Wine Quality Analysis
Danica Fliss
12/12/2020

Introduction

It would be nice to determine whether a wine is good or bad before investing in it. This study looks to see if there are patterns between physiological characteristics of the wine and wine quality. Specifically we utilize least-squares linear regression, LASSO regression, singular value decomposition, principal component analysis, and k-means algorithms to look for patterns. This project has a lot of applications outside of wine quality control, but for this project we used the Wine Quality Dataset from the University of California – Irvine's Center for Machine Learning and Intelligent Systems, which was originally used in research done by Cortez et al. [1]. The Wine Quality Dataset consists of two sub-categories, red wines and white wines, of the vinho verde wine from northern Portugal. The red wine dataset has 1599 entries; the white wine dataset has 4898 entries. Each entry consists of 11 characteristics or "features" based on "physicochemical" attributes and one quality rating or "label" on the scale from 1 to 10 [1].

Method

The first step in our method was to take separate our data into 16 subsets. We then rotate through these subsets choosing one to be the test set to calculate average error, while the rest are used for training the weights used for linear regression. Average error is defined as the average distance from the predicted value from the true value. We also repeated this process using LASSO regression with 0.05 as our alpha. Because LASSO regression tends towards a sparse solution, we can see from the weights which wine characteristics are most important. We then proceeded to perform singular value decomposition in order to find the condition number, which tells us how strong of a correlation there is between wine quality and the principal component, or the most important feature.

Continuing, we ran a k-means algorithm to cluster wines with similar combination of characteristics. We used cross-validation to determine the value of k and defined error as the sum of distances between samples and their cluster center. The program converged when the error of k-means algorithm changed by less than 10 percent of the previous k-means algorithm error, which had a k value of k-1. The cluster numbers were then added to the feature vectors in the data. This was done by adding k more features, and 1 was placed in the ith cluster feature if that sample belonged to the ith cluster of k clusters. Otherwise that feature had a value of 0. Originally, the cluster numbers were added as a single feature. However, we realized that the number of the cluster is arbitrary depending the where the k-means algorithm began, so these numbers would have no association with wine quality.

Using our new set of features with the data, we repeated the training and testing steps and compared the results to the original dataset. In particular, if average error decreased, we could determine if certain combinations of characteristics were more important than any direct correlation of individual characteristics and wine quality.

Results

The average error between the predicted value using least-squares linear regression and the actual value was 0.582 for white wine and 0.437 for red wine, and using LASSO regression the average error was 8.104 for white wine and 5.011 for red wine.

White wine used an average of 5 clusters when calculated using our algorithm. As can be seen in Figure 3, the weights found using LASSO regression did not have nonzero values for any of the added k-means cluster features, thus they were not as important as some of the original features. This is reinforced when comparing the average errors between the original least-squares linear regression and LASSO regression and the errors using the new dataset. The new data set had an average error of 0.575 using least-squares linear regression and 8.104 using LASSO regression, showing the cluster features had little affect on the optimal solution.

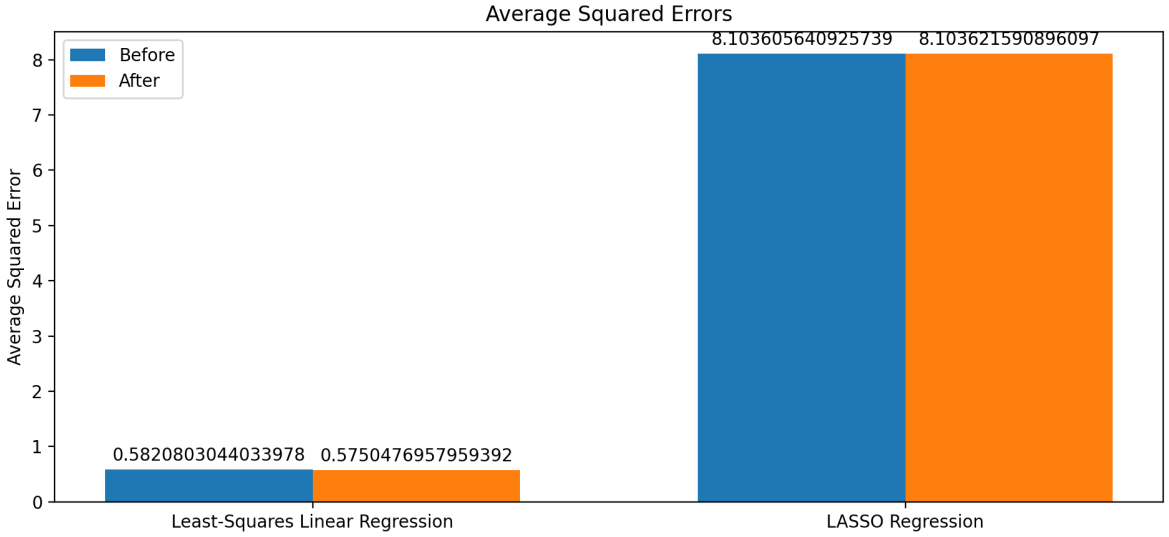


Figure 1: White Wine Error Analysis. This figure shows the average errors using the original dataset, labeled as "Before", and the average errors using the dataset updated with the k-means cluster features, labeled as "After". We see a little change in both linear and LASSO regression from "Before" to "After".

Red wine showed a similar pattern as the white wine. The data usually had 6 k-means clusters when calculated using our algorithm. Again, these added features had zero weights using LASSO regression, and the red wine dataset had an average error of 0.439 using least-squares linear regression and 5.011 using LASSO regression.

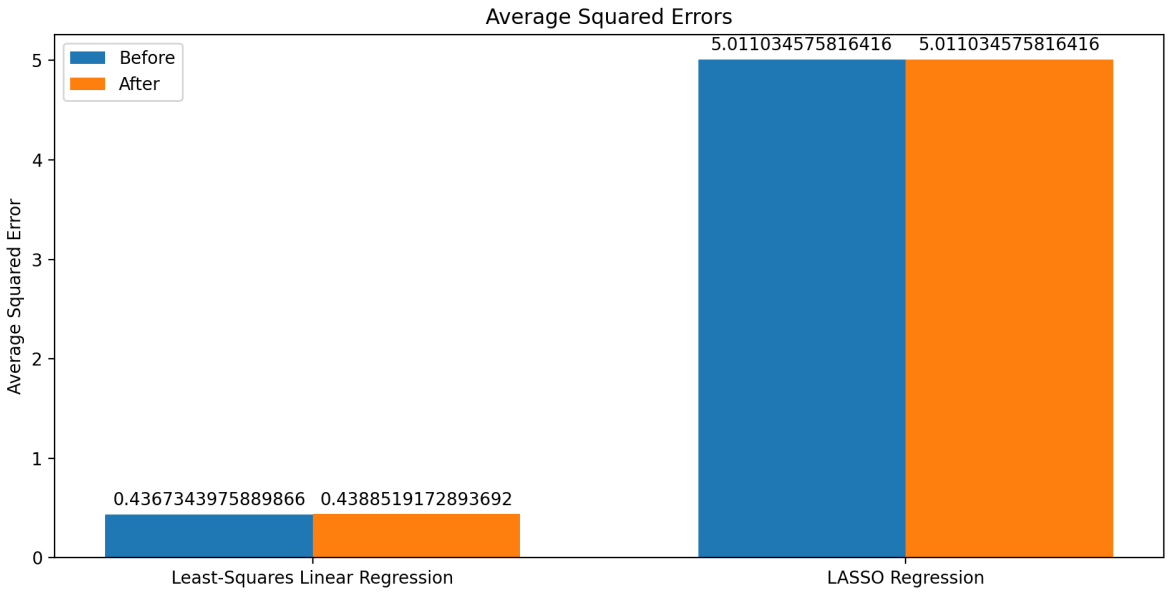
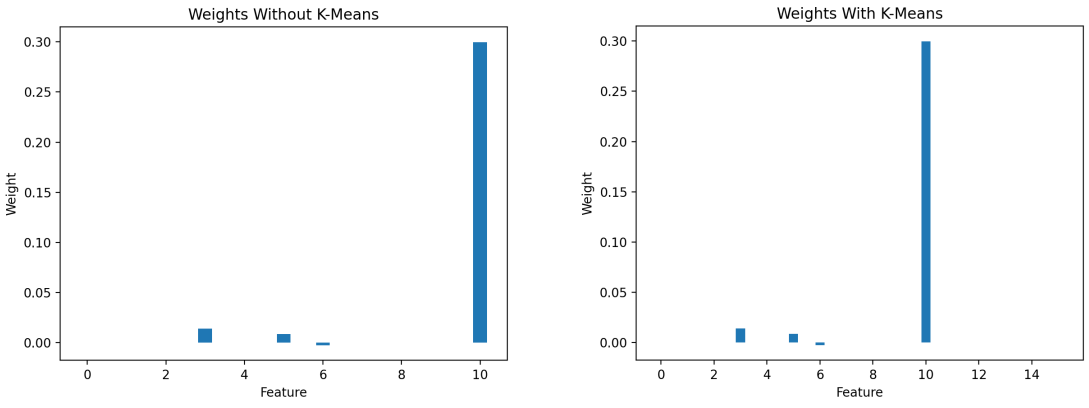


Figure 2: Red Wine Error Analysis. This figure shows the average errors using the original dataset, labeled as "Before", and the average errors using the dataset updated with the k-means cluster features, labeled as "After". We see a little change in both linear and LASSO regression from "Before" to "After".

We see in Figure 3 that the most important feature in determining wine quality is alcohol content, and it is a positive correlation. The condition numbers express the strength of the principal component alcohol, which are 6.42 for red wine and 11.55 for white wine. The condition numbers virtually do not change with the addition of k-means cluster features. The next important features for white wine are residual sugar, free sulfur dioxide, and total sulfur dioxide, the last being a slight negative correlation. The red wine data set also had nonzero values for free sulfur dioxide and total sulfur dioxide. However, red wine does not have any strong correlation to residual sugar but does have a correlation with fixed acidity.

White



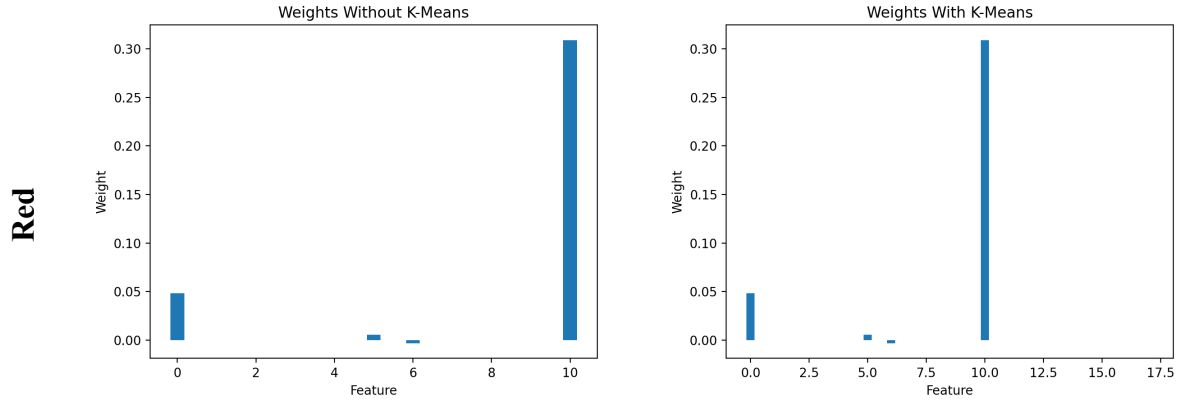


Figure 3: LASSO weights. White wine (top) and red wine (bottom). This figure shows the sparse solution to LASSO regression both for the original dataset (left) and with the dataset updated with k-means cluster features (right). Order of features are (0) fixed acidity, (1) volatile acidity, (2) citric acid, (3) residual sugar, (4) chlorides, (5) free sulfur dioxide, (6) total sulfur dioxide, (7) density, (8) pH, (9) sulphates, (10) alcohol, (11+) k-means cluster.

Table 1: Results from Wine Analysis

	White Wine	**Red Wine**
Average Number of K-means Clusters	5	6
Condition Number	11.55	6.42
Condition Number After K-means	11.55	6.41
Average LS Linear Regression Error	0.582	0.437
Average LASSO Regression Error	8.104	5.011
Average LS Linear Regression Error After K-means	0.575	0.439
Average LASSO Regression Error After K-means	8.104	5.011

Discussion

Using least-squares linear regression to train a model to predict wine quality based on 11 characteristics produces fairly accurate results. We could have made a clearer error analysis if we quantized the predictions of our model to be integers from 0 to 10, the values that wine quality could take. Instead, we included any distance from the true label in the error whether or not it would quantize to the correct value.

Originally we had not planned to include LASSO regression, but rather we would rely on principal component analysis to tell us the strength of the most important feature. However, we would not have been able to identify which feature was the principal component. Our method

also did not take into account the fact that the values of the features were on different scales (i.e. chlorides are on a scale ~10e-2 and alcohol is on a scale ~10e0).

Our cross-validation algorithm for k-means converges if the distances from the cluster centers change by less than 10 percent. However, this threshold was chosen arbitrarily, so it is unclear whether this produced a good number of k-means clusters. Still, the lack of success with the k-means cluster features we did use implies that more clusters would not have helped in predicting wine quality, but we cannot say for certain without further investigation.

Conclusion

The results from this experiment show that wine quality is strongly correlated to alcohol content. The average error using the least-squares linear regression was 0.582 for white wine and 0.437 for red wine. Considering if the difference between expected value and the predicted value is less than 0.5, then translating the predicted values to possible values of wine quality would be the same as the expected value. The LASSO regression was not accurate using an alpha of 0.05. LASSO regression had an average error of 8.104 for white wine and 5.011 for red wine. These results show most predictive values would not correspond to the true value.

The addition of k-means cluster features was supposed to see if particular combination of characteristics were a better indication of wine quality than linear correlations of each feature. Instead of decreasing error, the cluster features did not change error by any significant amount for both regression methods. Further investigation is needed to determine the optimal number of clusters to be certain that cluster features would not increase the accuracy of the model.

Link

https://github.com/flissd/ECE532_Final_Project

References

[1] P. Cortez, A. Cerdeira, F. Almeida, T. Matos and J. Reis. Modeling wine preferences by data mining from physicochemical properties. In Decision Support Systems, Elsevier, 47(4):547-553, 2009.