Investigation for the performance of Logistic Regression and Linear Discriminative Analysis on wine quality prediction and breast cancer diagnosis

Zhenghua Chen, Hehuimin Cheng, Xiaobin Shang

Abstract

In this project, we investigated two classification models, logistic regression with gradient descent (LR) and linear discriminative analysis (LDA), and how they predict the binary outcome on two specific datasets. We applied the heatmap correlation analysis on the different datasets and create a new subset of features, which slightly increased 2% of accuracy in the final test. We also showed that for both datasets with different feature data types, LDA model predicts more precisely and efficiently on both datasets, while LR model predicts the categorical dataset well when choosing a large learning rate. Meanwhile, we compared the different running time of the LDA model and LR model. We also analysed the prediction result in depth by interpreting confusion matrix and discussed findings which would be studied in future research.

1 Introduction

The aim of our project is to evaluate which of the two classification models performs better on the predictions of the binary dependent outcomes of the two different datasets by comparing and analysing their accuracies.

Logistic Regression model is one of the commonly used generative models, which is to build a non-linear association in a linear way using a logarithm transformation. It is efficient to train, easy to implement, but it requires significant amount of data compared to other linear models to achieve stable, meaningful results.

Linear Discriminative Analysis is a generative learning classification which predicts based on Bayes' rule and the assumption that both classes have normal distributions and same covariance matrix. Therefore, the model gives more flexibility for computing but also has more requirements and is more complicated to implement. However, with all its requirements met, it often classifies better than Logistic Regression. Our task is to investigate how these two models perform on two different sets of data.

We took our datasets from two collected data files: one is to predict the quality of wine based on its chemical properties, the other is to predict whether a tumour is malignant or benign based on various properties. By building effective predictive models for these datasets, we can tell whether a bottle of wine is of good favor without tasting it and anticipate the risk of breast cancer.

The overall task is divided into four parts: firstly, we preprocessed the datasets; secondly, we implemented the two models; thirdly, we set up the k-cross validation to train the models and validate which model better suits each dataset; lastly, we analysed the result and tried to improve the prediction accuracy by dropping the weak features of wine set and using regularization on gradient descents.

After analyzing the accuracies of the predictions, we found that mostly for wine dataset, LDA performs much better than LR under the same condition. For cancer dataset, LR with learning rate of 0.001 performs nearly as good as LDA. By comparing the prediction accuracy under different learning rates in LR, we learned that for wine data LR gives a higher accuracy when using the smallest learning rate we chose, 1e-06, for cancer data it’s the opposite: the biggest learning rate, 0.001, works the best. On the aspects of accuracy and runtime, LDA performs better than LR. Also, comparing to train with the original features, selecting relevant features to study improved the outcomes.

2 Datasets

We were given two datasets: one is a red-wine quality set contains 11 features and 1599 samples from the north of Portugal, collected by professor Paulo Cortez in University of Minho. Another data set is the breast cancer dataset created by *Dr. William H. Wolberg, W. Nick Street* and *Olvi L. Mangasarian*, all from University of Wisconsin. It contains 11 features and 699 samples. To conduct a binary classification, we transformed the quality ratings of wine from 0-10 to 0/1 by defining the ratings less than 6 as negative (class 0) and others as positive (class1) and we adjusted the “Class” column of breast-cancer-dataset which contain 2 as benign tumor and 4 as malignant tumor into 0 (benign class) or 1 (malignant class). Besides to the mapping result to binary outcome, we also found that the breast-cancer dataset has missing features, unimportant data, and some wrong types of data. By removing malformed data, unimportant features, such as ID number, and converting the char elements to integers, we collected 638 valid data and 10 features from the cancer set. We conducted some statistics on the data and conclude that generally the features satisfy the assumptions of models: they are relatively independently identically distributed.

Then we conducted the heatmap analysis on the correlation between different features and actual classes. For the wine quality dataset, we filtered out the features, which are not closely related to the outcome.

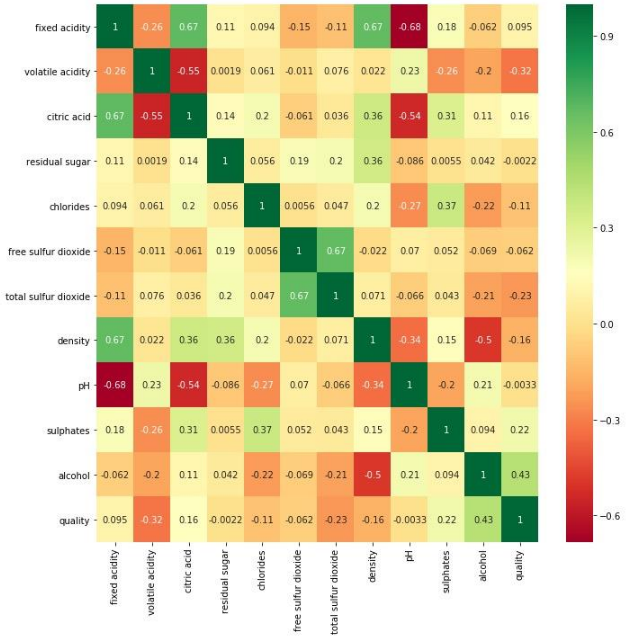

Figure 1 Heatmap for Corelation between attributes in Wine Set

From the figure of correlation between features, we ordered the features based on the absolute value of their correlations with the wine quality, from smallest to largest. The five smallest correlation features are residual sugar, pH, free sulfur dioxide, fixed acidity and chlorides. Utilizing this information, we created four combinations of features: first combination (F1) includes all the features except residual sugar and pH; second (F2) includes all except residual sugar, pH, and free sulfur dioxide; third (F3) includes all except residual sugar, pH, free sulfur dioxide, and fixed acidity; and fourth (F4) includes all except residual sugar, pH, free sulfur dioxide, fixed acidity, and chlorides.

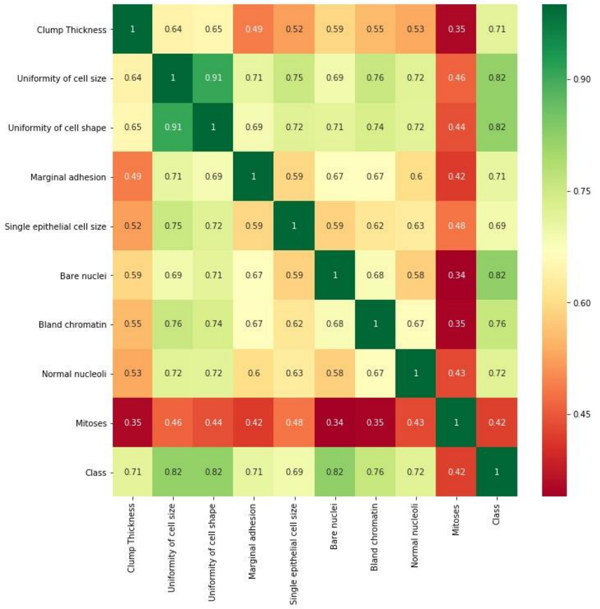

Figure 2 Heatmap for Corelation between attributes in Cancer Set

For the cancer set, we also checked the dependence between the variables and the outcomes. Each feature has a relatively strong relationship with the outcome class, which may be a good herald for later training.

After feature selection, we separated data as training set and testing set with a ratio, then implemented the model and run 5-fold-cross-validation on the training set. We tried 2 different ratios and compared the accuracies. At the beginning of each experiment we shuffled the data to ensure the randomness.

Note that in breast cancer dataset, there's one column named "ID number" which records ID information of study objects. This type of data is completely irrelative to our research on breast cancer and may leak study objects' personal information unintentionally without their permissions. Thus, for ethical concern, in this situation we remove this feature from the dataset at the beginning of feature processing.

3 Results

After running the experiment under different conditions, we filled the following charts:

			LR	
			Wine	Cancer
Accuracy (time)	R1	L1	64.03 (4.8160)	35.12 (2.2120)
		L2	57.50 (4.8135)	86.50 (2.2303)
		L3	58.47 (4.8202)	91.54 (2.2398)
		L4	58.61 (4.8060)	90.73 (2.2251)
	R2	L1	64.38 (4.3547)	35.48 (2.0031)
		L2	57.19 (4.3021)	85.20 (1.9935)
		L3	57.34 (4.2859)	91.23 (2.0111)
		L4	56.88 (4.2521)	94.15 (2.0026)

Table 1 Accuracy and time of LR model running 5-fold-cross-validation under ratio R and learning rate L on two datasets: L1=1e-06; L2=1e-05; L3=1e-04; L4=0.001; R1=1:9; R2=2:8. Accuracy is of % and time is of seconds

		LDA	
		Wine	Cancer
Accuracy (time)	R1	73.20 (0.1241)	95.98 (0.0507)
	R2	74.69 (0.1190)	95.61 (0.0512)

Table 2Accuracy and time of LDA model running 5-fold-cross-validation under ratio R: R1=1:9; R2=2:8 Accuracy is of % and time is of seconds

Accuracy	F1	65.52
	F2	66.46
	F3	64.26
	F4	62.26

Table 3 Accuracy of testing LR model on combinations of features under R1 and L1 for wine dataset. F1, F2, F3, F4 are combinations of features described in feature selection in dataset part

We found that for logistic regression, different dataset needs different learning rate to achieve a better prediction. For wine dataset, the accuracy reached its best at 64.38% when using L1 with learning rate of 1e-06 under ratio of 2:8, while for cancer dataset the accuracy gradually increased as learning rates increased from L1 to L4 and achieved its best at 94.15% when using L4 with learning rate of 0.001 under the same ratio. Our tentative idea is that continuous numerical features like in wine dataset, need a smaller learning rate than categorical features (which in cancer dataset, are all represented by integers from 0-10) in order to converge to a more precise value.

When comparing the performances of the two models on these datasets, we discovered that for wine dataset, LDA model's accuracy under R2, 74.69%, was higher than LR model's best accuracy under R2, 64.38%, by over 10%. This indicates that LDA should be a better model for predicting a wine's quality based on its numerical chemical properties. However, for cancer dataset, when choosing L4 learning rate under R2, LR performed nearly as good as LDA. To choose a most suitable model, we then look at their running time. In each situation LDA took significantly less time than LR. Therefore, by comprehensive analysis of both accuracy and time cost, LDA is an appropriate model for both datasets.

Furthermore, we improved the performance of LR on wine dataset by dropping irrelevant features. After testing the four combinations of features which we derived in dataset processing, we found that the accuracy improved around 2% at most when we dropped residual sugar, pH, free sulfur dioxide and fixed acidity factors. Besides, we also apply the regularization on the LR for wine dataset, we found that the accuracy improved.

In addition to analysis using accuracy, we also implemented the confusion matrix for visualizing the performance of our models. For an example, we presented the confusion matrix for the logistics regression on the cancer set.

44	0
2	22

Table 4 Confusion matrix for best LDA model on Cancer set

From this matrix, we acquired precision of 1.0, recall of 0.96, and the F1 score of 0.978. We also observed that the model had no type I error but 2 type II error, which is the prediction that the tumor is malign when it is not. Concerning the practicability of model in the reality, we achieved strong evidence that the model is well-constructed.

4 Discussion and Conclusion

In conclusion, we constructed and analyzed the performance of the logistic regression model and linear discriminative analysis model on two datasets. In the experiments, the heatmap analysis was used to detect the correlation between attributes. We noticed that the LR model made better predictions when we removed the attributes that are unrelated to the output variable, as well as attributes that are very similar (correlated) to each other. Also, for different datasets, the learning rates for LR models with the best predictions were different. We explored four different learning rates for the LR model. The most proper learning rate for the wine dataset is L1, and for the cancer dataset is L4.

For linear discriminative analysis model, it made better predictions on the smaller dataset than the larger dataset. But, since we only did the experiment on two datasets, we cannot conclude the tendency, which requires further investigations on more datasets with different size. Besides, the runtime was much less than the LR model on the both datasets since LDA do not need the gradient descent iterations to fit the data.

Both models performed better when the split ratio of the testing set and training set was 2:8 than when it was 1:9. A probable reason is that the testing sets were too small under ratio 1:9, and it is not a representative sample of the whole data. Moreover, both models performed worse on the wine dataset than the cancer dataset. Other than the limited functioning of models, the possible problem with the dataset itself might be that the wine quality is a factor that decided by human experts based on their tastes which may not be objective and precise, and the attributes of wines have relatively weaker correlations with their quality than the correlation between the attributes of cancer cells and their class.

5 Statement of Contributions

Zhenghua Chen works on establishing cross validation and writing report. Hehuimin Cheng contributes on processing data, building up the LR model, writing data section, and the model improvement and analysis. Xiaobin Shang is responsible for building up and analyzing the LDA model and performing the experiments.

References

P. Cortez, A. Cerdeira, F. Almeida, T. Matos and J. Reis. (2009). Modeling wine preferences by data mining from physicochemical properties. In Decision Support Systems, Elsevier.

Dr. William H. Wolberg, W. Nick Street, Olvi L. Mangasarian. (1995). General Surgery Dept. University of Wisconsin, Clinical Sciences Center Madison, Computer Sciences Dept. University of Wisconsin.

Niklas Donges. (2018). CODE University of Applied Sciences, Berlin. https://machinelearning-blog.com/2018/04/23/logistic-regression-101/

Adi Bronshtein. (2017). Train/Test Split and Cross Validation in Python. General Assembly, D.C. https://towardsdatascience.com/train-test-split-and-cross-validation-in-python-80b61beca4b6