Gradient Descent Algorithms

Types of Gradient Descent

1. Batch Gradient Descent

- **Description**: Uses the entire dataset to compute the gradient of the cost function.
- **Advantages**:
    - Converges to the global minimum for convex functions.
    - Stable updates.
- **Disadvantages**:
    - Computationally expensive for large datasets.
    - Requires a lot of memory.
- **Performence**:
    - with Zero Initialization Test Metrics: MSE: 66.6767, R2 Score: 0.8394
    - with Random Initialization Test Metrics: MSE: 66.4981, R2 Score: 0.8398
    - There is not much difference. Going with zero initialisation from now on

2. Stochastic Gradient Descent (SGD)

- **Description**: Uses one training example per iteration to compute the gradient.
- **Advantages**:
    - Faster convergence for large datasets.
    - Requires less memory.
- **Disadvantages**:
    - Updates can be noisy.
    - May not converge to the global minimum but rather oscillate around it.
- **Performance**:
    - MSE: 64.7104, R2 Score: 0.8441

PROFI

3. Mini-Batch Gradient Descent

- **Description**: Uses a small random subset of the dataset to compute the gradient.
- **Advantages**:
    - Balances the efficiency of SGD and the stability of Batch Gradient Descent.
    - Reduces variance in the updates.
- **Disadvantages**:
    - Still requires tuning of batch size.
    - May require more iterations to converge compared to Batch Gradient Descent.
- **Performance**:
    - MSE: 64.6181, R2 Score: 0.8444

Lasso and Ridge Regularization

Influence on the Model

- **Lasso Regularization**:
    - Adds a penalty equal to the absolute value of the magnitude of coefficients.
    - Can shrink some coefficients to zero, effectively performing feature selection.
    - MSE: 63.6213, R2 Score: 0.8468
- **Ridge Regularization**:
    - Adds a penalty equal to the square of the magnitude of coefficients.
    - Shrinks coefficients but does not set any to zero, retaining all features.
    - MSE: 113.3294, R2 Score: 0.7270

Optimal Lambda

- **Lasso**: The optimal lambda is the one that minimizes the test Mean Squared Error (MSE). Based on the chosen lambdas, the optimal value can be determined from the plot of test MSE vs. lambda.
- **Ridge**: Similarly, the optimal lambda for Ridge is the one that results in the lowest test MSE.

Scaling of Features

Effect on Model Performance

- **Improves Convergence**: Scaling features ensures that the gradient descent algorithms converge faster and more reliably.
- **Prevents Dominance**: Prevents features with larger scales from dominating the cost function, leading to better model performance.
- **Consistency**: Ensures that regularization terms (like Lasso and Ridge) penalize all features equally.