Logistic Regression and Linear Discriminant Analysis modelling on Binary Classification for Wine Quality and Breast Cancer

Yang Gao[1]**, Mingzhe Zhang**[1]**, and Yifei Zhao**[1]

[7]Department of Electrical Engineering, McGill University, Montreal, Canada

Logistic Regression (LR) and Linear Discriminant Analysis (LDA) are frequently used in data classification. In this project, we implemented and optimized the two algorithms and compared their performance on two UCI Dataset: *Wine Quality* and *Breast Cancer*. Model optimization was achieved through the implementation of feature selection, Grid Search, Random Search and L1 regulation. We showed that: i) Before feature engineering, LDA achieved higher prediction accuracy and shorter training time on both datasets with prediction accuracy at 74.90% for *Wine Quality* and 96.70% for *Breast Cancer Prediction*, comparing to LR with the accuracy at 56.70% for *Wine Quality* prediction and 84.70% for *Breast Cancer Prediction*. ii) Implementing data processing during LR modelling improved the prediction accuracy of wine quality from 56.70% to 74.9% iii)In the same experiment, large learning rate of LR model may contribute to low accuracy.

Logistic Regression | Linear Discriminant Analysis | UCI Dataset| Wine Quality | Breast Cancer | K Fold Validation | Grid Search | Random Search

1 Introduction.

Linear Discriminant Analysis (LDA) and Logistic regression (LR) are approaches adopted in data categorization with linear model. Both techniques obtain the linear decision boundary between classes, however based on different assumptions on the variable distribution. Linear discriminant Analysis assumes normal distribution variable with equal covariance matrix contrary to no assumption of the data in LR. LR deploys the method of directly calculating logistic function to estimate the probability of a sample belonging to a class. Compared with directly modelling, LDA applies Bayes rule to approximate the probability of a class possess certain features.
To further study LDA, LR properties and analyze the variants affecting performance, we implemented these two approaches with Python on the real-world data sets,wine and Breast cancer diagnostic data to achieve binary classification,evaluating the developed models with k-fold cross-validation and testing on the separated test set. The goal is to classify wine based on the quality and define a tumour is malignant or benign based on multiple properties. To build the LR model, gradient descent minimizing the negative log loss function is implemented. Before modelling, we performed data cleaning to filter out the samples with missing feature values and invalid inputs, which is a critical process for the predication accuracy of the trained model. Besides, Feature selection and engineering are also introduced to reduce unnecessary information processed to enhance the efficiency and the accuracy of the model. We proved that hyperparameters such as the learning rate and the iteration time in the LR model also have impacts on the performance of the developed model by griding search and random search.
In summary, we came up with a method which improves LR model accuracy from 0.56 to 0.78 maximum and 0.74.9 on average by feature selection and data processing. By using griding search and random search, we optimize the hyperparameters of LR model and learned the negative correlation between the learning rate and the predication accuracy. Compared to LDA model, the performance of LR model is unstable in terms of the fluctuation of the prediction accuracy and longer running time.

2 Dataset.

Missing values and invalid input were spotted in the sample features, which we chose to drop in the data preprocessing procedure, such as *Wine Quality* data sample with zero feature value and *Breast Cancer* data sample with '*?*' inputs are not being included in the training dataset. To interpret feature values and associate them to the classification categories, the correlation between each feature and each class has been analyzed.

2.1 Red Wine Quality. Figure 1 shows the correlation between input features and output feature. It is noticeable that '*alcohol*', '*sulphates*' and '*citric acid*' have strong positive correlation (greater than 0.2) with output variable '*quality*', and '*volatile acidity*', '*total sulfur dioxide*' have strong negative correlation (less than - 0.2) with '*quality*'. Additionally, we can see that some input features have strong correlation with other input features which means these features can 'represent' other features to some extent. So the correlation between features provides a very good reference for us to select features.
In this work, we used four features and data to predict wine quality, and made comparison with the results when all features were used in prediction. And to improve accuracy, data processing was used. Since four features ('*alcohol*', '*sulphates*', '*volatile acidity*', '*fixed acidity*') have the greatest correlation value to the quality feature, we only used them to construct the input features matrices. Based on the algorithm employed by python to compute correlation and loss function which involves the terms of centralized mean and variance. By using this property, we process the feature data by cen-

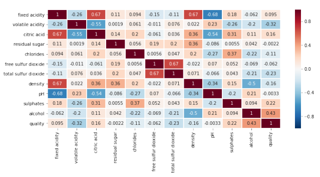


Fig. 1. Correlation between features in *Wine Quality*

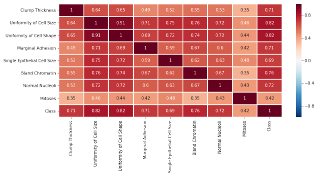


Fig. 2. Correlation between features in *Breast Cancer*

tralizing the data distribution through subtracting mean values and dividing the squared correlation coefficient to allow the gradient descent converging faster than the one without squared term, fixed learning rate and iteration times.

2.2 Breast Cancer Diagnosis. Figure 2 shows the correlation between input variables and output features. It is clear that all input variables are important for prediction, since all of them have strong correlation with output variables, so we used all features in this dataset.//

2.3 Ethical Consideration. The ethical issues surrounding machine learning is not simply about the algorithms themselves but more about the way the data is used. As algorithms expand their ability to organize society, politics, institutions, and behavior, sociologists have become concerned with the ways in which unanticipated output and manipulation of data can impact the physical world.

In the case of *Red Wine Quality Prediction*, we only use several wine attributes to predict its quality with high accuracy, it seems good enough to judge the quality of a good red wine only by estimate these attributes. However, in a real case, wine tasters take too many factors into consideration, and distilleries can never make good wines if they only attach importance on several attributes. And in the case of Breast Cancer Diagnosis, machine learning algorithms bring insights to diagnose cancers, but it can never replace doctors in the same way.

Additionally, algorithmic bias raised many social issues in recent years, and privacy violations is heated topic among general public, sociologists, government, celebrities and so on.

There's still much more about ethical concerns regarding machine learning, such as job automation, AGI alignment, AI rights and so on, what we need to do is to find appropriate legislation for AI in these fields, and be responsible for algorithms.

3 Results.

3.1 Task I. Three main tasks with LR and LDA models were performed to study Wine and Breast cancer data set, and showed that the LDA achieved higher prediction accuracy on both datasets. The performance of LDA model was estimated by investigating the correlation between learning rate and specificity, sensitivity, running time, and accuracy in *Wine Quality* dataset. 70 learning rates ranging from 10^{-5} to 100 were generated and Grid Search method was used on shuffled data set with 2000 iterations during the process.

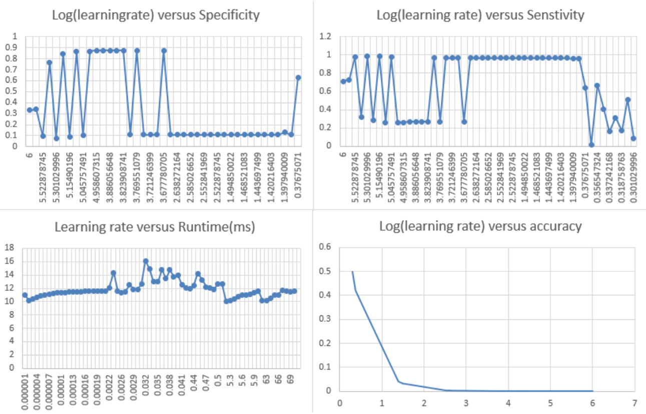


Fig. 3. The influence of learning rate on LDA performance

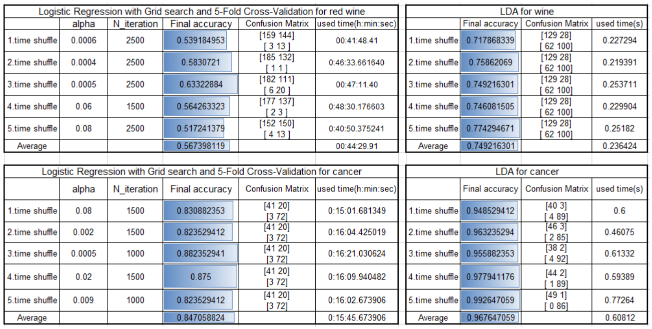
Figure 3 shows that high and stable specificity can be achieved when log learning rate ranging from 3.8 to 5, but low and stable when the learning rate is relatively high. While the situation is just the opposite for the correlation between log learning rate and sensitivity. The learning rate has little influence on running time which is closely related to number of iterations.

The Figure 3 also demonstrates that the prediction accuracy decreases with learning, showing negative exponential correlation.

3.2 Task II.. The main difference between LR and LDA is that LDA has no hyperparameters to be tuned. Therefore only one LDA model could be generated from a single training dataset. The results from the training phase would be used to do the prediction, so 5 Fold Cross Validation is not appropriate for LDA. On the other hand, LR has two different hyperparameters during the Gradient Descent Process:

- α_k, so called Learning Rate which represents the step size during every iteration and follows the direction to reach local minimum.
- N, the iteration times which represents how many times going down following the descent direction.

In our experiment, we use Grid Search to implement 5-Fold Cross Validation. From the first task we knew when learning rate α is located in range from 0.5 to 1 and the value of N is around 2000, we could possibly get the highest accuracy. Therefore we set in total 27 α and 3 N_k in those ranges. Two sets are created, one of which contains the values of α from 10^{-4} to 0.1 by 1 dB difference, and another set contains different values for iteration times N_k from 1000 to 2500 with difference of 500. In total 81 different pairs are tuned with 5 Fold Cross Validation, so the LR model will run 405 times in tuning process and 1 time for final test phase. After each experiment ends, we shuffled the whole dataset to start the new experiment. For each experiment, different training and testing datasets were used. From LR for wine quality in the Table 4 we can see that different dataset with same shape has huge influence on final accuracy. For example, the accuracy of the 5^{th} experiment is 12% lower than 3^{rd} experiment in LR for wine dataset.



Logistic Regression with Grid search and 5-Fold Cross-Validation for red wine					
	alpha	N_iteration	Final accuracy	Confusion Matrix	used time(h:min:sec)
1.time shuffle	0.0006	2500	0.539184953	[159 144] [3 13]	00:41:48.41
2.time shuffle	0.0004	2500	0.5830721	[185 132] [1 1]	0:46:33.661640
3.time shuffle	0.0005	2500	0.63322884	[182 111] [6 20]	00:47:11.40
4.time shuffle	0.06	1500	0.564263323	[177 137] [2 3]	0:48:30.176603
5.time shuffle	0.08	2500	0.517241379	[152 150] [4 13]	0:40:50.375241
Average			0.567398119		00:44:29.91

LDA for wine			
	Final accuracy	Confusion Matrix	used time(s)
1.time shuffle	0.717868339	[129 28] [62 100]	0.227294
2.time shuffle	0.75862069	[129 28] [62 100]	0.219391
3.time shuffle	0.749216301	[129 28] [62 100]	0.253711
4.time shuffle	0.746081505	[129 28] [62 100]	0.229904
5.time shuffle	0.774294671	[129 28] [62 100]	0.25182
Average	0.749216301		0.236424

Logistic Regression with Grid search and 5-Fold Cross-Validation for cancer					
	alpha	N_iteration	Final accuracy	Confusion Matrix	used time(h:min:sec)
1.time shuffle	0.08	1500	0.830882353	[41 20] [3 72]	0:15:01.681349
2.time shuffle	0.002	1500	0.823529412	[41 20] [3 72]	0:16:04.425019
3.time shuffle	0.0005	1000	0.882352941	[41 20] [3 72]	0:16:21.030624
4.time shuffle	0.02	1500	0.875	[41 20] [3 72]	0:16:09.940482
5.time shuffle	0.009	1000	0.823529412	[41 20] [3 72]	0:16:02.673906
Average			0.847058824		0:15:45.673906

LDA for cancer			
	Final accuracy	Confusion Matrix	used time(s)
1.time shuffle	0.948529412	[40 3] [4 89]	0.6
2.time shuffle	0.963235294	[46 3] [2 85]	0.46075
3.time shuffle	0.955882353	[38 2] [4 92]	0.61332
4.time shuffle	0.977941176	[44 2] [1 89]	0.59389
5.time shuffle	0.992647059	[49 1] [0 86]	0.77264
Average	0.967647059		0.60812

Fig. 4. Results of LR and LDA for both data sets a) 81 pairs for Grid Search **b)** For Grid Search, before every test begins, the training data sets and testing data set will be randomly shuffled. **c)** Elements for Confusion Matrix: [tp, fp] [fn, tn]

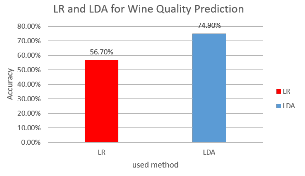


Fig. 5. Results of LR and LDA for Wine Quality

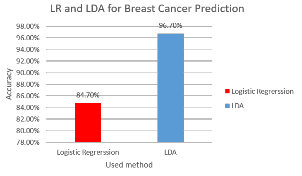


Fig. 6. Results of LR and LDA for Breast Cancer

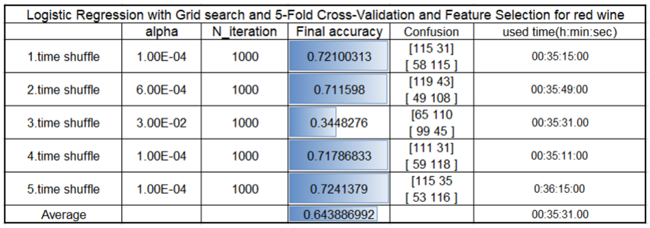
We also developed a function to calculate Confusion Matrix to show the trend of prediction. LR tends to predict positive values for all samples from wine dataset, and for samples from cancer dataset LR predicted back to balance. The total number of predicted positive values are similar with predicted negative values. LDA on wine dataset tended to do 50 percent predictions, and on cancer it can make the right prediction most of the time.

The process costs about 44 minutes in average. For LR with wine data, α varies from 6×10^{-4} to 6×10^{-2}. Iteration times, N, is more stable than learning rate. The best iteration times for wine data set is 2500 and for cancer dataset is 1500. Because of 5-Fold Cross Validation, the time complexity of LR is $O(5MN^2)$. And the time complexity of LDA is only $O(M^3N)$.

From the Bar Chart 5 and Bar Chart 6 we can easily see the comparison of accuracy for two data sets with different methods.

In summary, LDA has huge edge over LR in accuracy and also in running time with both cancer and wine quality prediction. The accuracy of LDA is about 10% higher than LR and thousands times faster. The performance of both hypothesis are highly related to the training data set and the testing data set.

3.3 Task III. After we analyzed the statistical attributes of each wine features , such as mean, maximum and minimum, standard derivation and correlations as shown in Figure 1, we selected 'alcohol', 'volatile acidity', 'citric acid', 'sulphates' and divided these feature values by its squared correlation coefficient to construct the input feature vector of the LR model.This algorithm is evaluated by 5 fold cross validation and tested on the separate data sets with random search and grid search respectively. The results are shown in Table 7. With random search executed, the average accuracy of prediction is varied between 0.73 to 0.75 which is 20 percent higher than the accuracy 0f 0.56 tested by the LR model with the grid search method on the unmodified data sets and slightly higher the accuracy of grid search from 0.71 to 0.72.



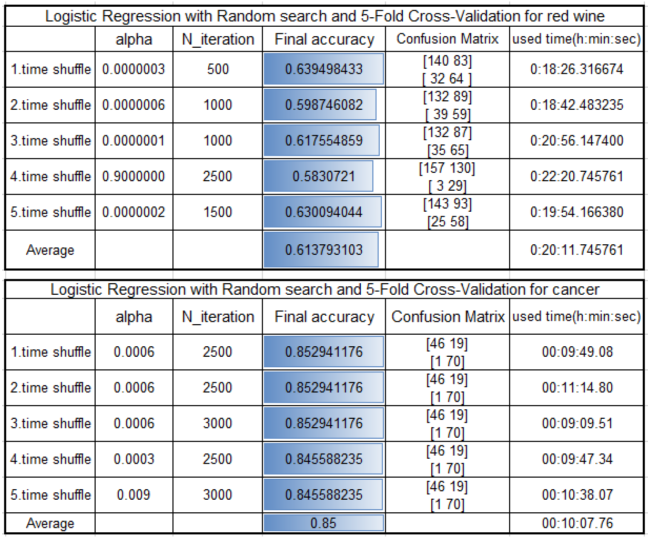
Logistic Regression with Grid search and 5-Fold Cross-Validation and Feature Selection for red wine					
	alpha	N_iteration	Final accuracy	Confusion	used time(h:min:sec)
1.time shuffle	1.00E-04	1000	0.72100313	[115 31] [58 115]	00:35:15:00
2.time shuffle	6.00E-04	1000	0.711598	[119 43] [49 108]	00:35:49:00
3.time shuffle	3.00E-02	1000	0.3448276	[65 110 [99 45]	00:35:31.00
4.time shuffle	1.00E-04	1000	0.71786833	[111 31] [59 118]	00:35:11:00
5.time shuffle	1.00E-04	1000	0.7241379	[115 35 [53 116]	0:36:15:00
Average			0.643886992		00:35:31.00

Fig. 7. Results of LR with Grid search on selected features[1,2,9,10] a) 81 pairs for Grid Search **b)** For Grid Search, before every test begins, the training data sets and testing data set will be randomly shuffled. **c)** Elements for Confusion Matrix: [tp,fp] [fn,tn]

4 Discussion and Conclusion.

4.1 Discussion. This project investigated LR and LDA modelling on binary classification of two classic data sets. Compared with LDA model, LR behaved unstable and predicted with lower accuracy. It reflects the mathematical characteristics of the LR and LDA model. To make gradient descent converges and achieve the maximum likelihood, the LR model in some cases, for example, the label is binary and some input variables have large deviation from average, has to give the weight vector a large value to prevent from overflow. Centralizing data is an option to reduce the possibility of overflow, extreme low sensitivity or extreme low specificity. Additionally, extra unnecessary data in the training data set add computation cost and lower the accuracy, such as LR has lower accuracy and longer computation time on the unprocessed data, therefore feature selection and data preprocessing are valuable. In our project, we processed the data based on the statistical attributes. For future research reference, more advance techniques such as PCA can be used to find the independent features. In order to improve accuracy and reduce the running time by mean of tuning parameters, we developed two approaches to optimize LR as listed below.

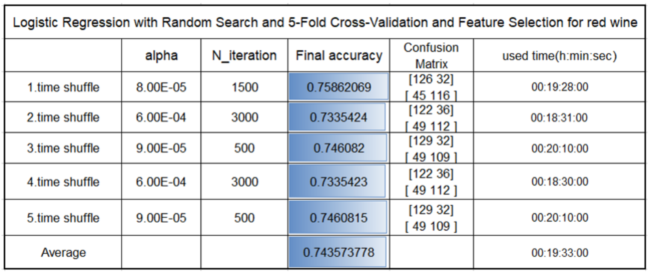
4.1.1 Random Search We noticed that with the use of LR from Scikit-Learn library, the accuracy could reach to 70%, but in task 2 the average accuracy is only about 54%. We found that the highest accuracy with grid search was 63%. That means the used pairs of hyperparameter which is chosen by the Grid Search is not good enough. We hypothesis the reason is that sice the set of hyperparameter pairs were so small, one of the best hyperparameter pairs is not included in our subset. Apparently we can easily extend the sets of hyperparameter to hundred or thousand level, however it means the running time would also be compromised. 81 hyperparameter pairs with 5 fold cross validation costs already 44 minutes in average and for thousand pairs, it may need ten hours in the tuning process. In order to overcome this problem, we developed Random Search instead of Grid Search. We chose 63 values for α from 10^{-7} to 1 by step 1 dB and 6 N from $[500, 3000]$ by step 500. Then we randomly chose 40 pairs of



Logistic Regression with Random search and 5-Fold Cross-Validation for red wine					
	alpha	N_iteration	Final accuracy	Confusion Matrix	used time(h:min:sec)
1.time shuffle	0.0000003	500	0.639498433	[140 83] [32 64]	0:18:26.316674
2.time shuffle	0.0000006	1000	0.598746082	[132 89] [39 59]	0:18:42.483235
3.time shuffle	0.0000001	1000	0.617554859	[132 87] [35 65]	0:20:56.147400
4.time shuffle	0.9000000	2500	0.5830721	[157 130] [3 29]	0:22:20.745761
5.time shuffle	0.0000002	1500	0.630094044	[143 93] [25 58]	0:19:54.166380
Average			0.613793103		0:20:11.745761

Logistic Regression with Random search and 5-Fold Cross-Validation for cancer					
	alpha	N_iteration	Final accuracy	Confusion Matrix	used time(h:min:sec)
1.time shuffle	0.0006	2500	0.852941176	[46 19] [1 70]	00:09:49.08
2.time shuffle	0.0006	2500	0.852941176	[46 19] [1 70]	00:11:14.80
3.time shuffle	0.0006	3000	0.852941176	[46 19] [1 70]	00:09:09.51
4.time shuffle	0.0003	2500	0.845588235	[46 19] [1 70]	00:09:47.34
5.time shuffle	0.009	3000	0.845588235	[46 19] [1 70]	00:10:38.07
Average			0.85		00:10:07.76

Fig. 8. Results of LR with Random Search without feature selection a) The experiments which generated of this table used all features from wine dataset **b)** The experiment of cancer classification used all features without selection



Logistic Regression with Random Search and 5-Fold Cross-Validation and Feature Selection for red wine					
	alpha	N_iteration	Final accuracy	Confusion Matrix	used time(h:min:sec)
1.time shuffle	8.00E-05	1500	0.75862069	[126 32] [45 116]	00:19:28:00
2.time shuffle	6.00E-04	3000	0.7335424	[122 36] [49 112]	00:18:31:00
3.time shuffle	9.00E-05	500	0.746082	[129 32] [49 109]	00:20:10:00
4.time shuffle	6.00E-04	3000	0.7335423	[122 36] [49 112]	00:18:30:00
5.time shuffle	9.00E-05	500	0.7460815	[129 32] [49 109]	00:20:10:00
Average			0.743573778		00:19:33:00

Fig. 9. Results of LR with Random Search with feature selection for Wine Dataset a) This table's experiments used only four features [1,2,9,10]

hyperparameter from 63×5 pairs. Random search can test more possible combinations of hyperparameter than Grid Search, and we can control the number of hyperparameter pairs easily and explicitly.In Figure 8 we can clearly see the improvement after training with Random Search. With 40 pairs, the accuracy of LR for wine with all features and random search is improved from 56.7% to 64.3%. For tumor, the accuracy stays in the same level, but the variance of accuracy from each experiment is reduced. With the subset of feature which is found in Task III as shown in Table 9, the accuracy could be improved to 74.35%. As reference, the accuracy of LR from Sci-kit Learn with same feature selection as input data is about 72% in average. The variance of the accuracy for each fold validation are also reduced. That means we have a more stable and reliable way to select hyperparameter pairs. Because we test with only 40 pairs, which are half of the pair number used in the Grid search, the corresponding running time is cut down by half. With Random Search we reduce the running time and also improve accuracy.

4.1.2 Lasso regularization To improve the performance of LR, we decided to add a penalty attribute in gradient descent function. As discussed before, several features could be deleted, so we choose L1-Regularization. We tested λ from 0.00 to 0.99 and from 1 to 150. From Figure 10 we know that λ works well in range of [7, 35]. Then we add a $lambda$ in range [0, 1,..., 50] as hyperparameter with $\alpha = 0.01, N = 500$, the accuracy could be improved by 4%. In the future, we could also implement a new LR model with 3 hyperparameters.

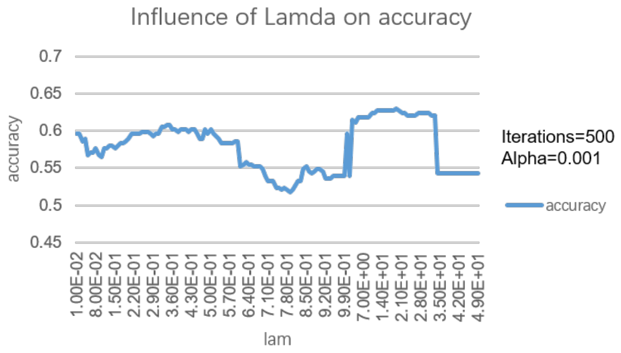


Fig. 10. Course with different Lambda in range[0,49] a) The values of Lambda have two parts. The first is [0, 0.01, 0.02,..., 0.99]. The second part is [1,2,3,...,49]

4.2 Conclusion. Through this project, we compared logistic regression algorithm and linear discriminant analysis algorithm by analysing the difference of prediction accuracy and showed that the LDA achieved much higher accuracy and requires less running time on two datasets and the LR method is more likely influenced by the statistic attributes of dataset, while the performance of LDA is more stable.

Then we applied some methods to improve the accuracy on the prediction of wine quality when applying LR. The test results shows that the choice of input features is crucial to achieve a meaningful model with high accuracy. Through the analysis of correlation between input variables and output variable, we chose 4 features and three of them are divided by its squared correlation coefficient. With random search applied, the accuracy improved by almost 20 percent from 0.56 to 0.75, and the implementation of gird search method improved the accuracy from 0.56 to 0.72. While the implementation of Lasso-Regularization method can improve the accuracy by 0.03 with λ ranging from 7 to 35.

5.Statement of contributions.

In this project we distributed tasks in several parts and met all the requirements listed on the instruction of Miniproject I.

Yang Gao: Developed algorithm for Feature selection . Collected and analyzed data of experiments results (learning rate). Coded LR model for only testing feature selection model.

Mingzhe Zhang: Developed code for LR, LDA for 3 tasks. Developed Random Search and Lasso Regularization to select and optimize models' performance. Analyzed and evaluated hyperparameters of implemented models.

Yifei Zhao: Analyzed feature distribution, feature selection, and tested evaluated test results. Participated in the development of another version of LDA.

Report is contributed by each of group member