EXPLORING LOGISTIC REGRESSION AND LINEAR DISCRIMINATIVE ANALYSIS FOR BINARY CLASSIFICATION

COMP 551, Applied Machine Learning
McGill University

Anton Gladyr, Saleh Bakhit, Vasu Khanna

September 2019

Abstract

In this paper we investigated the accuracy and runtime of Logistic Regression and Linear Data Analysis (LDA) for binary classification on red wine quality dataset as well as breast cancer dataset. We have found that vectorized Logistic Regression performs much faster than element-wise summation when updating weights. Our Logistic Regression model was able to reach 85% accuracy in 0.13 seconds on the cancer dataset and 72.66% in 17 seconds on the wine dataset. LDA on the hand reached an accuracy of 96% accuracy in 0.003 seconds on the cancer dataset and 74.23% in 0.005 seconds on the wine dataset. We can therefore conclude that LDA has better performance when compared to Logistic Regression's accuracy and runtime.

1 Introduction

This paper examines two linear classification techniques for binary classification problems. Namely, Logistic Regression and Linear Discriminant Analysis (LDA). For training and validation purposes, we use two datasets: red wine dataset to classify good quality red wine, and breast cancer dataset to classify malignant or benign tumors based on the given features. The paper explores different implementations of the algorithms and compares their accuracy and runtime.

Several past works have been done on both the data sets[2]. For instance, "Modeling wine preferences by data mining from physicochemical properties" is worth mentioning because of their promising results using Support Vector Machine which outperforms regression and neural network methods[1].

2 Datasets

In the first stage of this project, the datasets were investigated and cleaned from any missing or malformed examples, e.g. question marks in the cancer dataset. In addition, patients' id numbers have been removed from the cancer dataset because it has no significance in the generalization of the model. Since the goal of the project is to examine binary classification algorithms, initial target values have been changed into zeroes and ones. Therefore, in the wine dataset we convert reported quality readings $<= 0.5$ to 0 (bad quality) and readings > 0.5 to 1 (good quallity). In the cancer dataset, benign has been changed from 2 to 0 and malignant from 4 to 1. This step is important since further predictions will report a binary classification as 0 or 1.

The distribution of good and bad wines is 46% and 54% respectively, whereas cancer class distribution is 65% of benign vs. 35% of malignant tumors. Figures 1 and 2 show a heatmap of each dataset. This is useful in identifying highly co-related features for possible removal.

In terms of ethical concerns, as machine learning practitioners, we must recognize that our algorithms are often capable of working on information that users did not intend to make public. If for example wine data comes from private-sector sources, intellectual property issues as well as potential conflicts of interest issues may arise. There is also a risk of misinterpretation if the data were not appropriately documented by the original collector. For the breast cancer data set, it is necessary to ask for patients' approval before using their information.

Function *split_dataset*() splits the data into training and validation dataset based on the percentage

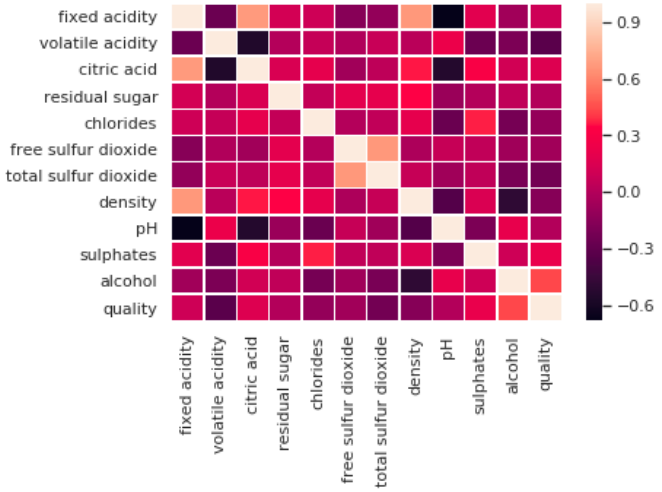

Figure 1: Heatmap of wine dataset, 1 represents similarity and 0 represents

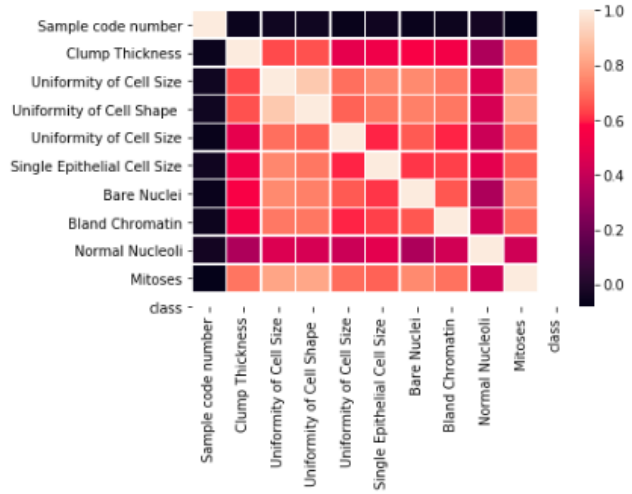

Figure 2: Heatmap of cancer dataset, 1 represents similarity and 0 represents

entered. Cancer dataset has 683 examples and 9 features, whereas wine dataset has 1599 examples and 11 features.

3 Results

All the results shown down below are run on the datasets after cleaning as described in section 2. The implemented models have been run using 5-fold cross validation to estimate performance in all of the experiments. The results of running implemented models are shown in Table 1.

3.1 Logistic Regression

Logistic Regression is a probabilistic discriminative learning approach for classification problems. It estimates the probability of class y given a set of features x, where $y \in 0, 1$ for binary classification. More specifically, it estimates the log-odds function α with a linear one according to the following set of equations:

	Element-wise LR	Vectorized LR	Binary LDA	Multiple LDA
Accuracy	85	85	96.03	96.03
Runtime (seconds)	18.97	0.16	0.0042	0.0046
Number of iterations	4000	4000	-	-
Learning rate	0.001	0.001	-	-

Table 1: results for running different models on the cancer dataset

	Element-wise LR	Vectorized LR	Binary LDA	Multiple LDA
Accuracy	84.41	84.41	91.30	91.30
Runtime (seconds)	19	0.1	0.0048	0.0040
Number of iterations	4000	4000	-	-
Learning rate	0.001	0.001	-	-

Table 2: results for running a new subset on the cancer dataset

$$\Pr(y = 1 \mid X) = \frac{1}{1 + e^{-\alpha}} \tag{1}$$

$$\alpha = \ln(\frac{\Pr(y = 1 \mid x)}{\Pr(y = 0 \mid x)}) \approx w_0 + w_0 x_1 + ... + w_m x_m \tag{2}$$

$$\alpha = \approx w_0 + w_0 x_1 + ... + w_m x_m = -W^T X \tag{3}$$

Finally, we approximate $\Pr(y = 0 \mid x)$ by minimizing the Cross-entropy loss using gradient descent according to the following update rule:

$$w_{k+1} = w_k + \alpha_k \sum_{i=1}^{n} x_i (y_i - \sigma(w_k^T x_i)) \tag{4}$$

$$w_{k+1} = w_k + \alpha_k (X \cdot (Y - \sigma(W_k \cdot X))) \tag{5}$$

Optimal Number of Iterations

In our implementation of Logistic Regression, we have two varying parameters: learning rate and number of iterations. In order to find the number of iterations at

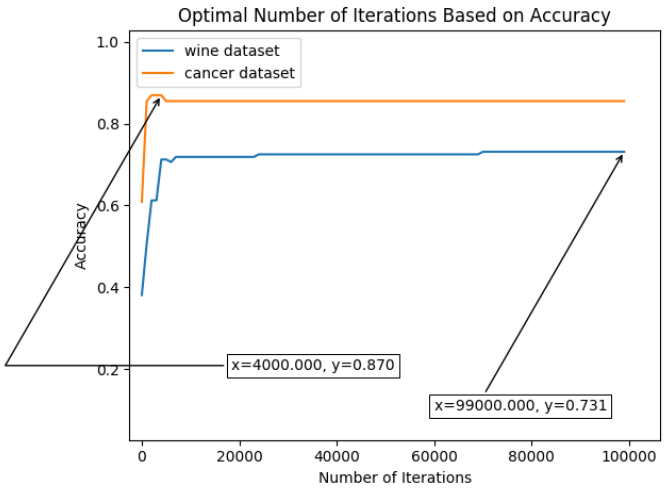

Figure 3: Optimal Number of Iterations

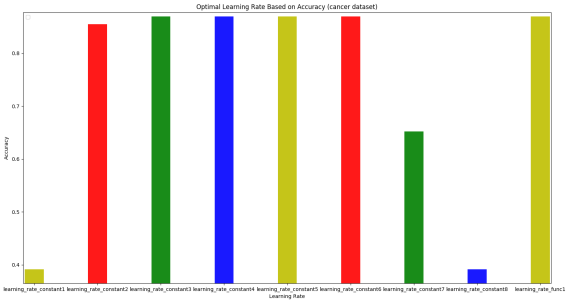

Figure 4: Optimal Learning Rate (Cancer Dataset)

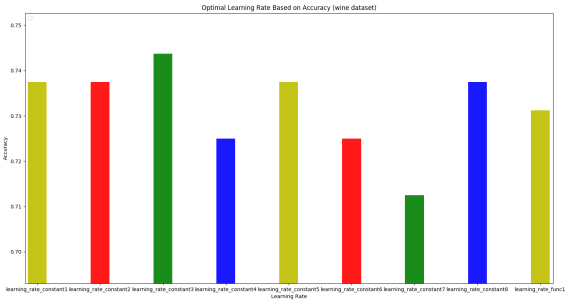

Figure 5: Optimal Learning Rate (Wine Dataset)

highest accuracy, we fix learning_rate to $\frac{1}{1+k}$, where k is the iteration number, and vary the number of iterations until the accuracy plateaus. As seen in Figure 3, **4000** iterations is optimal for cancer dataset and **99000** for wine dataset.

Optimal Learning Rate

In order to obtain the optimal learning rate function, we fix the number of iterations according to Figure 3 and vary the learning rate. Since learning rates in practice are small constants, we mainly test different constant values in addition to $\frac{1}{1+k}$. As seen in Figures 4 and 5, **learning_rate_func1 (last bar)**, corresponding to 0.001, is the best learning rate amongst the ones tested. These are 1.0, 0.01, 0.001, 0.0001, 0.00001, 0.000001, 0.0000001, 0.00000001 and $\frac{1}{1+k}$ respectively to Figures 4 and 5.

Element-wise Operations

Initially, we implemented Logistic Regression with gradient descent by updating the weights according to Equation 4 element-wise.

```
for itr in range(maxitr):
    for row in features:
        ex = features[row]
        t = targets[row]
        p = sigmoid(np.dot(ex, parameters.T))
        sum += ex*(tar - p)
    parameters += learning_rate*sum
```

Listing 1: Element-wise snippet

Looking at Listing 1, we see that we iterate over all features element-wise to compute the gradient. This proved to be extremely time-consuming.

Vectorized Operations

An alternative to element-wise summation is taking advantage of vectorized operations. As seen in Equation 4, each parameter in the summation is independent from the others. Equation 5 illustrates how we can employ vectorization to get rid of the inner for-loop. Table 1 shows detailed comparisons of both methods, and we can see significant runtime improvements.

3.2 Linear Discriminant Analysis

Another approach for linear classification is generative learning, which models the distribution of the different classes. In frameworks of this project, linear discriminant analysis (LDA) has been used as a generative model. LDA assumes that every class in a dataset is a

normally distributed cluster of datapoints. For training an LDA model for two-classes case, it is needed to find probabilities of target classes $P(y = 0)$ and $P(y = 1)$, mean vectors μ_0 and μ_1, and shared covariance matrix Σ for the entire dataset. For making predictions the model uses a linear decision boundary, which returns log-odds ratio between two classes. To predict values, the model takes a set of features x and classify each sample as a class 1 if the log-odds ratio is greater than 0. Otherwise the sample will be predicted as a class 0. In the developed implementation, the model can be trained for binary classification as well as for predicting multiple classes. Training and validating the model on wine dataset gives the following performance results: **74.23% of accuracy** and **0.0062 seconds of runtime**. The results of training the model on cancer dataset are shown in the Table 1.

4 Discussion and Conclusion

Throughout this project, we learnt that different number of iterations and learning rates can have a high impact on the accuracy and runtime of logistic regression model. We have also seen first hand that feature engineering can have a great effect on a models' performance. It was very helpful to visualize the data and know the correlation between different variables as well as knowing different distribution of the features in the datasets. This helps identifying possible redundant features the removal of which would reduce noise and prevent overfitting of the model.

For future work and potential enhancements, it might be helpful to use normalization and regularization techniques on the datasets to improve results. It would be a good idea to remove highly correlated features, i.e features that have correlation above 0.9 in the correlation matrix (or heatmap). For example, Figure 2 shows that "uniformity of cell size" and "uniformity cell shape" features in the cancer dataset are highly correlated. Deleting one of the features would possibly increase the accuracy or at least improve runtime as there are less features to take into account.

5 Statement of Contributions

Project tasks were discussed and equally distributed among the team members during team meetings, such as:

- cleaning the data sets
- acquiring, analyzing and visualizing of data sets
- implementing logistic regression model
- implementing linear discriminant analysis model
- comparing the execution time and accuracy of the models
- finding a new subset of features that improve the accuracy
- writing-up of the results obtained during the project

Every team member made a significant contribution to the project and was actively involved in tackling the issues.

References

[1] F. Almeida T. Matos P. Cortez, A. Cerdeira and J. Reis. Modeling wine preferences by data mining from physicochemical properties. In *Decision Support Systems*, volume 47, pages 547–553. Elsevier, 2009.

[2] W.H. Wolberg W.N. Street and O.L. Mangasarian. Nuclear feature extraction for breast tumor diagnosis. In *IST/SPIE 1993 International Symposium on Electronic Imaging: Science and Technology*, volume 1905, pages 861–870. San Jose, 1992.