ECSE551 - Mini Project 1

Aly Mohamed
260902616
aly.mohamed@mail.mcgill.ca

Robert Aprahamian
260889522
robert.aprahamian@mail.mcgill.ca

Abstract

In this first project, the goal was the implement a logistic regression binary classifier. The final goal was to classify all the data. Therefore, we went ahead and used two different data sets that were provided to us. We started by downloading the data, then doing some statistical analysis to achieve a better understanding and clarification of the data, then we needed to find the best logistic regression model that accomplishes our need. To achieve that we made 10-fold cross validation runs as well as testing to determine if Gradient Descent (GD) works better than Linear Discriminant Analysis or not.

1 Introduction

During this first mini-project, we were asked to build a logistic regression binary classifier from scratch. We were asked to apply this classifier to two different sets of data: White wine quality and Kidney Disease. This was done through several steps: We first had to download the data that was given to us. We then had to perform some operations that would serve as a statistical analysis of the data sets just to have an idea about the data and if there are any notes to make about the two data sets in general. Next, we implemented the linear classifiers logistic regression from scratch. We used two different approaches: Gradient Descent (GD) and Linear discriminant analysis (LDA) to fit and predict labels for the input points. Moreover we measured the models accuracy and made a 10-fold cross validation to estimate the performance. In addition to that we implemented a correlation map that helps us eliminate an irrelevant feature to increase accuracy. We then finally found that the difference between them was mild. We found that LDA was sligthly more accurate generally in our experience.

2 Datasets

Let us note that were given two sets of data. One set represents white wine quality and the other represents Kidney disease. They are both given in the form of CSV files. They also are completely independent. They are formed of columns of numbers and that is it. We were also given the titles of the columns separately (what each column means) so that we can manipulate our data the correct way.

2.1 Information:

2.1.1 White Wine Quality:

The dataset for white wine quality consists of 1599 samples. The goal we need to achieve is classifying a sample into low-quality (Class 0) and good-quality (Class 1). Each sample is represented in the data set by 10 features that affect wine quality. These features are Alcohol, Malic acid, Ash, Alkalinity of

Submitted to 34th Conference on Neural Information Processing Systems (NeurIPS 2020). Do not distribute.

ash, Magnesium, Total phenols, Flavanoids, Nonflavanoid phenols, Proanthocyanins, and Hue. The last column shows the class label for each data point.

2.1.2 Kidney Disease:

The dataset contains 330 observations. The goal we need to achieve is classifying a sample into whether blood is in the kidney (Class 1) or not (Class 0). Each sample is represented in the data set by 9 features which are Pregnancies: Number of times pregnant, Glucose: Plasma glucose concentration a 2 hours in an oral glucose tolerance test, BloodPressure: Diastolic blood pressure (mm Hg), Heart Rate, SkinThickness: Triceps skin fold thickness (mm), Insulin: 2-Hour serum insulin (mu U/ml), BMI: Body mass index, DiabetesPedigreeFunction: Diabetes pedigree function, Age: Age (years). The last column shows the class label for each data point.

2.2 Statistical analysis:

2.2.1 Classes distribution:

For White wine distribution, the class distribution was quite inequal. This can be observed in Figure1, where there is around 860 rows counted in class 1 as well as around 740 classed in class 0.

On the other hand, if we look at Kidney disease, we can see in figure 2 that they are more equal than in white wine distribution. Both classes have around 160 elements in it as seen in Figure 2.

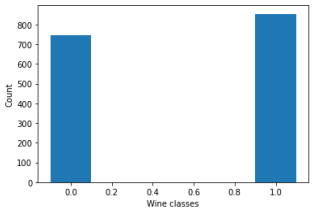

Figure 1: Wine classes

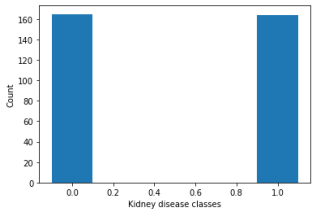

Figure 2: Kidney classes

2.2.2 Distribution of features:

For wine quality, we can see from Figure 3 how the features are distributed and that have a normal distribution.

Similarly for Kidney disease, here is a distribution of the features in Figure 4:

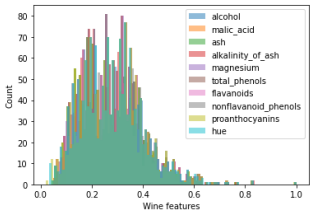

Figure 3: Wine features

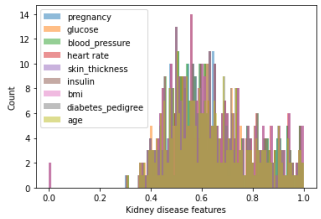

Figure 4: Kidney features

2.2.3 Average of data:

For white wine quality we can observe the average of data for white wine quality from Figure 5.

On the other hand, we can observe the average of data for kidney disease from Figure 6.

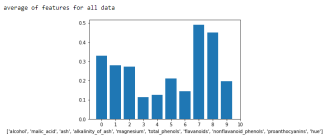

Figure 5: Wine averages

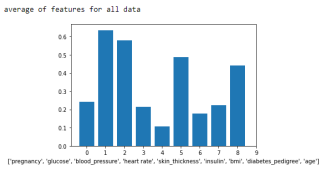

Figure 6: Kidney averages

2.2.4 Standard deviation:

Here, we can observe from this chart the standard deviation of the data for white wine quality (Figure 7).

On the other hand, we can observe the same data for kidney disease (figure 8):

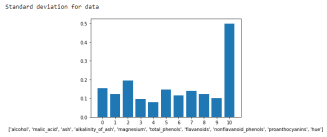

Figure 7: Wine standard deviation

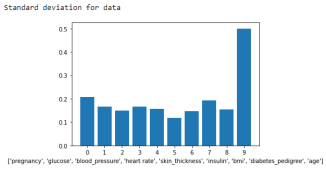

Figure 8: Kidney standard deviation

3 Results

3.1 White wine quality:

3.1.1 Accuracy using learning rates:

The idea we need to discuss in this subsection is that a learning rate yields to a certain accuracy. We used Gradient Descent for that. This happens when we run 10-fold cross validation which loops on while changing the learning rate. In one iteration, the cross validation will call the fit and predict functions, will return a predicted output, and with that output call the function accu_eval which will determine the accuracy of the output given to it. We keep track of the accuracy level outputted at each learning rate and then give the result in form of a chart (figure 9). In this plot we found that the highest accuracy is achieved at a learning rate of 0.001 with a percentage of 72.7%.

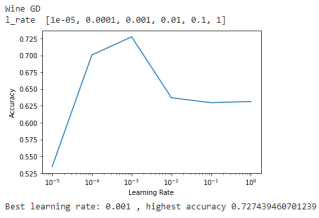

Figure 9: Wine, Learning rate & Accuracy

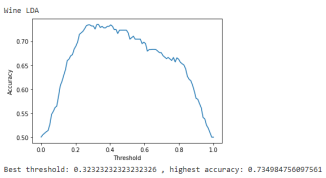

Figure 10: Wine, Threshold & Accuracy

3.1.2 Correlation map and feature removal:

Looking at Figure 11, we observe that features such as magnesium, malic acid for wine are the least correlated features to the result class. We decided to try removing each of the features of the datasets, and analyze the change in the accuracy of the model. As we can see in Figure 12, after removing such features, we found out that for wine dataset, the best improvement in accuracy of the model would be achieved by removing the fifth feature (magnesium).

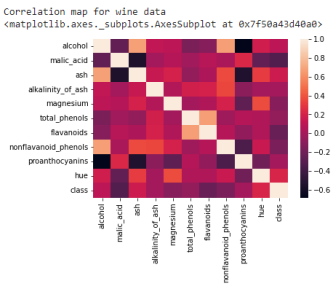

Figure 11: Wine correlation map

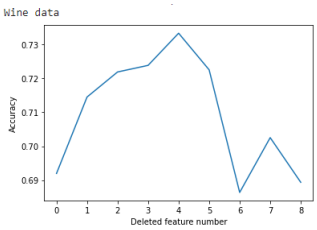

Figure 12: Wine Feature deletion chart

3.1.3 Contrast with LDA:

To have something to compare with and to show a variety of techniques, we also implemented a model for LDA and from it we applied the same idea (finding the accuracy) with LDA to be able to find the best accuracy with different thresholds. We plotted that, as seen in Figure 10, from it we find that the accuracy of LDA turned out to be 73.5%. To compare between the two, we plotted Figure 13 that shows that the accuracy of the test data with GD and LDA is the same (76.76%)!

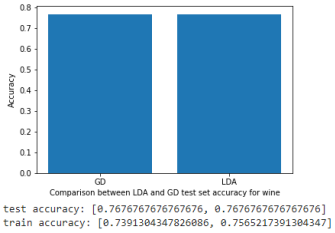

Figure 13: Wine GD vs LDA test data accuracy

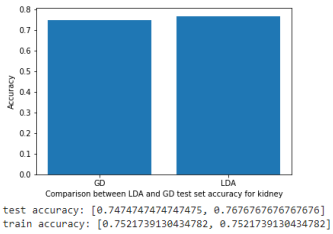

Figure 14: Kidney GD vs LDA test data accuracy

3.2 Kidney disease:

3.2.1 Accuracy using learning rates:

Just like we did for white wine quality data, we plotted the results in terms of accuracy as well for the kidney disease while following the same method in the 10-fold cross validation, the graph is in Figure 17 where we can see that the highest accuracy is achieved with a learning rate of 0.001.

3.2.2 Correlation map and feature removal:

Looking at figure 15, we observe that features such as age, heart rate for kidney are the least correlated features to the result class. We found out, after looking at Figure 16, that for the kidney disease dataset, removing the fourth feature (heart rate) would make the accuracy of the model as high as it can get (in the context of deleting a feature).

3.2.3 Contrast with LDA:

Similarly as we did for white wine, we implemented the LDA model and found its accuracy as seen in Figure 18. We then had to compare both GD and LDA by using the graph in Figure 14. The figure shows that the accuracy of LDA with a percentage of 76.8% is higher than the accuracy of GD with percentage of 74.7% on the test data.

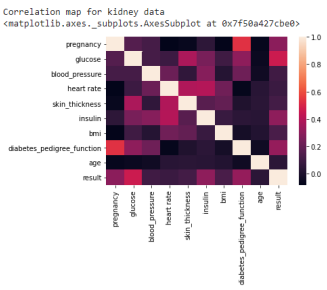

Figure 15: Kidney correlation map

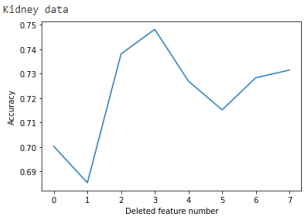

Figure 16: Kidney Feature deletion chart

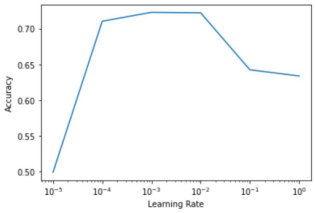

Figure 17: Kidney, Learning rate & Accuracy

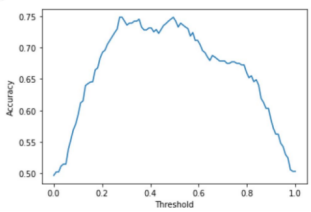

Figure 18: Kidney, Threshold & Accuracy

4 Discussion and Conclusion:

During this very interesting project, we worked on clarifying the data and its distribution in several forms including distribution of classes, of features, averages and standard deviation for both data sets. More importantly, we worked on building a logistic regression binary classifier. We built 2 models: one for Gradient Descent (GD) and one for Linear Discriminant Analysis (LDA). From there we implemented a 10-fold cross validation and found the accuracies of both models using different learning rates (for GD) and thresholds (for LDA) to determine which works best. Each model was formed by 2 functions. The first one is fit that takes the training data X and its corresponding labels vector y as well as other hyperparameters and executes the model training through modifying the model parameters. The second is predict which takes a set of test data as input and outputs predicted labels for the input points. In addition to that, we implemented a correlation map that helped us identify how some of the features are more irrelevant than others and that when we remove those irrelevant features the accuracy increases, sometimes by a lot.

5 Statement of contribution:

We would like to mention that both members of the group worked equally. Both members contributed to the code and report. Aly worked on the Gradient descent functions (fit and predict) and Robert worked on the LDA functions (fit and predict too). Both contributed to the report in a way that each member explained their work in the report. Both worked on the dataset code and report part. The abstract, introduction and conclusion were also split between the two members of the group.