109550198, 卜銳凱

PM2.5 Prediction Model Evaluation Report

Introduction

This report assesses the efficacy of a Linear Regression model developed to predict PM2.5 levels, which are indicators of air quality. The model provides predictions based on a set of features taken from a given dataset.

1. Feature Selection and Preprocessing

Approach

Extracted features from a CSV dataset containing time-series data on air quality.
Handled non-numeric values by converting them to NaN and then imputing with zero values.
Constructed feature vectors by flattening the relevant columns from the dataset.
Preprocessing Steps

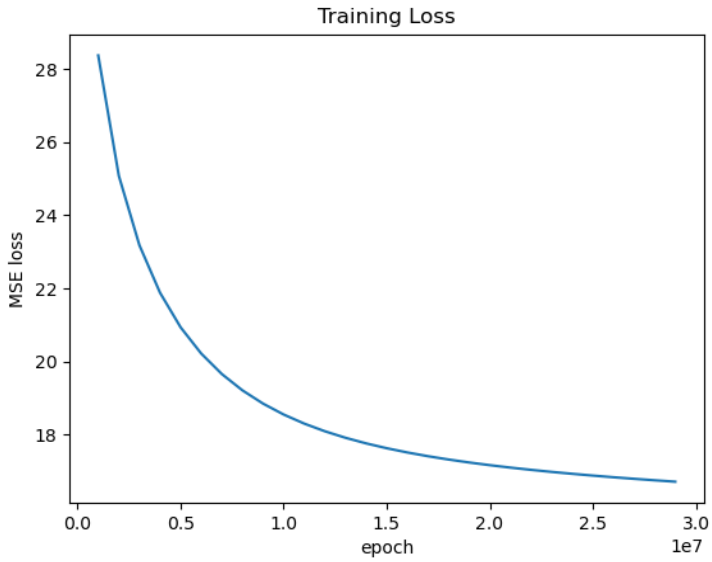
Numeric Conversion: Applied pd.to_numeric with the errors='coerce' parameter.
Imputation: Filled NaN values with zero, assuming missing data does not contribute to PM2.5 levels.
Data Flattening: Transformed multi-dimensional arrays into one-dimensional arrays for model consumption.

Observations

Feature selection is crucial for model performance. It was considered that the selected columns were directly related to predicting PM2.5 levels. The use of zero imputation may have an influence on the model's performance if the missing data have a non-zero impact on PM2.5 levels.

2. Impact of Training Data Volume

Methodology
The model was trained over a range of epochs to minimize the Mean Squared Error (MSE) loss function, as visualized in the provided plot.



Observations
The plot shows a decreasing trend in training loss with increasing epochs, indicating learning and improvement in model accuracy over time. However, the plot does not directly show the impact of varying training data volume on accuracy. Typically, more training data leads to better model generalization but can also introduce complexity and noise.

Recommendations
To comprehensively evaluate the impact of training data volume, perform experiments with varying sizes of the training set, and plot training and validation loss against the amount of data used.

Analysis

Larger datasets typically enable the model to learn more nuanced patterns, hence improving prediction accuracy.
However, without varying the dataset size and visualizing its direct impact on MSE, it is challenging to conclude the precise effect.

3. Regularization and Prediction Accuracy

L1 (lasso) and L2 (ridge) regularization approaches were identified as viable additions to the loss function during model training.

Observations
Regularization can prevent overfitting by punishing the loss function, deterring the model from becoming overly complicated. This is especially beneficial when working with high-dimensional or tiny datasets.

Recommendations
Conduct a set of trials that include L1 and L2 regularization. Evaluate the model's performance

under various regularization strengths and compare the MSE loss on a validation set. This will yield the best regularization parameter for minimizing overfitting while preserving prediction accuracy.

Conclusion

The linear regression model has the capacity to predict PM2.5 values, with a decrease in MSE loss across training epochs. However, more testing is needed to determine the effect of training data amount and the advantages of regularization on prediction accuracy.