Assignment 1 Comp 551

Angad Verma, Angela Moskal, Kai Turanski

October 1, 2024

Abstract

In this assignment, we implement and evaluate the performance of two machine learning models—linear regression and logistic regression—on two datasets: the Infrared Thermography Temperature dataset for regression and the CDC Diabetes Health Indicators dataset for classification. Both models are trained using mini-batch stochastic gradient descent, and in the case of linear regression, we compare results to the closed form analytical solution. We investigate the effect of training size, batch size, learning rate and momentum on model performance. Our results show that small batch gradient descent with a high momentum has optimal performance.

1 Introduction

Our team analyzed the performance of two machine learning models subject to a variety of hyperparameters, data distributions, and loss-reduction techniques. A linear regression model was trained on a dataset of demographic information and temperature readings for predicting oral temperature, a continuous variable [1]. A logistic regression model was trained on a dataset of health predictors and used to predict the presence of diabetes, a binary variable [2]. The models were trained with analytic, gradient-descent, and mini-batch-stochastic methods. Additionally, the application of class-weight and momentum techniques to the training process was tested. The speed of convergence, mean squared error (MSE), and cross entropy error (CEE) were compared to the regression type, the size of the training set, the batch size, the use of class-weight, and momentum. It was found that training set size had a variable effect on the error, depending on which training model was used. As expected, training error convergence was improved by smaller batches. Overall, our linear model achieved a lower MSE than others developed from the same data [3]. The results demonstrate that faster convergence methods might result in worse performance during testing, and that more complex training methods don't guarantee better performance.

2 Datasets

2.1 Infrared Thermography Temperature

This dataset contains 1020 instances of 33 features consisting of temperatures read from various locations of infrared images of patients, along with demographic data [1]. To predict the oral temperature, we use linear regression, which assumes that the features are independent of each other. Looking at figure 1a, we see that this assumption may not apply to the temperature features, as T Max1 and T RC1 are strongly correlated. We also see that there is an imbalance in gender statistics, with 606 female instances and 414 male instances.

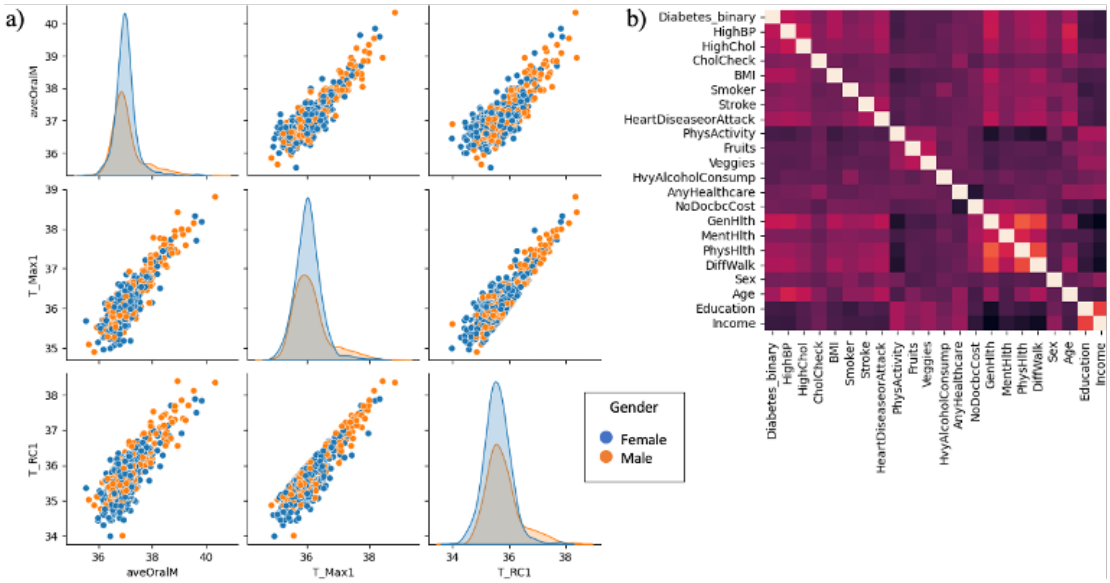

Figure 1: Statistics of both datasets. **a)** Dataset 1: Pairwise plot of the target variable (aveOralM), maximum temperature (T Max1) and right canthus temperature (T RC1) separated by gender. **b)** Dataset 2: Correlation matrix of all features, where general health (GenHth) shows the highest correlation with the target variable (Diabetes binary).

2.2 CDC Diabetes Health Indicators

This dataset contains 35 features including demographics, lab test results, and answers to survey questions for 253680 patients [2]. We use logistic regression to predict the binary target variable indicating whether a patient has diabetes or not. The target variable is strongly imbalanced with 86% instances belonging to the '0' class (not diabetic). Figure 1b reveals some correlation between health indicators such as poor general health (GenHth) and difficulty walking (DiffWalk).

Both datasets underrepresent marginalized groups: in the thermography data, ethnicities, and in the diabetes data, those with low income and education. This must be considered, since underrepresentation can result in worse algorithmic prediction and outcomes for those groups.

3 Results

3.1 Infrared Thermography Temperature

We perform a linear regression using analytically calculated weights, stochastic gradient descent (SGD), fully batched SGD and minibatch SGD. The main indexes used to analyze our linear regression models were mean squared error (MSE) vs epoch, and test/training size. These indices were compared on models varying along three main axes: training set size, learning rate, and batch-size. Additionally, we explore the role of the momentum hyperparameter in reducing weight oscillations. Table 1 displays the analytically calculated weights, with T Max1 exerting the most influence, which makes sense because the maximum face temperature should be a strong predictor of oral temperature. In our base 80/20 train/test analytical model, a MSE of 0.058 and 0.084 were achieved for training and test sets, respectively.

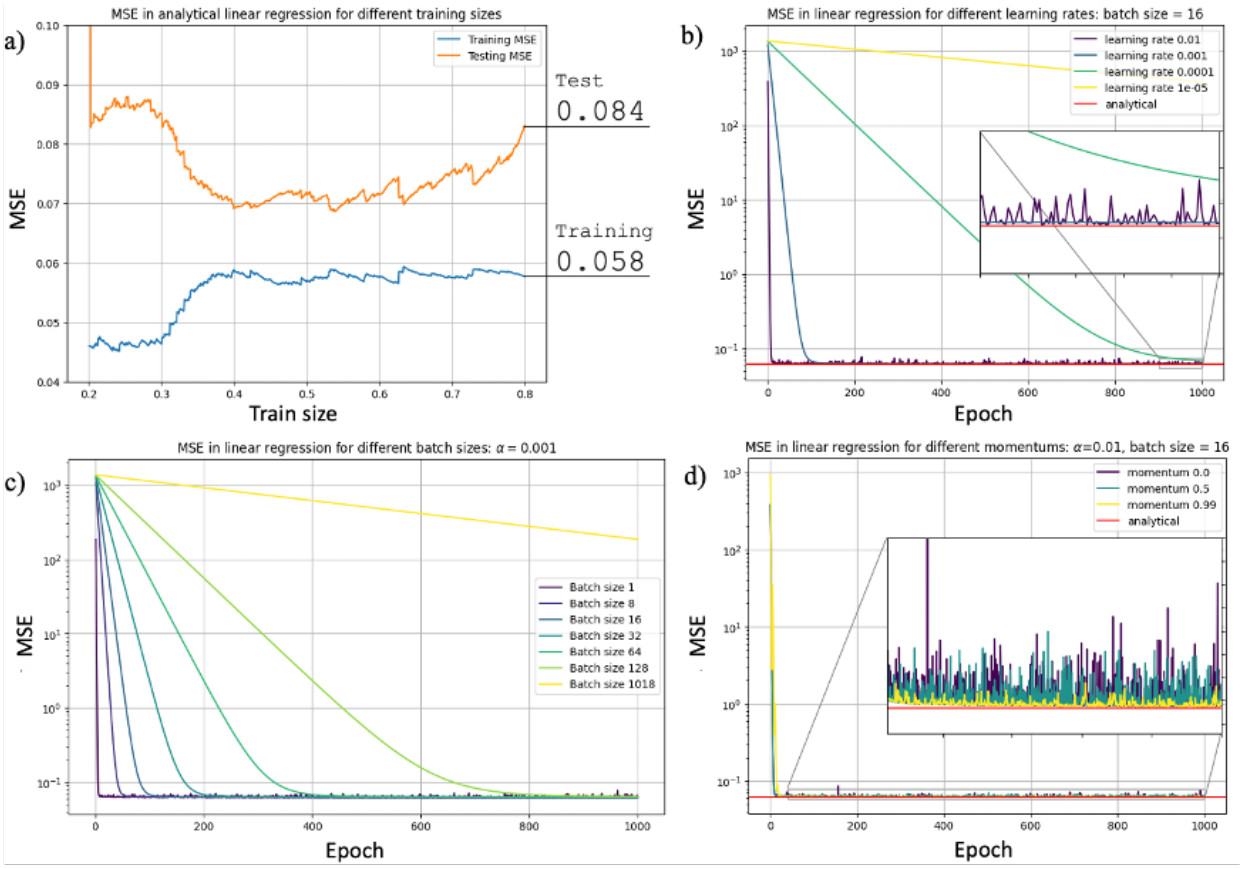

Figure 2: **a)** MSE for analytical logistic regression performed in a test set and training set as the training set size grows. **b)** Evolution of MSE in minibatch SGD for different learning rates: batch size = 16. **c)** Evolution of MSE for different batch sizes in minibatch SGD: $\alpha = 0.01$. **d)** Evolution of MSE in minibatch SGD for different momentums: batch size = 16, α=0.01

Table 1: Analytically calculated weights for linear regression performed on the infrared thermography temperature dataset.

bias	37.028	T_RC_Wet1	0.034	T_FHRC1	-0.015	21-30	0.01
T_atm	-0.091	T_RC_Max1	0.23	T_FHLC1	-0.067	26-30	0.003
Humidity	0.001	T_LC1	0.725	T_FHBC1	0.061	31-40	-0.004
Distance	0.003	T_LC_Dry1	-0.091	T_FHTC1	0.004	41-50	0.01
T_offset1	0.023	T_LC_Wet1	-0.072	T_FH_Max1	0.057	51-60	0.003
Max1R13	-0.204	T_LC_Max1	-0.499	T_FHC_Max1	0.056	>60	0
Max1L13	-0.073	RCC1	0.037	T_Max1	0.28	Asian	0.001
aveAllR13	-0.024	LCC1	0.085	T_OR1	0.128	Black or African-American	0.015
aveAllL13	-0.037	canthiMax1	-0.202	T_OR_Max1	-0.061	Hispanic/ Latino	-0.009
T_RC1	-0.122	canthi4Max1	0.157	Gender	-0.019	Multiracial	-0.012
T_RCDry	0.144	T_FHCC1	-0.065	21-25	0.007	White	-0.026

Growing subsets of training data: Figure 2a shows how a model trained on a small training set is not representative of a larger test set, overfitting to the training data and resulting in a high test MSE. As the training size grows, it becomes more representative, and the test MSE correspondingly decreases.

Changing learning rates: Figure 2b shows how the convergence speed decreases as the learning rate decreases. However, large step sizes have the danger of overshooting, as evidenced

by the oscillations in MSE for $\alpha = 0.01$.

Changing batch sizes: Figure 2c shows how small batch sizes converge rapidly because there are more weight updates in an epoch. However, the small batches have more noise.

Comparing gradient descent methods to the analytical solution: Zooming into the insets of figure 2b,d, we see that the MSE from SGD converges to the analytical solution, but never actually reaches it.

Extending the model to include momentum: We can minimize the aforementioned oscillations by introducing the momentum hyperparameter, as shown in figure 2d. However, our tests show that a large momentum (>0.99) has the danger of converging more slowly if an early outlier instance significantly skews the trajectory.

3.2 CDC Diabetes Health Indicators

We perform a similar analysis to section 3.1 except we characterize the performance of our logistic regression model on the binary target variable using the cross entropy error (CEE), the F1 score, and the accuracy. The optimal weights as determined from fully batched gradient descent in logistic regression with a learning rate of 0.01 are shown in table 2, where poor general health is shown to be the strongest indicator of diabetes.

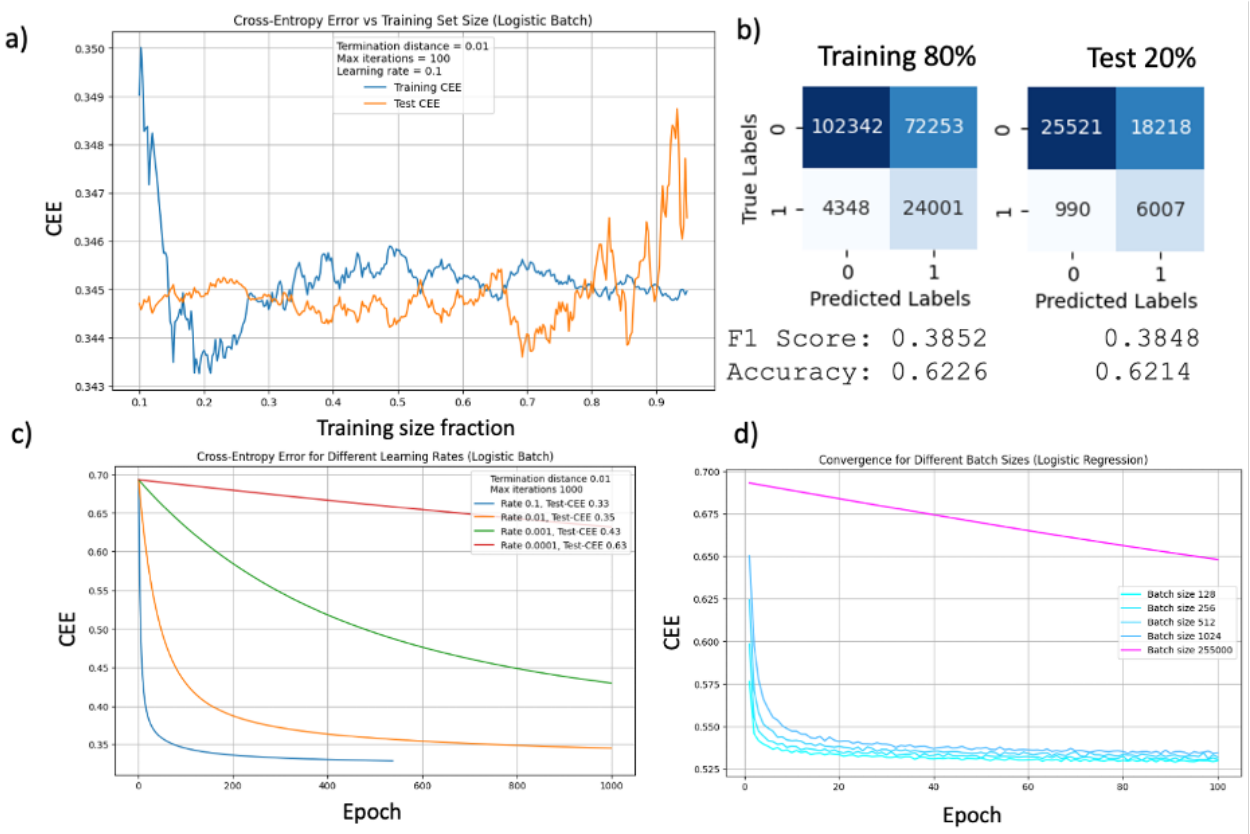

Figure 3: **a)** CEE in fully batched logistic regression with a learning rate of 0.1 between training and test set as the training size grows. **b)** Confusion matrices of the training and test set for the 80% train/ 20% test split.**c)** Evolution of CEE for varying learning rates in fully batched logistic regression.**d)** Evolution of CEE for different batch sizes in minibatch SGD logistic regression.

Table 2: Weights determined from fully batched gradient descent on a logistic regression model with a learning rate of 0.01. The heaviest weight is highlighted in red.

bias	-0.0003				
HighBP	0.00904	PhysActivity	-0.0037	MentHlth	0.00474
HighChol	0.00677	Fruits	-0.0015	PhysHlth	0.01177
CholCheck	0.00059	Veggies	-0.0018	DiffWalk	0.0057
BMI	0.01526	HvyAlcoholConsump	-0.0009	Sex	0.00103
Smoker	0.00202	AnyHealthcare	-3E-05	Age	0.01249
Stroke	0.00147	NoDocbcCost	0.00059	Education	-0.0086
HeartDiseaseorAttack	0.00365	GenHlth	0.02046	Income	-0.0114

Growing subsets of training data: 3a,b shows that the difference in CEE between training and test set is less pronounced than what we observed in figure 2 from the previous section. This could be because this dataset is significantly larger and therefore, training on a smaller subset of the data is already representative. Focusing on the 80% / 20% train/test split in figure 3b, we see that the F1 score and accuracy are marginally better for the training set.

Changing learning rates: Figure 3c shows similar trends to 2b in the previous section, with larger learning rates converging more quickly than smaller learning rates.

Growing minibatch sizes: Figure 3d shows how small batch sizes converge rapidly to similar values, although the smallest batch sizes oscillate greatly once near the minima. Large batch sizes oscillate somewhat less, but take much longer to converge.

Applying a class-weight Throughout the implementation of our logistic regression model, we have been applying a class-weight which helped increase the F1 score of the model from 0.23 to 0.43. This is a great improvement and shows that the non weighted model over-predicts the majority-class.

4 Discussion and Further Improvements

In this assignment, we observed that smaller batch sizes in mini-batch SGD led to faster initial convergence but oscillated more, slowing down the final convergence. Increasing the training set size consistently improved model performance. A smaller learning rate resulted in slower but more stable convergence, while momentum helped stabilize the noisy updates. In logistic regression, using inversely weighted class weights improved predictions on the imbalanced dataset. Our results indicate a medium batch, medium learning rate model with momentum, class-weights, and a large enough training set offers the best performance. Future work could explore the impact of regularization techniques like L1 and L2, adaptive gradient methods, cross validation and the use of non-linear bases to enhance both model generalization and performance.

Statement of contributions:

All team members independently completed the entire assignment, including data preprocessing, model implementation, and experiments. Afterward, we collaborated to compile the results, discuss key findings, and write the final report together.

References

[1] Q. Wang, Y. Zhou, P. Ghassemi, D. McBride, J. P. Casamento, and T. J. Pfefer, "Infrared thermography for measuring elevated body temperature: clinical accuracy, calibration, and evaluation," *Sensors*, vol. 22, no. 1, p. 215, 2021.

[2] N. R. Burrows, "Incidence of end-stage renal disease attributed to diabetes among persons with diagnosed diabetes—united states and puerto rico, 2000–2014," *MMWR. Morbidity and mortality weekly report*, vol. 66, 2017.

[3] P. Razmara, T. Khezresmaeilzadeh, and B. K. Jenkins, "Fever detection with infrared thermography: Enhancing accuracy through machine learning techniques," *arXiv preprint arXiv:2407.15302*, 2024.