Analysis of Wine Quality

Yi Han, Zhaoyang Shi

Department of Statistics, UC Davis

1 Introduction

Wine is very important in our everyday life not only as a kind of drink but stands for significant meanings in social activities. Good wine quality depends on many kinds of its features. In this report, based on the red wine quality data with some features of red wine as the predictors and quality as their response, we propose different statistical models from the view of regression and classification for predicting wine quality and further introducing Bootstrap methods for statistical inference.

All works of this project are jointly and collaboratively done by Yi Han and Zhaoyang Shi together including codes, presentation and report. In general, we work independently and then discuss for modification and improvements

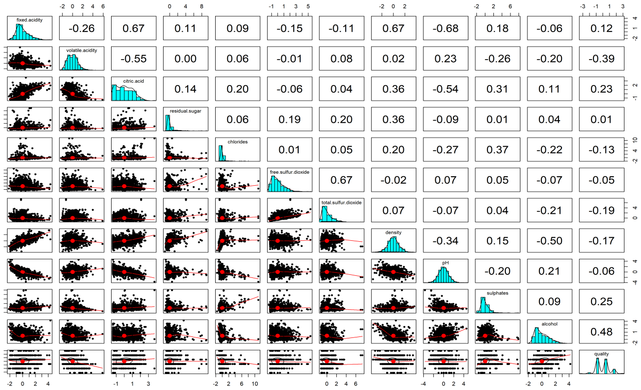
2 Data Description

This data[1] set has 11 predictors which represent different substantial in red wine such as acidity, sugar, pH an sulfur dioxide together with one response which represents the quality of wine.

The predictors include: fixed.acidity, volatile.acidity, citric.acidity, residual.sugar, chlorides, free.sulfur.dioxide, total.sulfur.dioxide, density, pH, sulphates, alcohol. The response is quality. The size of this data set is 1599 × 12 with no missing value. The value of response is discrete ranging from 3 to 8. Therefore, we can view it from two perspectives. One is to view it as a regression problem and another as a classification problem.

We first draw the paris graph between each random variable to have a whole perception of the data set.

Figure 1: The correlation of variables.



The pairs graph (figure 1) suggests that there are several variables that are strong correlated such as: fixed.acidity and citric.acidity, fixed.acidity and pH, free.sulfur.dioxide and total.sulfur.dioxide and etc.

3 Regression

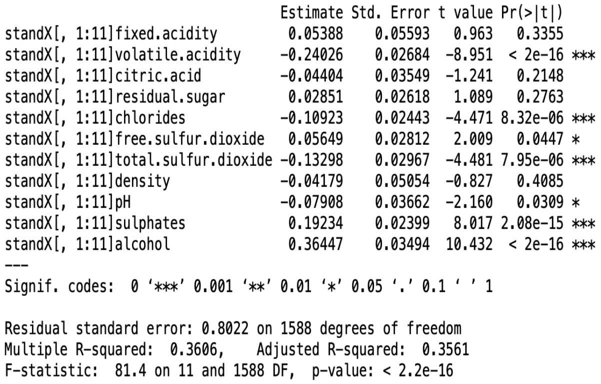
Before building regression model, data is standardized with zero mean and unit variance.

3.1 Ordinary Least Squares

Here we first try the Ordinary Least Squares. And the formula of this model is:

$$Y = X\beta + \epsilon, \epsilon \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma^2\mathbf{I}). \tag{1}$$

According to figure 2, the result shows that due to multicollinearity, the R^2 is small and some predictors are not significant.



```
                                    Estimate Std. Error t value Pr(>|t|)
standX[, 1:11]fixed.acidity         0.05388    0.05593   0.963   0.3355
standX[, 1:11]volatile.acidity     -0.24026    0.02684  -8.951  < 2e-16 ***
standX[, 1:11]citric.acid          -0.04404    0.03549  -1.241   0.2148
standX[, 1:11]residual.sugar        0.02851    0.02618   1.089   0.2763
standX[, 1:11]chlorides            -0.10923    0.02443  -4.471 8.32e-06 ***
standX[, 1:11]free.sulfur.dioxide   0.05649    0.02812   2.009   0.0447 *
standX[, 1:11]total.sulfur.dioxide -0.13298    0.02967  -4.481 7.95e-06 ***
standX[, 1:11]density              -0.04179    0.05054  -0.827   0.4085
standX[, 1:11]pH                   -0.07908    0.03662  -2.160   0.0309 *
standX[, 1:11]sulphates             0.19234    0.02399   8.017 2.08e-15 ***
standX[, 1:11]alcohol               0.36447    0.03494  10.432  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.8022 on 1588 degrees of freedom
Multiple R-squared:  0.3606,    Adjusted R-squared:  0.3561
F-statistic:  81.4 on 11 and 1588 DF,  p-value: < 2.2e-16
```

Figure 2: The result of Ordinary Least Squares.

We also do residual analysis here. The result is shown in figure 3.

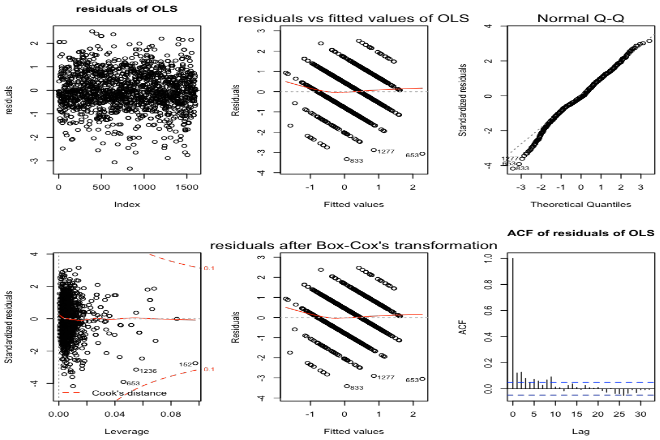


Figure 3: The residuals plot of OLS.

From the first plot we find that the residual points scatter randomly around the zero line so we assume that they have zero mean value. The third plot shows that the normal assumption is satisfied. The forth plot using Cook's distance to show there are no outliers in this data set. The second and fourth plot are the fitted value versus residuals before and after Box-Cox's transformation. There is very little difference between them which means that Box-Cox's transformation cannot help us. The sixth plot suggests that this data set may have autocorrelation problem.

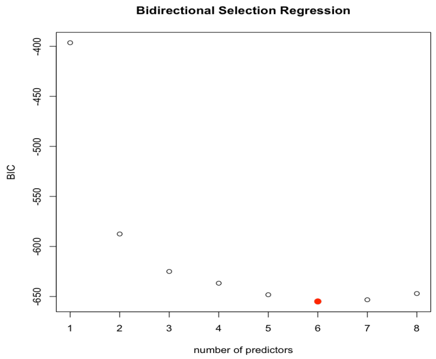
3.2 Stepwise Regression

In order to solve the multicollinearity problem, we first use the stepwise method. The formula of this model is:

$$Y = X_{(k)}\beta_{(k)} + \epsilon,\ \epsilon \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma^2\mathbf{I}). \tag{2}$$

Here we use the Bayesian Information Criterion to help us choose the number of predictors and their coefficients. The result is shown according to figure 4.

Figure 4: BIC vs number of predictors.



From the figure 4, we can see that the Bayesian Criterion Information suggests choosing 6 predictors and the final result is:

$$\begin{aligned} quality = &-0.23 volatile.acidity - 0.12 chlorides \\ &- 0.08 pH - 0.10 total.sulfur.dioxide \\ &+ 0.19 sulphates + 0.38 alcohol. \end{aligned} \tag{3}$$

3.3 Principle Component Regression and Partial Least Squares Regression

When there exists multicollinearity among predictors, the design matrix degenerates into a lower dimensional subspace. By PCA, PCR tends to project data into some subspaces that can reserve most of the information and meanwhile increase the signal-noise ratio.

Apply SVD to the design matrix X:

$$X = U\Sigma V^T, \tag{4}$$

where Σ is a diagonal matrix of singular values in descending order: $\sigma_1 \geq \sigma_2 \geq ... \geq \sigma_r \geq 0$. We form a low-rank approximation with the first $k < r$ principle components:

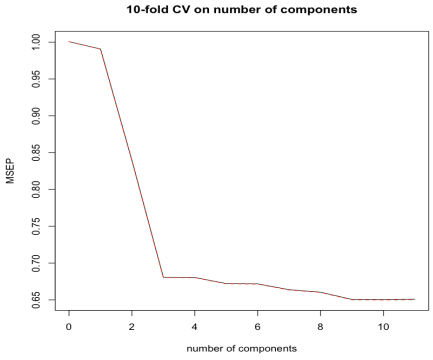
$$X_{new} = (U\Sigma)_k V_k^T = \sum_{i=1}^{k} \sigma_i U_i V_i^T, \tag{5}$$

where $(\cdot)_k$ denotes the $k-$th column of the matrix. Then fit the new regression model:

$$Y = X_{new}\Theta + \epsilon \tag{6}$$

We use 10-fold CV to find the optimal number of components in terms of mean square error of prediction (MSEP). According to figure 5, we choose the first three components since it is within one standard error of the minimal MSEP and has a minimum dimension.

Figure 5: 10-fold CV on number of components of PCR.



The regression result of PCR is:

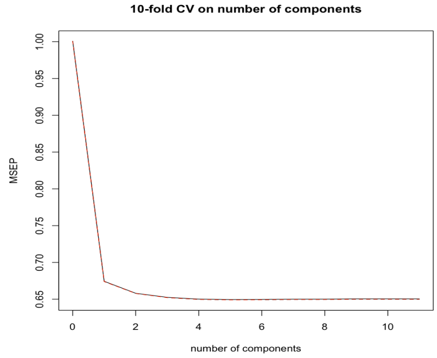
$$\begin{aligned} quality = {} & 0.022 fixed.acidity - 0.236 volatile.acidity \\ & + 0.148 citric.acid - 0.034 residual.sugar \\ & - 0.058 chlorides - 0.008 free.sulfur.dioxide \\ & - 0.054 total.sulfur.dioxide - 0.011 pH \\ & + 0.115 sulphates + 0.252 alcohol. \end{aligned} \tag{7}$$

While PCR makes projection on X, which dose not guarantee good regression effects on Y, PLSR tends to find projection basis by maximizing the covariance of X and Y.

$$w_1 := \underset{w:\ w^T w = 1}{\arg\max}\, Cov(Xw, Y). \tag{8}$$

Similarly, we use 10-fold CV to find the select the optimal number of components. In figure 6, we choose the first two components within one standard error of the minimal MSEP. And the regression result is:

Figure 6: 10-fold CV on number of components of PLSR.



$$\begin{aligned} quality = {} & 0.043 fixed.acidity + 0.308 alcohol \\ & + 0.082 citric.acid + 0.023 residual.sugar \\ & - 0.100 chlorides + 0.005 free.sulfur.dioxide \\ & - 0.102 total.sulfur.dioxide - 0.006 pH \\ & + 0.156 sulphates - 0.216 volatile.acidity. \end{aligned} \tag{9}$$

3.4 Penalized Linear Regression and Bootstrapping Elastic Net

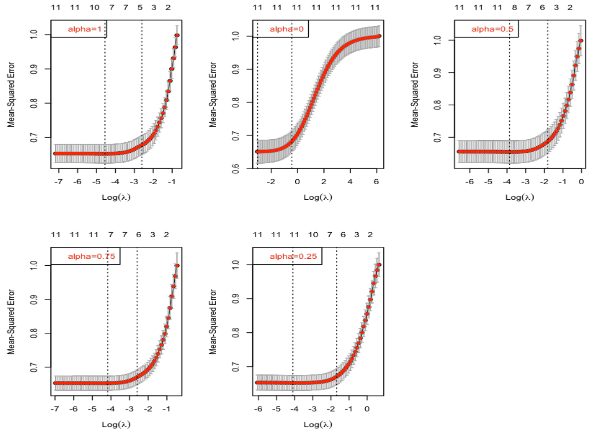
Another way to deal with multicollinearity is to impose some penalties on parameters, for example, the elastic net and meanwhile it prefers more sparse model.

$$\hat{\beta}^{e-net} := \underset{\beta \in \mathbb{R}^p}{\arg\min} \frac{1}{2n}\|Y - X\beta\|_2^2 + \lambda(\frac{1-\alpha}{2}\|\beta\|_2^2 + \alpha\|\beta\|_1), \tag{10}$$

where $\alpha = 0$ means Ridge and $\alpha = 1$ means LASSO.

For turning regularization parameters, we choose five α and use 10-fold CV. From figure 7, within one standard error of the minimal MSEP, we choose more sparse model, which is shown in the fourth picture ($\alpha = 0.75, \lambda = 0.108$). With this setting, the regression result of the elastic net on the data is:

Figure 7: 10-fold CV on the elastic net with $\alpha = 0, 0.25, 0.5, 0.75, 1$.



$$\begin{aligned} quality = &-0.213 volatile.acidity + 0.082 sulphates \\ &+ 0.330 alcohol - 0.022 total.sulfur.dioxide. \end{aligned} \tag{11}$$

The elastic net selects four predictors: volatile.acidity, total.sulfur.dioxide, sulphates and alcohol.

As for inference on the elastic net, we introduce Bootstrapping elastic net[2] to compute bias, standard error of estimators and construct confidence intervals. Before each step of Bootstrap, we need do hard-thresholding on the coefficients. Determine a sequence of real numbers:

$$t_n = cn^{-\delta}, \tag{12}$$

where $c > 0$ and $0 < \delta < \frac{1}{2}$ therefore cut off with this sequence:

$$\hat{\beta}_{n,j}^{ht} = \hat{\beta}_{n,j}\mathbf{1}(|\hat{\beta}_{n,j}| \geqslant t_n), \ j = 1, 2, ..., p. \tag{13}$$

We set the number of replicates of Boostrap $B = 500$ and the sequence $t_n = 0.001$. The table 1 shows the bias, the standard error and the confidence intervals of estimators of coefficients.

4 Classification

The above are all the methods regarding to regression and we also use two classification methods here. And before we build classification models, we first do some transformation to the response variable in order to get better results. If the value of wine quality is 3, 4 and 5 then we define it as low quality wine. If the value of wine quality is 6 then we define it as medium quality wine and if the value of wine is 7 or 8 then we define it as high quality wine.

4.1 Support Vector Machine

The first classification method is Support Vector Machine. The main idea of SVM is to try to find a boundary line between different data that has maximum margin based on different kernel functions. The formula of this model is[3]:

$$\textbf{classifier}: \ f(x, w, b) = sign(K(w, x) - b), \tag{14}$$

Table 1: Bootstrapping bias, standard error and confidence intervals ($\alpha = 0.05$).

Predictors	Bias	Standard Error	Lower Bound	Upper Bound
fixed.acidity	4.736634×10^{-6}	1.059144×10^{-4}	9.473268×10^{-6}	9.473268×10^{-6}
volatile.acidity	5.614928×10^{-2}	2.036924×10^{-2}	-0.1955180	-0.1127194
citric.acidity	8.273131×10^{-4}	4.054676×10^{-3}	-0.011000920	0.001654626
residual.sugar	0	0	0	0
chlorides	0	0	0	0
free.sulfur.dioxide	0	0	0	0
total.sulfur.dioxide	2.219938×10^{-2}	2.199153×10^{-3}	-0.0004942784	0.0012972997
density	-2.941979×10^{-6}	6.578465×10^{-5}	-5.883958×10^{-6}	-5.883958×10^{-6}
pH	0	0	0	0
sulphates	-6.043260×10^{-2}	1.773293×10^{-2}	-0.01735357	0.04352981
alcohol	-6.785598×10^{-2}	2.110711×10^{-2}	0.2176877	0.3015352

$$\max \sum_{k=1}^{n} \alpha_k - \frac{1}{2}\sum_{k=1}^{n}\sum_{l=1}^{n} \alpha_k \alpha_l Q_{kl}, \qquad (15)$$

$$s.t.\ 0 \le \alpha_k \le C,\ \sum_{k=1}^{n} \alpha_k y_k = 0,\ Q_{kl} = y_k y_l K(x_k, x_l). \qquad (16)$$

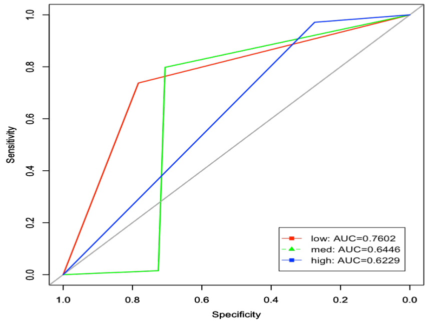
We first split the whole data set into train data set and test data set whose ratio is 3:1. Then we also use 10 fold cross validation here to help us choose the best kernel function. The result shows the best kernel function is:

$$\begin{gathered} Kernel : e^{-\gamma \|u-v\|_2^2} \\ \gamma = 0.0909 \end{gathered} \qquad (17)$$

It also shows that the total accuracy on train test is 0.6535. The total accuracy on test set is 0.6535.

Figure 8 shows another criterion names AUC (Area Under the Curve) to evaluate the accuracy of classification model. The larger AUC stands for better classification performance. From the figure 8, we can draw the conclusion that SVM performs better when classifying low quality wine with no low quality wine and perferms not so well in other two groups.

Figure 8: ROC curve and their AUC



4.2 Random Forest

Random Forest is generally better than SVM while SVM tends to overfit. The idea of Random Forest is when a node of the tree goes to different branches, it is based on two criteria: the information gain in terms of Kullback-Leibler divergence and the Gini index. Each branch will give scores to determine classification. There are two main parameters: m_{tree} the number of branches and n_{tree} the number of trees.

We use 10-fold CV on the train set turning parameters with the smallest OOB (out of bag) error. That is, $m_{tree} = 4$ and $n_{tree} = 500$. The figure 9 shows the variable importance and the figure 10 shows AUC of Random Forest. According to the figure 10, it suggests Random Forest has better AUC than SVM even within all types of wine quality ($\text{AUC}_{low} = 0.8249$, $\text{AUC}_{medium} = 0.887$, $\text{AUC}_{high} = 0.6999$).

Figure 9: Variable importance of Random Forest.

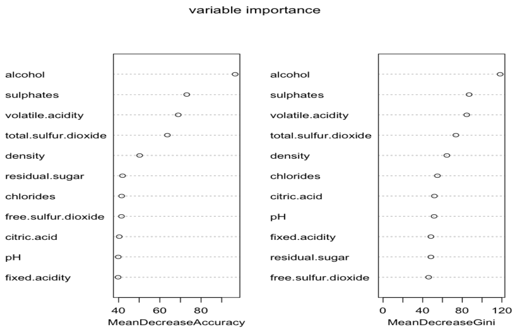
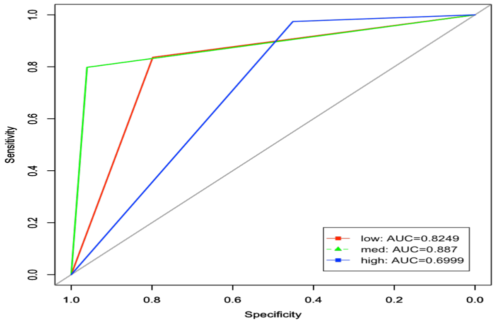


Figure 10: AUC of Random Forest.



5 Conclusion and Discussion

After comparing different models we finally choose two best regression model and one classification model.

For the regression part we choose stepwise method and penalized linear regression.

- Stepwise Regression

$$\begin{aligned} quality = &-0.23volatile.acidity - 0.12chlorides \\ &- 0.08pH - 0.10total.sulfur.dioxide \\ &+ 0.19sulphates + 0.38alcohol \end{aligned}$$

- Penalized Linear Regression:

$$\begin{aligned} quality = &\,0.004fixed.acidity + 0.34alcohol \\ &- 0.02chlorides - 0.04total.sulfur.dioxide \\ &+ 0.10sulphates - 0.22volatile.acidity \end{aligned}$$

For the classification part, we choose Random Forest with parameters $m_{tree} = 4$ and $n_{tree} = 500$.

As for statistical inference on the penalized linear regression model, according to the table 1, we can conclude coefficients of the predictors volatile.acidity and alcohol are significantly nonzero and for some zero-valued coefficients, Bootstrap can still give valid confidence intervals like citric.acidity.

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.

[2] Zhaoyang Shi, Lei Shi. Residual-based Bootstrap Methods on the Elastic Net. Unpublished transcripts and undergraduate essay.

[3] Gareth J. An introduction to statistical learning: with applications in R[M]. Springer Verlag, 2010.

[4] Zhao, Qi, Lei Su, Zhaoyang Shi, Ping Ling, Nannan Yan, Chunjie Gu, and Zhixiong Shi. "The Analysis of Features Importance in Electrical Infrared Images Faults Diagnosis." In Proceedings of the 1st International Workshop on Internet of People, Assistive Robots and Things, pp. 60-65. ACM, 2018.