Predicting the Quality of Wine

Leo Hu and **Natnael Mulat**
{lehu,namulat}@davidson.edu
Davidson College
Davidson, NC 28035
U.S.A.

Abstract

In this paper, we used polynomial regression techniques to predict quality rating of the Portuguese "Vinho Verde" wine based on physio-chemical features. We tuned regression models ranging from degree of 1 to degrees of 4 with either L1 (LASSO) or L2 (RIDGE) regularization based on our training data. It turns out that polynomial regression models with a degree of 2 gave the best prediction based on our testing data. Degree 2 model using L2 regularization performs slightly better according to the adjusted R^2 metrics than its L1 counterpart.

1 Introduction

Background and Related Work

Unprecedented data analysis emerges in the era of big data. Various forms of online reviews are available through data mining. These massive online databases of reviews expose information about reviewer's criteria based on some features of a product. Industry professionals are interested in discovering these criteria, which can be used to predict the performance of their products.

However, traditional analysis of online reviews focuses on building recommending systems instead of predicting the rating. For example, there is an hotel recommendation algorithm that imitates a user that favors reviews written with the same trip intent and from people of similar background (nationality) and with similar preferences for hotel aspect (Levi and Mokryn 2012).

Although similar to recommending systems, rating prediction using machine learning techniques is an area not actively researched, except for a few on bonds rating. For example, optimal neural network structures are used to predict the risk of some bonds, which affects they credit rating(Utans and Moody 1991). However, less researches are available when it comes to predicting the rating of a consumer product using polynomial regression technique.

Research Question

Under this direction, our research aims to predict experts' quality ratings of the Portuguese "Vinho Verde" wine based on its physio-chemical features using polynomial regression technique. Specifically, we are interested in knowing whether we can predict the quality rating of a wine based on some of its physio-chemical features even though perceiving the flavor of a wine might be a very complicated chemical process. Additionally, we want to find what the best polynomial regression model is and its corresponding metrics of goodness of fit if it is possible to predict such ratings.

Organization

The rest of the paper is organized as follows: Section 2 describes the dataset and how relevant each feature is for prediction. Section 3 describes the experiment performed on the dataset. Section 4 discusses the results of the models for prediction. Section 5 points out the broader impact of our model to the wine industry as well as social justice concerns brought by rating prediction algorithms. Section 6 summarizes our conclusion. Section 7 summarize the contributions from our group members. Section 8 is the acknowledgements.

2 Data Preparation

Dataset

We used the dataset titled "Wine Quality", which is public available for research (Cortez et al. 2009). The dataset consists of two sub-datasets created using red and white wine samples of the Portuguese "Vinho Verde" wine. Our dataset contains 1599 samples of red wine and 4898 samples of white wine. The inputs are features of a wine from objective tests (e.g. volatile acidity, alcohol, and PH values), while the output is a score of sensory review, taken from the median of at least 3 evaluations made by wine experts. Each expert graded the wine quality between 0 (very bad) and 10 (very excellent). There is no missing data in our datasets.

Here is a full list of the 11 features based on physicochemical tests:

1. fixed acidity
2. volatile acidity

3. citric acid
4. residual sugar
5. chlorides
6. free sulfur dioxide
7. total sulfur dioxide
8. density
9. pH
10. sulphates
11. alcohol

Reprocessing

In each red and white wine dataset, we generated a binary variable named "type" which equals 1 if the sample is a red wine and 0 otherwise. Then, we concatenated these two datesets. We shuffled all samples because there might be a systematic bias in rating different types of wine, which might potentially affect our machine learning outcome. We also split our dataset into training set and testing set. We standardized the training set, which helps to reduce multicollinearity that is produced by higher-order terms, leading to imprecise coefficients. The testing set is also standardized, which is prepared for testing the fit of our model. Thus, all following analysis are based on the standardized data.

Feature Selection

We were not sure if all 12 input features are relevant to the quality rating. Thus, we created scatter plots for each of the 11 features versus the quality score with the corresponding line of best fit. We also computed the corresponding Pearson correlation coefficients ranging from 0 to 1, which measure the level of correlation between two variables (Anderson and Williams 2015).

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$$

As shown in **Figure 1,2,3** in **Appendix section**, alcohol, density, chlorides, and volatile acidity and type are the most relevant features (with Pearson correlation coefficients greater than 0.1) while sulphate, pH, free sulphur dioxide, and residual sugar are the least relevant features (with Pearson correlation coefficients less than 0.05). However, we did not have enough evidence to exclude sulphate, pH, free sulphur dioxide, or residual sugar from our model because we did not know if the correlation will get stronger with high degrees. We have no information about what causes the significant correlation between quality and alcohol, density, chlorides, or volatile acidity. For example, a few outliers in **Figure 1 (c)** might have a significant influence on its correlation coefficients. We did not delete those outliers in case that our action might affect learning outcome. Thus, we kept all observations in our training set and their all features as the input for our model.

3 Experiments

In order to predict quality ratings of the Portuguese "Vinho Verde" wine, we trained 5 models with linear regression using the closed form solution with OLS. Each model used cross validation with k-fold of 10, and each prediction is computed with an estimator fitted on the corresponding training set.

Our models uses linear regression with multiple degrees and cross validation in order to choose the best fit. For polynomial regressions with degree up to 4, we run $l1$ and $l2$ regularization, where $l1$ adds absolute value of magnitude of coefficient as penalty term to the loss function, while $l2$ adds squared magnitude of coefficient as penalty term to the loss function. Overall, our models are categorized into three groups, where the first model represents the linear regression with no polynomial features, $l1$ regularization for the degree of 2,3, and 4 polynomial features, and we run similar experiment for $l2$ regularization.

With these different models, we used a train set of 70% and test set of 30% out of the 6497 wine observations in the data set. Standardization across instances is done after splitting the data between training and test set, using only the data from the training set. Since using any information coming from the test set before or during training is a potential bias in the evaluation of the performance. Standardization uses the $z - score$ formula below since an observation is scaled by the training set data by seeing how far away the observation is from the mean in terms of the standard deviation(Anderson and Williams 2015).

$$z = \frac{x_i - \bar{x}}{\sigma} \tag{1}$$

The regression models defined in this paper to predict quality ratings of wine use the form :

$$Y_i = \beta_0 + \beta_1 X_{i1} + \cdots + \beta_p X_{ip} \tag{2}$$

where $i = 1, 2, \ldots, n$ for n observations, while Y represents the quality rating observation, and X represents the p features the model is controlling for. Each feature has a parameter β in the model to account for the the degree of change in the outcome variable, *i.e.* quality rating, for every 1-unit of change in the feature variables

For the models defined, we solve the regression parameters using a closed-form solution instead of using optimization algorithms such as gradient descent, stochastic gradient, etc.. One of the major reasons we chose to used a closed-form solution is the size of the data set we are using. For the first model with only 12 features and 6497 observations, it is better to use the closed form solution since solving for the model parameters is not as costly as using a big data. For our experiment, at most the features are less than a 2000, and the number of observation are less than 10,000. Therefore, given these factors, the different models' parameters, β is solved as below, where X' is the transpose of all the

features, and $(X'X)^{-1}$ is the inverse of $X'X$

$$\beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \\ . \\ . \\ \beta_n \end{bmatrix} = (X'X)^{-1}X'Y \quad (3)$$

For $l1$ and $l2$ regularization of polynomial terms to address the problem of over-fitting, alpha value that is in the log-space of -6 to 6 is chosen depending on which alpha value better fits the different models using cross validation.

To measure the performance of the models, accuracy score, r-squared, adjusted r-squared, and mean-squared-error(MSE) are reported. These different ways of evaluating the models ultimately measure how closely the models can predict the true ratings of the quality of different wines. The accuracy score gives a percentage of how many of the model-predicted ratings *exactly* match the true ratings of the wines. Thus, the formula below is used to compute the accuracy where y represents the true quality that is recorded in the data set, and $\hat{y}$ is the predicted ratings that the models give with n number of observation. *i.e*

$$accuracy(y, \hat{y}) = \frac{1}{n}\sum_{i=0}^{n-1} y_i = \hat{y}_i \quad (4)$$

Another measure of our model is mean squared error(MSE)[1]. Essentially MSE measures how good a model is by calculating mean of the sum of squared difference between actual values and the predicted / estimated values. Mathematically, it is expressed as

$$MSE = \frac{1}{n}\sum_{i=0}^{n}(y_i - \hat{y}_i)^2 \quad (5)$$

R-squared [2], on the other hand, is the fraction of the sample variance of the true ratings that is explained by a model. Mathematically, R^2 is 1 minus the fraction of the variance of the true ratings not explained by the models.

$$R^2 = 1 - \frac{\sum_{i=0}^{n}(y_i - \hat{y})^2}{\sum_{i=0}^{n}(y_i - \bar{y})^2} \quad (6)$$

Although R^2 gives us a good estimate of goodness-of-fit for the different models, when we have multiple regressions, the R^2 increases whenever we add more features. But the increase in R^2 does not mean that adding variable actually improves the fit of the model. In this case, the R^2 gives an inflated estimate of how well the model fits the data. In order to correct this R^2 by a factor that can sensibly account for the amount of features, we will instead use adjusted r-square[3], which is represented by $\bar{R}^2$. The factor will penalize added features that do not improve the model. Mathematically, it is expressed as below where n is the number of observation, and p is the number of features.

$$\bar{R}^2 = 1 - (\frac{n-1}{n-p-1})\frac{\sum_{i=0}^{n}(y_i - \hat{y})^2}{\sum_{i=0}^{n}(y_i - \bar{y})^2} \quad (7)$$

In order to determine which model best estimates ratings, we will use all four measurements to compare the goodness of fit and decide which model we should use for the test set. After choosing a model for which we have a relatively good goodness of fitness, test data set is used to get an unbiased and consistent parameter values for the model.

4 Results

To assess the goodness-of-fit of our models, we are using accuracy score, mean-squared-error, r-squared, and adjusted r-squared. The measurement results are reported on table 1.

The first row has the results for the first model which uses linear regression with no polynomial features. In order to find the predicted quality ratings, we are using cross-validation on the training data set with k-fold of 10. The accuracy score for that model is 0.45, in other words the model is able to predict quality ratings *exactly* with an accuracy of 45%. The predicted quality ratings from model 1 that match the true quality ratings of wine is 45%, but this measurement does not account for the other predicted ratings that may be off by a small or large factor. In order to get a larger picture of the fit, we also consider MSE. Model 1 has a MSE of 0.53, which tells us that model 1's prediction error squared is on average is 0.53. In model 1, 30% of the variation in quality rating of wine is explained by the 12 features that the model included. Overall a good model should have a higher accuracy score, a lower MSE, and a larger $\bar{R}^2$ in order to get a good fit of the data set.

According to our measurements, Model 2 with degree 2 and $L2$ regularization has a better fit than the other models with different degrees and regularization. Model 2 has an accuracy rate of 45%, which is as good as model 1. However, Model 2 has a better fit in terms of minimizing the mean squared errors, since it has MSE of 0.5. The R^2 measurement of fit is the best among the other models we run, with an R^2 of 0.329, in other words 33% of the variation on quality ratings of wine is explained by the features included in the model, which is slightly better than model 1. But in order to account for the added features, we also considered adjusted R^2, which is 31%, a little lower than the R^2 fit since it penalises features that are not relevant for the model.

[1] Introduction to Econometrics [see][p.533] Stock & Watson
[2] [see][p.196] Stock & Watson
[3] [see][p.197] Stock & Watson

Model	Accuracy	MSE	R^2	$\bar{R}^2$
$degree = 1$	0.45	0.53	0.297	0.295
$L2$ ($degree = 2$)	0.45	0.50	0.329	0.316
$L2$ ($degree = 3$)	0.42	0.56	0.251	0.168
$L2$ ($degree = 4$)	0.35	1.29	−0.715	−1.859
$L1$ ($degree = 2$)	0.45	0.51	0.322	0.320
$L1$ ($degree = 3$)	0.45	0.53	0.297	0.295
$L1$ ($degree = 4$)	0.47	0.85	−0.122	−0.125

Table 1 : Models' Measurement Results

Interestingly, model 4 and model 7 have a negative R^2 and adjusted R^2. The reason is because the formula for R^2 uses the sum of error terms squared, *i.e*. sum of predicted quality ratings minus quality ratings squared, as the numerator. And if that value is greater than the denominator, which is the sum of square of quality ratings distance in the distribution from the mean, then the value of R^2 can be negative. This is an instance where the model that was trained does not follow the trend of the data, so it fits worse than a horizontal line, which is why sum of predicted quality ratings minus quality ratings squared is larger than the sum squared deviation from the mean of the true quality ratings. Model 4 with $l2$ regularization and model 7 with $l1$ regularization are the anomalies of the models developed as they do not fit the data-set as well as the other models. Overall, from the table we can see a trend that the models that have a higher degree than 2 tend to not have a good fit as the models detract from the true quality ratings of wine.

With the models we developed, we expect the test set to perform as similar to model 2 with degree 2 and $l2$ regularization. The measurements should also match the model 2's metrics. Accuracy scores should be around 45%, meaning 45% of the model predictions match the true quality ratings, but since we are using our test set the predictions should be unbiased.

Model	Accuracy	MSE	R^2	$\bar{R}^2$
$L2$ ($degree = 2$)	0.44	0.53	0.315	0.312

Table 2 : Test set model results with degree=2 and $l2$ regularization.

The model with the test set that has a polynomial degree of 2 with $l2$ regularization performed similar to model 2 and model 1. Accuracy scores are 1 *percentage* points off with 44%. The R^2 and $\bar{R}^2$ are very similar, which implies that the added features are relevant and do have an effect on the prediction of the model. The MSE of the test set is similar to model 1, but the variation of predicted ratings of wine is as good as model 2.

Overall, the model given the 12 features and the type of wine, either white or red, it can *exactly* predict the quality rating of the wine with an accuracy of 44%. There are a lot of features, observed and unobserved, that affect the quality rating of a wine that are not included in this model. For instance, a wine quality is affected by the age of the wine, although that can be present in the physio-chemical features, just that fact that the wine is old can affect the opinion of the expert who rates the wines, which affects the quality rating of the wine. Features such as aroma can partially be expressed by physio-chemical features, but unobservable features like the wine rater's preference of aroma do affect the quality rating of the wine. But the model is able to predict quality ratings from only the physio-chemical features, and the fact that it does omit features that may affect quality ratings makes the model liable to omitted variable bias.

5 Broader Impacts

As machine learning has become more ubiquitous, concern is growing in society at large about the unintended side-effects of this technology. Under the context of predicting wine rating, the polynomial regression technique we applied might cause unintended feedback loop. For example, our model might wrongly predict that the a certain new wine should be rated highly. In reality, the wine might be actually rated highly by some experts if they trust our model and know our model gives a high prediction. As a result, our model will *reinforce* its believe and continue its wrong prediction, posing challenges when we are trying to assess the quality of a new wine.

Moreover, making new wine creatively is harder if the wine industry accepts such prediction system. Our model assumes only some physio-chemical features determine the quality of a Portuguese "Vinho Verde" wine. However, determining flavor is a complicated chemical process, which might be simplified by the data available for our model. Therefore, some wine manufacturer might reject a wine maker' creative proposal because of a low predicted quality score if the manufacturer adopt our model.

The concern is aggravated particularly when similar rate prediction machine learning technique is deployed in domains where it has the potential to reinforce or exacerbate existing inequities. For example, machine learning systems determining whether to deny mortgage loans to an individual can be considered as systems predicting credit scores. If the model is trained on datasets that is biased to individuals of some demographic characteristics, the trained model will continue to assign low credit score to those individuals. The bias will be reinforced if the individual's mortgage loan is denied. The reason is that machine learning models focus on predictions instead of causal reference. What causes an individual to have a low credit score remains unclear in machine learning models. Therefore, machine learning developers should focus more on causal reference when trying to deploy machine learning techniques in domains where it has the potential to reinforce bias or existing inequities.

6 Contributions

Leo Hu wrote the Introduction, Data Preparation, and Broader Impacts section, while Natnael Mulat prepared the data for the models. Leo Hu used the prepared data to pre-

pare figure 1 that maps out the features and their correlation with quality ratings. Both Leo Hu and Natnael Mulat worked on developing the models using Scikit-learn's built-in regression solver, specifically Leo Hu prepared the $l2$ regression while Natnael Mulat carried that work to include more polynomial features and $l1$ regressions. Natnael Mulat wrote the Experimentation, Result, and Conclusion section and prepared the tables that summarize the metric of the models. Both partners proof-read each sections.

7 Acknowledgements

We want to thank Prof. Raghuram Ramanujan for his generous advice about our drafts and coding.

References

Anderson, David R., S. D. J., and Williams, T. A. 2015. *Modern Business Statistics with Microsoft Excel*. Cengage Learning.

Cortez, P.; Cerdeira, A.; Almeida, F.; Matos, T.; and Reis, J. 2009. Modeling wine preferences by data mining from physicochemical properties. In Decision Support Systems, Elsevier, 47(4):547-553. ISSN: 0167-9236.

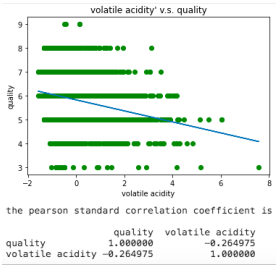
Levi, A., and Mokryn, O. 2012. Finding a needle in a haystack of reviews: Cold start context-based hotel recommender system abstract. *RecSys '12: Proceedings of the sixth ACM conference on Recommender systems* 115–122.

Utans, J., and Moody, J. 1991. Selecting neural network architectures via the prediction risk: application to corporate bond rating prediction. *Proceedings First International Conference on Artificial Intelligence Applications on Wall Street*.

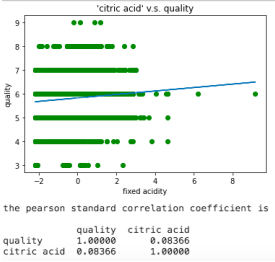
8 Appendix

Due to their length, Figure 1, 2 and 3 begin on the next page.

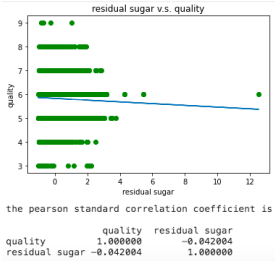
(a) fixed acidity



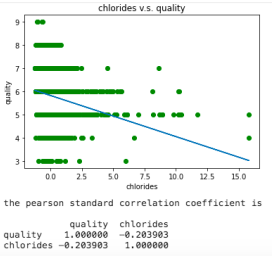
(b) volatile acidity



(c) citric acid

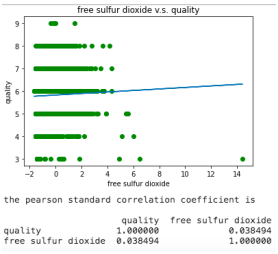


(d) residual sugar

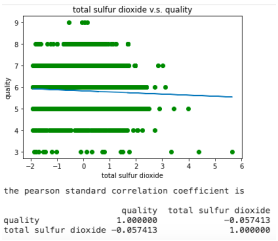


(e) chlorides

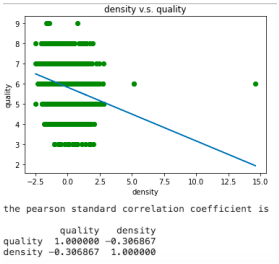
Figure 1: Correlation between quality and each feature (part 1)



(a) free sulfur dioxide

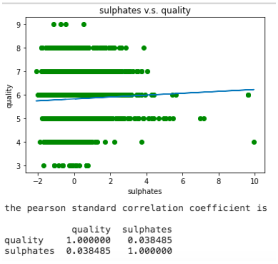


(b) total sulfur dioxide



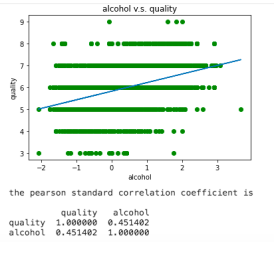
(c) density

(d) pH

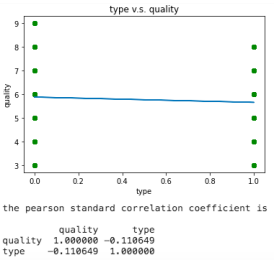


(e) sulphates

Figure 2: Correlation between quality and each feature (part 2)



(a) alcohol



(b) type

Figure 3: Correlation between quality and each feature (part 3)