The Efficiency of Logistic Regression Compared to Linear Discriminant Analysis in Linear Classification Problem

Geng Shao, 260898952, Yimeng Hu, 260795862, Yueyang Wan, 260884782,

***Abstract*—In this project, two linear classification techniques, logistic regression and Linear Discriminant Analysis(LDA), are performed to analyze two datasets, *Wine Quality Dataset* and *Breast Cancer Wisconsin (Diagnosis) Dataset*. First, the characteristic of the input features of both datasets are analyzed and appropriate data pre-processing techniques are applied accordingly. Furthermore, the difference in performance of Logistic Regression and LDA on the two benchmark datasets are investigated. In addition, the effect of change in input features and adjustment of hyperparameters on the efficiency, as well as accuracy, of the linear classification model is carefully reviewed. Lastly, a new stopping criteria found for the gradient descent in logistic regression is discussed. Through the experiment, we found that both LDA and Logistic Regression achieves the same level of accuracy however LDA takes much lesser time. Moreover, we noticed how new subsets of input features can effect the performance of the model. New interaction terms of *Wine Quality Dataset* and improvement of the model's accuracy are also discussed in this report.**

***Index Terms*—logistic regression, linear discriminant analysis, binary classification.**

I. Introduction

A. Background

LOGISTIC regression and linear discriminant analysis (LDA) are both widely used techniques to linear classification problem in machine learning. One of the difference between these two approaches is that LDA assumes Gaussian distribution on the input data, whereas *Logistic Regression* makes no such assumptions on the input. Therefore it is reasonable to expect that LDA would give more accurate prediction over *Logistic Regression* when the input features are in fact normally distributed. [1]

It is interesting to note that the quality of the dataset and the information that can be derived from it can easily affect the ability of the model to learn, hence data pre-processing is a substantial step in machine learning. Data pre-processing refers to the transformation of the data that is not feasible for the analysis into a clean dataset. Since we will be comparing the performance between two algorithms, it is important that the dataset is formatted in such a way that it can be fed into both two models. The two datasets used in these tasks will be pre-processed to eliminate the malformed values and to be better fit into our models. Since it was required to convert the task into binary classifications, which means the output should be categorized into only two classes, the output data is transformed into binary representation before being fed into the algorithms.

It is noteworthy that, apart from the dataset and the structure of the model itself, the hyperparameters also takes in a considerable impact on the performance of the model. They directly control the behaviour of the training algorithm. In general, the hyperparameters of a machine learning model includes *learning rate* and *Number of iterations*. For example, if the learning rate is too low, the model will miss the important patterns in the data, if it is too high, it may have collisions. There are many *Hyperparameters Optimisation Techniques* available, however they will not be the focus of this project.

One of the great concern when training a model is *overfitting*. If the algorithm is trained more on the training set, it will overfit the training set with relatively small loss. However, it may incur a large loss when predicting unknown new data. Different stopping criteria can be applied to efficiently cope this problem[2].

B. Objectives

The goals of this project are as following: 1) Carefully analyze the datasets, examine different data pre-processing techniques; 2) Implement the above two algorithms from scratch, using the methods and equations discussed in the lectures; 3) Run each model over two datasets with different hyperparameters and review their performance; 4) Explore how different subsets of input features can effect the efficiency of the model. 5) Investigate different stopping criteria for the gradient descent in logistic regression.

C. Results at a Glance

It is confirmed that in general *LDA* behaves better than *Logistic Regression* on both two datasets. Nonetheless, the difference is more evident on the *Breast Cancer Wisconsin (Diagnostic) Dataset* over *Wine Quality Dataset*. This result is expected since the input data of *Breast Cancer Wisconsin (Diagnostic) Dataset* has a better distribution than the one of *Wine Quality Dataset*. In addition, with the newly found subset of input features of *Wine Quality Dataset*, the accuracy of *Logistic Regression* model is increased from 70.2% to75.6%. Finally, with the new stopping criteria, the *Logistic Regression* model is able to achieve the same accuracy with much less time.

II. Datasets

A. Overview

The two datasets used in this project are *Breast Cancer Wisconsin (Diagnostic) Dataset* [3] and *Wine Quality Dataset*[4].

The *Wine Quality Dataset* consists of 11 chemical components found in red wine as input features and wine quality as output. *Breast Cancer Wisconsin (Diagnostic) Dataset* contains 9 real-valued features computed for each tumor as input and the condition of the tumor. Both datasets are each fed into the *logistic regression* model and the *LDA* model to evaluate the performance of the algorithms.

	Inputs (features)			Output (class)
Red Wine Quality	fixed acidity	chlorides	PH	
	volatile acidity	free SO2	sulphates	**Quality**
	citric acid	total SO2	alcohol	
	residual sugar	density		
Breast Cancer Diagnosis	Clump Thickness	Marginal Adhesion	Normal Nucleoli	**Benign or malignant**
	Uniformity of Cell Size	Single Epithelial Cell Size	Mitoses	
	Uniformity of Cell Shape	Bland Chromatin		

Fig. 1. Input Features and Output Classes of The Two Datasets.

B. Features and Distributions

1) Wine Quality Dataset: The output of this dataset are integers from 1-10 with 1 indicating the worst quality and 10 being the best. The input features are typical chemical components found in red wines with different measuring scales. For instance, density and pH obviously have different units, and the values of chlorides are distributed between 0.07 and 0.12, whereas the values of total sulfur dioxide can reach up to 289. This result in the fact that the entire dataset has relatively large variance. The immense gap in the scale could become an issue when fed into the model. Therefore pre-processing *Wine Quality Dataset* is essential for the success of our models.

2) Breast Cancer Wisconsin (Diagnostic) Dataset : The output of the dataset is binary classified where '2' stands for benign and '4' refers to malignant. Its input features are 9 real-valued features computed for each tumour, each attribute is assigned an integer number indicating its degree of malignancy. However, it is worth noting that, as opposed to the *Wine Quality Dataset*, all its data points are scattered between 1 to 10. Therefore, the input of the *Breast Cancer Wisconsin (Diagnostic) Dataset* is much more closed to a Gaussian Normal distribution.

For both datasets, the distribution of the positive versus negative classes are plotted as bar charts, as shown in *Figure 2*. The probability density function of the *Wine Quality Dataset*'s input data are visualized in *Figure 3* to provide a more intuitive view of how the data is distributed.

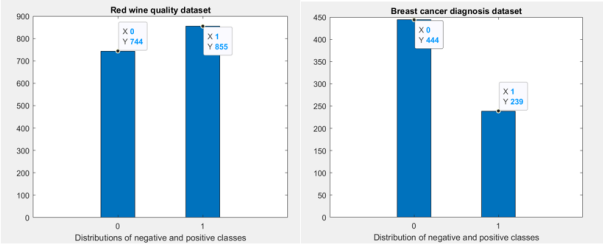

Fig. 2. Distributions of negative and positive classes of both datasets

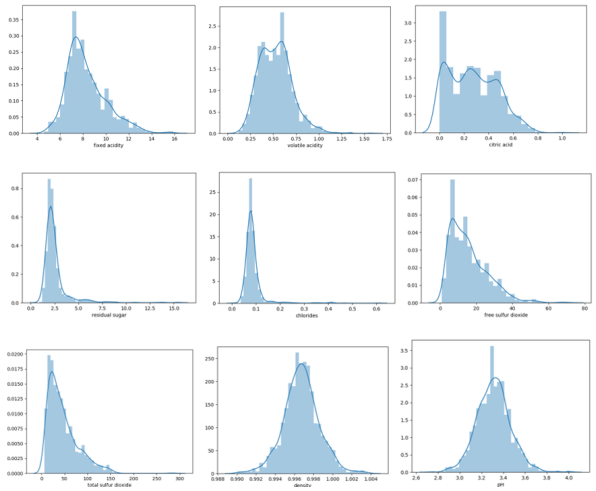

Fig. 3. PDF of the input features in the *Wine Quality Dataset*.

C. Data Pre-processing

1) Wine Quality Dataset: By looking into the dataset, we notice that there is no special characters and, most importantly, no missing attributes. The output values are binarized and grouped into two classes. Then we applied the method of normalization on the input features to make the data closer to a Gaussian Distribution, without causing any information distortion or loss. The method is especially helpful when gradient descent is used in the algorithm as it can sometimes increase the speed of convergence and make the model run much smoother.[5] Normalization on a column of feature is described in Equation 1:

$$x' = \frac{x - min(x)}{max(x) - min(x)} \tag{1}$$

where x is the original feature, x' is the normalized feature, min(x) and max(x) are the minimum and maximum values of that specific features respectively.

2) Breast Cancer Wisconsin (Diagnostic) Dataset : Although the output of the dataset is binarized, we further mapped '2' into '1' and '4' into '0' to facilitate the calculations. Differs from *Wine Quality Dataset*, it has some special characters and missing attributes in its dataset. Hence a special pre-processing algorithm was done in order to eliminate those invalid data points and made sure that the dataset contains only meaningful information. As for the rest of the input features, as mentioned before, all input data are shown to have a good distribution. Hence, we determined that the data normalization is not necessary here.

In order to evaluate the final accuracy of our model, 10% of the data was held out from each dataset in advance to be used as test sets. The 5-fold Cross Validation technique was employed to pick the best model, therefore the training set was further divided into five batches.

D. New Features Added

There are many ways to improve the accuracy of a model. One of the most intuitive way is to apply some transformations

over the input data. Apart from the data pre-processing, it is also possible to explore new iteration terms of two features. We will use the *Wine Quality Dataset* as an example in the demonstration.

Research shows that free sulfur dioxide and the pH value are two highly correlated parameters that affect the quality of red wine. The interaction between these two features will be more useful on our model than themselves being used separately[6]. Intuitively, the product of these two features are added as a new feature of the dataset.

Inspired by this idea, we evaluated other feature combinations and two additional input terms were found that improves the accuracy of the model. One is the product of volatile acidity and density, the other is the product of residual sugar and alcohol.

In addition to the introduction of new features, we eliminated the so-called dummy/intercept term, since it cannot be normalized and it has a negative impact on the overall performance.

After all these operations, the accuracy of the *Logistic Regression* model on the *Wine Quality Dataset* has increased by approximately 5%.

E. Ethical Concerns

The breast cancer database was collected from real hospital with real diagnostic data from hundreds of patients. In other words, this dataset contains private information since each data point is associated with a patient's ID number. Abundant personal information, such as name, address or contact number, may be maliciously obtained using these ID numbers. Hence it is very important for us to keep these data confidential and not to expose any of these ID numbers to public.

III. Results

A. Overall Performance

Figure 4 shows the best accuracy each model can get on each dataset. According to the experimental result, Logistic Regression algorithm achieves slightly better accuracy in general. However, it is worth noting that, as shown in *Figure 6*, time needed to train a LDA model is much more less than the one needed to train a Logistic Regression model. On the other hand, both algorithms generate better results on cancer diagnosis set. This implies that the accuracy and efficiency not only depend on the algorithm of the model but also on the dataset. It shows the importance of the data pre-processing.

B. Learning Rate and Iteration Times vs. Accuracy

Hyperparameters also play an important role in the performance of machine learning model. We will take Logistic Regression model as an example to show how different combinations hyperparameters can affect the overall performance of the model. Typically, two datasets reflect respective characteristics. *Figure 5* and *Figure 6* indicate how learning rate and iteration times will affect the accuracy of logistic regression model on two datasets. In generally, given the a constant learning rate, the higher it is the number of iterations, the higher it is the accuracy. However the rule does not apply to learning rate. In *Figure 4*, the accuracy peaks when the learning rate is 0.0001 while in *Figure 5* the learning rate of 0.1 contributes to the highest accuracy. Interestingly, in *Figure 4*, the difference of 500 iterations times does not have much impact when the values of learning rate are 0.0001, 0.001 and 0.01, for the light blue marks nearly cover the red ones.

Another point worth mentioning is that these two datsets response conversely to the tendency of learning rate. Logistic regression algorithm will predict the outcome of breast cancer when the learning rate is small while requires large learning rate when deals with wine quality.

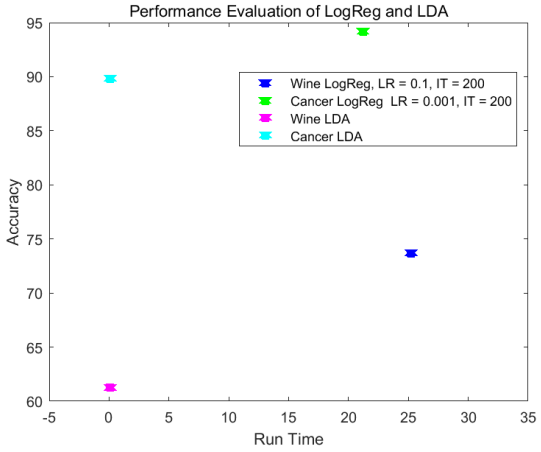

Fig. 4. Performance Evaluation of Two Linear Regression Algorithm Based on Two Datasets .

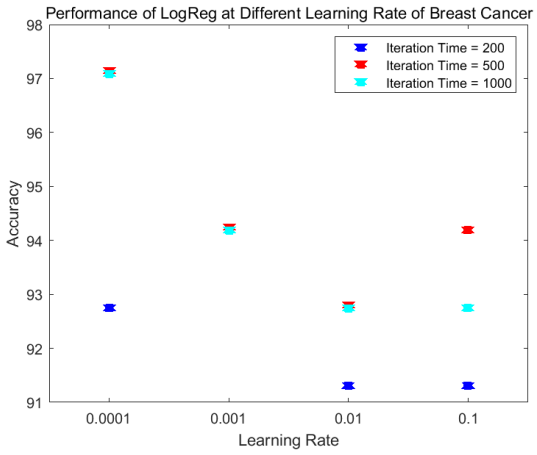

Fig. 5. Performance of Logistic Regression at Different Learning Rate and Iteration Times of *Breast Cancer Wisconsin (Diagnosis) Dataset*.

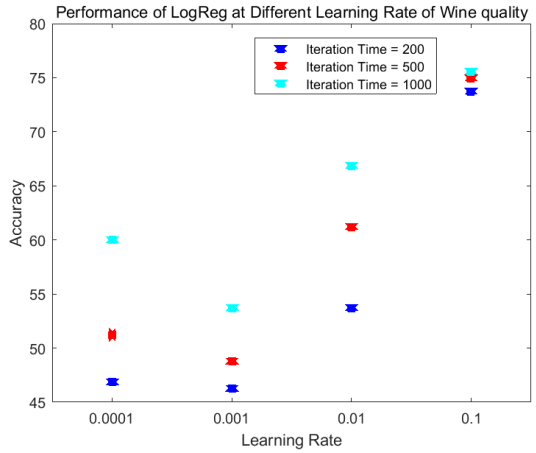

Fig. 6. Performance of Logistic Regression at Different Learning Rate and Iteration Times of *Wine Quality Dataset*.

C. New Features

Again we will use Logistic Regression model and *Wine Quality Dataset* as an example to demonstrate the improvement of overall performance with the introduction of new features in the input dataset. As shown in Figure 7, given a set of hyperparameters, the model performs better with the datasets after introducing new features. The improvement is surprisingly large with learning rate of 0.01 and 200 numbers of iterations. Notably, Consider the accuracy of the model with improved dataset trained with 200 numbers of iteration and the one with original dataset trained with 1000 numbers of iteration, the model trained with the improved dataset achieves the same level of accuracy with much lesser time.

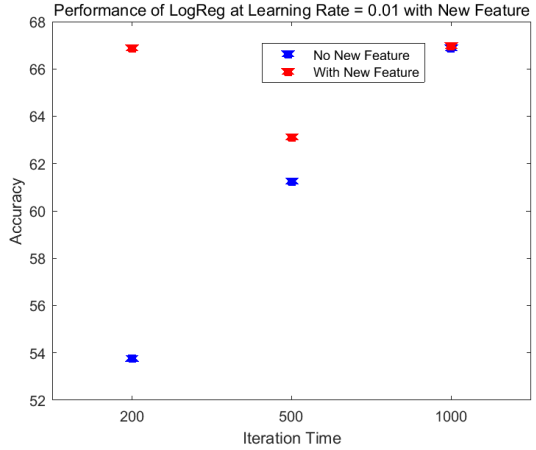

Fig. 7. Comparison of Performance of Logistic Regression model on *Wine Quality Dataset* with and without new features

D. Adaptive Stopping Criteria

The predetermined iteration times cause inconvenience, such as how to decide the proper iteration times to prevent both underfitting and overfitting. The repetitive adjustments of iteration times lengthen the time to train a good model and undermine the consistency of one machine learning algorithm. Meanwhile, the low change rate of the gradient indicates the possibility of the convergence of the gradient. Intuitively, we think about stopping the iteration according to the derivative of gradients. We define the percentage of gradient difference (PGD) in Equation 2:

$$PGD(i) = \frac{gradient(i-2) - gradient(i-1)}{gradient(i-1) - gradient(i)} (i >= 3) \tag{2}$$

It will continuously monitor the change of the gradient count the number of points where PGDs are smaller than 0.001. When number of these points is greater than 3, we break the iteration and finish the training. This is to prevent sudden change in the gradient and avoid stopping at the flat area between two decreasing interval.

Once again we take the Logistic Regression model and *Wine Quality Dataset* as an example, at learning rate of 0.01, the model trained with adaptive stopping criteria can maintain the similar level of accuracy while the running time is much more shortened.

Table1: Comparison between Adaptive Stopping Criteria and Predetermined Iteration Time

	Accuracy	Total Runtime
Adaptive Stopping Criteria	66.87	54.24
Predetermined Iteration Time	66.88	201

IV. Discussion and Conclusion

A. Source of Error

One lesson we learnt from this project is that data pre-processing is a crucial part of machine learning task, bad data can result in very abnormal performance and sometimes even crush the entire algorithm, especially when the number of input feature is large. In our case, the Logistic Regression model can not work properly on the raw *Wine Quality Dataset*. Nonetheless data pre-processing may not always be necessary. In the case of the*Breast Cancer Wisconsin (Diagnosis) Dataset*, it does not require normalization since its input data is already clean and stable. In most circumstance, data pre-processing should be considered as an substantial part of the machine learning task. It is also important to apply the correct pre-processing techniques.

One other source of error is that both datasets used in this project are considerably small. *Breast Cancer Wisconsin (Diagnosis) Dataset*has less than 700 samples, and they become even smaller after splitting out the test set, so the outcomes from these insufficient training are imperfect due to the lack of experience of the model.

Overfitting is also a minor problem that we encountered during the experiment. Although in our cases, the degree of overfitting is not very bad, the difference between test accuracy and the average validation accuracy in within 5% - 10%.

B. Improvement Suggestion

First of all, the importance of data pre-processing should be reaffirmed here. In fact, data pre-processing is not only about adjusting the value of the inputs, there are many other aspects

associated with the term ”data cleaning”. For example, in some circumstances, some sets of data may be unintentionally recorded for more than once, then it is very important that these duplicates be removed from the dataset before the training process[7]. In the *Breast Cancer Wisconsin (Diagnosis) Dataset* case, each set of the tumour information is linked to a unique ID number, so if two lines of exact data appear, one of them can be discarded without any doubt. In addition, the method of normalization that we used on the wine data may not be the best way to deal with these complicated inputs. There are many other pre-processing approaches that one can explore, such as taking the logarithm of all the features, or applying standardization method on the features, etc.

Another way to improve the algorithm quality can be gaining some knowledge of the dataset in advance. For example, in the *Breast Cancer Wisconsin (Diagnosis) Dataset*, if one has some relevant knowledge in the medical field, the non-numerical values in the dataset may be replaced with some reasonable and empirical predictions instead of being deleted, the dataset is pretty small already so it is good to keep as much reliable information as possible. Same thing would help in the red wine dataset, if one knows the relationship between these chemicals and wine quality, some data that are obviously wrong or irrelevant may be deleted, unless these false data can significantly affect the accuracy of the model.

C. Summary

In conclusion, the main objectives of this project have been fulfilled. We successfully implemented Logistic Regression and LDA model, and tested their performance on *Wine Quality Dataset* and *Breast Cancer Wisconsin (Diagnostic) Dataset*. The resulting accuracy of our two models on two distinct datasets seem to be reasonable, although they may still be further improved with the suggestions stated above. From the experimental results, we can conclude that in terms of accuracy, the performance of LDA model does not differ too much from the one of the Logistic Regression model. The difference of prediction accuracy of LDA model and Logistic Regression model is within 10%. However, LDA model has much less computational complexity and consumes much lesser time. Other than that, we demonstrated the possibility of improving the accuracy by manipulating with existing input features. Lastly, with the new adaptive gradient stopping criteria, we are able to shorten the run-time while maintaining the same level of accuracy.

STATEMENT OF CONTRIBUTION

Yueyang was in charge of data pre-processing and data distribution analysis(Task 1). Geng was responsible for the implementation of the two models and 5-fold cross validation(Task 2). Yimeng formulated the code, ran the experiments and collect/analyzed the result(Task 3).

REFERENCES

[1] M. Pohar et al., ”Comparison of Logistic Regression and Linear Discriminant Analysis: A Simulation Study ,” Metodoloki zvezki, Vol. 1, No. 1, 143-161, 2004. [online]. Available: https://www.stat-d.si/mz/mz1.1/pohar.pdf

[2] N. Bao, Experiments on logistic regression.[online]Available: https://pdfs.semanticscholar.org/6d46/152c2e2b9a62d00f19316b8baaf95aa9fa28.pdf

[3] P. Cortez, A. Cerdeira, F. Almeida, T. Matos and J. Reis. Modeling wine preferences by data mining from physicochemical properties. In Decision Support Systems, Elsevier, 47(4):547-553. ISSN: 0167-9236. [online]Available: [@Elsevier] http://dx.doi.org/10.1016/j.dss.2009.05.016

[4] O. L. Mangasarian and W. H. Wolberg: ”Cancer diagnosis via linear programming”, SIAM News, Volume 23, Number 5, September 1990, pp 1 18.

[5] S. Ioffe and C. Szegedy, ”Batch normalization: Accelerating deep network training by reducing internal covariate shift,” in 32nd International Conference on Machine Learning, ICML 2015, July 6, 2015 - July 11, 2015, .

[6] M. Dwicahyo, Improve Wine Prediction with Feature Engineering [online]Available: https://www.kaggle.com/mahendrimd/improve-wine-prediction-with-feature-engineering

[7] O. Elgabry, The Ultimate Guide to Data Cleaning [online]Available: https://towardsdatascience.com/the-ultimate-guide-to-data-cleaning-3969843991d4