Tingran Yang, Zihao Zhou

Professor Predrag Jelenkovic

EECS6690 Statistical Learning

8 May 2018

Modeling Wine Preferences by Data Mining from Physicochemical Properties

1. Introduction

Wine is one of the most favorite drinks all over the world, and Portugal is one of the most largest wine exporting country, with 3.17% of the market share in 2005 [1]. As wine industry grows, wine assessment industry become more and more important since it could prevent adulteration and assure wine quality [1]. In the meantime, wine assessment could distinguish wine quality to rate premium brands [1].

However, wine assessment by physicochemical properties is difficult. Taste is the most complex human sensory to understand and it is hard to set for a universal standard. Assessment can vary a lot due to oenologist's health situation or mood.

Recently, as the information technologies develop, technology to assess wine quality has been invested on the base of oenologist's evaluation [1]. In supervised machine learning, we could put physicochemical properties as input and apply classic machine learning model, such as Multilinear Regression (MR), Neural Network (NN), and Support Vector Machine (SVM). Wine assessment technology could provides reference to oenologist and facilitate to make a quick judgement for wine quality [1].

2. Original Data Set and Paper

In order to model wine quality based on physicochemical properties, we chose a data set from UCI Machine Learning Repository. It is composed of two datasets, one related to red variants of the Portuguese “Vinho Verde” wine, and the other related to white variants of the Portuguese “Vinho Verde” wine [1]. There are 1599 entries in red wine dataset and 4898 entries in white wine dataset. Each dataset has a 12 input variables: (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, (12) quality. In these variables, the last one, quality, is the response variable and the others are predictors. The predictors are all numeric. The quality can either be seen as a numeric or classified variable. So we could either do regression or classification. The data is ordered and not balanced. Also, the variables could be irrelevant. So some techniques could be applied. [1]

The original paper Modeling Wine Preferences by Data Mining from Physicochemical Properties by Cortez et al which we referred proposed a data mining method to predict human taste preference for wine [1]. In the paper variable selection is used to filter out irrelevant variables and MR, NN and SVM were applied to predict wine quality.

3. Reproduce

Firstly, we used the same datasets that were used in the paper. Regression approach was used, and the responses preserve the order of the wine quality. MR, NN and SVM model were applied in the paper.

Generally, we reproduce the following methodology and results:

a. Variable Selection

As variables might be irrelevant to wine quality. We need to know which variables are the most important factors in determining and which are less important. We would like to preclude the irrelevant predictors to pruning our model.

To calculate the importance of the variables, we tested each predictor by taking mean of other predictors and then train model and get predictions. Then we calculate the variance of predictors. Higher variance indicates higher importance. The importance calculation can be given by the following formula:

$$V_a = \sum_{j=1}^{L} \left(\hat{y}_{a_j} - \overline{\hat{y}_{a_j}}\right)^2 / (L-1)$$
$$R_a = V_a / \sum_{i=1}^{I} V_i \times 100(\%)$$

Where Va is the variance of the responses, Ra is the percentage of one variance that takes account.

So here is the algorithm of variable selection [1]:

1) Using holdout to split training data and test data, in our trial, we randomly split ⅔ of data as training data and ⅓ as test data.
2) While training data has at least one predictor:
    a) Test each parameter in hyperparameter until the generalization estimate decreases
    b) For this parameter, calculate the least relevant predictors (least important), and then discard it from the training data
    c) Test if generalization estimate gets better. If better, repeat step 2. Else, return the previous model and the generalization estimate.

In this algorithm, we regard the accuracy of prediction as our generalization estimate (But later in the section 5 we will talk about why this is not a good criterion). To specify for the hyperparameter, as we are going to use SVM and NN model. The number H of nodes in hidden layer, and the γ in the kernel function of kernel SVM: $K(x,x') = \exp(-\gamma||x-x'||2)$, $\gamma > 0$ are very crucial for the performance of NN and kernel SVM model. So we also need to find the best parameter during the variable selection algorithm. MR does not have hyperparameters.

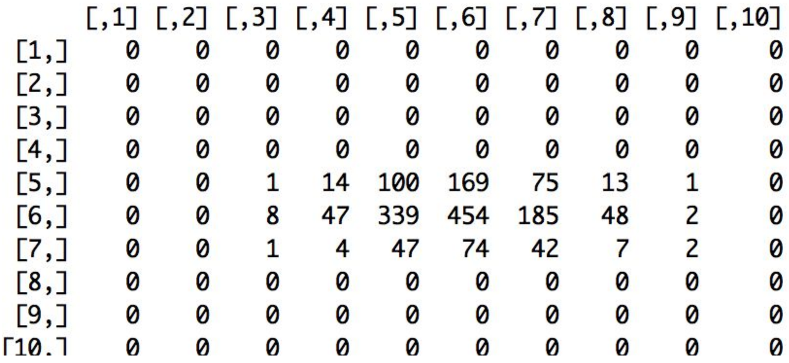
b. MR, SVM and NN model construction and test.

We used functions in library Rminer as were used in paper, which is a very convenient library to facilitate the use of Data Mining techniques in classification and regression tasks [1]. In the paper, MR, SVM and NN were implemented.

We implemented these model on both Red Wine and White Wine based on variable selection. First we scaled the predictors, making them have 0 mean and 1 standard deviation. Then we applied the functions and here is the result for Red Wine:

a) NN

For NN, H = 2 is the best parameter. This is the Confusion Matrix.

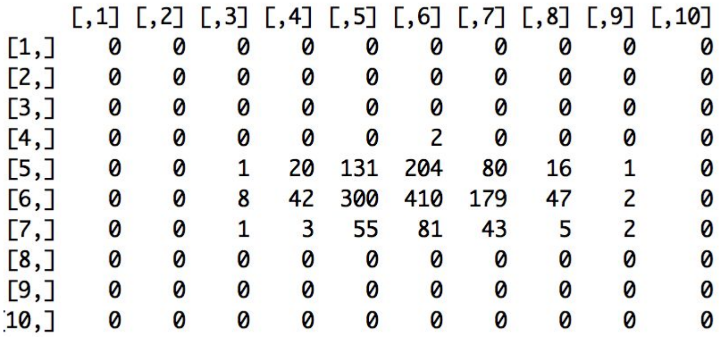


```
      [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8] [,9] [,10]
 [1,]    0    0    0    0    0    0    0    0    0     0
 [2,]    0    0    0    0    0    0    0    0    0     0
 [3,]    0    0    0    0    0    0    0    0    0     0
 [4,]    0    0    0    0    0    0    0    0    0     0
 [5,]    0    0    1   14  100  169   75   13    1     0
 [6,]    0    0    8   47  339  454  185   48    2     0
 [7,]    0    0    1    4   47   74   42    7    2     0
 [8,]    0    0    0    0    0    0    0    0    0     0
 [9,]    0    0    0    0    0    0    0    0    0     0
[10,]    0    0    0    0    0    0    0    0    0     0
```

Fig. 1

b) SVM

For SVM, γ = 0.00003 is the best parameter. This is the confusion matrix.

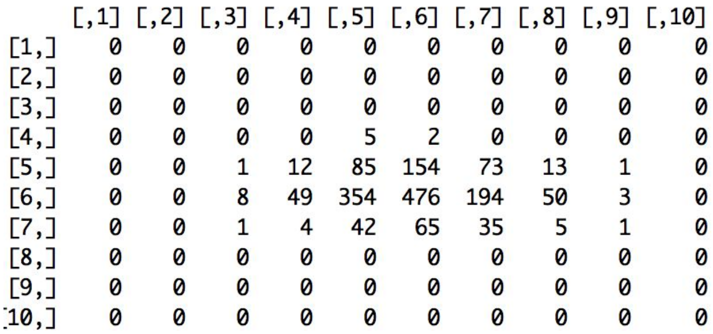


```
      [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8] [,9] [,10]
 [1,]    0    0    0    0    0    0    0    0    0     0
 [2,]    0    0    0    0    0    0    0    0    0     0
 [3,]    0    0    0    0    0    0    0    0    0     0
 [4,]    0    0    0    0    0    2    0    0    0     0
 [5,]    0    0    1   20  131  204   80   16    1     0
 [6,]    0    0    8   42  300  410  179   47    2     0
 [7,]    0    0    1    3   55   81   43    5    2     0
 [8,]    0    0    0    0    0    0    0    0    0     0
 [9,]    0    0    0    0    0    0    0    0    0     0
 10,]    0    0    0    0    0    0    0    0    0     0
```

Fig. 2

c) MR

For MR, This is the confusion matrix.



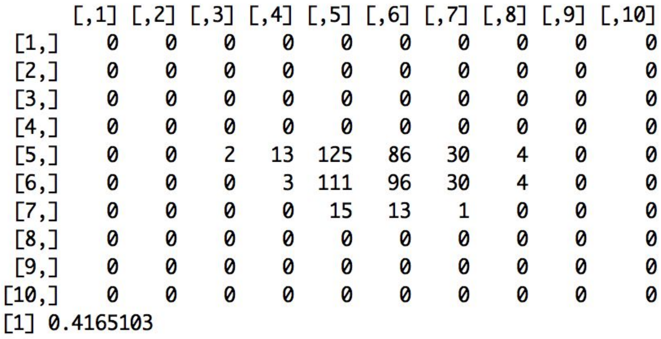
```
      [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8] [,9] [,10]
 [1,]    0    0    0    0    0    0    0    0    0     0
 [2,]    0    0    0    0    0    0    0    0    0     0
 [3,]    0    0    0    0    0    0    0    0    0     0
 [4,]    0    0    0    0    5    2    0    0    0     0
 [5,]    0    0    1   12   85  154   73   13    1     0
 [6,]    0    0    8   49  354  476  194   50    3     0
 [7,]    0    0    1    4   42   65   35    5    1     0
 [8,]    0    0    0    0    0    0    0    0    0     0
 [9,]    0    0    0    0    0    0    0    0    0     0
 10,]    0    0    0    0    0    0    0    0    0     0
```

Fig. 3

And this is the results for White Wine.

d) NN

For NN, H = 2 is the best parameter. This is the Confusion Matrix.

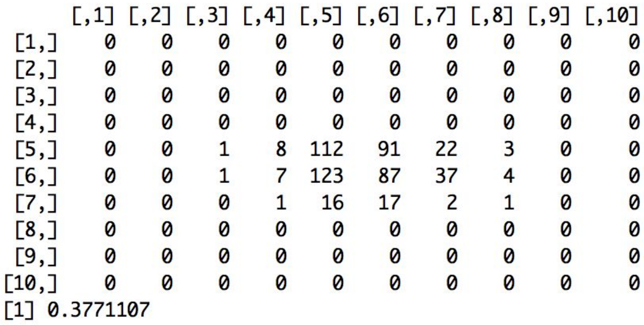


```
      [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8] [,9] [,10]
 [1,]    0    0    0    0    0    0    0    0    0     0
 [2,]    0    0    0    0    0    0    0    0    0     0
 [3,]    0    0    0    0    0    0    0    0    0     0
 [4,]    0    0    0    0    0    0    0    0    0     0
 [5,]    0    0    2   13  125   86   30    4    0     0
 [6,]    0    0    0    3  111   96   30    4    0     0
 [7,]    0    0    0    0   15   13    1    0    0     0
 [8,]    0    0    0    0    0    0    0    0    0     0
 [9,]    0    0    0    0    0    0    0    0    0     0
[10,]    0    0    0    0    0    0    0    0    0     0
[1] 0.4165103
```

Fig. 4

e) SVM

For SVM, γ = 0.0078 is the best parameter. This is the confusion matrix.

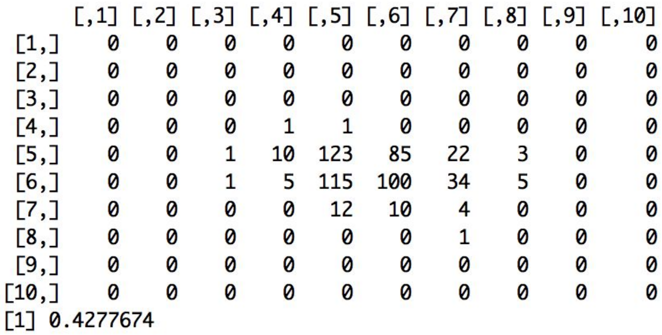


```
      [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8] [,9] [,10]
 [1,]    0    0    0    0    0    0    0    0    0     0
 [2,]    0    0    0    0    0    0    0    0    0     0
 [3,]    0    0    0    0    0    0    0    0    0     0
 [4,]    0    0    0    0    0    0    0    0    0     0
 [5,]    0    0    1    8  112   91   22    3    0     0
 [6,]    0    0    1    7  123   87   37    4    0     0
 [7,]    0    0    0    1   16   17    2    1    0     0
 [8,]    0    0    0    0    0    0    0    0    0     0
 [9,]    0    0    0    0    0    0    0    0    0     0
[10,]    0    0    0    0    0    0    0    0    0     0
[1] 0.3771107
```

Fig. 5

f) MR

For MR, This is the confusion matrix.



```
      [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8] [,9] [,10]
 [1,]    0    0    0    0    0    0    0    0    0     0
 [2,]    0    0    0    0    0    0    0    0    0     0
 [3,]    0    0    0    0    0    0    0    0    0     0
 [4,]    0    0    0    1    1    0    0    0    0     0
 [5,]    0    0    1   10  123   85   22    3    0     0
 [6,]    0    0    1    5  115  100   34    5    0     0
 [7,]    0    0    0    0   12   10    4    0    0     0
 [8,]    0    0    0    0    0    0    1    0    0     0
 [9,]    0    0    0    0    0    0    0    0    0     0
[10,]    0    0    0    0    0    0    0    0    0     0
[1] 0.4277674
```

Fig. 6

For better visualization, we also plotted Regression Error Characteristic (REC) curve [2] for Red Wine. REC curves generalize ROC curves to regression [2]. The x-axis is the tolerance of absolute deviation from the target y and y-axis is the accuracy. When tolerance = 0, which means every prediction must be exactly equal to target y to be labeled as correct. Under this hypothesis, accuracy will be 0. As tolerance increases, accuracy increases. According to be definition, the closer the curse is near the top-left corner, the better the regression model is, which has a similar interpretation of ROC curves. So these are REC curve of NN, SVM and MR model respectively.

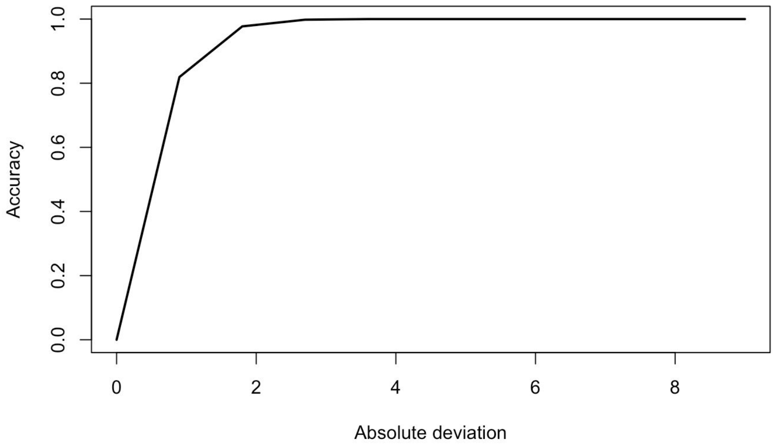


Fig. 7

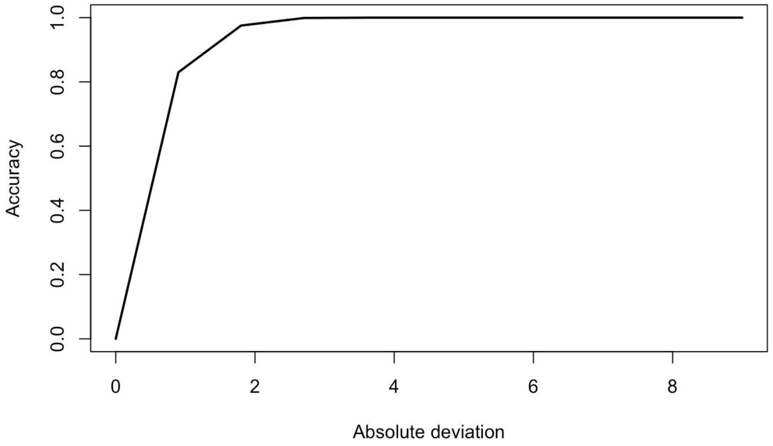


Fig. 8

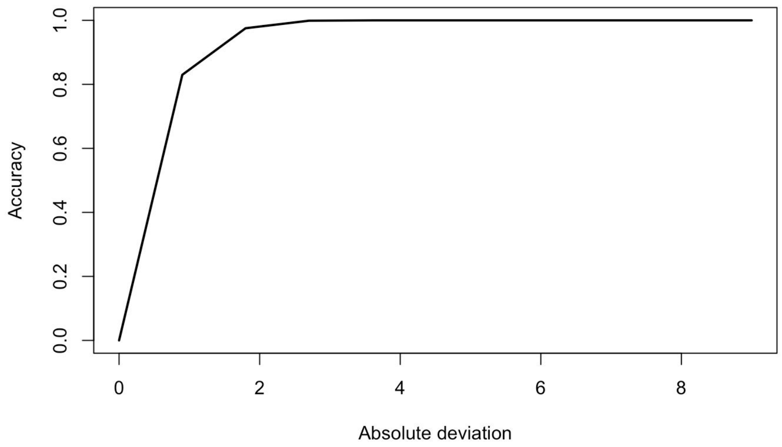


Fig. 9

As we could observe that the REC curves are almost the same. So according to our trial the three models have similar performances for Red Wine dataset.

c. Variable Importance Diagram

During the variable selection algorithm, we also want to see the importance of the predictors in determining the wine quality and discard the least relevant variables to get a better model. We plotted the variable importance of Red wine dataset using NN model.

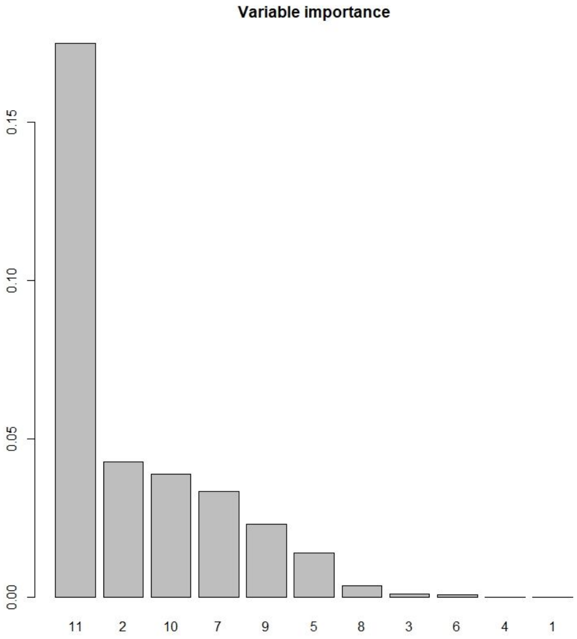


Fig. 10

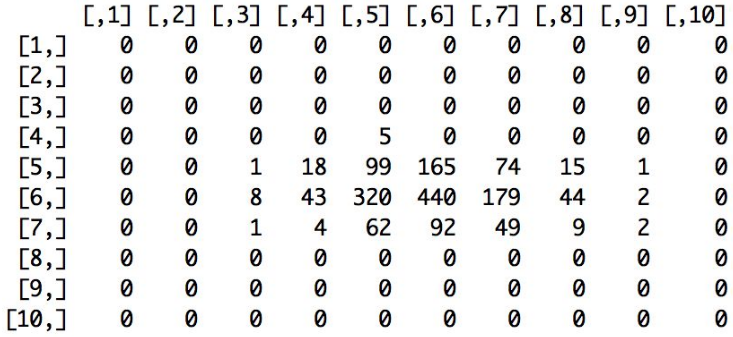
As we could see that the top 3 important variables are alcohol (11), volatile acidity (2), sulphates (10), and the least 3 important variables are fixed acidity (1), residual sugar (4), free sulfur dioxide (6), which is the similar conclusion on the paper.

4. New Approach

In addition to the reproduction of the methods on the paper, we tried several new methods.

a. Decision tree model:

RMiner support a number of models from machine learning. Here we tried to use another popular model, Decision tree. The following figure shows a confusion matrix using this model using red wine dataset. It’s accuracy is 36.0%, much lower than those previous mentioned models.



```
      [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8] [,9] [,10]
 [1,]    0    0    0    0    0    0    0    0    0     0
 [2,]    0    0    0    0    0    0    0    0    0     0
 [3,]    0    0    0    0    0    0    0    0    0     0
 [4,]    0    0    0    0    5    0    0    0    0     0
 [5,]    0    0    1   18   99  165   74   15    1     0
 [6,]    0    0    8   43  320  440  179   44    2     0
 [7,]    0    0    1    4   62   92   49    9    2     0
 [8,]    0    0    0    0    0    0    0    0    0     0
 [9,]    0    0    0    0    0    0    0    0    0     0
[10,]    0    0    0    0    0    0    0    0    0     0
```

Fig. 11

b. Implement cross-validation in the variable selection:

A holdout method was used in the paper during variable selection. We think it can be improved by using cross-validation instead of holdout because cross validation can be more stable. Here is the algorithm we created:

```
1) Use cross validation to get the precision as evaluation,
2) While training data has at least one predictor:
    a) Test each parameter in hyperparameters
    b) Record the best parameter according to evaluation, calculate the least
       relevant predictors (least weight), and delete it from training data.
3) Find the best evaluation value in all, and return the corresponding parameter,
   precision and variable names.
```

Using this algorithm is rather successful. For NN, we get a better accuracy of 60% and the optimal number of nodes is 3 in contrast to 2 using holdout method.

c. Plotted ROC curve for binary classification

We attempted to apply a binary classification trial to evaluate the models by plotting ROC curves. So we labeled target and predictions that are 6 as 1 and the others as 0. Then we plotted the ROC curve of red wine dataset.

For NN,

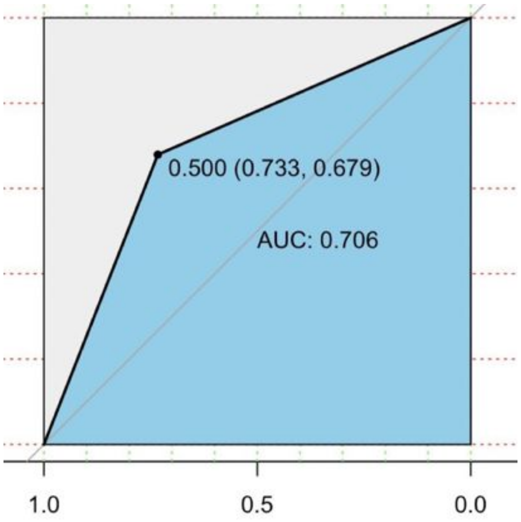


Fig. 12

For MR,

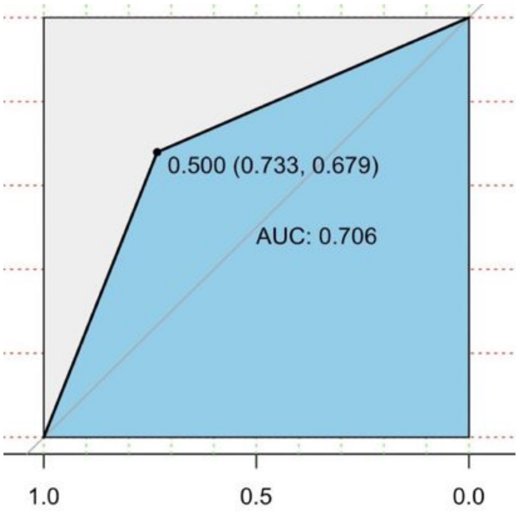


Fig. 13

For SVM

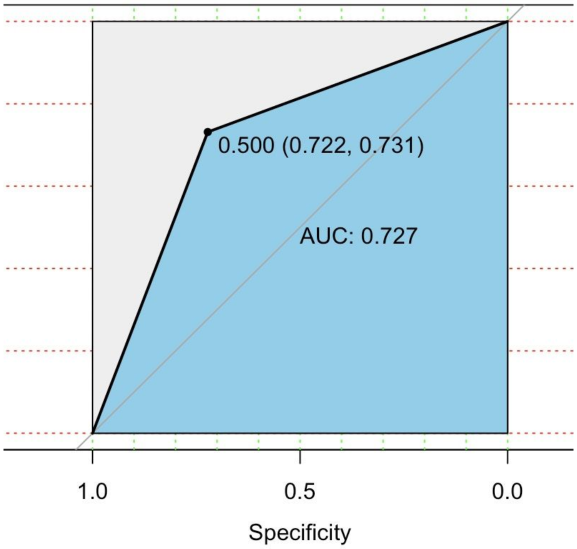


Fig. 14

In this trial, we could see that AUC of SVM is slightly better than MR and NN, which indicates SVM is a good choice over the other ones. However, we later think this is not a valid evaluation (discussed in section 5).

5. Discussion and Conclusion

We have tried many methods to analyze the datasets, both by reproducing the methods in papers and developing our own approach.

We reproduced the variable selection and applied NN, SVM and MR model using rminer. The way we dealt with prediction is that we rounded the prediction into a classification and then Compared with the target, which could result in low accuracy because the wine quality is an ordinal variables [3], so it can be tricky to choose whether ceiling or floor to round. So simply analyzing the accuracy by comparing the rounded prediction with the target is not a good method. A better way to analyze is to calculate the MAD [1]:

$$MAD = \sum_{i=1}^{N} |y_i - \hat{y}_i| / N$$

As a regression method to deal with ordinal responses, calculating MAD is viable and can reflect the how predictions deviate from target.

Then we chose to draw the REC curves as were done in the paper. We could see that in our REC curves, when T = 0.5 the accuracy is around 50% and when T = 1.0 the accuracy is around 80%. Which is similar to the results in paper. However, our results do not distinguish the difference among SVM, NN, MR while it indicates that SVM has a slightly better performance over NN and MR in paper [1].

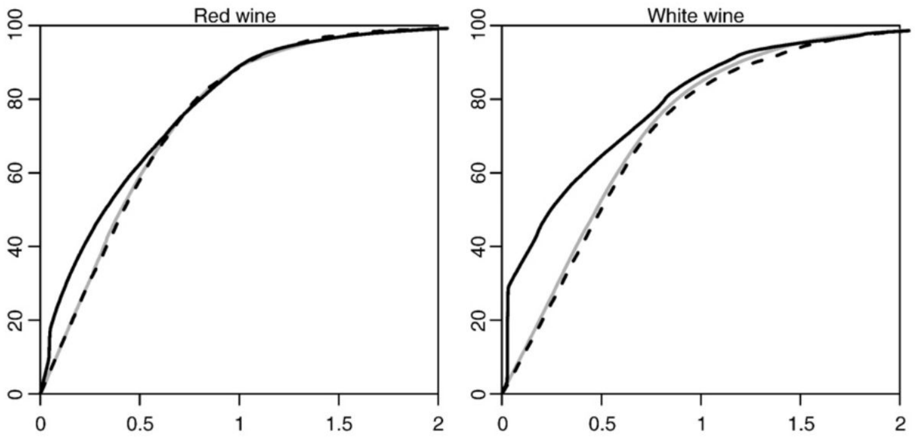


Fig. 15

The reason that we did not get that result might be that the hyperparameter we chose and the seed we set are not different. Another important reason might be that we chose accuracy as generalization estimates while the paper chose MAD. We can try more seeds and do MAD in the future.

We also did a K-fold cross-validation to find the best hyperparameter and the important variables. And our trials shows that it is a good approach to increase accuracy. However, it will take much more time since cross-validation can be time consuming if K is big. So here is a trade-off to wisely choose how many rounds of cross-are supposed to do.

We attempted to draw nice ROC curves but we did not find a way to do this for multiple classification. So we tried to create a meaningful binary classification subquestion. However, the probability of the acceptance of classification was not available in this question. So we only found one valid point in ROC curve. So at last we think this is not a valid approach to analyze the datasets.

In the future, we still find a lot to discover and analyze on the datasets. Firstly we could reproduce the results using MAD as generalization estimate, which we believe might be better than using accuracy. Secondly, we could do more statistical test and applying more models.

Works Cited

[1] P. Cortez, A. Cerdeira, F. Almeida, T. Matos and J. Reis. Modeling wine preferences by data mining from physicochemical properties

[2] J. Bi, K. Bennett, Regression error characteristic curves, Proceedings of 20th Int. Conf. on Machine Learning (ICML), Washington DC, USA, 2003.

[3] WHAT IS THE DIFFERENCE BETWEEN CATEGORICAL, ORDINAL AND INTERVAL VARIABLES? From: https://stats.idre.ucla.edu/other/mult-pkg/whatstat/what-is-the-difference-between-categorical-ordinal-and-interval-variables/