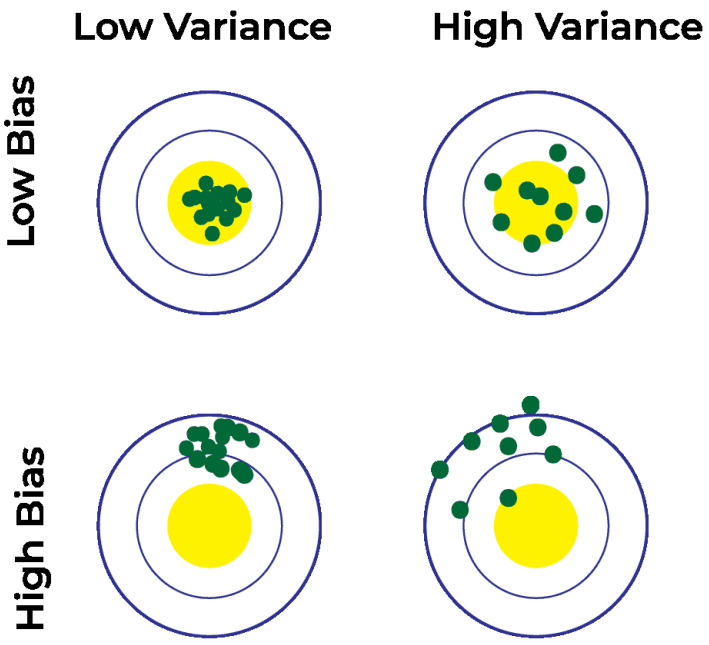
Variance is the model's sensitivity to small fluctuations in the training data. It measures how much the predictions would change if we used a different training dataset.

High variance:

- Model is too complex (overfit)
- Captures noise in the training data
- Predictions vary significantly with small changes in the training set

Low variance:

- Model is simpler and more stable
- Less sensitive to small changes in the training data

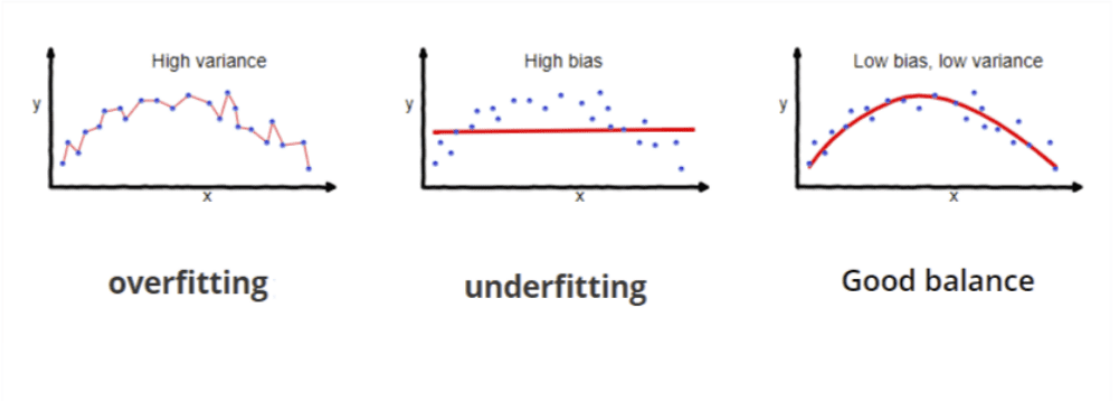


More Accurate Description:

1. Training Data:
    - Shows effects of both bias and variance
    - High bias: Poor performance on training data

- High variance: Very good performance on training data

2. Test Data:
    - Also shows effects of both bias and variance
    - High bias: Poor performance on test data (similar to training data)
    - High variance: Much worse performance on test data compared to training data



Common Regularization Techniques:

a. L1 Regularization (Lasso):

- Adds the absolute value of the magnitude of coefficient as penalty term to the loss function.
- Formula: Loss = Error(Y, Y_pred) + λ * Σ|w|
- Tends to produce sparse models (feature selection).
- Good for feature selection when you have a high number of features.

b. L2 Regularization (Ridge):

- Adds the squared magnitude of coefficient as penalty term to the loss function.
- Formula: Loss = Error(Y, Y_pred) + λ * Σ(w^2)
- Tends to shrink coefficients for less important features, but doesn't eliminate them.

- Generally preferred when you want to keep all features.

c. Elastic Net:

- Combines L1 and L2 regularization.
- Formula: Loss = Error(Y, Y_pred) + λ1 * Σ|w| + λ2 * Σ(w^2)
- Balances the benefits of both L1 and L2.

Lasso Regression (L1)

A regression model which uses the **L1 Regularization** technique is called **LASSO(Least Absolute Shrinkage and Selection Operator)** regression. **Lasso Regression** adds the *"absolute value of magnitude"* of the coefficient as a penalty term to the loss function(L). Lasso regression also helps us achieve feature selection by penalizing the weights to approximately equal to zero if that feature does not serve any purpose in the model.

The change from linear regression is added primarily as L1:

$$cost = 1/m * [\Sigma(y(i) - h(x(i))^2 + \lambda * \Sigma|wj|]$$

Ridge Regression (L2)

A regression model that uses the **L2 regularization** technique is called **Ridge regression**. **Ridge regression** adds the "*squared magnitude*" of the coefficient as a penalty term to the loss function(L).

The change from linear regression is added primarily as L2:

$$cost = 1/m * [\Sigma(y(i) - h(x(i))^2 + \lambda * \Sigma wj^2]$$