Comparative Analysis between Linear Discriminant Analysis and Logistic Regression for Classifying Red Wine and Breast Cancer.

Michael Li(260869379), Li Zhang(260743980) and William Zhang(260865382)

September 29, 2019

Abstract

Various machine learning models are used to classify datasets. In this paper, logistic regression and linear discrimninant analysis, two linear classification models will be compared on a wine and breast cancer dataset. Feature selection and generation will also be explored as they directly impact model performance. Logistic regression's hyperparameters tuning will also be evaluated to optimize its performance. Finally, both models will be compared based on their run time and their accuracy.

1 INTRODUCTION

Logistic regression (LR) and linear discriminant analysis (LDA) are widely used in classification tasks. These two models were compared in their performance and accuracy using two datasets, notably the Wine dataset and Breast Cancer Diagnosis dataset.

For the wine data set, various feature generation and evaluation methods were explored based on Aich et al. [1] and Gupta's [2] findings regarding feature selection with this dataset.

The logistic regression model was trained using gradient descent while the linear discrimnant model was fitted using matrix multiplication.

The logistic regression model was found to be performing slightly better than the linear discriminant analysis model for both datasets (see Table 1). This is due to the fact that LDA assumes that the classes follows a normal distribution.

	Wine dataset	Cancer dataset
Logistic regression	75.11%	96.76%
LDA	75.24%	94.56%

Table 1. Comparison of model accuracy on different datasets

2 Datasets

Some of the various datasets used for this analysis are discussed below.

2.1 Wine Dataset

The wine data set was comprised of 11 input and 1 output variable. The input variables, acidity, residual sugar and pH among others, were based on physicochemical tests while the output variable was simply the quality of the wine, with a score between 0 and 10, based on sensory data.

In order to create a binary classification problem, the wine dataset was separated into two main classes, with the boundary set at a quality rating of 5, such that wine with quality of 6 or above were rated as 1, and wines rated 5 and below were rated 0. After this transformation, the dataset was comprised of 744 wines with a rating of 0, and 855 wines with ratings of over 1. The probability density histogram for each of the features was also plotted to have a visual representation of the data and identify any major outliers(see submitted files, 'breast cancer plots' and 'wine plots').

Furthermore, to account for the difference in units of the features, min-max normalizing was done on each of

the feature to allow the model to compare data from different units. The formula for min-max normalization is below.

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

In order to determine the best subset of features for the model, findings from Aich et al. were used to determine that only 7 of the 11 given features were necessary to capture most of the variance in the data.

Feature generation was done on a trial-and-error basis, where features would be created based on various interaction terms and tested with cross-validation to see if they improved accuracy. If no tangible benefit was seen, the newly created feature would be removed in order to reduce the complexity of the model. Mahendrimd's [3] article also mentioned the importance of molecular SO^2 in determining the quality of a wine, with the increase in accuracy validating his claim.

2.2 Cancer Dataset

The cancer data set has 11 variables, with 9 input variables, 1 output variable and 1 identification variable(the patient ID) that was removed in the preprocessing. The output variables were also converted to simplify the classification task. The previous classification labels were 2 for benign tumor and 4 for malignant tumor, which were transformed to 0 and 1 respectively. The resulting dataset was comprised of 444 patients with benign tumors and 239 patients with malignant tumors.

Some patients also had missing data, which was simply removed. It can be noted that the values could have been replaced by the mean of the feature, or the mean of the feature with respect to other features, but for the sake of the project, the simplest method was chosen.

Feature creation was attempted through various interaction terms, however they did not result in substantial difference in the accuracy of either model. Therefore, the initial set of features were kept.

For this dataset, z-score normalization was applied to each of the features. This was done to preserve the meaning of various outliers in the dataset.

Medical datasets may cause ethical concerns due to their confidential nature. Most health datasets will remove explicit information about patients such as name, address, telephone number and so forth, but other data can still allow patients to be identified. Combination of information, such as recurring medical condition in the family, date of birth, and more can be used to be traced back to a patient. Furthermore, when training models to predict a medical condition, a collaboration needs to be made with health professionals, as a model cannot be used as the only source of diagnosis.

3 Results

3.1 Logistic Regression Model

In order to evaluate the logistic regression model, various hyper-parameters were tested to evaluate their effects on the model's accuracy. The most important ones were epoch, learning rate and error threshold. As the learning rate decreased, the convergence speed increased. This was demonstrated in figure X. When the learning rate is very low, the model has difficulty converging within a reasonable number of epochs, as opposed to a very high learning rate, which caused a high variance in the loss, which was fluctuating.

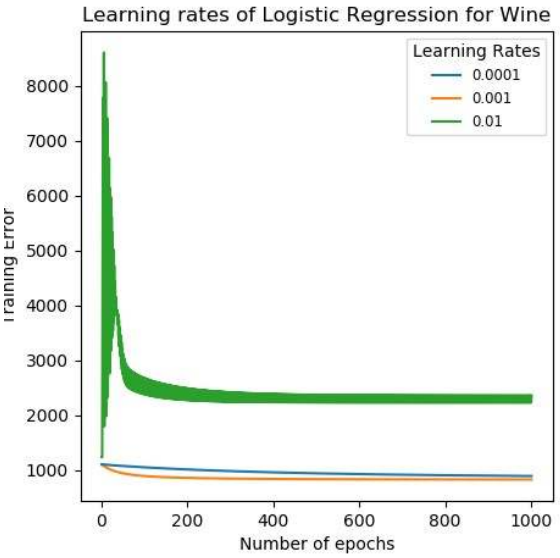

Figure 1. various learning rates for the wine dataset

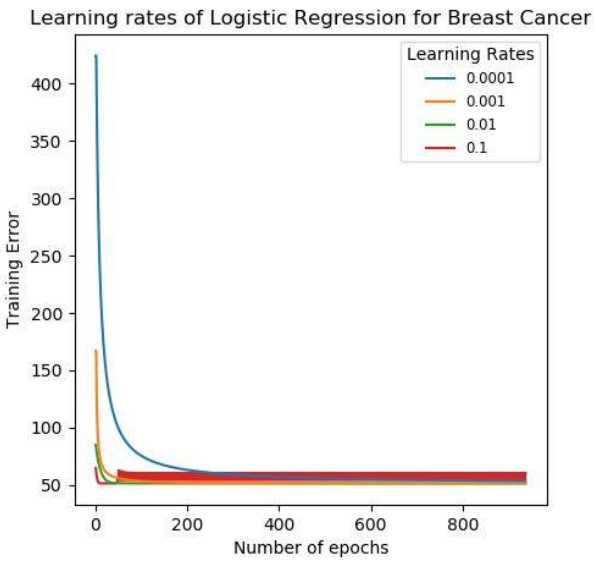

Figure 2. various learning rates for the breast cancer dataset

Another hyper-parameter that affected the performance of the model was the error threshold. A very low error threshold would cause the model to keep training despite having converged. On the other hand, a very high threshold would cause the model to stop training before convergence, leading to an underfit model.

The number of epoch is used as a ceiling to the number of times gradient descent is run. However, based on the results, the model would consistently converge before having reached this ceiling. It was concluded that as long as the maximum number of epochs was set higher than what was needed for convergence, the model was unaffected by it as it would stop training.

Since gradient descent always converges at a local minimum for this cross-entropy loss function, the model was at risk to overfit the dataset. To mitigate this risk, L2 regularization was applied to the model, lowering the general coefficient of the weights.

The results show that linear discriminant analysis consistently runs faster than logistic regression. This is due to their implementation, with linear discriminant analysis being matrix multiplications and logistic regression having to run multiple iterations for gradient descent.

Comparing their accuracy, both models were found to perform similarly for the cancer dataset. While for the wine data set, logistic regression was shown to barely outperform LDA with additional features. With the baseline features, both models showcase an accuracy of 74%, however with additional features created from various non-linear interactions, the LR increased by 1% while LDA only increased by 0.5%. It is important to note that LDA was significantly faster at training, which might make it a better model despite its slightly lower accuracy.

	Wine dataset	Cancer dataset
Logistic regression	79.24 s	4.32 s
LDA	0.35 s	0.29 s

Table 2. Comparison of model speeds

In order to verify the relevance of the newly generated features, it was ensured that these new features would increase the accuracy of the model and that the weight associated to these new features were high relative to other features.

Using the new subset of features for the wine dataset, the accuracy improved from 74% up to 75%. This subset was created by dropping some features that did not impact the performance of the model.

When dealing with medical data such as the breast cancer dataset, it is important to have a high recall(sensitivity), as it is important to detect all malignant tumors, to the expense of labelling some benign tumors as malignant. To verify the performance of the model, a confusion matrix was generated. The rate at which the model was predicting type I and type II errors could then be tracked. The confusion matrix was also applied to the wine dataset as it could also produce more insights to the predictions. These insights could then be used change the decision boundaries of the models.

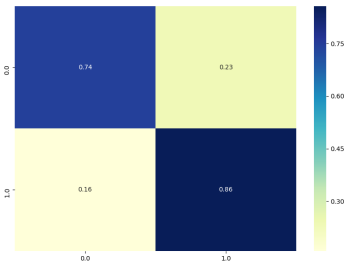

Figure 3. confusion matrix for the wine dataset using LDA

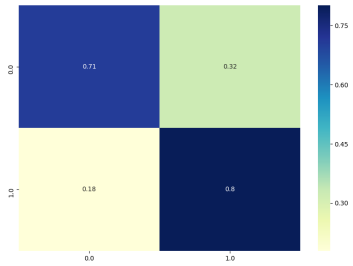

Figure 4. confusion matrix for the wine dataset using logistic regression

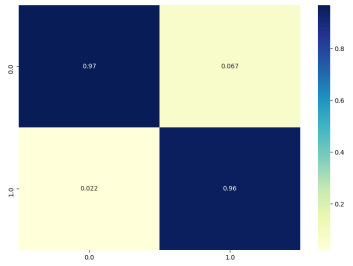

Figure 5. confusion matrix for the breast cancer dataset using LDA

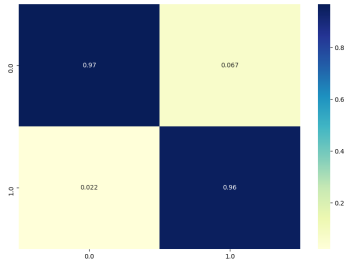

Figure 6. confusion matrix for the breast cancer dataset using logistic regression

4 Discussion and Conclusion

Both linear classification models were shown to display very similar results in both data sets despite their different fundamental assumptions on the data. However, they showed tremendous differences in their run time, with LDA consistently performing orders of magnitude faster than LR. Nonetheless, if compared purely on accuracy, LR performed marginally better on the wine data set, and both models performed similarly with classifying breast cancer, with around 96.5% accuracy.

Moving forward, it would be beneficial to automate the process of feature selection by using different feature selection techniques such as genetic algorithm and simulated annealing as proposed by Aich et al[2], where features would be generated, implemented and tested to determine their value.

For logistic regression, it was determined that the optimal hyperparameters to use were learning rate = 0.001, error threshold = 0.001, $\max_e poch = 1000$.

5 Statement of Contributions

With two models to implement and two data sets to train and validate on, the workload had to be separated. Li handled the initial pre-processing, the implementation of LDA and the confusion matrix. Michael implemented logistic regression, fine tuned the hyper parameters, applied regularization and aided in the processing. William worked on pre-precessing, feature selection and generation, evaluation function and k-fold validation function.

References

[1] S. Aich, A. Absi, K. Hui, and M. Sain, "Prediction of Quality for Different Type of Wine based on Different Feature Sets Using Supervised Machine Learning Techniques". 2019, pp. 1122-1127.

[2] Y. Gupta, "Selection of important features and predicting wine quality using machine learning techniques," Procedia Computer Science, vol. 125, pp. 305-312, 2018/01/01/ 2018, doi: https://doi.org/10.1016/j.procs.2017.12.041.

[3] Mahendrimd, Improve Wine Prediction with Feature Engineering, Kaggle, 07-Jul-2018. [Online]. Available: https://www.kaggle.com/mahendrimd/improve-wine-prediction-with-feature-engineering. [Accessed: 29-Sep-2019].

[4] M. Perme, M. Blas, and S. Turk, "Comparison of Logistic Regression and Linear Discriminant Analysis: A Simulation Study," Metodoloki Zvezki, vol. 1, pp. 143-161, 01/01 2004.