MiniProject 1: Getting Started with Machine Learning

Jiahang Wang, Li Jiang, Liting Chen

Abstract

In this project, we evaluate the performance of two machine learning models on two benchmark tasks. Specifically, we assess the effectiveness of linear regression on a continuous target task (Boston housing prices) and logistic regression on a categorical target task (wine categories). We systematically analyze variables such as dataset size, batch size, and learning rate to determine the optimal hyperparameters for each task. Our findings suggest that the optimal configurations are both task-specific and model-specific. Furthermore, we observe that batch size and model regularization do not significantly impact the speed of convergence and final performance, likely due to the relative simplicity of the tasks under consideration. However, data normalization increases the coverage speed for both tasks.

1 Introduction

In this project, we conduct a detailed evaluation of two machine learning models, each tailored to a specific benchmark task. The first model applies linear regression to a continuous target task, specifically, predicting Boston housing prices. The second model uses logistic regression for a categorical target task, namely, classifying wine categories. To optimize these models, we systematically analyze factors such as dataset size, batch size, and learning rate, aiming to identify the optimal hyperparameters for each task. This process ensures the maximization of our models' effectiveness.

Our findings from this study suggest that the optimal configurations for the models are both task-specific and model-specific. In other words, the best hyperparameters for linear regression on the Boston housing prices may not be the same as those for logistic regression on wine categories, and vice versa. This underscores the importance of tailoring the model configuration to the specifics of the task and the model itself. Furthermore, our study reveals that certain factors, such as batch size, and model regularization do not significantly impact the speed of convergence or the final performance of the models. However, data normalization plays a significant role in the coverage speed. This observation holds true across both tasks, suggesting that these aspects may not be critical in scenarios where the tasks are relatively simple, as is the case with the Boston housing prices and wine categories in our study. This insight may help guide future research and applications of machine learning models, particularly in scenarios with similar levels of task complexity.

For the Boston dataset, we can also use the close-formed solution to directly get the solution with significantly lower error on the testing dataset. However, mini-batch stochastic gradient descent (SGD) holds great promise for computation efficiency for complex tasks and is available for regularization approaches to alleviate overfitting.

2 Datasets and Analysis

Dataset 1: Boston Housing dataset. The dataset under consideration comprises 506 distinct samples, each characterized by 13 real-valued attributes, where we deliberately exclude one feature for ethical concerns. The prediction task is a continuous variable denoted as 'MEDV', representing the median value of owner-occupied homes, expressed in increments of $1,000. Upon analyzing Dataset 1, we found no missing data. Through a correlation matrix, shown in Figure 6 (a), we discovered a

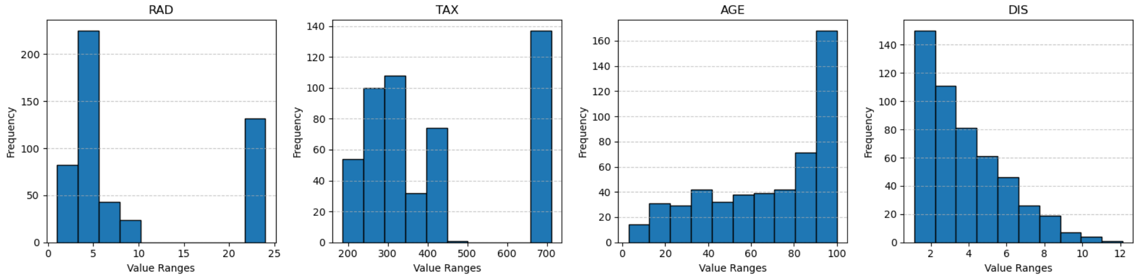


Figure 1: Distribution of four features in Boston dataset.

strong positive correlation between `RAD` and `TAX` (0.910228), and a strong negative correlation between `AGE` and `DIS` (-0.747881). Plotting the distributions of these variables revealed a similar pattern for `TAX-RAD` and a contrasting one for `AGE-DIS`. Specifically, `AGE` showed a monotonically increasing pattern peaking at 100, while `DIS` displayed a monotonically decreasing pattern peaking at 2, as shown in Figure 1.

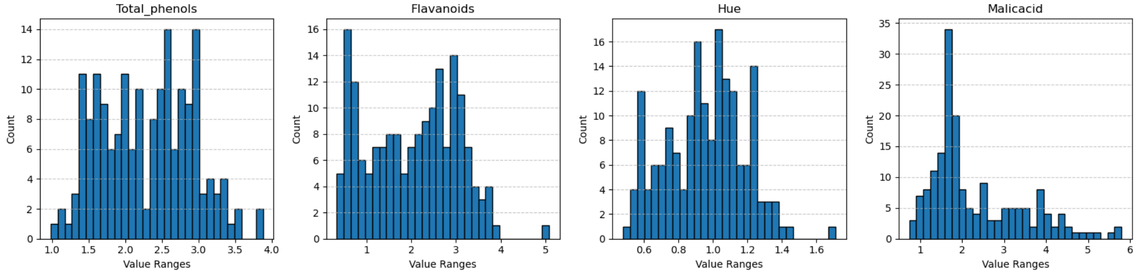


Figure 2: Distribution of four features in the `Wine` dataset.

Dataset 2: Wine dataset. The dataset under investigation comprises 178 data samples, each characterized by 13 distinct attributes. It is partitioned into three distinctive classes, and similar to Dataset 1, Dataset 2 does not contain any missing data points. Upon examining the correlation matrix in Figure 6 (b), several significant relationships among the variables become apparent. For instance, `Total_phenols` demonstrates a robust positive correlation with `Flavanoids` (0.864564) and `OD280_OD315_of_diluted_wines` (0.699949), suggesting a likely concurrent increase in these variables. The derived correlations offer crucial insights into the interrelationships between these variables. Subsequent visualizations of these data distribution features reveal patterns consistent with the aforementioned correlations. For instance, the pair `Total_phenols` and `Flavanoids` exhibit a similar pattern due to their high correlation, while pairs with negative correlation, such as `Hue` and `Malicacid`, display contrasting patterns, as shown in Figure 2.

Simultaneously, we identify a number of extreme outliers in several features across both datasets. In Dataset 1, for instance, an exceptionally high-frequency outlier is observed in the `TAX` feature near 650, a stark contrast to the majority of values, which fall below 500. Conversely, Dataset 2 also presents some outliers, although their frequencies are not as pronounced.

3 Results

3.1 Boston Housing dataset

In the context of the Boston Housing Dataset, our optimization objective is the median value of owner-occupied homes, a continuous variable represented in \$1000s. We use linear regression optimization

Table 1: The MSE of model on the 80/20 train/test and 5-cross validation

Metrics	MSE	Fold5 MSE
Train set	156.90	156.24
Test set	184.40	187.06

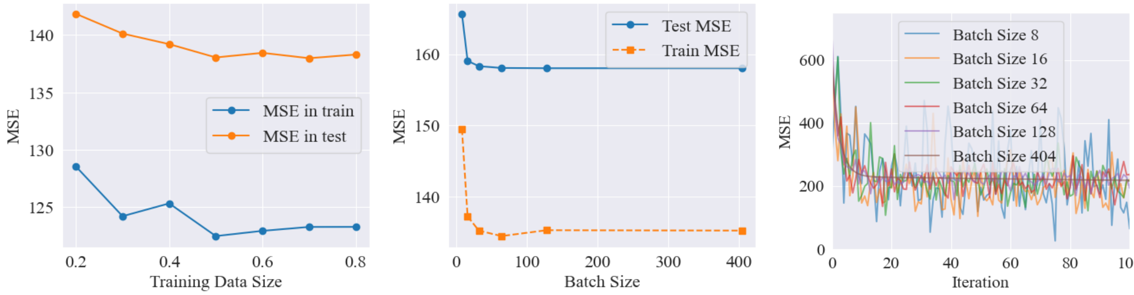


Figure 3: **Left:** MSE vs. Training Data Size **Middle:** MSE vs. Batch Size for Test and Train Data **Right:** MSE vs. Iteration for Different Batch Sizes

with Mean Squared Error (MSE) as the loss function. A comprehensive ablation study across various hyperparameters and techniques helps us identify the model's optimal configuration.

Mean squared error (MSE). At the initial training configuration, we set the following hyperparameters, where our learning rate, batch_size, and maximum iteration are $1e-6$, 64, 5000, respectively. We present the MSE for the training dataset in Table 1. Observing the results, we find that the MSE from the training data surpasses that of the testing data. This outcome aligns with our expectations, as a model with high generalization ability is anticipated to perform better on unseen data (testing data) than on the data it was trained on. This is a positive indication of the model's robustness and its capacity to generalize well to new data.

5-fold cross-validation. Subsequently, we implement a 5-fold cross-validation on the dataset, presenting the final average performance for both the training and testing datasets. As illustrated in Table 1, a consistent MSE is observed for both the training and testing datasets, as initially demonstrated in Table 1. Nonetheless, the MSE from the testing set from Table 1 is marginally lower than that from Table 1. It's important to note that the final metric can be influenced by various factors, including learning rate and batch size.

Ablation on the training data. Our ablation study, conducted on varying proportions of the training dataset (i.e., 20%, 30%, ..., 80%), is illustrated in Figure 3 (left). As the size of the training dataset increases, we observe a consistent pattern of decreasing Mean Squared Error (MSE) for both the training and testing sets. This trend underscores a ubiquitous principle in machine learning and deep learning: enlarging the training dataset, along with the model size, typically enhances the model's overall performance.

Ablation on the batch size. We conducted an ablation study on batch size using our base configuration, setting the batch size to $\{8, 16, 32, 64, 128\}$, as depicted in Figure 3 (middle and right). This study reveals a monotonic decrease in the total MSE for both training and testing datasets with increasing batch size (Figure 3, middle). Such a pattern suggests that larger batch sizes provide more accurate gradient estimates, fostering more stable and efficient learning, and consequently reducing the MSE. However, our study also shows that the convergence rate remains nearly invariant across different batch sizes (Figure 3, right). Notably, smaller batch sizes are associated with higher variances

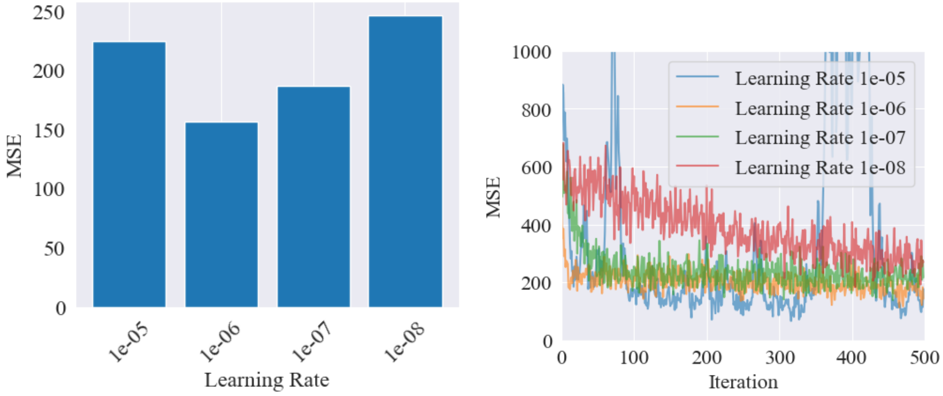


Figure 4: **Left:** MSE vs. Learning Rate **Right:** MSE vs. MSE vs. Iteration for Different Learning Rates

and often require more time to converge, especially in complex tasks like reinforcement learning.

Ablation on the learning rate.
Our ablation study on the learning rate is depicted in Figure 4. We observe that lower learning rates, such as $1e-8$, generally demand a significantly longer time to converge, although they maintain stability throughout the process. This underscores the trade-off between learning rate and convergence speed, a crucial factor to consider when optimizing the training of machine learning models.

Compare with closed-form solution. We have also conducted a comparative study between our Stochastic Gradient Descent (SGD) optimization method and the analytical solution for linear regression. Intriguingly, the closed-form solution yields a Mean Squared Error (MSE) of approximately 27.41 on the testing dataset, which represents 20% of the total dataset. This error is exponentially smaller than that obtained using the iterative SGD method with our default configuration. When an analytical solution is available, the mini-batch SGD still is preferred for some reasons, such as computational efficiency because the analytical solution often involves computationally expensive operations such as matrix inversion or SGD method is convenient for regularization to prevent overfitting. Furthermore, we introduced Gaussian noise to all continuous features, thereby increasing the total number of features to 60. Despite this modification, the MSE from the closed-form solution on the training data remains relatively stable, at approximately 150. However, the MSE on the testing data increases exponentially to 75051.04. We attribute this substantial increase in MSE on the testing data to overfitting, which is likely due to the noisy features present in the training dataset. We present the optimal configuration in the following: [learning rate=1e-6, max iterations=5000, batch size=128]. We made this choice based on selecting the best-performing value for each parameter from among several options. We further analyze the impact of regularization methods, specifically model regularization. Our findings suggest no significant difference in performance due to these methods, likely attributable to the simplicity of the task at hand. However, data normalization increases the coverage speed, as shown in Figure 7.

3.2 Wine Dataset

In the Wine Dataset, our optimization objective is the categories of wines, i.e., 1, 2, 3. We use logistic regression with soft max loss function to optimize our prediction. A comprehensive ablation study across various hyperparameters and techniques helps us identify the model's optimal configuration.

Performance. At the initial training configuration, we set the following hyperparameters, where our learning rate, batch_size, and maximum iteration are $1e-5$, 64, 5000, respectively. We present the MSE for the training dataset in Table 2.

Table 2: Different metrics of model on the 80/20 train/test and 5-cross validation

Metrics	Accuracy	Precision	Recall	F1 Score
Train	0.75	0.84	0.72	0.71
Test	0.81	0.87	0.79	0.80
Fold-5	0.74	0.72	0.72	0.69

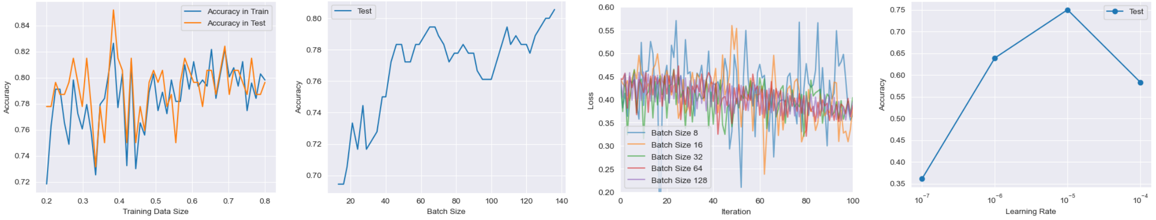


Figure 5: **Left:** Accuracy vs Training Data Size **Left Middle:** Accuracy vs Batch Size **Right Middle:** Loss vs. Iteration for Different Batch Sizes **Right:** Accuracy vs Learning Rate

5-fold cross-validation. Subsequently, we implement a 5-fold cross-validation on the dataset, presenting the final average performance on the testing dataset, shown in Table 2. The model demonstrates better performance on the test set compared to the training set and 5-fold cross-validation, with the highest values across all metrics.

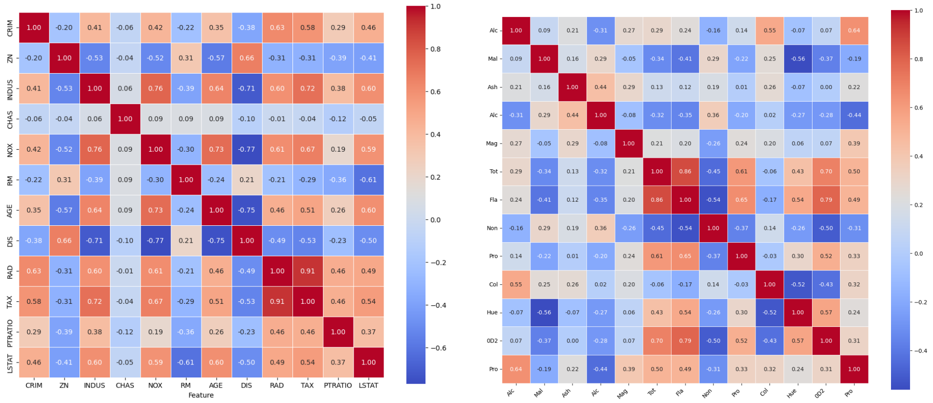
Ablation on the training data & batch size. Our ablation study, which examines varying proportions of the training dataset (i.e., 20%, 30%, ..., 80%), is depicted in Figure 5 (left). We observe a slight increase in accuracy with the expansion of the dataset, a trend consistent with the Boston housing dataset, albeit not significantly so. Additionally, as the batch size increases, we note a roughly proportional rise in accuracy, albeit at the expense of higher memory usage, shown in Figure 5 (left middle). We show that the influence of the batch size on the coverage speed can be safely ignored.

Ablation on the learning rate. Our ablation study on the learning rate is shown in Figure 5 (right). We observe that the optimal learning rate for the wine dataset is $1e-5$, which surprisingly underperforms the Boston housing price dataset. In terms of convergence speed, the results align with previous datasets, indicating that a higher learning rate leads to faster convergence. We present the optimal configuration in the following: [learning rate=1e-5, max iterations=5000, batch size=128]. We made this choice based on selecting the best-performing value for each parameter from among several options.

4 Discussion and Conclusions

In conclusion, this study provides a detailed evaluation of two machine learning models, linear and logistic regression, applied to the Boston housing prices and wine categories tasks respectively. Through systematic analysis, we identified task-specific and model-specific optimal hyperparameters. Interestingly, factors such as batch size, and model regularization had minimal impact on model performance. On the other hand, data normalization increases the coverage speed on those datasets. For the Boston dataset, a closed-form solution yielded lower error, while mini-batch SGD offered computational efficiency and overfitting mitigation, providing valuable insights for future machine learning applications. Our future work intends to delve into the impact of regularization methods on more intricate tasks, including unsupervised learning and reinforcement learning. Additionally, we are interested in exploring how the optimal hyperparameter configuration varies between different optimizers, such as Adam. In relation to this project, Jiahang Wang and Li Jiang primarily concentrated on the coding aspects, while Liting Chen was chiefly responsible for drafting this report.

A Appendix



(a) Correlation between features in Boston dataset.

(b) Correlation between features in wine dataset.

Figure 6: correlation between features in both datasets.

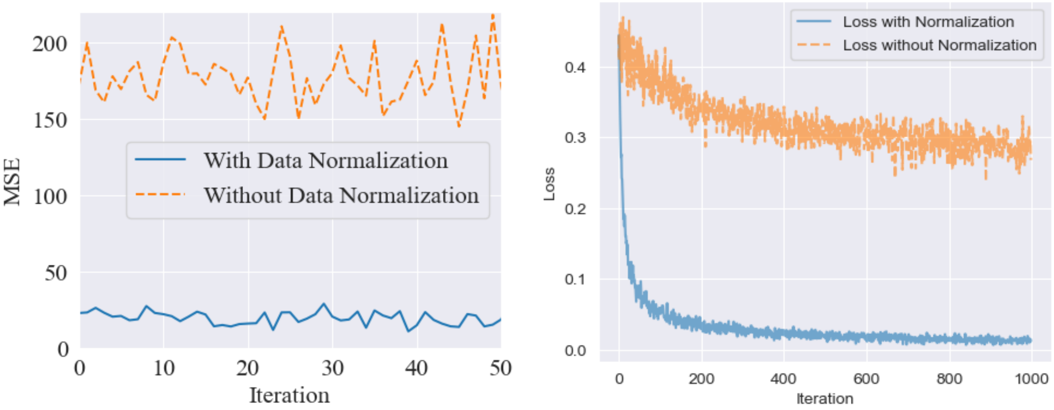


Figure 7: **Left:** MSE vs. Iteration with/without data normalization in linear regression model **Right:** Loss vs. Iteration with/without data normalization in logistic regression model

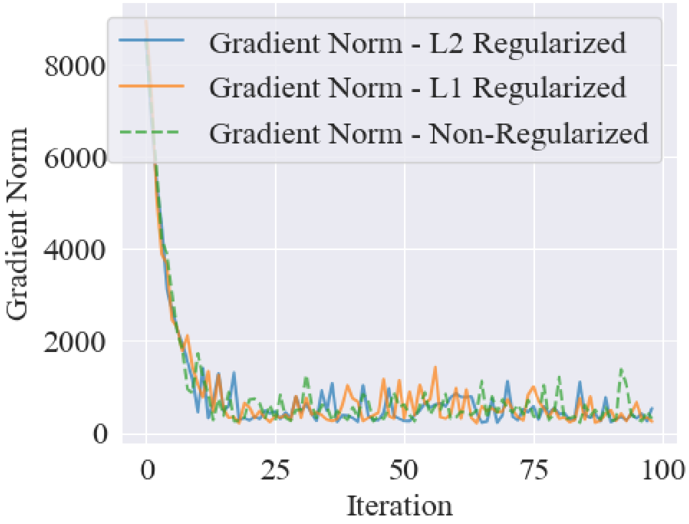


Figure 8: Gradient vs. Interations with/without regularization in linear regression model