A2: Regression Model Selection in Predicting Wear Rate of Mechanical Component

Discussion:

## Introduction

In the field of mechanical engineering, predicting the wear rate of components is crucial for ensuring the reliability and longevity of machinery. To tackle this challenge, we embarked on a journey of data analysis and predictive modeling. Our goal was to create accurate models for estimating wear rates based on three fundamental features: rotational speed (RPM), normal load (Load), and material hardness (Hardness). We employed polynomial regression models of varying degrees and introduced Ridge regularization to enhance predictive accuracy. Additionally, we adopted 5-fold cross-validation to assess model generalization.

## Part 1: Polynomial Models

### Data Preprocessing

Our journey began by importing the training dataset, "training\_data.csv." This dataset consisted of RPM, Load, Hardness, and Wear\_Rate columns. Fortunately, the data was relatively clean and did not require extensive preprocessing, which allowed us to focus on the modeling process.

### Polynomial Regression with Cross-Validation

To understand the relationship between our input features and wear rate, we constructed polynomial regression models with degrees ranging from 1 to 6. For each degree, we generated polynomial features based on RPM, Load, and Hardness. The pivotal step was applying 5-fold cross-validation to evaluate model performance. We leveraged the Root Mean Square Error (RMSE) as our performance metric—a crucial measure of predictive accuracy.

### Results

Our exploration yielded a significant insight: the polynomial order that minimized the average RMSE across the 5 folds was 4. This finding suggests that a polynomial regression model of degree 4 provides the best representation of the intricate relationship between input features and wear rate. To visualize this relationship, we created a graph depicting the correlation between polynomial order and average RMSE.

![Polynomial Order vs. Average RMSE](polynomial\_order\_vs\_rmse.png)

### Cross-Validation

The principle of 5-fold cross-validation entails splitting the data into five subsets, known as folds. During each iteration, one fold is designated as the validation set, while the remaining four folds serve as the training data. RMSE is calculated for each fold, and the average RMSE across all folds is reported. This technique ensures that our models' predictive performance is evaluated comprehensively and helps us assess their generalizability to unseen data.

## Part 2: Ridge-regularized Polynomial Models

### Polynomial Models with Ridge Regularization

In this phase, we introduced Ridge regularization to our polynomial regression models. We considered polynomial orders ranging from 1 to 6 and explored a range of penalty coefficients (alphas) to control the degree of regularization.

### Results

Interestingly, the combination that yielded the lowest average RMSE across the 5 folds differed from the previous results. In this case, a polynomial order of 5 with a Ridge alpha of 0.1 emerged as the best-performing model. Ridge regularization introduces a penalty term into the model's loss function, encouraging smoother and more generalizable models. This shift towards a higher-degree polynomial underscores the importance of regularization in preventing overfitting and enhancing predictive accuracy.

## Saving Models for Evaluation

To ensure that our models can be evaluated on a separate test dataset, we used Python's joblib library to save the best-performing models from both Part 1 and Part 2. These saved models can undergo rigorous testing to validate their real-world predictive performance.

## Conclusion

In conclusion, our data analysis and modeling endeavors have culminated in the selection of two top-performing models:

1. A polynomial regression model of degree 4 without regularization.

2. A Ridge-regularized polynomial regression model of degree 4 with a regularization alpha of 0.001.

The choice between these models may depend on specific requirements, such as the desired balance between model complexity and regularization strength. To ensure their practical utility, we recommend evaluating these models on an independent test dataset.