Using Logistic Regression and Linear Discriminant Analysis to predicting wine quality and breast cancer class

Faizan Khan
Electrical and Computer Engineering
McGill University
faizan.khan3@mail.mcgill.ca
260783869

Anthony Ho
Computer Science,
McGill University
anthony.ho@mail.mcgill.ca
260501840

Maruthi Rangappa
Electrical and Computer Engineering
McGill University
maruthi.rangappa@mail.mcgill.ca
260819491

***Abstract*—In this project we implemented two main Linear Classification (LC) techniques of machine learning - the Logistic Regression (henceforth referred to as LR) and the Linear Discriminant Analysis(LDA). We also evaluated their performance on two unique data-sets and found that LDA had a better accuracy compared to LR and also was significantly faster to train. We also implemented L1 and L2 regularisation for LR and found out that L2 improved the accuracy of our Logistic Regression model but at the cost of runtime. In addition to that we found that Logisitic Regression performed significantly better on the dataset where the output classes were skewed towards one class. On the other hand, LDA performed better irrespective of the distribution of resulting class .**

I. Introduction

Linear classification is a commonly used supervised learning machine learning technique that uses information from an object's features to predict the class to which it belongs to. There are two types of linear classifications techniques viz Discriminative and Generative classification. Both types differentiate in the way they arrive at their results.In Discriminative Learning we model the boundaries between the classes that an object might belong to, in Generative Learning we find the estimate the probability of an object belonging to a certain class by using the distribution information of the features. In a way the Discriminative approach is calculating the classification probability directly, whereas the Generative approach is calculating it indirectly by first calculating the probability distribution of its features.

In this project we implemented and compared the runtime performances of two classification techniques, Linear Classification (LR) and Linear Discriminant Analysis (LDA). Logistic Regression is a Discriminative Learning approach that makes certain assumptions about the distribution of an objects features. It assumes that the features are Independant and Identically Distributed (IID). Linear Discriminant Analysis is a Generative Learning approach that calculates the probability of the objects belonging to a certain class, by first calculating the probability of its features given the probability of the class. LDA assumes that the features given they belong to a class are Gaussian distributed with the same covariances. This assumptions makes LDA faster to train. From empirical data in the past it has also be proven that LDA works better for small datasets.

A. Relevant Work

The inspiration for the dataset came from similar research done on these datasets before hadn. For example, Cortez et al [2] used this dataset to predict wine preference of people using Support Vector Machine (SVM). Similarly, breast cancer data has also been used to classify tumors into malignant and benign. Mangasarian and Wolberg [4] applied linear learning techniques to diagnose breast cancer from fine-needle aspirates. Wolberg and Mangasarian [5] [3] used multi-surface separation techniques to classify breast cancer. Similarly, Bennet and Mangasarian [1] used the same dataset to propose linear techniques to separate two linearly inseparable sets.

II. Methodoloy and Terms

The project can easily be divided into three main tasks viz. Acquiring/Cleaning Data, Implementing Models, Calculating Results of the Models.

A. Data Acquisition

We used two different data-sets for this project, wine-quality[kp] data and breast-cancer[kp data from University of California Irvine's online data repository.

Wine data didn't require much cleaning, but the column/feature "bare nuclei" of breast cancer data contains "?", so we considered that value to be malformed and removed the row altogether. More details on this can be found in section II.

B. Model Implementation

We implemented two models from scratch viz Logistic Regression and Linear Discriminant Analysis (LDA).

1) Logistic Regression: Logistic Regression uses a gradient descent to train the the models. Gradient descent was implemented using the following update rule.

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

Where w is the param vector, x is the feature vector, and σ is the sigmoid function that results in a 1 or a zero. The prediction in Logistic Regresion is calculated by simply calculating the log odds ratio of the given classes, and then giving the input to the sigmoid function. As shown in the following formula.

$$\hat{P}(y = 1|x) = \sigma(w^T x) = \frac{1}{1 + e^{-w^T x}} \quad (2)$$

2) Regulariations: In addition to the basic error optimization for Gradient Descent in logistic regression, we also implement L1 and L2 regularisations for Logistic Regression.

L2 regularization also known as ridge regression constraints the weights by imposing a penalty on the sizes. Therefore the higher the weights the more the penalty. By adding the L2 factor, the update equation in 1 is changed to the following equation.

$$w_{k+1} = w_k + \alpha_k((\sum_{i=1}^{n} x_i(y_i - \sigma(w_k^T x_i)) + 2\lambda w) \quad (3)$$

Where λ is the regularization factor. It can be found manually or by using k-fold validation. We do this using k-fold validation. The values are given in Table VII.

L1 regularization also known as Lasso regularization, does the same thing but instead of adding squared of the paramters to the gradient descent equation we add an absolute value. Taking the derivate of the equation updates the equation in 1 to the following.

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

where $sign$ is a function given shown in the following image

3) Linear Discrimant Analysis: LDA uses Bayes rule to estimate distribution of two class, and the distribution is used in computing log odd ratio for each value. For 2 classes, we can easily compute the log-odds ratio of the two classes, making the decision boundary at 0. In addition to that, LDA assumes that the features have a gaussian distribution with the same covariances but can have different means.

Therefore, for the final calculation, all that is needed is calculating the mean of each class, the shared covariance and relative probabilities of the two classes and we can calculate the log odds ratio of the two classes. Which can then be used to in predicting whether the objects belongs to one class or the other.

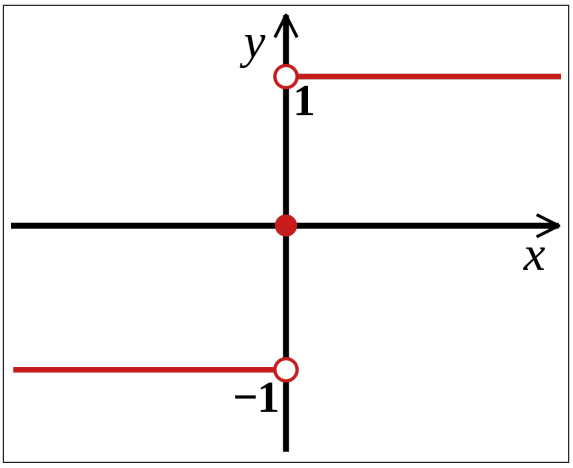

Fig. 1. Sign function

C. Runtime Calculations

We performed 5-fold validation to evaluate the performance of our implementations. We found out that our logistic regression implementation performed better at higher learning rates (0.1, 0.01) giving us an accuracy of 0.54 in predicting the wine quality with an average run-time of about 6s, and an accuracy of 0.79 with an average run-time of 10s for the breast-cancer dataset. In addition to that L2 regularization with a lambda factor of 1, improved the performance to 0.61 on the wine dataset with the same runtime of around 6s.

Similarly, our LDA implementation performed better than Logistic Regression on both data sets, with an average accuracy of 0.69 for wine data set and 0.89 for breast cancer dataset respectively. The training time was also shorter than the Logistic Regression in both cases.

III. Experiment

A. Dataset Pre-Processing

The datasets are filtered for any possibly ill-formed feature vector to avoid wrongly training the prediction models. In this step we identified that 16 of 698 records in the column (*Bare Nuclei*) of breast cancer data with have a question mark "?" as data. These records were removed from the dataset.

For the wine quality dataset, we created a new feature representing the target variable wine quality by mapping the quality values (0,5) to 0 which represents bad wine quality and by mapping values (6,10) to 1 to represent good wine quality. Similarly for cancer dataset we created a new feature representing the target variable of cancer data, by converting class-4 in the data to 1 representing malign, and converting class-2 to 0 representing benign.

Learning Rate	Accuracy	Time(s)
0.1	0.509375	6.410499095916748
0.01	0.50625	6.231107711791992
0.001	0.315625	6.278414011001587
0.0001	0.39375	6.084130048751831
0.00001	0.50625	6.041225910186768

TABLE I
PERFORMANCE AND ACCURACY OF LOGISTIC REGRESSION FOR WINE DATASET

Learning rate	accuracy	Time (s)
0.1	0.79562044	9.389914751052856
0.01	0.79562044	10.34795904159545
0.001	0.79562044	10.60004520416259
0.0001	0.79562044	11.59649491310119
0.00001	0.79562044	9.325804948806763

TABLE II
PERFORMANCE AND ACCURACY OF LOGISTIC REGRESSION FOR BREAST CANCER DATASET

B. k-fold cross validation

We use k-fold validation model to validate our theoretic models. The k-fold validation method divides the training data into k chunks, taking 1 chunk as a validation chunk and the remaining k-1 chunks as test chunks. We used a 5 fold validation method, so the chunk size in our case is 0.2. We take the average of accuracy resulting from each chunk for each theoretic model. By comparing the average accuracy scores of two models, we select the one with the higher accuracy

Once the best model is validated, we then train our model on the entire training set. Then we use this trained model to test it on the test data set.

IV. RESULTS

We performed 5 fold validation of logistic regression on both datasests with different learning rates. Table I lists the performance and accuracy of logistic regression at various learning rate for the wine dataset. Our implementation of logistic regression running with wine dataset did not achieve a good accuracy. At learning rates of 0.1, 0.01, 0.00001 we achieved a maximum accuracy of 50 percent at an average runtime average of 6s at 1000 training iterations. But for the breast cancer data set, our logistic regression model achieved a good accuracy and runtime as shown in Table II at all learning rates with an accuracy of 79 percent with a runtime average of 10s at 1000 training iterations.

Table III provides a comparison of accuracy and runtime of our linear regression model with the LDA model. We see that in terms of accuracy and runtime, Linear Discriminant Analysis out-performed Logistic Regression with wine dataset. A similar comparison of results for the breast cancer data set is shown in Table IV and we see that here also Linear Discriminant Analysis out-performed Logistic Regression in terms of accuracy and runtime. We also compared the accuracy and runtime and our implementation with the sklearn implementations. As shown in Table V, in terms of accuracy and runtime, sklearn implementation of LR out-performed our implementations by a fair amount. On the other hand, our implementation of LDA has very similar accuracy and runtime as sklearns implementation of LDA as shown in Table VI

Model	accuracy	Time (s)
LR($alpha$=0.1, $iter$=1000)	0.50	6.41
LDA	0.69	0.042
LR($alpha$=0.1, $iter$=1000, λ=1)	0.62	21.30

TABLE III
LR, LDA AND LR WITH L2 FOR WINE DATASET

Model	accuracy	Time (s)
LR($alpha$=0.1, $iter$=1000)	0.79	9.3
LDA	0.89	0.024
LR($alpha$=0.1, $iter$=1000, λ=1)	0.84	10.59

TABLE IV
LR, LDA AND LR WITH L2 FOR BREAST-CANCER DATASET

Based on the results, breast cancer dataset are much easier to train and to obtain a good result than wine dataset. In terms of runtime performance, LDA is 10 times faster than logistic regression. However, in terms of accuracy, both LDA and logistic regression have almost the same accuracy.

For breast cancer dataset training with linear regression model, changing of learning rate doesn't seem to have an impact on both accuracy and runtime. However for the wine dataset, changing of learning does have an impact on the accuracy of the classifier where a learning rate of 0.01 and 0.001 has decreased the accuracy from 50 percent to 30 percent and a learning rate of 0.00001 will increase the classifiers accuracy back to 50 percent.

V. DISCUSSION AND CONCLUSION

In this project, we implemented two main linear classification techniques - linear regression using gradient descent and linear discriminant analysis - and evaluated their performances. We learnt that linear regression and LDA are

Model	accuracy	Time (s)
sklearn($solver$=lbfgs, $iter$=1000)	0.69	0.23
Custom ($lr = 0.1$, $iter$=1000)	0.50	6.41
Custom_L2 ($alpha$=0.1, $iter$=1000, λ=1)	0.62	21.30

TABLE V
CUSTOM, SKLEARN, AND CUSTOM WITH L2 IMPLEMENTATION OF LR ON WINE DATASET

Model	accuracy	Time (s)
sklearn	0.96	0.013
Custom	0.79	0.92
Custom_L2	0.84	10.59

TABLE VI
CUSTOM, SKLEARN, AND CUSTOM WITH L2 IMPLEMENTATION OF LR ON BREAST CANCER DATASET

lambda	accuracy	Time (s)
5	0.5866026645768025	6.246491861343384
4	0.5491026645768025	6.850890302658081
3	0.5578526645768025	7.043605470657349
2	0.6159776645768025	6.050075531005859
1	0.6166026645768025	6.20238790512084
0.1	0.6047198275862069	5.834296464920044
0.01	0.5440967868338558	6.0814002513885494
0.001	0.5846943573667711	6.0814002513885494
0.0001	0.5090615203761756	5.869805860519409

TABLE VII
L2 REGULARIZATION RESULTS WITH WINE DATA WITH 300 TRAINING ITERATIONS, 0.1 LEARNING RATE

lambda	accuracy	Time (s)
1	0.5384776645768025	5.753042125701905
0.1	0.5771963166144201	6.658836603164673
0.01	0.5659463166144201	5.635331439971924
0.001	0.5659463166144201	5.8134012699127195
0.0001	0.5659463166144201	5.712115001678467

TABLE VIII
L1 REGULARIZATION RESULTS WITH WINE DATA WITH 300 TRAINING ITERATIONS, 0.1 LEARNING RATE

very effective on binary classification problems in terms of prediction accuracy and runtime provided a clean dataset that is nicely distributed and is free from ill formed records. We also conclude that in comparison to logistic regression, LDA is more efficient in terms of both accuracy and runtime on the dataset that we used.

Using L2 regularization improved the accuracy of logistic regression but increases the training time. In addition to that using different combination of features, i.e. squaring features gave us a better results on than just using using plain features in combination with square features.

VI. STATEMENT OF CONTRIBUTIONS

All groups members contributed equally in this project. Table IX lists contribution from different group members.

VII. ETHICAL CONSIDERATIONS

The chemical properties of the wine in the wine dataset could be miss used by one company if they know that this data is coming from a competition that is doing well in a particular segment of wine market. Datasets like breast cancer dataset might have personal information about the patient which must be treated very confidential by the developers working on these datasets.

Faizan Kahn	Anthony Ho	Maruthi Rangappa
Dataset processing	Dataset processing	Dataset processing
Implement LR	Implement LDA	Implement cross validation
run tests	run tests	run tests
report	report	report

TABLE IX
STATEMENT OF CONTRIBUTIONS

REFERENCES

[1] K. P. Bennett and O. L. Mangasarian, ""Robust linear programming discrimination of two linearly inseparable sets"," in *Optimization Methods and Software 1*, vol. 23-34, no. 5, 1992, pp. 1,18.

[2] P. Cortez, A. Cerdeira, F. Almeida, T. Matos, and J. Reis, "Modeling wine preferences by data mining from physicochemical properties," vol. 47, no. 4, pp. 547 – 553. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0167923609001377

[3] Mangasarian, Setiono, and Wolberg, ""Pattern recognition via linear programming: Theory and application to medical diagnosis"," in *Large-scale numerical optimization*, vol. 23, no. 5. SIAM Publications, 1990, pp. 22–30.

[4] O. L. Mangasarian and W. H. Wolberg, ""Cancer diagnosis via linear programming"," in *SIAM News*, vol. 23, no. 5, 1990, pp. 1,18.

[5] W. H. Wolberg and O. Mangasarian, ""Multisurface method of pattern separation for medical diagnosis applied to breast cytology"," in *Proceedings of the National Academy of Sciences*, vol. 87, no. 5, 1990, pp. 9193–9196.