Prediction of wine quality from its physicochemical properties

Xiaoying Wang and Sohom Mandal
Dept. of Applied Math, Dept. of Civil and Environmental Engineering
Western University

1 Introduction

Wine is a luxary alcoholic beverage.A significant amount of business revenue is invloved with this industry. The economical value depends on it's quality and quality of wine depends on physicochemical properties of ingredients. This analysis is important for vineyard to improve or predict wine quality based on its physicochemical characteristics.

For this study one red wine and one white wine datasets have been used were obtained from UCI Machine Learning Repository ([4]). The each dataset contains 12 variables, 11 among of them are explanatory variable and 1 variable is integer catagorical variable. There are 1599 samples for the red wine and 4898 samples for the white wine, no missing data was found.

2 Preliminary Data Analysis

The variables summary statistics for red wine are given in Tables 1, 2 and 3. The summary statistics for white wine data are shown in Tables 4, 5 and 6. From summary statistics, we can see that the quality of red wine ranges from 3 to 8 and the quality of white wine data is between 3 and 9.

	fixed.acidity	volatile.acidity	citric.acid	residual.sugar
1	Min. : 4.60	Min. :0.1200	Min. :0.000	Min. : 0.900
2	1st Qu.: 7.10	1st Qu.:0.3900	1st Qu.:0.090	1st Qu.: 1.900
3	Median : 7.90	Median :0.5200	Median :0.260	Median : 2.200
4	Mean : 8.32	Mean :0.5278	Mean :0.271	Mean : 2.539
5	3rd Qu.: 9.20	3rd Qu.:0.6400	3rd Qu.:0.420	3rd Qu.: 2.600
6	Max. :15.90	Max. :1.5800	Max. :1.000	Max. :15.500

Table 1: Input variables 1 to 4 for the red wine data

	chlorides	free.sulfur.dioxide	total.sulfur.dioxide	density
1	Min. :0.01200	Min. : 1.00	Min. : 6.00	Min. :0.9901
2	1st Qu.:0.07000	1st Qu.: 7.00	1st Qu.: 22.00	1st Qu.:0.9956
3	Median :0.07900	Median :14.00	Median : 38.00	Median :0.9968
4	Mean :0.08747	Mean :15.87	Mean : 46.47	Mean :0.9967
5	3rd Qu.:0.09000	3rd Qu.:21.00	3rd Qu.: 62.00	3rd Qu.:0.9978
6	Max. :0.61100	Max. :72.00	Max. :289.00	Max. :1.0037

Table 2: Input variables 5 to 8 for the red wine data

	pH	sulphates	alcohol	quality
1	Min. :2.740	Min. :0.3300	Min. : 8.40	Min. :3.000
2	1st Qu.:3.210	1st Qu.:0.5500	1st Qu.: 9.50	1st Qu.:5.000
3	Median :3.310	Median :0.6200	Median :10.20	Median :6.000
4	Mean :3.311	Mean :0.6581	Mean :10.42	Mean :5.636
5	3rd Qu.:3.400	3rd Qu.:0.7300	3rd Qu.:11.10	3rd Qu.:6.000
6	Max. :4.010	Max. :2.0000	Max. :14.90	Max. :8.000

Table 3: Input variables 9 to 12 for the red wine data

	fixed.acidity	volatile.acidity	citric.acid	residual.sugar
1	Min. : 3.800	Min. :0.0800	Min. :0.0000	Min. : 0.600
2	1st Qu.: 6.300	1st Qu.:0.2100	1st Qu.:0.2700	1st Qu.: 1.700
3	Median : 6.800	Median :0.2600	Median :0.3200	Median : 5.200
4	Mean : 6.855	Mean :0.2782	Mean :0.3342	Mean : 6.391
5	3rd Qu.: 7.300	3rd Qu.:0.3200	3rd Qu.:0.3900	3rd Qu.: 9.900
6	Max. :14.200	Max. :1.1000	Max. :1.6600	Max. :65.800

Table 4: Input variables 1 to 4 for the white wine data

	chlorides	free.sulfur.dioxide	total.sulfur.dioxide	density
1	Min. :0.00900	Min. : 2.00	Min. : 9.0	Min. :0.9871
2	1st Qu.:0.03600	1st Qu.: 23.00	1st Qu.:108.0	1st Qu.:0.9917
3	Median :0.04300	Median : 34.00	Median :134.0	Median :0.9937
4	Mean :0.04577	Mean : 35.31	Mean :138.4	Mean :0.9940
5	3rd Qu.:0.05000	3rd Qu.: 46.00	3rd Qu.:167.0	3rd Qu.:0.9961
6	Max. :0.34600	Max. :289.00	Max. :440.0	Max. :1.0390

Table 5: Input variables 5 to 8 for the white wine data

	pH	sulphates	alcohol	quality
1	Min. :2.720	Min. :0.2200	Min. : 8.00	Min. :3.000
2	1st Qu.:3.090	1st Qu.:0.4100	1st Qu.: 9.50	1st Qu.:5.000
3	Median :3.180	Median :0.4700	Median :10.40	Median :6.000
4	Mean :3.188	Mean :0.4898	Mean :10.51	Mean :5.878
5	3rd Qu.:3.280	3rd Qu.:0.5500	3rd Qu.:11.40	3rd Qu.:6.000
6	Max. :3.820	Max. :1.0800	Max. :14.20	Max. :9.000

Table 6: Input variables 9 to 12 for the white wine data

3 Materials and Methods

3.1 LDA, KNN and Multinomial Logistic Regression

In this section, we apply three different classification methods, LDA, KNN, Multinomial Logistic Regression to both red and white wine data. LDA classifier is a common choice when the response variable has more than two categories. LDA method assumes each predictor x_i, where $i = 1 \text{ to } 11$, has the Gaussian distribution and $X = (x_1, x_2, ..., x_{11})$ follows the multivariate Gaussian distribution. It then uses the linear discriminant functions to approximate the Bayes classifier. Next, we try KNN method. Unlike LDA method, KNN method is a non-parametric method. KNN classifier first identifies K nearest points to the certain point and then predict the conditional probability based on these points. We also explore Multinomial Logistic Regression to our wine data. Multinomial Logistic Regression is similar to binary-case logistic regression. The difference is that instead of predicting probability of certain observation in each category in binary-case logistic regression, multinomial logistic regression takes one category as the base one, and then calculates the relative probability compared with the base category. In addition, we combine ridge regression with multinomial logistic regression in order to make better predictions. Ridge regression uses the L_1 penalty to shrink the parameter values and reduce dimensions. For this part, we employ glmnet R package to do analysis. Before we apply any method, the data are scaled first.

3.2 K-means clustering, PCA and Regression Tree

Acid is a measure of H^+ ions concentration in a solution, on the other hand pH of a solution measured by $-\log[H^+]$ ion concentration. Actually if pH of a solution is less than 7, it's consider as acid or more than 7 is consider as base. PH and acidity are correlated among each

other. 3 among 11 explanatory variables are acid and other variables are also correlated with pH. Therefore for this analysis pH has used as a dependent variable or predictand variable where other variables (e.g. fixed acidity, volatile acidity, density etc) were independent variable or predictors. In this study, 2/3 of red wine data is used as a training data and rest is for testing purpose.

According to Kodur (2011) ([1]), if pH is low ($e.g. < 3.0$), quality of wine will be better than if pH is high ($e.g. > 3.8$) in the grape juice. On account of this condition, K-means clustering is used for clustering data. The clusters are made on basis of pH. As predictor data and response data are highly co-related with each other, to avoid multi-dimensionality and multi-collinearity, Principal Component Analysis (PCA) has applied in this work. After PCA analysis, regression tree is used for predicting wine quality. The categorical variable data is not used for this part of analysis.

4 Results

4.1 Results of LDA, KNN and Multinomial Logistic Regression

In this section, we present the results of LDA, KNN and Multinomial Logistic Regression on wine data. After applying lda classier on red wine data, we obtain the confusion matrix as below Table 7. The horizontal axis is the true classification and the vertical axis is the prediction.

	3	4	5	6	7	8
3	2	4	6	2	0	0
4	0	1	5	4	0	0
5	7	30	504	202	13	0
6	1	16	158	370	108	9
7	0	2	8	59	78	9
8	0	0	0	1	0	0

Table 7: Red Wine: Confusion Matrix of LDA Method

We use 10-fold cross-validation to calculate the prediction error on test data. As shown below, the test error rate of LDA method applying on the red wine data is:

```
[1] 0.4027517.
```

That's saying, by using LDA classifier, we can make around $1 - 0.4027517 = 0.5972483$ correct prediction on test data. Similarly, we get results of white wine data. The confusion matrix of LDA classifier on white wine data shows as Table 8.

	3	4	5	6	7	8	9
3	4	2	3	1	0	0	0
4	2	34	44	18	0	2	0
5	5	66	725	393	40	4	0
6	9	58	671	1598	595	103	1
7	0	3	13	185	245	66	4
8	0	0	1	0	0	0	0
9	0	0	0	3	0	0	0

Table 8: White Wine: Confusion Matrix of LDA Method

Again, we use 10-fold cross-validation to estimate the test error of LDA classifier on white wine data, which is shown below:

```
[1] 0.4687628.
```

For KNN method, first, we choose $k = 1$, where the first nearest point is identified to make prediction. We use 10-fold cross-validation method to estimate test errors. The test error for KNN classifier on red wine data is

```
[1] 0.1163227
```

and on white wine data is

```
[1] 0.1075949.
```

Varying the value of k, we get cross-validation errors as listed below. The first list shows the test errors for red wine and the second one indicates test errors for white wine. Here, we choose k values from 1 to 20. We visualize the test error with respect to different k values, as indicated in Figures 1 and 2. Based on Figures 1 and 2, we can say that increasing the value of k will not improve the prediction too much. Cross-validation errors of red wine data are:

```
 [1] 0.1163227 0.1663540 0.1500938 0.1638524 0.1644778 0.1694809 0.1682301
 [8] 0.1663540 0.1776110 0.1726079 0.1682301 0.1732333 0.1757348 0.1763602
[15] 0.1719825 0.1726079 0.1732333 0.1751094 0.1751094 0.1719825
```

Cross validation errors of white wine are:

```
 [1] 0.1075949 0.1455696 0.1380155 0.1351572 0.1380155 0.1443446 0.1463863
 [8] 0.1496529 0.1518987 0.1584320 0.1606778 0.1608820 0.1649653 0.1686403
[15] 0.1692528 0.1678236 0.1665986 0.1698653 0.1753777 0.1757860
```

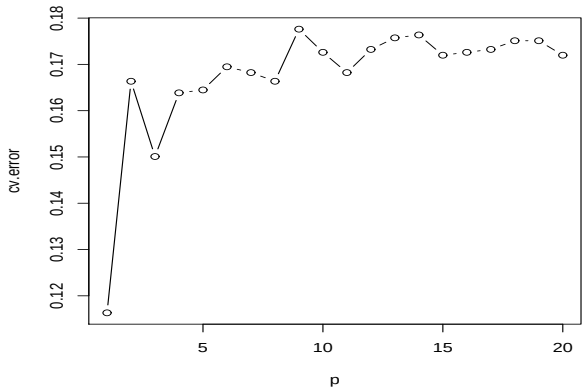

Figure 1: Red Wine: Varying the value of $p = k$, the cv.error will change.

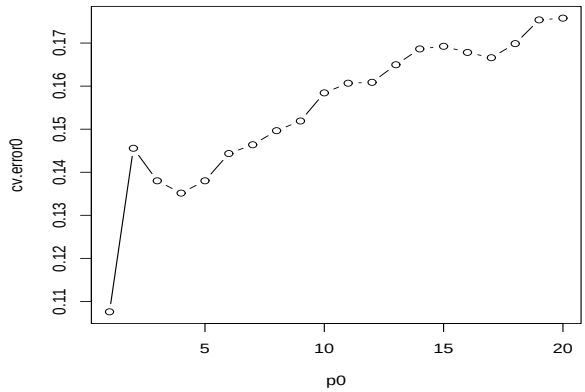

Figure 2: White Wine: Varying the value of $p_0 = k$, the cv.error will change.

For Multinomial Logistic Regression with L_1 penalty, the value of the tuning parameter λ will affect the prediction outcomes. Here, we use cross-validation to estimate the best λ value. As indicated in Figure 3 and Figure 4, we can see that when the value of λ changes, the cross-validation error will change. We choose the best λ value by choosing the one such that the cross-validation error is minimized. For red wine data, the best λ is

```
[1] 0.01891325
```

and for white wine data, the best λ is

```
[1] 0.01675522.
```

The confusion matrix of red wine data and white wine data by using multinomial logistic regression are presented as Table 9 and Table 10 respectively. The correct prediction rate for red wine data and white wine data are

```
[1] 0.63125
```

and

```
[1] 0.5357289
```

respetively.

	3	4	5	6	7	8
4	1	0	0	0	0	0
5	3	13	264	101	4	0
6	1	5	74	217	73	6
7	0	1	0	9	24	4

Table 9: Red Wine: Confusion Matrix of Multinomial Logistic Regression with Penalty

	3	4	5	6	7	8	9
5	4	49	396	210	21	3	0
6	6	25	359	856	345	56	0
7	0	0	3	45	60	9	2

Table 10: White Wine: Confusion Matrix of Multinomial Logistic Regression with Penalty

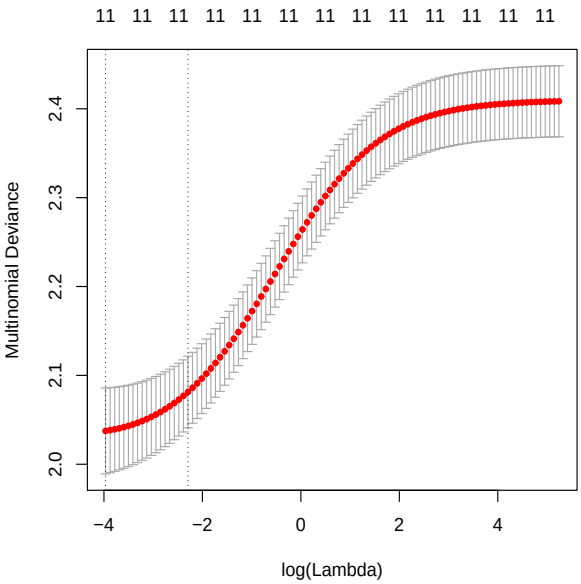

Figure 3: Red Wine: Varying the value of the tuning parameter λ, the deviance will change

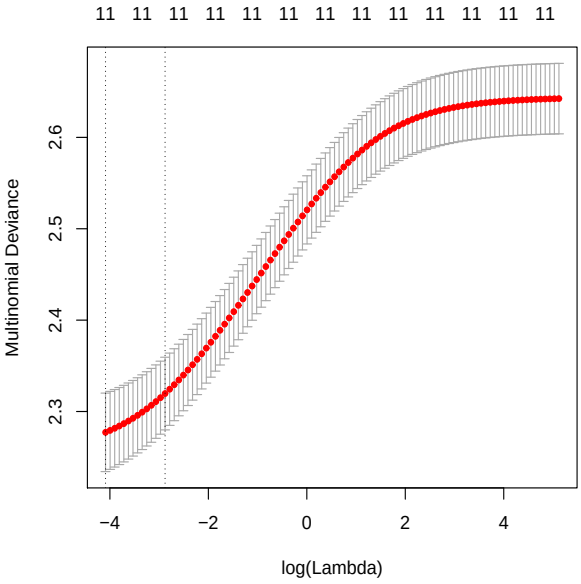

Figure 4: White Wine: Varying the value of the tuning parameter λ, the deviance will change

4.2 Results of K-means clustering, PCA and Regression Tree

K-means Clustering: The thumb rule of finding number clusters is equal to $\sqrt{\frac{n}{2}}$ where n is the number of data points. For red wine data set which is 28 and 50 for white wine data. But optimum number of cluster is not known. The optimum number of clusters is decided on basis of cluster validity index. This cluster algorithm gives classified cluster identification (ID) in each runs which are used for computation of cluster validity index. In this case study, internal validity measures are used which measures compactness, connectedness, and separation of clusters ([2]). Three validation measures Dunn Index, Silhoutte Index and connectivity are computed for each clusters for internal validity measures. Figure 5 shows plot of various validity indices computed against the number of clusters used for clustering.

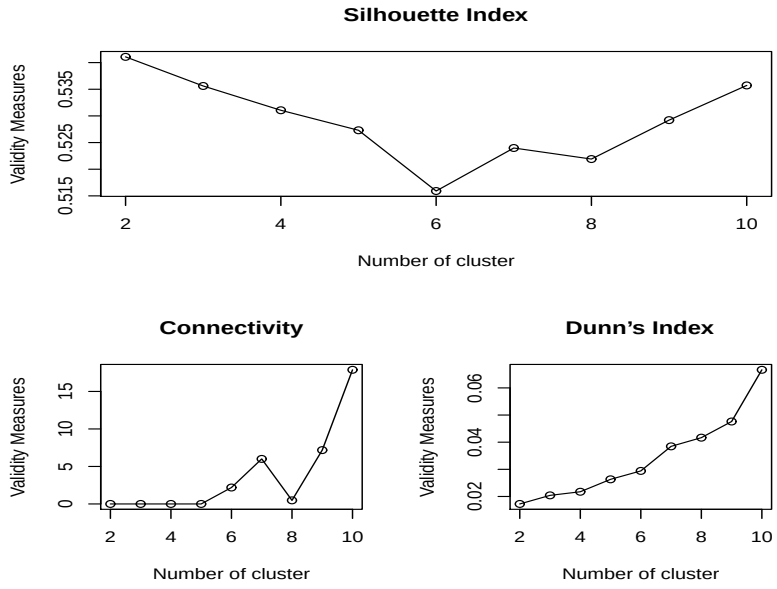

Figure 5: Cluster validity indexes

For optimum clusters, Dunn Index and Silhouette Index should be maximized where connectivity should be minimized. From the Figure 5 and on account of Kour's theory ([1]) optimal clusters for this case is taken as 3.

```
> k2=matrix(k1$center)

      [,1]
1 3.524732
2 3.130541
3 3.321922
```

After computing k=3 clusters, means of clusters are 3.52, 3.32, 3.13. It can be concluded, vectors in clusters means 3.13 indicate excellent quality of wine than other two clusters because wine quality will be good if pH is less.

Principal Component Analysis (PCA): To reduce multicollinearity and dimensionality PCA is performed. PCA is used separately for training and testing data. The benefit

of PCA, it is possible to represent variability among dataset by using a small number of principal components. Number of principal component is decided based on percentage of total variability in principal components. The percentage of total variance (w_k) explained by kth principal component is given by:

$$w_k = \frac{100\,\lambda_k}{\sum_{m=1}^{M} \lambda_m},$$

where λ_k is kth eigenvalue and M is dimensionality of original dataset. In this study we consider 97% as threshold (i.e., the total percentage of the variability carried by the selected principal components should be more than or equal to 97% of the original data). Figure 6 shows percentage of variance for training and testing data. It has found that first eight principal components represent more than 97% variability of both training and testing datasets.

```
> per_var_tr

 Comp.1  Comp.2  Comp.3  Comp.4  Comp.5  Comp.6  Comp.7  Comp.8  Comp.9
  25.95   44.46   59.32   72.48   81.97   88.16   93.68   97.03   98.77
Comp.10
 100.00

> per_var_test

 Comp.1  Comp.2  Comp.3  Comp.4  Comp.5  Comp.6  Comp.7  Comp.8  Comp.9
  26.47   46.92   63.28   73.77   82.37   88.53   93.15   96.79   98.75
Comp.10
 100.00
```

Therefore we build regression trees based on first eight principal components of red wine data. It is found from summary of regression tree fitting; five principal components (comp.1, comp.4, comp.5, comp.6, comp.7) are used in constructing tree.

```
> summary(tree.fit)

Regression tree:
tree(formula = pH ~ da[, 1] + da[, 2] + da[, 3] + da[, 4] + da[,
    5] + da[, 6] + da[, 7] + da[, 8], data = da)
Variables actually used in tree construction:
[1] "da[, 1]" "da[, 5]" "da[, 7]" "da[, 4]" "da[, 6]"
Number of terminal nodes:  10
Residual mean deviance:  0.01251 = 13.21 / 1056
Distribution of residuals:
     Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
-0.388600 -0.075260 -0.004615  0.000000  0.069150  0.369100
```

The tree indicates that lower value of component 1 ($da[,1] < 1.039$) (Figure 7) correspondent to worst wine quality.

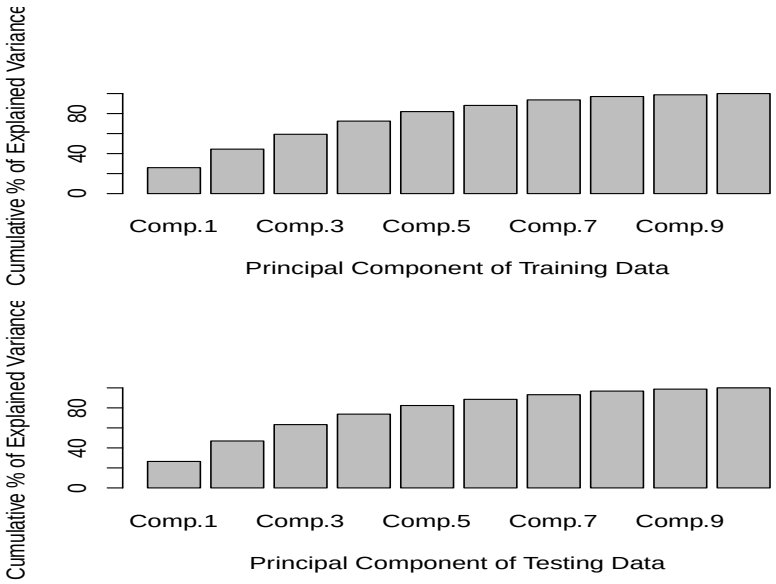

Figure 6: Percentage of Total Variance Vs Principle Components

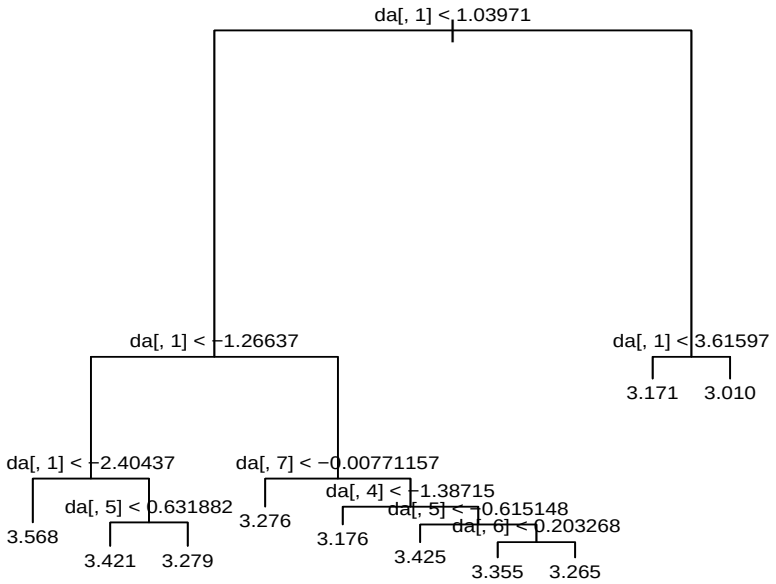

Figure 7: Regression Tree of Test Data

```
> mean((tree.pred-da1)^2)

[1] 12.09435
```

We use this regression tree to predict wine quality from test data. MSE associated from tree is 12.099. The square root of MSE is 3.47 which indicates the best wine quality we can get below or equal to this pH value. Above this ($pH > 3.47$) the quality of wine will be deteriorated.

References

[1] Kour S., Effects of juice pH and potassium on juice and wine quality , and regulation of potassium in grapevines through rootstocks (Vitis): a short review, J. Grapevine Res., 50(1) 1-6, 2011.

[2] Brock G, Pihur V, Datta S, Datta S. clValid , an R package for cluster validation. J. Stat. Softw., 1-32, 2011.

[3] Gareth James, Daniela Witten, Trevor Hastie and Robert Tibshirani, *An Introduction to Statistical Learning with Applications in R*, Springer, 2013.

[4] 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.