Logistic Regression and Linear Discriminant Analysis Models for Predicting Wine Quality and Breast Cancer Diagnosis

Nikhil Krishna, Brendan Furtado, and Chenqing Hua

I. ABSTRACT

This work applies two linear classification models, logistic regression and linear discriminant analysis, to the task of predicting wine quality and breast cancer diagnosis. The breast cancer dataset is relatively small, with 699 data points, compared to the dataset of wine which contains 1599 data points. Within these two benchmark datasets, we investigate certain patterns and trends and describe their importance. Next, we analyze and compare the performance of both classification models on the same sets. We have found that Linear Discriminant Analysis runs more efficiently than Logistic Regression and that both models provided similar accuracy levels. Further, we discovered that a higher learning rate, for logistic regression, leads to a steeper initial change in the cross-entropy loss than lower rates. In terms of accuracy, we also show that an alpha value of .001 gives the highest accuracy calculation.

II. INTRODUCTION

A. Preliminaries

Logistic regression and linear discriminant analysis (LDA) are two simple mathematical models used in machine learning to predict values on a continuous function (Logistic regression) and discrete classes (LDA) based on given features and datasets. In the continuous case, we implement the logistic regression model, where we are given n data points, each with m features $\{x_1, ..., x_m\}$. The task focuses on optimizing the weight vector $W = \{w_0, w_1, ..., w_m\}$ through gradient descent method so that the predicting model $y^i := w_0 + w_1x_1 + ... + w_mx_m$ can be used as the argument of the sigmoid function to give a positive or negative value corresponding to its predicted class. In discrete cases, since we are given two classes (known as binary classes) $\{0, 1\}$, we are then accurately predicting and classifying a datapoint by the linear decision boundary

$$\ln \frac{P(y=1)}{P(y=0)} - \frac{1}{2}\mu_1^T\Sigma^{-1}\mu_1 + \frac{1}{2}\mu_0^T\Sigma^{-1}\mu_0 + x^T\Sigma^{-1}(\mu_1 - \mu_0)$$

that returns a positive or negative value corresponding to its predicted class.

K-fold cross-validation is a procedure applied in both models, which allows us to mathematically, fairly and precisely validate the predictions. The dataset is evaluated in k folds after randomly split into k subsets. During k-fold cross-validation, on the i^{th} fold, the i^{th} subset is going to be used for prediction validation while other subsets (dataset excluding i^{th} subset) are used to train the model.

B. Task Description

In this work, we implement both the logistic regression model and the LDA model. However, their performance varies depending on features, learning rates, iterations, and size of datasets. In this work, we compare their performance, in terms of prediction accuracy and run time, over predictions of two datasets: Red Wine (1599 data points) and Breast Cancer (699 data points). We demonstrate how different weights and learning rates can affect the prediction accuracy and run time; k-fold cross-validation is used to validate the predictions from the models to further test our results. In particular, we standardized data and visualized important comparisons and findings in order to interpret and understand the two models in a clear way. The further details are illustrated in section 4, where we have graphs to explain findings, to discuss the performance and differences of two models. And in sections 2 and 3, we discuss how and what we selected for different features and learning rates so that the model returns a higher accuracy (i.e. better performance). In other words, the model is improved to perform a better job by selecting a reasonable learning rate and subset of weights.

C. Important Findings

We tested different learning rates for logistic regression on both datasets. The results show us that a decreasing learning rate, which converges to a small decimal improves the prediction for logistic regression (i.e. the accuracy improves as the learning rate decreases). As the learning rate converges, the accuracy stops improving at some point (i.e. accuracy stops changing or only with some unnoticeable changes). Different learning rates bring different convergence rates of cost. The cost actually converges to a certain value with high speed when we implement a relatively high learning rate (0.01), while a small learning (0.00001) rate brings a low convergence rate of cost.

By adding or subtracting different features, the prediction and accuracy vary significantly. Choosing proper features definitely improves the prediction and accuracy for the models.

We also noticed that even though, the wine dataset is larger compared to the breast-cancer dataset (1599 vs 699), logistic regression and LDA models have really close accuracy for prediction and classification. The noticeable point is that the LDA runs significantly faster compared to what logistic regression takes, even though they return the almost the same accuracy for prediction for both datasets.

III. DATASETS

A. Red Wine Data

One of the datasets used in this work focuses on certain features of 1599 samples of Italian red wine. The original data ranked the quality of the different wines on a scale from 0 to 10; however, to simplify classification into a binary task, we reranked the qualities as either a 0 or a 1. A "0" was assigned to those data whose quality ratings were less than 6, and "1" otherwise. 13 features were originally provided ranging from the alcohol content of the wine to its color intensity. One can see from Figure 1 that certain features, such as density and pH, do not change mean values significantly depending on wine quality. On the other hand, the average levels of sulfur dioxide across different red wines change dramatically when varying wine quality. Furthermore, an interaction term was added to the data: the interaction between total sulfur dioxide and sulfates. Multiple subsets of two features were selected and the logistic regression model was fit to the data including the interaction term between these two features. The interaction term chosen produced the highest increase in the accuracy of the model.

To select a subset of features, scatter plots between every pair of feature was constructed. The idea was to remove the features that were correlated with others, as they add no new information to the data. However, after removing the linear relationships, it was found that this decreased the accuracy of the model significantly. Therefore, all original features were kept.

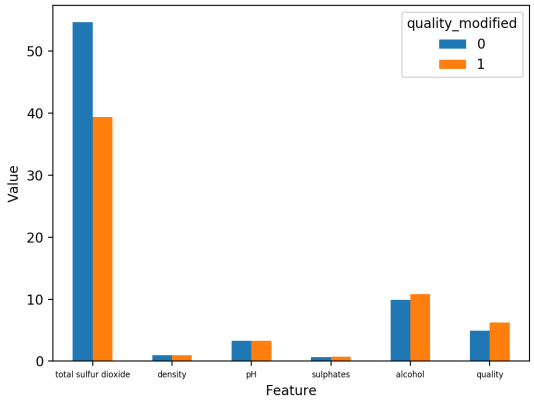


Fig. 1. Average values of selected features of the red wine dataset, separated by class

B. Breast Cancer Data

The second dataset comprises of 699 clinical cases of medical diagnosis applied to breast cytology. Occurrences of tumors within these clinical cases are classified as Benign or Malignant. Out of the 11 features given in the dataset, 9 of them are used to determine whether a tumor is benign or malignant. The ID number and classes itself are completely ignored for our predictions. The 9 features are scored on a scale from 1 to 10, 1 being a normal state and 10 being the most abnormal state. Additionally, there were 16 training examples with at least one feature that had missing data. Because of this, we decided to remove these examples from the set and proceed with one with 683 training examples and no missing information.

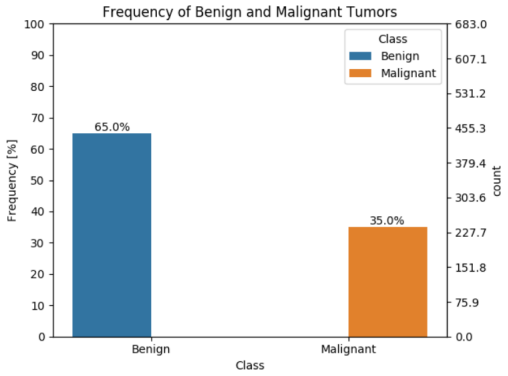


Fig. 2. Frequency of Benign and Malignant Tumors from Breast Cancer dataset

Figure 2 shows the frequency and number of occurrences for both classes in the dataset. We immediately see that most of data comprises of cases with benign tumors at 65.0% and malignant at 35.0%. This fact can be connected to the distribution of each of the features. As shown in the histograms in the Appendix (figure 8), we can see that much of the data is skewed to the right. For almost all of these features (with perhaps a few ambiguous exceptions), a higher value (>3) assigned to a feature tend classify the tumor as malignant. In other words, higher values within the features reveal higher percentages in malignant tumors.[1] From this, we concluded that the right-skewness of the features can be explained by the higher quantity of benign cases and that lower feature values tend to be found in the same cases.

When plotting certain sets of two features against one another, a decision boundary between the two classes can be found. In the Appendix (figs 9,10,11), we show that when Clump Thickness is paired with other features (Bland Chromatin, Marginal Adhesion, Uniformity of Cell Size), there are clear regions where the classes lie. Benign occurrences tend to be in the lower value region. Whereas Malignant cases have a lot more variance as the values of the features increase. This is significant because even though there are more benign cases in the dataset, the plots show that higher feature values (>3) reveal more malignant cases as expected.

C. Ethical Concerns

The only ethical concerns that arise with the red wine dataset have to do with getting permission from the wine makers to use the data. This dataset is publically available, however. As for the breast cancer dataset, more ethical questions arise. This is actual medical data collected from individuals in a study. The data was collected by researchers at the University of Wisconsin, and is widely used throughout the machine learning community. This coupled with the fact

that it is easily available for the public and proper permissions were taken by the researchers makes this concern moot. Additionally, participants of the study were only identified by ID number, no other information is included. Lastly, there are other outside variables that can potentially contribute to the diagnosis of a tumor. The given features used in the data should not be considered as the definitive variables to completely predict benign and malignant tumors. Caution has to be considered when analyzing data related to human illnesses.

IV. RESULTS

To test how the runtime of our logistic regression model changed with our learning rate, the parameters were fit with learning rates of 0.00001, 0.0001, 0.001, 0.01, 0.05, 0.1, and 0.5. Using python's time package, the runtime of 5-fold cross-validation was calculated for each learning rate and 10 gradient descent iterations. Both datasets were analyzed with this approach. We can see from Figure 3 below that, for the wine data, as the learning rate increased, first there was an increase in time, but after a certain threshold was crossed, the runtime started to decrease. Interestingly, a learning rate of 0.001 gives the highest accuracy, but also the highest runtime for the wine dataset when using logistic regression.

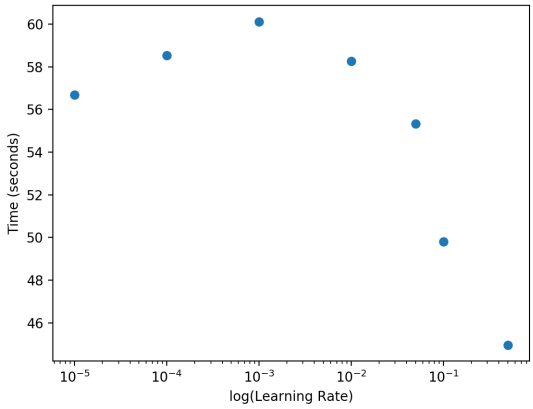


Fig. 3. Plot of runtime versus log(learning rate) for the wine dataset

Running the same setup as above for the breast cancer dataset reveals a similar downward trend in execution time as the learning rate increases. Unlike the wine dataset, this seems to happen more in a negatively linear fashion. We also determined that a learning rate of 0.001 and 0.01 give the best and equivalent accuracy of 0.9707. In addition, it can be seen on the graph that the same two learning rates give nearly the same execution time.

To compare the performance of linear discriminant analysis and logistic regression models, it is necessary to fix some variables. In order to fairly and precisely compare two models, we selected and fixed the learning rate(0.001), iterations(20), and a set of features that return the highest accuracy for wine and breast-cancer datasets.

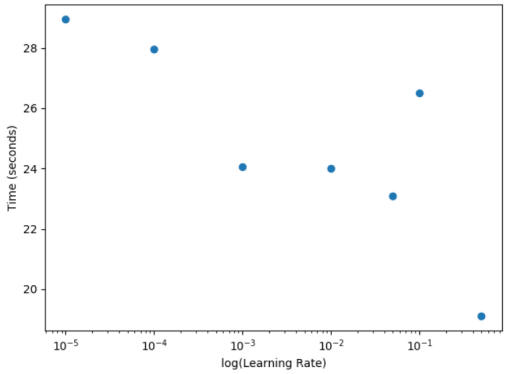


Fig. 4. Plot of runtime versus log(learning rate) for the breast cancer dataset

TABLE I
RUNTIME FOR MODELS WITH BOTH DATASETS

second(s)	LDA-WINE	LR-WINE	LDA-BC	LR-BC
AVG	6.877	111.022	1.846	44.203
MIN	5.994	107.039	1.641	42.234
MAX	8.010	120.987	2.298	48.247

TABLE II
ACCURACY OF MODELS BASED ON DATASET

Dataset	LDA	LR
Wine	0.763	0.752
BC	0.974	0.974

From Table II, the two models return close accuracy for prediction with the difference being less than 0.1 for both datasets. Even though the difference in accuracy comparison is not significant or noticeable (e.g Wine: 0.763 vs 0.752; BC: 0.974 vs 0.974), we can find significant difference in run time as shown in Table I. For prediction in breast cancer data, LDA takes 1.846s while logistic regression takes approximately twenty times of it, 44.203s. And for the prediction in the wine dataset, the average run time of logistic regression is also approximately twenty times of what LDA takes, 6.877s vs. 111.022s. After seeing the significant difference in run time, we can conclude that the LDA model performs a better job in predicting and classifying objects (in both wine and breast-cancer datasets) in terms of efficiency.

TABLE III
CHOSEN METRICS COMPARED WITH OR WITHOUT INTERACTION TERM

Value	Without Interaction Term	With Interaction Term
Accuracy	0.724	0.752
Precision	0.719	0.749
Recall	0.749	0.780

Before adding the new interaction term to the wine data, 5-fold cross-validation using logistic regression was performed. After each fold, a confusion matrix was calculated. At the

end of all 5 folds, the confusion matrices were all added together. This same process was repeated after adding the new interaction term to the data. This resulted in the two confusion matrices shown below in Equation 1. Table III shows how accuracy, precision, and recall all change when this new term is added. This change is a 3-4% increase in the three metrics. Note that during all 5-fold cross-validations with logistic regression, a learning rate of 0.001 was used and gradient descent was run for 20 iterations.

$$C_{\text{noint}} = \begin{bmatrix} 610 & 238 \\ 204 & 547 \end{bmatrix}$$

$$C_{\text{int}} = \begin{bmatrix} 640 & 215 \\ 181 & 563 \end{bmatrix}$$

Eqn. 1. Confusion matrices for logistic regression

V. Discussion & Conclusions

In this work, we focused on using two different prediction models, logistic regression and linear discriminant analysis, to predict wine quality and seriousness of breast cancer tumors. We also analyzed these models further, specifically the logistic regression model, determining how many iterations of gradient descent should be used and what learning rate is optimal. Initially, we analyzed features of both datasets in an attempt to understand the data better. It was found that graphing certain features against other features and graphing the distribution of individual features gave us valuable insight as to why the model behaved the way it did when fitting data. It also told us whether certain independent variables were correlated with each other.

In the runtime analysis we performed, it was seen that, in all cases, LDA was significantly faster than logistic regression. This is most likely due to the closed-form nature of LDA compared to a numerical gradient descent approach to logistic regression. Additionally, the learning rate and number of gradient descent iterations was varied in the logistic regression model. It was noted that there is a "peak" in runtime when varying the learning rate. Also, as learning rate increased, the drop in the cost after a few iterations of gradient descent became steeper as shown in Figures 5, 6, and 7 below.

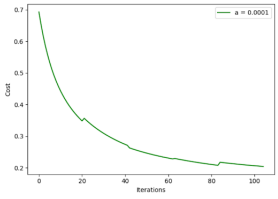


Fig. 5. Plot of cross-entropy loss versus iterations of gradient descent with $\alpha = 0.0001$ for breast cancer dataset

We expected LDA to produce better results for smaller datasets when compared to logistic regression. This is due to the fact that the LDA is a closed-form solution. However, we experienced no significant differences in terms of differences of accuracies between the logistic regression and the LDA models.

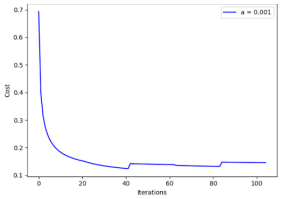


Fig. 6. Plot of cross-entropy loss versus iterations of gradient descent with $\alpha = 0.001$ for breast cancer dataset

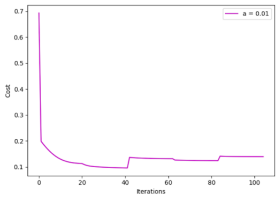


Fig. 7. Plot of cross-entropy loss versus iterations of gradient descent with $\alpha = 0.01$ for breast cancer dataset

It is also worth noting that the k-fold cross validation that was applied to models was incredibly beneficial to model performance. We were able to compare several different logistic and LDA models (varying the hyperparameters, subsets of features, etc.) to determine which one gives the best accuracy.

This work, however, can be improved in the future. For example, one could vary the learning rate with a fixed number of gradient descent iterations, making it approach 0 as the number of iterations approached infinity. This would most likely lead to better accuracy on the wine dataset. Furthermore, many of the weights calculated each dataset (more-so the wine data) were not significant. We could employ L1 or L2 regularization to optimize weights, allowing a better fit for the data by promoting bias over variance. Furthermore, sophisticated feature selection tools could be employed in the future to improve accuracy, especially for the wine dataset.

VI. Statements of Contribution

Nikhil Krishna: Analyzed and produced graphs the red wine dataset, explored which features would increase model accuracy, compared how runtime changed with different hyperparameter values, wrote k-fold function for logistic regression, helped implement logistic regression model's fit and predict functions, and wrote part of the final report.

Chenqing Hua: Designed and build two models from scratch including all the required methods and computations(e.g gradient descent, cross entropy, covariance matrix, linear decision boundary etc.), fit, predict, and evaluate functions, plus confusion matrix. And wrote part of the report.

Brendan Furtado: Prepared, analyzed, and produced graphs for the breast cancer dataset. Experimented with different graphs and distributions. Ran runtime and accuracy calculations for both models on the same dataset. Modified part of the logistic regression to include a cost list, this list

was used to graph number of iterations vs cost given different learning rates. Wrote part of the final report.

REFERENCES

[1] Alaa. M. Elsayad, H. A. Elsalamony "Diagnosis of Breast Cancer using Decision Tree Models and SVM," International Journal of Computer Applications (0975 – 8887), Volume 83. Cairo, Egypt: December 2013

[2] Hastie, Trevor "An Introduction to Statistical Learning with Applications in R," 8th ed. New York: Springer, 2013

APPENDIX

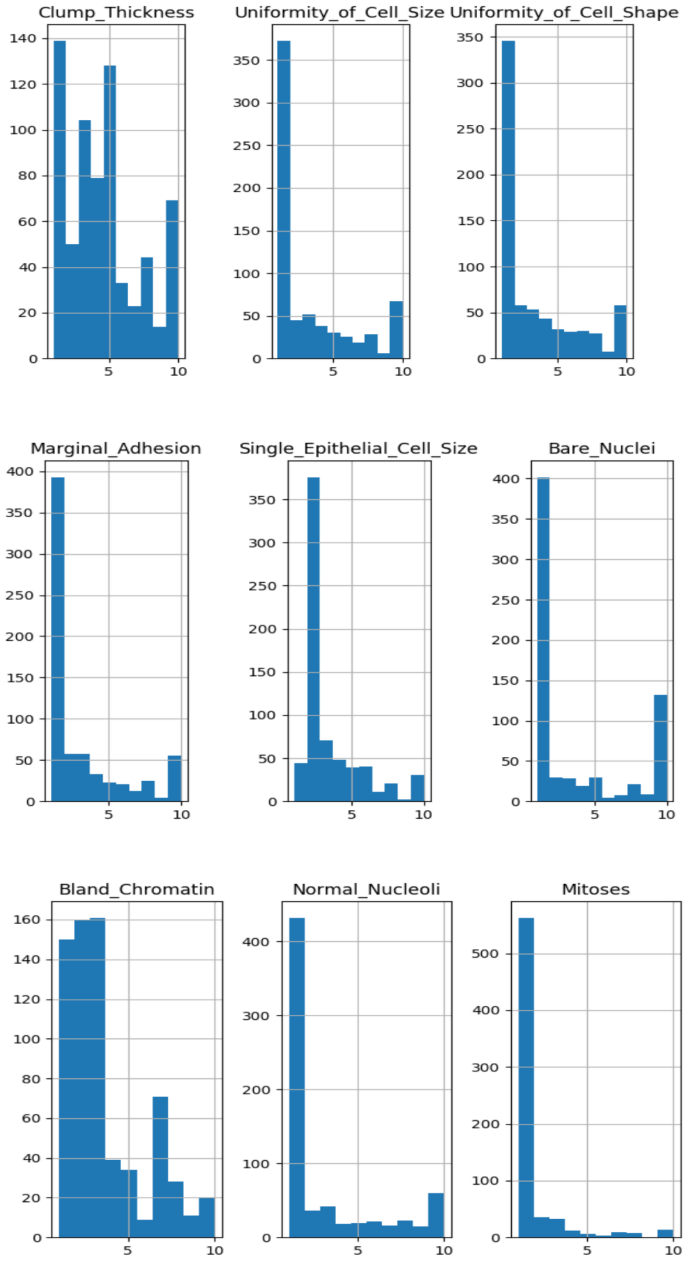
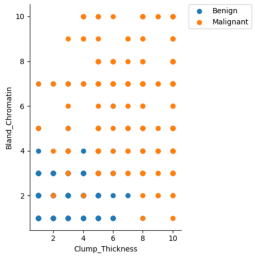


Figure 8: Distribution of Breast Cancer Features



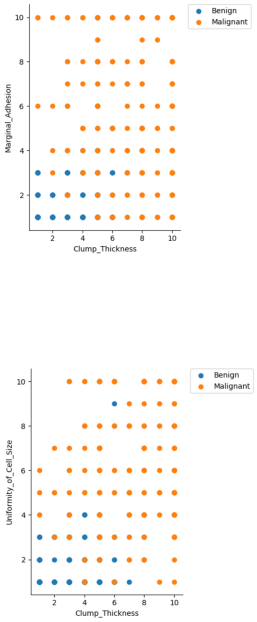


Figure 9: Several scatter plots between selected features to show a decision boundary between benign and malignant tumor cases.

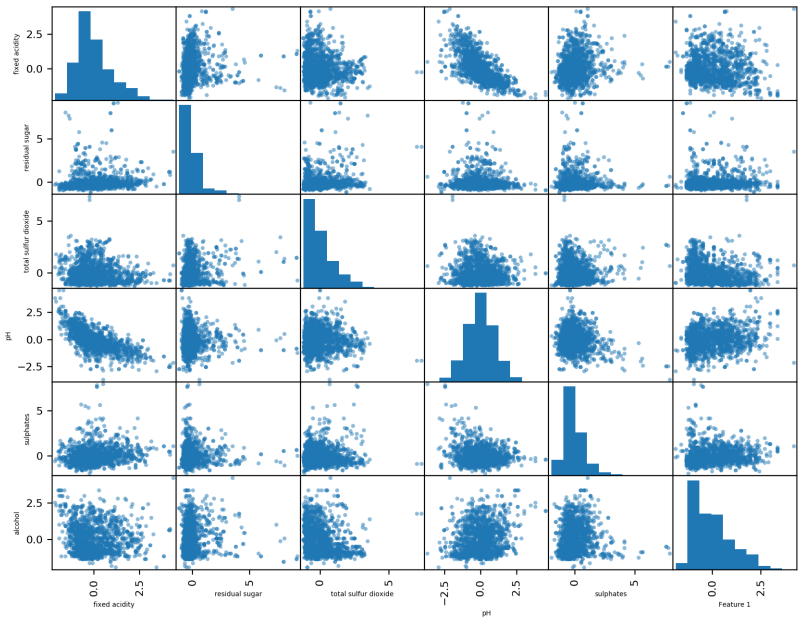


Figure 10: A scatter matrix showing scatter plots between selected features of the wine dataset. Note the diagonal of the matrix shows the distribution of each selected feature.