ECEN 689
Challenge 4
Exercises with Wine Quality Data

Brandon Thayer, Harish Kumar, and Mason Rumuly

Abstract

We use a wine dataset for both red and white wines to predict white wine quality via linear regression, create a classification model for differentiating between red and white wines, and apply the white wine quality regression to the red wines. All of these tasks accomplished a high degree of accuracy. We then analyze the implications of classification accuracy on linear regression generalizability.

I. Introduction

GENERALIZATION of learned models is a cornerstone of machine learning. It is important to know whether a model trained on one set of data will be applicable to new data, as well as to data collected about a distinct but similar subject. To explore this concept, we apply regression and classification techniques to a quality data set of white and red wines and analyze the implications of the results.

A. Data Set

We use a wine quality data set [1] which contains information on a number of white and red wine samples. For each sample, there are eleven objective chemical measurements (e.g. pH, free sulphur dioxide, density) in addition to subjective quality estimates produced by wine-tasting experts.

B. Goals

We focus on three concrete goals for the quantitative analysis:

1) Use only the white wines in the data set to train and verify a linear model to predict the quality of the wine.
2) Train and verify a classifier to distinguish between red and white wines in the data set.
3) Apply the white wine quality model to predict the quality of the red wines.

II. Methods

A. Linear Regression

We explored three variations of linear regression: unregularized least squares, least absolute shrinkage and selection operator (LASSO), and Ridge. The primary purpose of regularization is to reduce over-fitting in the trained model. Additionally, LASSO regularization drives coefficients corresponding to less-important features towards zero for feature reduction, making it useful for general analysis.

B. Classification

We explored Binary Decision Trees (BDT) to distinguish between red and white wines. We conducted classification on the raw data, on data with a Principle Component Analysis (PCA) transformation applied, and on data with a Linear Discriminant Analysis (LDA) applied. Additionally, several binary tree parameters (e.g., maximum depth, minimum samples per leaf, etc.) were explored. Ultimately, the best performance came from training a tree with the raw data and limiting tree depth to seven.

III. Results

A. Quality Prediction

The below table provides a summary of the performance of three variations of linear regression on the quality prediction task. We noted that lasso regression performed best, with the coefficient corresponding to density being driven to zero.

Variation of Linear Regression	**MSE**	**Optimal hyperparameter value**
Unregularized	0.6159	N.A.
Lasso regression	0.5943	0.00015
Ridge regression	0.6050	0.00015

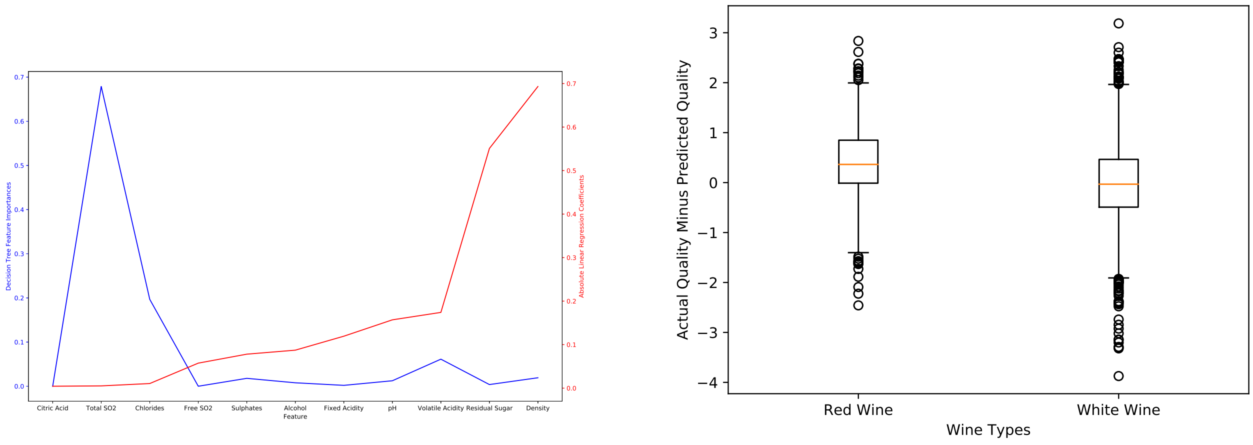
B. Classification

The BDT's performance using the raw data was near perfect on this task, achieving a classification accuracy of 99.14% on our validation set. **Fig: 1a** shows the feature importance values from the decision tree in blue. Use of PCA or LDA did not improve the overall performance of the BDT.

C. Model Transfer

We conjecture here that in order for our white wine linear regression model to generalize well to red wines, it must be largely *independent* of features which allow the BDT to differentiate between red and white wines. If a feature differs significantly between white and red wines, making it useful for classification, then the quality estimation model will be significantly skewed if it strongly depended on that feature. If the BDT performed poorly, it would imply that no features varied significantly between the varieties of wine and that the quality model would generalize well; however, the BDT performed well, allowing the possibility for the model to generalize well or poorly. Empirically, the model generalized well, indicating that the classifier and regressor relied on separate information; see **Fig: 1b** .

To further interpret these results and confirm our conjecture about the reliance on separate features, we obtained normalized linear regression coefficients and compared them to the importance of each feature in the BDT as shown in **Fig 1a**. This plot fits well with our conjecture as we note that features like Chlorides and Total Sulfur Dioxide which are useful in the BDT, are assigned small coefficients in the normalized regression model. Conversely, the features that greatly help in predicting the quality in the normalized linear regression model (residual sugar and density) are unimportant in differentiating red and white wines.



(a)

(b)

Fig. 1. (a): Binary decision tree feature importance (blue) and the absolute normalized coefficients from linear regression (red). (b): Boxplot for actual quality minus predicted quality for red and white wine. Whiskers are at the 1st and 99th percentiles.

IV. Conclusion

Classification and model transfer both performed well on the wine data because each process used different features to achieve its results. These insights indicate a qualitative property of model generalization based on the similarity of data sets which future work may quantify.

References

[1] https://archive.ics.uci.edu/ml/datasets/Wine+Quality