Zhaoyang Gu

STAT 617

11/1/2020

Wine Quality Classification Model Outcome with LDA and QDA

Abstract

After applied PCA, both LDA and QDA methods were used on the wine quality dataset. we found out that LDA has a better precision rate on successful classifications. However, the highest precision is lower than 80% which means there is still improvement needed for the model.

Introduction

Wine is one of the most consumed alcohol among all kinds of alcohols in the world. The quality of wine can be affect by many factors such as the color, the taste and the chemical component in each kind of wine. Under this premise, our group wanted to train a classification model that can classified the quality of wine by using its chemical components and chemical properties. The dataset that we were using was downloaded from UCI Machine Learning Repository .(https://archive.ics.uci.edu/ml/datasets/wine+quality) The original folder contains two datasets, one is for red wine and the other is for white wine. Our group decided to use the one for red wine. The datasets have 12 variables and 1599 observations. The variables, which are made up by chemical components ("Citric Acid"," Residual Sugar" etc.), chemical properties ("fixed acidity", "alcohol" etc.) and quality, which is a score given between 0 – 10.

Methods

Our group decided to use Principle Component Analysis (PCA), Linear Discriminant Analysis (LDA) and Quadratic Discriminant Analysis(QDA) to help the classification of the dataset. PCA is very useful to reduce high dimensional data into fewer dimensions while preserving as much variation as possible by projecting current data onto a few principle components. Both QDA and LDA are approximating the Bayes classifier for the boundary to separate different features. LDA provides a linear decision boundary while QDA provides a quadratic decision boundary. Both algorithms hold under normality assumption. Comparing to QDA, LDA also need to satisfy the common variance assumption.

Exploratory Data Analysis

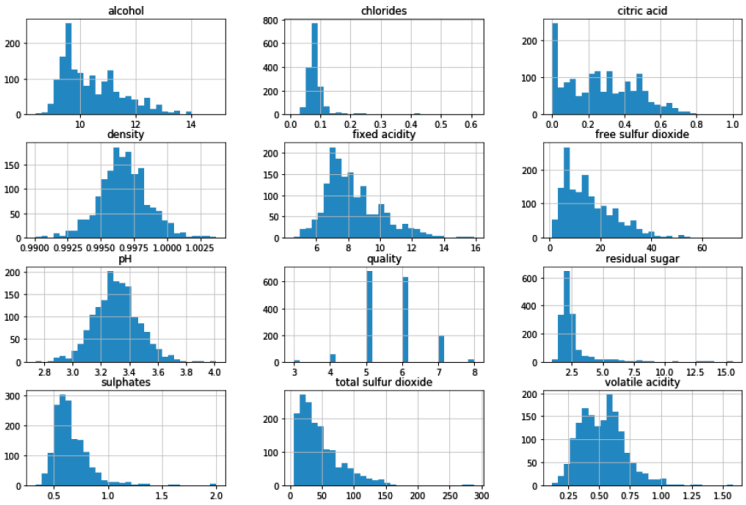

Figure 1. Histograms of each variables of the original dataset

From the first look of the frequency histograms, we can tell that the some of the variables are failed at the normality assumption and are very right skewed.

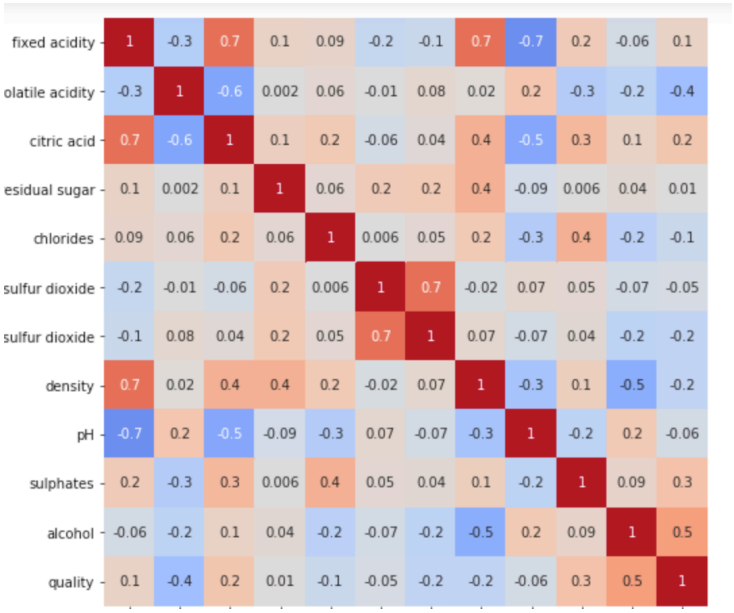

Figure 2. Correlation matrix of the original data

By checking the correlation matrix, we can find there are some variables that are also highly correlated to each other. This also confirms that we need to use PCA to reduce the dimension of our data. According to the accumulation variance data (supplementary **Figure 1.**), we can find there are almost 80% of the variation can be explained by using 5 principle components. Therefore, we reduced the dimensionality of the dataset to 5.

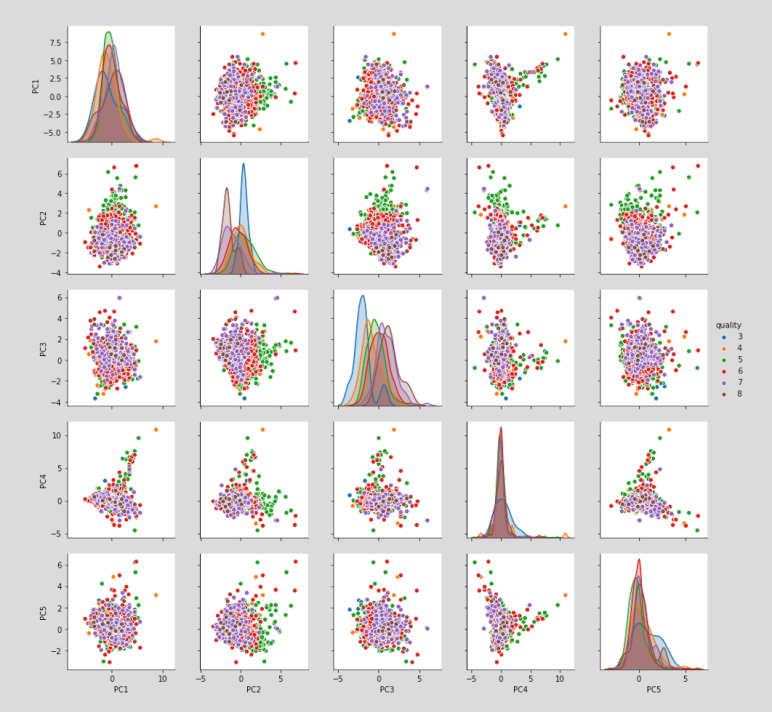

Figure 3. Scatter plot of the reduced data

After the PCA procedure, each of the principle components seemed following the normality assumption.

Statistical Analysis/Modeling

At the beginning we split our data into 70% for training data and 30% for test data. Then we tried to use LDA and QDA with the original PCA dataset since both methods are capable of multi-category classification. However, the results were not very promising. Both methods were having accuracies around 50% (supplementary **figure.2,3,4,5**) One reason could be the points standing for different quality scores are mixed together and caused difficulties to draw a border line. Therefore, we decided to re-coded to quality into three categories:

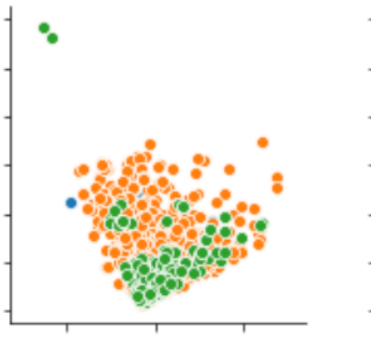

Figure 4. A sample plot from the pair plots

After recoding, we checked the pair plots again and this time the border line between different quality-categories were more obvious.

After getting our data ready for modeling, we started modeling procedure in modeling in following order: QDA, LDA, QDA cross-validation and LDA cross-validation.

QDA Results:

	precision	recall	f1-score	support
0	0.000	0.000	0.000	23
1	0.866	0.930	0.897	402
2	0.432	0.345	0.384	55
accuracy			0.819	480
macro avg	0.433	0.425	0.427	480
weighted avg	0.775	0.819	0.795	480

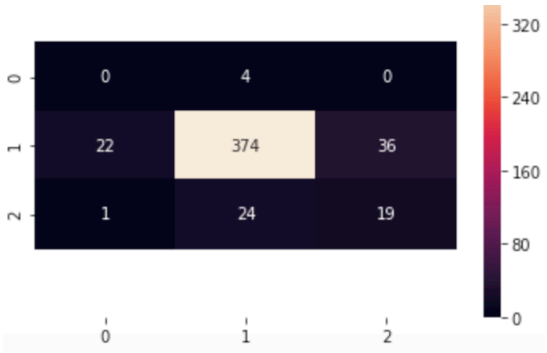

Figure 5. Scores and Confusion Matrix for QDA results

LDA Results:

	precision	recall	f1-score	support
0	0.000	0.000	0.000	23
1	0.866	0.950	0.906	402
2	0.474	0.327	0.387	55
accuracy			0.833	480
macro avg	0.447	0.426	0.431	480
weighted avg	0.780	0.833	0.803	480

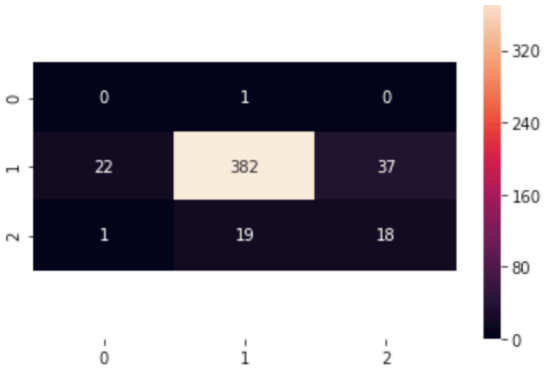

Figure 6. Scores and Confusion Matrix for LDA results

QDA Cross-Validation Result:

	precision	recall	f1-score	support
0	0.077	0.016	0.026	63
1	0.852	0.883	0.867	1319
2	0.330	0.332	0.331	217
accuracy			0.774	1599
macro avg	0.420	0.410	0.408	1599
weighted avg	0.750	0.774	0.761	1599

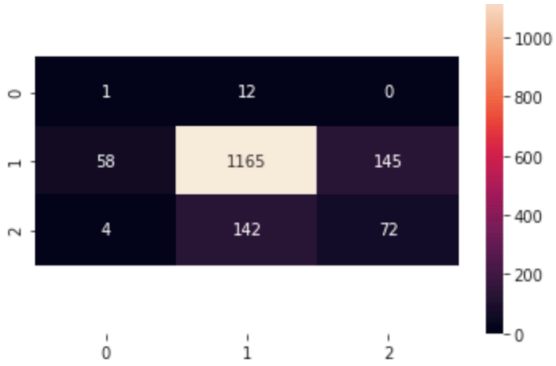

Figure 7. Scores and Confusion Matrix for QDA CV results

LDA Cross-Validation Result:

	precision	recall	f1-score	support
0	0.000	0.000	0.000	63
1	0.851	0.937	0.892	1319
2	0.418	0.281	0.336	217
accuracy			0.811	1599
macro avg	0.423	0.406	0.409	1599
weighted avg	0.758	0.811	0.781	1599

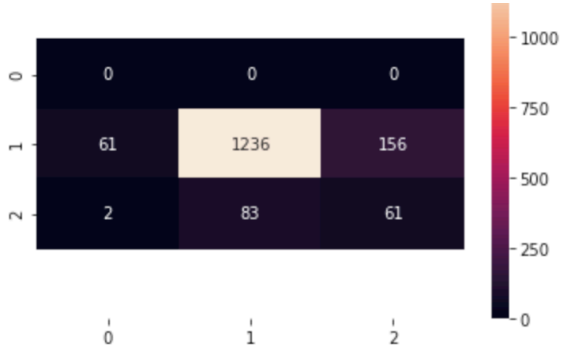

Figure 8. Scores and Confusion Matrix for LDA CV results

Discussion:

Our results show that LDA algorithm has the highest precision, which is the accuracy of the classification. While at the same time, LDA always has better precision than QDA under both conditions.

However, our highest precision is still lower than 80%. One possible reason is that our dataset has a very unbalanced sampling after re-coded into three categories. This can cause the problem that one category might not be selected in training data at all.

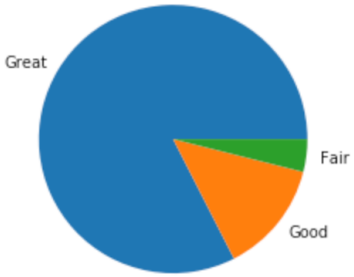

Figure 9. Pie chart for the frequency of each category

One possible way we can come up is using SMOTE (Synthetic Minority Over-Sampling Technique) to remedy this class-imbalance.

We once coded our data into just two categories. However, the pair plots showed that the boundaries between categories are not so obvious. After changed the data to three categories, the logistic regression part was removed since there are more than two categories. However, the information can be found in supplementary document.