**The Importance of Data Standardization in Machine Learning: A Study Using Linear Regression**

**Abstract**

Data standardization is a very important preprocessing step in machine learning workflows. This study analyses the impact of standardization on the convergence and stability of gradient descent algorithm in linear regression. Through simulations and visualizations, this study shows how standardized datasets and loss functions mitigate exploding gradients and accelerate convergence. Findings of this study underline the importance of standardization techniques to ensure convergence during the data training phase.

**Introduction**

Machine learning models often perform sub optimally when trained on datasets with features that vary significantly. The presence of vastly different magnitude of data (non-standardized data) can lead to issues such as vanishing or exploding gradients, making gradient descent ineffective as the gradients either keep oscillating between very big values, or barely changes at all. This study investigates the theoretical and practical benefits of data standardization, particularly focusing on its impact during linear regression training.

**Methods**

**Data Generation**

We generated a dataset with two features (“x” and “y”) derived from a linear equation with added Gaussian noise. The dataset was split into training and testing subsets to avoid data leakage during standardization.

**Standardization Techniques**

1. **Feature Standardization:**
   * Features were standardized to have a mean of zero and a variance of one before training.
2. **Loss Standardization:**
   * The squared error loss was scaled by the number of samples to achieve stability during training. This is supposed to be done either way, as the formula by default scales it.

**Gradient Descent Analysis**

An implementation of gradient descent was used to demonstrate the impact of standardization. The algorithm’s convergence was tracked by logging the gradient of loss with respect to weight over iterations. Hyperparameters such as learning rate and initialization were kept consistent across experiments to isolate the effects of standardization.

**Results**

**Convergence Analysis**

1. **Without Standardization:**
   * Gradients exhibited instability, leading to oscillations and poor convergence.
2. **With Standardization:**
   * Loss and gradients stabilized, enabling smooth convergence even at higher standard deviation of data.

**Visualizations**

The gradient of loss with respect to weight was plotted for varying standard deviation of the synthetic data, to see how the gradient changes. This was done for variants of the model where one variant using standardization, while the other kept the data non-standard, and another variant where the loss function in itself had the loss unscaled (spoiler: it performed terribly).

The visualizations can be found in the plots folder in the GitHub repository of this project.

Github : https://github.com/AdityaKulkarni2706/regression-analysis

**Practical Implications**

Standardizing features before training mitigates exploding gradients and ensures consistent optimization. Loss standardization further complements feature scaling by enhancing numerical stability during gradient computation.

**Common Pitfall: Data Leakage:**

Standardizing data before splitting into training and testing sets makes the training data dependent on the testing data, which basically introduces a bias, which also is classified as data leakage, although by indirect means as no one ever intends to leak their testing data into the training data.

**Conclusion**

This research underscores the importance of data standardization in machine learning workflows. By addressing scaling issues in both features and loss functions, standardization significantly improves convergence behavior and model effectiveness.