Assignment 1 Write Up: Getting Started with Machine Learning

Students Rob Li, Ian Sonoda McFarland, Finnley Howald
Professors Reihaneh Rabbany, Isabeau Prémont-Schwarz

1 Abstract

In this assignment, two machine learning models were implemented: linear regression and logistic classification. The performance of the models was evaluated on two datasets: the Infrared Thermography Temperature dataset for linear regression and the CDC Diabetes Health Indicators dataset for classification. In addition to implementing these models, we prepared the datasets through preprocessing steps such as handling missing data, transforming categorical variables, and scaling features.

We developed both an analytical solution for linear regression, a gradient descent-based approach for logistic regression, and an incorporated minibatch stochastic gradient descent for both models. The project investigates the effects of key factors such as batch size, learning rate, and training set size on model performance. After observing feature weights after fitting, we gained insights on the importance of different features on the prediction of our target features. All in all, this project shows the importance of dataset preparation and optimization strategies when implementing different machine learning models.

2 Introduction

The task of this assignment was to develop a Linear Regression model using an analytical and minibatch stochastic gradient descent solution and a Logistic Regression model using a full batch gradient descent and minibatch gradient descent solution.

The Linear Regression model was used on the Infrared Thermography Temperature dataset and aims to predict the average oral temperature of a patient given features such as age, gender, ethnicity and ambient temperature. Research in this area shows that older people tended to have the coolest and African-American women the hottest individual-level temperature (Mullainathan et al., 2017 [2]). Our findings showed that African-Americans tended to have higher average oral temperature when using the Analytical Linear Regression model but did not support elderly people being the coolest, this could be attributed to the use of different datasets.

The Logistic Regression model was used on the CDC Diabetes Health Indicators dataset. The model aims to predict whether or not a person has diabetes given features such as health history, income, age, and gender. A study in this field showed that elderly people with type 2 diabetes tended to have a smoking history and diastolic blood pressure (Zhang et al., 2022 [4]). Our work supports this as in our model, blood pressure and smoking history is correlated to diabetes.

We found that in Linear Regression, the analytical solution provides a more accurate model than using stochastic gradient descent; however, it is evident that it has a more expensive runtime. It was also found that varying minibatch size and different split ratios did not drastically affect converging rates in the Logistic Regression.

3 Datasets

3.1 Thermography Temperature Dataset

For Dataset 1: Infrared Thermography Temperature dataset, two rows contained null values. We changed those null values to be the average of all the other values of their respective features. Next, we scaled the numerical columns to standardize the range of the values for a more consistent model performance. The categorical variable ”Gender” was transformed into a binary column, while the variables ”Age” and ”Ethnicity” were one-hot encoded to ensure they were correctly represented in the regression model. During data preprocessing, we also identified an outlier in the ”Distance” feature, which we changed to the average of all other distance values.

3.2 CDC Diabetes Health Indicators Dataset

For Dataset 2: the CDC Diabetes Health Indicators dataset, we left the ordinal categorical variables ”Age,” ”Income,” and ”Education” in their original form, as these were already appropriately structured for classification. No null rows were present in this dataset, thus, no additional missing data handling was required. We also noticed a large class imbalance in the target feature as there are many more individuals in the Diabetes Dataset who do not have diabetes.

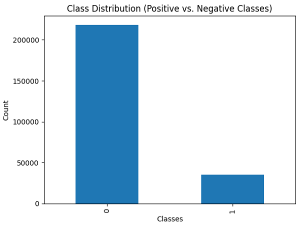


Figure 1: Diabetes Dataset: class imbalance

3.3 Dataset Analysis

An ethical concern with Dataset 2 arises from its representation of patient data. While the majority of patients in the dataset had average BMI, mental health, and physical health, there were heavy tails in these distributions, representing patients with severe mental illness or physical health issues. This skewed distribution means the dataset does not fully capture the range of health conditions, potentially leading to biased predictions. A similar concern exists for Dataset 1, where older age categories were underrepresented compared to younger individuals. Since these datasets involve healthcare, the inclusion of demographic features such as gender and ethnicity may be relevant. However, reckless applications of these categories especially in finance, law enforcement, and medical access, could raise civil rights concerns due to potential biases and legal implications.

4 Results

4.1 Linear Regression

We first ran both of our Linear Regression models to see how well each model predicted the target values in our test set. For Analytical Linear Regression, our r^2 value against the training set trended around 0.7 - 0.75, depending on the randomness after the shuffling of the dataset, and against the test set, the r^2 varied from 0.6 - 0.7. For Minibatch Stochastic Linear Regression, with the default hyperparameters of: learning rate = 0.01, max iterations = 10000, epsilon = 1e-8, and batch size = 32, our r^2 value was around 0.4 - 0.6 against the training set, and 0.3 - 0.45 for the test set, depending on the randomized instances selected in the training set. Comparing the r^2 values, it is evident that the Analytical Linear Regression model performed better than the Stochastic model on average. Furthermore, if we tune the hyperparameters of this model to the optimal values of learning rate = 0.1, max iterations = 10000, epsilon = 1e-8, and batch size = 16, we can see an r^2 value of up to 0.61 (see section 5).

Additionally, when looking at the weights that both of the models gave each feature after fitting, we observed an interesting variation. Notably, in the Stochastic Linear Regression, with the default hyperparameters, each ethnicity feature's weight is extremely high. This may mean that although there are more multiracial and white individuals, since every ethnicity's weight is high, the ethnicity does not correlate to the target's average oral temperature.

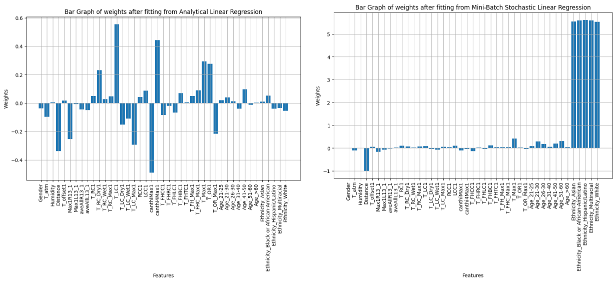


Figure 2: Linear Regression Weights

4.1.1 Varying the Training Set Sizes

When varying the training set sizes between 20 - 80%, we noticed the r^2 values for the test sets were proportional to the split size. This is because as the training set size increases, the model can be trained on more data, thus, giving more accurate predictions for the test set. This trend only applies to the Analytical model.

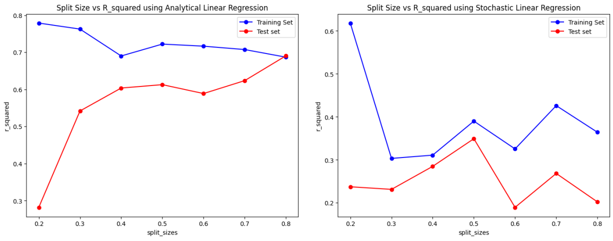


Figure 3: Linear Regression: Split Size vs R^2

4.1.2 Varying the Minibatch Sizes

We also examined the effect of changing batch sizes on the Minibatch Stochastic Gradient Descent Linear Regression model. Using minibatch sizes of 8, 16, 32, 64, and 128, we plotting the convergence speed of the model. We saw no significant change in the convergence speed using different batch sizes under the same fixed max iterations. However, a noticeable difference is that the lower batch sizes sees greater oscillations in their loss functions. We see the highest r^2 value for a batch size of 16.

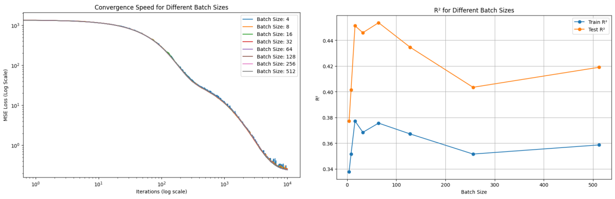


Figure 4: Linear Regression: Batch Size vs R^2

4.1.3 Varying the Learning Rate

Finally, we examined the performance of the stochastic model based on different learning rates (0.001 - 0.2). We observed that the range (0.01 - 0.1) gave the best r^2 values, with performance dropping before and after this range. The gradient descent diverged at a learning rate of 0.2, as seen by the r^2 value of 0.

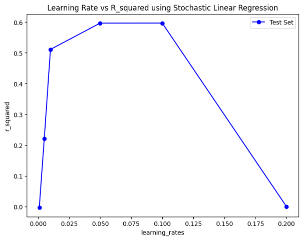
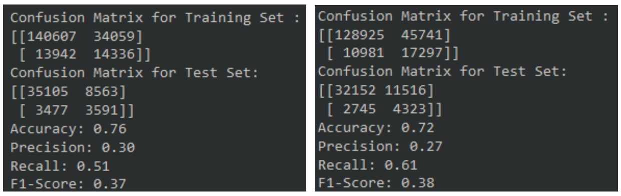


Figure 5: Linear Regression: Learning Rates vs R^2

Comparing both models, we see that the Analytical Linear Regression performs much better on average than the Stochastic Model. However, in separate runs of the same experiment, the learning rate of 0.2 on the Stochastic Model would outperform the other learning rates. This is due to the fact that a high learning rate added to the Stochastic Noise sometimes results in a very accurate prediction. However, this is not a consistent trend, and only occurs in one off cases.

4.2 Logistic Regression

First, we tested both Regular Gradient Descent, and Minibatch Stochastic Gradient Descent Logistic Regression models with the default hyperparameters of: learning rate = 0.01, max iterations = 1000, epsilon = 1e-8, and batch size = 32. After multiple rounds of testing, we also chose a threshold/decision boundary of 0.2 for a value to be positive. This is attributed to the fact that the class imbalance in our target feature is very large: there are many more people that do not have diabetes than people that do. The models performed as such:



```
Confusion Matrix for Training Set :
[[140607  34059]
 [ 13942  14336]]
Confusion Matrix for Test Set:
[[35105  8563]
 [ 3477  3591]]
Accuracy: 0.76
Precision: 0.30
Recall: 0.51
F1-Score: 0.37
```

```
Confusion Matrix for Training Set :
[[128925  45741]
 [ 10981  17297]]
Confusion Matrix for Test Set:
[[32152 11516]
 [ 2745  4323]]
Accuracy: 0.72
Precision: 0.27
Recall: 0.61
F1-Score: 0.38
```

Figure 6: Logistic Regression: Regular (on left) and Stochastic (on right) Gradient Descent performance

The weights found by each model were consistent with each other:

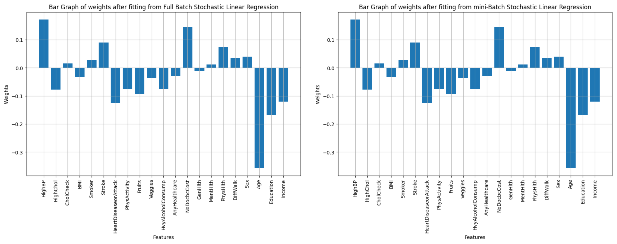


Figure 7: Logistic Regression: Weights

Features such as high blood pressure, stroke, and lack of doctor access due to cost were highly positively weighted which makes sense in diagnosing diabetes patients. Features such as age, education, and income were negatively weighted, demonstrating that the lower those features were, the more likely an individual was to have diabetes. We hypothesize that this is because people with a diabetes diagnosis on average have a shorter lifespan and that the lower one's education and income, the harder it is to eat healthy.

4.2.1 Varying the Training Set Sizes

Varying the test split sizes for both Regular Gradient Descent, and Minibatch Stochastic Gradient Descent Logistic Regression models showed interesting results. Even while changing the split sizes, the regular gradient descent model showed no significant positive or negative trends in accuracy, precision, and recall, while the stochastic model saw much more significant oscillations.

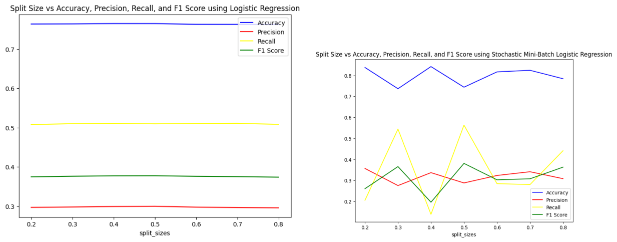


Figure 8: Logistic Regression: Split Size vs Accuracy, Precision, Recall, and F1 Score

4.2.2 Varying the Minibatch Sizes

Testing with different minibatch sizes (8, 16, 32, 64, 128, 256, 512), again shows that despite changing the size of the minibatches, the convergence rate of each model is not significantly affected. However, the small batch sizes resulted in severe oscillations which could undermine performance. This issue is resolved with larger batches.

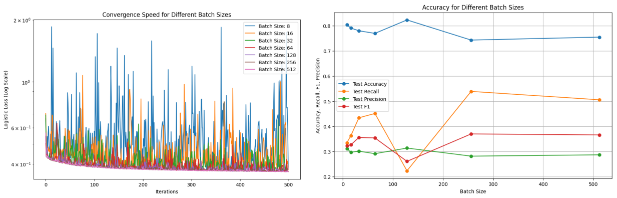


Figure 9: Logistic Regression: Convergence, and Accuracy, Recall, Precision, and F1 Score vs batch size

4.2.3 Varying the Learning Rate

Finally, the performance of both models based on learning rates. There seems to be little correlation between the learning rate and our performance metrics in both models. Other than a bit of randomness in the Stochastic Model, due to small batch sizes, it appears that due to the sheer size of the training data (250000+ values for each feature), the learning rate has little effect on performance.

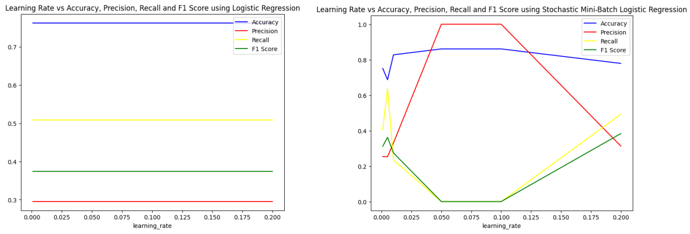
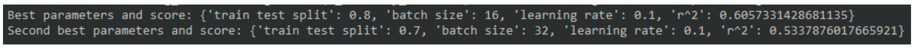


Figure 10: Logistic Regression: Learning Rate vs Accuracy, Precision, Recall, and F1 Score

5 Originality/Creativity

Rather than focusing on optimizing specific hyperparameters with other hyperparameters fixed, we sought to find the best combination of hyperparameters for our mini-batch stochastic linear model. We achieved this with a simple search method over a range of various parameters, and we discovered that the best hyperparameters were:



```
Best parameters and score: {'train test split': 0.8, 'batch size': 16, 'learning rate': 0.1, 'r^2': 0.6057331428681135}
Second best parameters and score: {'train test split': 0.7, 'batch size': 32, 'learning rate': 0.1, 'r^2': 0.5337876017665921}
```

Figure 11: Linear Regression: optimal hyperparameters

However, running this experiment many times resulted in the same observation as in section 4.2.3: a high learning rate coupled with a low batch size resulted oscillations that would sometimes preform better. This issue inspired us seek out other improvements to our mini-batch stochastic gradient descent methods.

We decided to try to implement momentum-velocity into both mini-batch stochastic models.

We discovered that our momentum implementation did not reduce our stochastic oscillations, rather high momentum often increased the severity of oscillations as the velocity amplifies oscillations with previous oscillations. Nevertheless, we did notice an improved convergence for the linear model.

The resulting graph shows a noisier convergence, however, as the convergence happens much faster. For instance, at 100 iterations, MSE is at less than 100, whereas in the analytical case, at 100 iterations, MSE is at least 200 - 300 (Figure 4.).

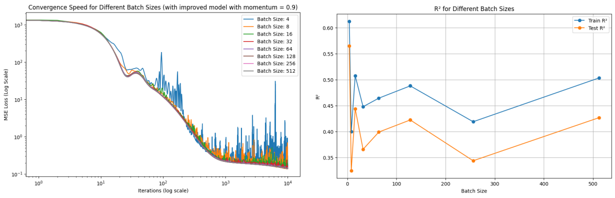


Figure 12: Linear Regression: Convergence rate with Momentum

Although we also implemented momentum for the logistic model, we were less interested in any improvement in convergence. The severity of the increased stochastic vacillations detracted from any improvement in convergence.

6 Discussion and Conclusion

In this assignment, we implemented and evaluated both Linear Regression and Logistic Regression models. Through these implementations, we explored the impacts of varying the training and test split sizes, as well as the impact of hyperparameters such as learning rate and batch size on model performance. A clear conclusion in Linear Regression was the superiority of the analytical solution compared to the stochastic gradient descent method without finely tuned hyperparameters, as seen from consistently higher r^2 values across different experiments. The performance gap between these two methods highlights the efficiency of the brute force least squares equation versus the gradient descent version.

For the Logistic Regression task, we observed consistent performance between regular gradient descent and mini-batch stochastic gradient descent. Analyzing the feature weights after fitting revealed that variables like high blood pressure, income, and access to healthcare played a significant role in diabetes classification. Additionally, the experiments showed that minibatch sizes and different split ratios did not drastically affect convergence rates, suggesting that the model is relatively stable across varying conditions.

Moving forward, further tuning of the stochastic gradient methods could improve their performance, particularly in Logistic Regression where batch sizes and learning rates may require more fine-tuning. Additionally, expanding the models to handle regularization could provide better generalization to unseen data.

Overall, this project highlighted the importance of both model design and data preparation in machine learning tasks. Hyperparameters can significantly influence a model's performance, making it crucial to thoroughly evaluate and experiment with these elements in any machine learning pipeline.

7 Statement of Contributions

Rob Li: Worked on Task 1, focusing on the Infrared Thermography Temperature dataset, contributed to Task 2, and worked on the experiments with batch size, tuning, and momentum.

Ian Sonoda McFarland: Handled Task 1 for the CDC Diabetes Health Indicators dataset, contributed to Task 3 (parts 1, 2, 3, 5, and 6), and was responsible for the Overleaf writeup.

Finnley Howald: Contributed to Task 3 (parts 1, 2, 3, 5, and 6), performed codebase cleanup, fixed bugs, and assisted with the Overleaf writeup.

References

[1] CDC diabetes health indicators. *UCI Machine Learning Repository.*
https://archive.ics.uci.edu/dataset/891/cdc+diabetes+health+indicators

[2] Obermeyer, Z., Samra, J. K., & Mullainathan, S. (2017, December 13). *Individual differences in normal body temperature: Longitudinal big data analysis of patient records.* BMJ (Clinical research ed.). https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5727437/

[3] Wang, Q., Zhou, Y., Ghassemi, P., Chenna, D., Chen, M., Casamento, J., Pfefer, J., & Mcbride, D. (2023). *Facial and oral temperature data from a large set of human subject volunteers (version 1.0.0).* PhysioNet. https://doi.org/10.13026/3bhc-9065.

Wang, Q., Zhou, Y., Ghassemi, P., McBride, D., Casamento, J. P., & Pfefer, T. J. (2022). *Infrared Thermography for Measuring Elevated Body Temperature: Clinical Accuracy, Calibration, and Evaluation.* Sensors, 22, 215.

Goldberger, A., Amaral, L., Glass, L., Hausdorff, J., Ivanov, P. C., Mark, R., ... & Stanley, H. E. (2000). *PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals.* Circulation [Online]. 101 (23), pp. e215–e220.

[4] Zhang, L., Yang, H., & Yang, P. (2022, January 22). T*he correlation between type 2 diabetes mellitus and cardiovascular disease risk factors in the elderly.* Applied bionics and biomechanics. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8800622/