COMP 551 Assignment 1 Report

Author Billy Zheng, Yunjie Xiao, Tailai Li

Abstract

This project investigated the performance of two machine learning models, which are linear regression and logistic regression (with gradient descent), on two benchmark datasets. Our goal is to apply linear regression to predict a specific target using a range of biomedical voice measurements from people with Parkinson's disease. The second task is to train a logistic regression to classify benign and malignant breast cancer tumors. Meanwhile, we seek to implement mini-batch stochastic gradient descent for both models. We evaluated and trained the models under various conditions, including different training data sizes, mini-batch sizes, and learning rates, and we compared the performance between models. Our results demonstrated the performance of linear and logistic regression and showed that learning rates, batch sizes, training size, and optimizer choices all affected the model to different degrees. While analytical linear regression performs better than mini-batch regression, the difference is minute.

Keywords: Parkinson's Disease, Breast Cancer, Linear Regression, Logistic Regression, Mini-Batch Stochastic Gradient Descent, Machine Learning, Dataset Analysis.

1 Introduction

With the growing population, the number of people affected by Parkinson's disease and breast cancer is also increasing. In those two aspects, tracking Parkinson's disease (PD) symptom progression often uses the Unified Parkinson's Disease Rating Scale (UPDRS), which requires the patient's presence in the clinic. Meanwhile, compared to a full biopsy, Fine Needle Aspirations (FNAs) can examine a small amount of tissue from the cancer tumor. We aim to apply machine learning techniques to these two medical technologies to make predictions on UPDRS and improve precision for identifying cancer from FNAs. Previous research [1] has used a linear-programming-based inductive classifier to analyze cancer images with an accuracy of 86%, which is fairly high. It highlights the importance of interpretable feature engineering, but it relies on manual initialization. Acquiring data from the UCI Machine Learning Repository, one composed of a range of biomedical voice measurements from 42 early stage Parkinson's patients in a six-month telemonitoring trial [2], the other composed of features computed from a digitized image of a fine needle aspirate (FNA) of a breast mass [3], we applied analytical linear regression to predict motor_UPDRS and logistic regression to classify cancers, both with mini-batch SGD. Experiments varied training size, batch size, and learning rate, and compared analytical vs. mini-batch solutions. Our key findings: (1) larger training sizes improve generalization, (2) batch size has little effect, (3) moderate learning rates work best, (4) analytical linear regression is slightly better, and (5) optimizer choice matters mainly for nonlinear models (sigmoid + SGD outperformed, Adam was unstable).

2 Dataset

Our datasets are from the UCI Machine Learning Repository. The first dataset is *Parkinsons Telemonitoring* [2] (5,875 instances, 16 biomedical voice characteristics, 3 subject information fields, and 1 target), where our target variable is motor_UPDRS. The second dataset is *Breast Cancer Wisconsin Diagnostic* [3] (569 instances and 30 characteristics), where the target is diagnosis (classify the tumor as benign or malignant).

Before training, we converted all "?" values into missing values (NaN) and dropped any rows that contain them. We separated features from the target variable and dropped the Subject ID/ID for both datasets. Moreover, we encoded the Diagnosis (categorical variable) as 1 for malignant and 0 for benign in the breast cancer dataset. We used the `StandardScaler` from `scikit-learn` to standardize the data to eliminate the obvious disparity between features. We then computed descriptive statistics for all features and targets, and we visualized their distributions. Specifically, we selected features with distinctive distributions and presented them. All selected distributions of features of both datasets nearly demonstrate a right-skewed tendency (in which the mean is greater than the median, also confirmed statistically). In the Parkinson's dataset, the target variable shows a clear departure from normality. For the Breast Cancer dataset, the target is categorical, with about 37% malignant and 63% benign cases.

Lastly, because all datasets include personal medical information, we highlight the importance of protecting privacy and confidentiality by removing all information on the subject's ID. Moreover, the data are collected from specific cohorts, and the resulting models may not generalize well across different populations, ethnic groups, or geographic regions.

3 Results

3.1 Experiment 1: Performance of Two Models (80/20 split)

Dataset	MSE	RMSE	R^2
Training Set	6.3164	2.5132	0.9052
Test Set	5.9958	2.4486	0.9061

Table 1: Linear regression performance on the Parkinson's dataset (training vs. test set).

Observations:

- **Relatively Low RMSE.** Using an 80/20 train/test split, we observed that RMSE for the training set and test set is about 2.5132 and 2.4486, respectively, which means that the prediction deviates from the actual motor_UPDRS value by about 2.51 and 2.45 units, on average.
- R^2 **Close to 1.** The R^2 for the training set (0.9052) and the test set (0.9061) are very close to 1, which means that the linear regression model explains about 90% of the variance in motor_UPDRS.

Dataset	**Accuracy**
Training Set	0.9824
Test Set	0.9912

Table 2: Accuracy of Logistic Regression model on the WDBC dataset.

Observations:

- **Relatively High Accuracy.** Using an 80/20 train/test split, the model achieves very high accuracy on both the training (98.24%) and test set (99.12%), suggesting strong generalization. The slightly higher accuracy on the test set indicates that the dataset is nearly linearly separable, and logistic regression is well-suited for this classification task.

3.2 Experiment 2: Report Weights of Each Feature

Linear Regression (Parkinson's)

Feature	**Weight**	**Feature**	**Weight**
age	-0.2146	Jitter:APQ3	-103.8326
sex	0.4164	Shimmer:APQ5	-1.3022
test_time	-0.0268	Shimmer:APQ11	0.8483
total_UPDRS	7.7321	Shimmer:DDA	103.8723
Jitter(%)	1.1383	NHR	-0.0130
Jitter(Abs)	-0.6639	HNR	0.0736
Jitter:RAP	-24.4180	RPDE	-0.2308
Jitter:PPQ5	-0.0352	DFA	-0.0823
Jitter:DDP	23.7253	PPE	0.4552
Shimmer	0.5914	**Bias**	21.2919

Logistic Regression (WDBC)

Feature	**Weight**	**Feature**	**Weight**
f0	0.3752	f15	-0.1136
f1	0.3601	f16	-0.0931
f2	0.3711	f17	0.0478
f3	0.3711	f18	-0.0757
f4	0.1559	f19	-0.1962
f5	0.1328	f20	0.4397
f6	0.2893	f21	0.4349
f7	0.3914	f22	0.4187
f8	0.0934	f23	0.4133
f9	-0.1497	f24	0.3078
f10	0.3293	f25	0.2095
f11	0.0135	f26	0.2911
f12	0.2753	f27	0.3980
f13	0.2999	f28	0.3089
f14	0.0031	f29	0.0938
Bias			-0.3265

Table 3: Compact model weights for Parkinson's and WDBC datasets

Discussion:

- **Features' Contribution.** Table 2 demonstrates the weights of different features for both models. For the Linear Regression model, the strongest predictors are total_UPDRS, Jitter:RAP, Jitter:DDP, Shimmer:APQ3, and Shimmer:DDA. For Logistic Regression, the decision boundary of logistic regression is linear in the feature space, and we observe that certain features (e.g., f20–f23) have the largest positive weights, indicating they strongly influence classification toward malignant, while others (such as f9 and f19) with negative weights tend to push toward benign.

3.3 Experiment 3: Performance Along Different Training Size

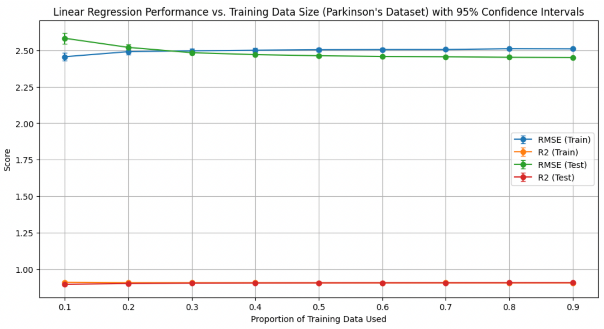

Figure 1: Effect of training size on model performance (Linear Regression RMSE)

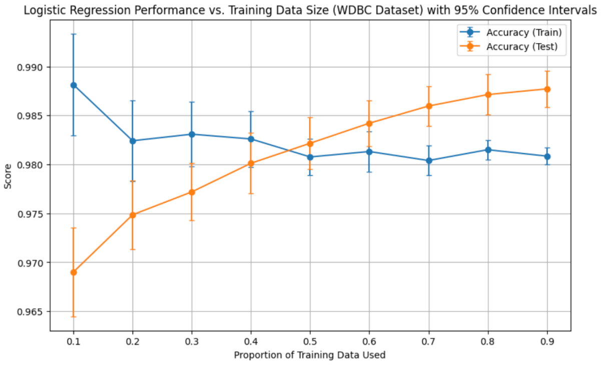

Figure 2: Effect of training size on model performance (Logistic Accuracy)

Observation:

- **RMSE for Linear Regression Model** We see from Figure 1 that RMSE increases for the training set but decreases for the test sets as the size of the training data increases. This is because for training sets, increasing the size of the training data increases the difficulty for the model to fit all data points (increases training error), while for test sets, more training data improves the precision on unseen data.
- **Accuracy for Logistic Regression Model** We see from Figure 2 that accuracy decreases for the training set but increases for the test sets as the size of the training data increases. For the same reason mentioned above, accuracy should show the opposite trend compared to RMSE (since it is not error but precision).

3.4 Experiment 4: Performance Along Different Mini-Batch Sizes

Linear Regression (Parkinson's Dataset)

Batch Size	MSE	RMSE
8	6.0933	2.4685
16	6.1061	2.4711
32	6.0484	2.4593
64	6.0499	2.4597
128	6.0385	2.4573
4700	6.0395	2.4575

Logistic Regression (WDBC Dataset)

Batch Size	Accuracy
8	0.9825
16	0.9912
32	0.9825
64	0.9825
128	0.9825
455	0.9825

Table 4: Performance of Linear and Logistic Regression with Different Mini-Batch Sizes

Observation:

- **Varying Mini-Batch Size For Linear Regression** The results show that changing the batch size from very small (8) to very large (full batch, 4700) has minimal effect on model performance. MSE and RMSE stay relatively stable. A batch size of 128 provides the lowest RMSE (2.4573).
- **Varying Mini-Batch Size For Logistic Model** Accuracy is generally high across all batch sizes (98.25%), but the batch size of 16 achieves the best performance (99.12%), so a moderate batch size can provide better prediction.

3.5 Experiment 5: Performance Along Different Learning Rates

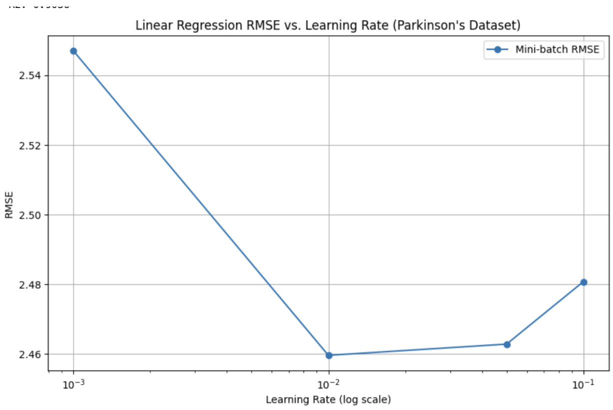

Figure 3: Linear Regression: RMSE across Learning Rates

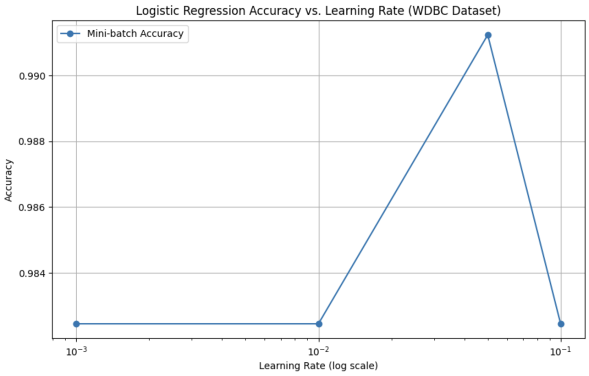

Figure 4: Logistic Regression: Accuracy across Learning Rates

Observation:

- **Varying Learning Rates for Linear Regression** Given 4 tested learning rates (0.001, 0.01, 0.05, 0.1), we want to avoid the learning rate being too small, which may take a very long time to converge. We also want to avoid the learning rate being too large, in which the model might overshoot or even diverge from the target. In Figure 3, we observed that the relative optimal learning rate (where the lowest RMSE occurred) is 0.01.

- **Varying Learning Rates for Logistic Model** For the same explanation as in the linear regression case, we observed in Figure 4 that the relative optimal learning rate (where the highest accuracy occurred) is 0.05.

3.6 Experiment 6: Compare analytical Linear Regression Solution with Mini-Batch Stochastic Gradient Solution

Metric	**Mini-Batch Linear Regression**	**Analytical Linear Regression**
Mean Squared Error (MSE)	6.0658	5.9958
Root Mean Squared Error (RMSE)	2.4629	2.4486
R-squared (R^2)	0.9050	0.9061

Table 5: Analytical and Mini-Batch Linear Regression Model Performance

Observation:

- **Similar Performance But Analytical is Slightly Better** From Table 4, we see that their performance achieves a similar level of accuracy, but analytical linear regression is slightly better since it has a lower RMSE (2.4486 vs. 2.4629) and R^2 value (0.9061 vs. 0.9050) closer to 1.

3.7 Originality: Compare the Results of Original and Using Different Optimizer Results of the First Dataset

Idea: We built a flexible regression framework that can handle both linear and nonlinear basis functions while supporting different optimization methods(Adam, SGD, etc.). Using the Parkinson's dataset, the framework compares four models: standard closed-form linear regression, linear regression trained with Adam, and regression models with sigmoid basis functions trained using either SGD or Adam.
Observation: We observed that the closed-form original linear regression remains as a stable baseline. Among the three combinations of optimizers tested, only nonlinear (in our case, sigmoid function) SGD shows better predictive power than the original linear one because it has the lowest MSE (4.8182), the R^2 value closest to 1, and the steepest MSE (from the figure).

Model	MSE	RMSE	R^2
Original Linear	6.0024	2.4500	0.9060
Linear + Adam	146.7402	12.1136	-1.2990
Sigmoid + SGD	4.8182	2.1950	0.9245
Sigmoid + Adam	39.9301	6.3190	0.3744

Table 6: Performance comparison of models on the Parkinson's dataset.

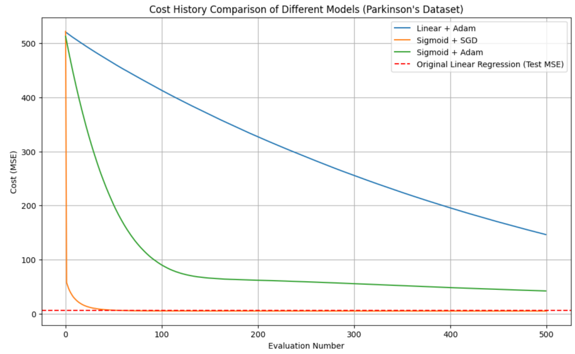

Table 7: Performance comparison of models on the Parkinson's dataset.

4 Discussion and Conclusion

4.1 About Linear&Logistic Regression Result

We see that both linear and logistic models achieve a high predictive model for biomedical tests. In the experiment performed, we observe that f20-f23, f9, and f19 are notable features in Logistic Regression. Then, increasing the training set size generally improves generalization, resulting in a reduction in test error. Specific batch-size and learning rates can provide better predictions for the model, although batch-size has a very subtle influence on the model. Also, compared to mini-batch linear regression, the analytical linear solution gives slightly better predictions. In future studies, we suggest applying L1/L2 regularization to prevent overfitting and to better understand which features are most important.

4.2 About Optimizer Experiments

The optimization choice mattered mainly with the non-linear basis functions: sigmoid + SGD improved performance, while Adam was unstable and underperformed. Future studies can work toward more combinations of optimizers or more complex nonlinear functions, such as Gaussian or polynomial bases. Furthermore, we can assess robustness across different data splits or using cross-validation for stronger generalization claims.

5 Statement of Contributions

Billy and Yunjie are responsible for organizing the data, implementing the model, and running the experiments. Tailai is responsible for cleaning the data, gathering the results, and writing the report.

References

[1] William Nick Street, William H. Wolberg, and Olvi L. Mangasarian. Nuclear feature extraction for breast tumor diagnosis. In *Proceedings of Electronic Imaging: Science and Technology*, 1993.

[2] Athanasios Tsanas and Max A. Little. Parkinsons telemonitoring [dataset]. UCI Machine Learning Repository, 2009.

[3] William H. Wolberg, Olvi L. Mangasarian, W. Nick Street, and William H. Street. Breast cancer wisconsin (diagnostic) [dataset]. UCI Machine Learning Repository, 1993.