ADA Project “White Wine Quality”

Page 1.

Hello everyone, today our project is about “White Wine Quality”. Our group members are Mengtian Song, Qing Dong, Wenxin Liang, Yuying Chen and

Yi Jiang.

Page 2.

Nowadays more and more consumers will chose white wine as their favorite luxury good. To safeguard human health and to develop the technologies of wine making, white wine quality evaluation become more and more important.

Page 3.

The objective of our “White wine quality” project is to find out which predictors have more influence on determining the quality of white wine and to find out the capable models of prediction quality of the white wine

Page 4.

Our data comes from UCI Machine Learning Repository, in the following website you can find the original data we use. From the original data, we can observe that the inputs, fixed acidity, volatile acidity, citric acid, residual sugar, chlorides, free sulfur dioxide, total sulfur dioxide, density, pH, sulphates, and alcohol are numeric and the output quality is scored between 0 and 10.

Page5.

We can saw the minimum of the quality is 3 and the maximum of the quality is 9 from the summary we obtained in R. For the inputs, mean is usually greater than the median. These observations indicate that there are outliers in the data set.

Page 6

For preparation, we check the outliers by boxplots and we check the correlation between predictors.

Page 7

Based on the boxplots we can conclude that first Fixed acidity, volatile acidity, citric acid Free sulfur dioxide and density have outliers. Second Residual sugar has a positively skewed distribution.

Page 8

For the correlation we marked the high correlations, which is greater or equal to 40% in absolute value. Based on the checking result for outlier and correlation, we can conclude that for our data, there are both outliers and correlations existing then we did data clean.

Page 9

Until now, the methods we use are multiple Regression, Ridge, Lasso, Support Vector Machines, Decision Tree and Random Forest.

Page 10

First we use multiple regression, we know that the predictors citric Acid, chlorides and total sulfur dioxide are not significant.

The R square for the predictions is nearly 27%, which means the model does not match the condition of the data so well. The error rate for the predictions is nearly 48%, which means there is a lot of error happened for prediction of quality.

(If Alem ask why round the prediction, the answer as followed. Since the quality is scored between 0 and 10 then when find the error rate we need to make the prediction of quality into integers then calculate the error rate)

Page 11

Based on the graphs showed for residual, we can conclude that Residuals have an approximately Normal Distribution. Since the response actually takes only integer values but has been assumed to be continuous, such pattern showed above.

Page 12

Since the result of multiple regression tells us that there are some predictors are not significant, we will then use ridge and lasso to do the regularization work.

From the result of ridge, none of the predictor is dropped, and there is no improvement from multiple regression.

The R square for the predictions is nearly 27%, and the error rate for the predictions is nearly 48%.

Page 13

Then we try lasso regression. In the result of lasso, fixed acidity, citric acid, total sulfur dioxide, density are dropped. These parameter dropped are the correlated predictors shown in the correlation part we have mentioned.

The R square for the predictions is nearly 26%, and the error rate for the predictions is nearly 48%, which are still not good. We can say that linear regression is not that appropriate for this data set.

Page 14

Since SVM is expected to be less sensitive to outliers and this effect results in a higher accuracy for low error tolerances so we use SVM method in our project. As to SVM model, the value of parameter epsilon is 0.1 and cost C is 2, and the radial basis function was used as kernel function. R2 here is nearly 29% and the error rate is about 47%.

(Support Vector Machines are very specific class of algorithms, characterized by usage of kernels, absence of local minima, sparseness of the solution and capacity control obtained by acting on the margin, or on number of support vectors, etc.

In SVM regression, the variables are transformed into a high m-dimensional feature space, by using a nonlinear mapping that does not need to be explicitly known but that depends of a kernel function. The aim of a SVM is to find the best linear separating hyperplane, tolerating a small error when fitting the data.

The SVM cost function gives a linear penalty to large errors, so it is expected to be less sensitive to outliers and this effect results in a higher accuracy for low error tolerances.)

Page 15

Decision Tree is a very suitable to multi-class classification problem. It’s very natural to apply decision tree algorithm to this dataset.

We can see the table of the result. The row names are the predicted class and the column names are the true class. The content of the table is the number of corresponding observations. Say, the number of Row 5 and Column 6 is 54, which means that there are 54 kind of wine with quality score 6 that are predicted as wine of quality score 5. The error rate of the random forest for classification is 49%, i.e. the sum of diagonal values over the sum of the whole table.

However, we found a problem. There are actually no observations that are predicted as wine of quality score 3,4 and 8,9. It means that the decision tree algorithm tend to give “conservative” prediction.

Page 16

For the classification part, we also use the random forest algorithm.

We can see the table of the result. The row names are the predicted class and the column names are the true class. The content of the table is the number of corresponding observations. Say, the number of Row 5 and Column 4 is 9, which means that there are 9 kind of wine with quality score 4 that are predicted as wine of quality score 5. The error rate of the random forest for classification is 31.3%, the sum of diagonal values over the sum of the whole table.

Through the table of the result, we know that there are still no observations that are predicted as 3 and 9, but some observations are predicted as 4 and 8. This means that the random forest algorithm has not been so “conservative” as the decision tree algorithm. But it is still a little different from the real situation.

Page 17

The error rate for random forest with regression prediction is almost 31.7%. The R squared here is nearly 52%.

Based on Random Forest we can conclude that alcohol, free sulfur dioxide and volatile acidity have more influence on white wine quality than other predictors.

（Since the Random Forest is one of the most accurate learning algorithms available, we should try this model. Based on decision trees, the algorithm nature makes Random Forest suitable to regression and classification. The advantages include efficiency on large data, Can handle thousands of input variables without variable deletion, Gives estimates of what variables are important in the classification, Generated forests can be saved for future use on other data.）

Page 18

We summaries the error rate for each method and the R2 for regression methods we find out that all the R2 are not big enough since the biggest one is only 52%. For error rate, the error rate of the random forest method is nearly 20% lower than other methods. Then we obtain the following results.

Page 19

First since the values of R2 are very small, regression models cannot appropriately fit the data. Second Alcohol, free sulfur dioxide and volatile acidity have more influence on white wine quality than other predictors. Third since error rate for random forest is lower than other methods, random forest is more capable model for predicting quality of the white wine.

Page 20

In the future, we will use more methods such as polynomial regression, Adaboost or Classified wine into three categories (low, median, high) to check our result.

(First, we have mentioned in the classification part, the two model we use did not give predictions of 3,4,8 and 9, that is simply because the data we use. The data we use, only little of wine quality is 3,4,8 and 9, so a classification of wine quality is appropriate. Second, when evaluating a wine, what we concerns is whether it is good or not, so 3 category for wine quality is enough for evaluation.)

Page 21

Are there any questions?

Page 22

Thanks for watching.