Red Wine Quality Evaluation Proposal

Qi Wang[1], Maryam Adamzadeh[2]

1. Introduction

Wine, maybe the oldest drink that its secret recipes, has been passed to us through centuries by our ancient. Like everything else, wineries are looking to evolve the way they are making wine and apply technology and innovations to this most popular drink in the world. To study more about this subject, the wine data collect from the north-west region, named Minho, of Portugal, and this data set is available from the UCI machine learning repository (UCI, 2015). It has been proposed for both, regression and classification, by Cortez et al. (2009). Cortez et al proposed a data mining approach to predict human wine taste preferences. Three regression techniques were applied, under a computationally efficient procedure that performs simultaneous variable and model selection(Cortez et al. 2009). Such model is useful to support the oenologist wine tasting evaluations and improve wine production. Later, Agyemang presents an analysis to extend what Cortez et al accomplished by using two logistic regression approaches to predict human wine taste preferences with the goal of better predictions(Agyemang 2010). Nebot et al used hybrid fuzzy logic techniques to predict human wine test preferences based on physicochemical properties from wine analyses (Nebot, Mugica, and Escobet 2015). The fuzzy technique result presents a better performance rather than other data mining techniques previously applied to the same data set, such are neural networks, support vector machines and multiple regression. Recently, Angus try to find out if it is possible to predict what score a wine would be given based on its chemical properties and wine testers' opinion on the wine quality (Angus, n.d.). The result opens up the possibility of assigning wine score without the use of wine testers.

2. Goal

Goal for Statistical Analysis

Here, in this paper, we are going to explore some methods to classify the wine into different quality levels. We do not use clustering methods here but we will use logistic regression to achieve our goal. And our goal is to explore which of these variables which I will talk about in the variable description section will have a significant effect on the red wine qualification and it is positive or negative, and how much will the significant variables affect the level of the red wine qualification. Also, we will use some criteria to find the best model fit for this case and then make it easier for people's further reference.

Goal for Salute

I once encountered this data in my undergraduate level in the class called categorical data analysis, but I did really a bad job at that time. However, my professor encouraged me a lot and gave me much confidence, although that class still seemed too hard for me. I saw this data set on the UCI machine learning data sets again at the beginning of this quarter, so I decide to do it again in honor of my old professor and show my determination to learn this modelling class and later all regression classes excellently as a young generation of statistician. I have really learned a lot from this class, now I have much more interest in the regression models than before. I used to get afraid of data analysis since I don't know what to do, I learned lots of methods but they are just such a mess in my head. Now, it is becoming more clearly due to the help of this class.

3. Variable Despription

There are 11 covariates and one categorical response variable.

Fix Acidity

It is the most acids involved with wine or fixed or non-volatile (do not evaporate readily).

Volatie Acidity

Volatile acidity is the amount of acetic acid in wine, which at too high of levels can lead to an unpleasant, vinegar taste. Citric acid is the most important organic acid produced in tonnage by fermentation, with a taste of sour like lemons (Soccol et al. 2006).

Residual Sugar

Residual sugar is the amount of sugar left in a wine, to some extent, it measures the sweetness of a wine.

Chlorides

Chlorides is a key role in the salty taste of a wine, which will make customers feel uncomfortable.

Sulfur Dioxide (Two Types)

Sulfur dioxide (SO_2) is important in the wine making process as it aids in preventing microbial growth and the oxidation of wine (Monro et al. 2012). The difference between free sulfur dioxide and total sulfur dioxide is the way of measuring them. Gaseous SO2 is released from the sample by addition of acid and swept into the ICP by an argon stream. The intensity of the sulfur atomic emission lines is measured in the vacuum UV region. Determination of total SO_2 is performed after hydrolysis of bound forms with sodium hydroxide ($NaOH$)(Čmelík et al. 2005). For sulfates, many experts believe that higher sulfurous content causes a duller taste in wine, and that high potency of sulfite ions presents a health risk and speeds up the wine's fermentation process.

Alcohol, pH and Density

The other covariates including alcohol, pH and density are basically simple indexes of a red wine. Our response variable is an ordered categorical variable indicating the quality of red wine, from 0 to 10.

All of the covariates are continuous, so we don't need to consider the interaction here. But some problems exists when analyzing the data

Response Variable: Quality

This is a categorical data range from 0 to 10 to describe which category this wine belongs to. It is ordinal since greater number means that the quality of wine is also better. It is hard to manage this response variable if we do not transform it into a better version since I am going to use the logistic regression so it would be easier to merge some categories together.

4. Challenges

1. To begin with, the response variable in this data set is categorical but have 11 degrees, although in the data there are only five levels, we still need to first merge some groups to get a 0-1 variable to continue the logistic regression.

2. Then, there are some inner colinearship among the covariates. Citric acid amount is strongly positively correlated with fixed acidity. And fixed acidity is also strongly positively correlated with the density of the wine. Researches have shown that the citric acid has an effect on the acidity of the liquid, and a more concentration of citric acid means a stronger acidity (Lustig et al. 2017).Also, the amount of free sulfur dioxide is positively correlated with the amount of total sulfur dioxide. However, there are still some negative correlations among variables. For example, the pH of the red wine is negatively correlated with the fixed acidity and the amount of citric acid in the wine. From common sense, we know a stronger acidity means a lower pH, that's why the pH is lower for those wine with more concentration of fixed acidity and citric acid. Also, larger concentration of alcohol gives a smaller density. As we know, the density of alcohol is smaller than water, so if more alcohol is included in the wine, the density must be lower than those without that much alcohol. I fixed this problem by first select variables for a base model which has weak relationships, then add variables to check the model selection criteria.

3. There are some outliers in the model, but after checking the resource of data, I don't think it is a measurement error, therefore, I just leave it there and still looking for further methods to handle it.

4. It is a logistic regression, so the residual analysis is tricky. I will begin this part after learning this method in the class.

References

Agyemang, Perpetual. 2010. "Modeling the Preference of Wine Quality Using Logistic Regression Techniques Based on Physicochemical Properties." PhD thesis.

Angus, Dale. n.d. "Modeling Wine Quality from Physicochemical Properties." *Red* 895 (384): 320.

Čmelík, Jiří, Jiří Machát, Eva Niedobová, Vítězslav Otruba, and Viktor Kanickỳ. 2005. "Determination of Free and Total Sulfur Dioxide in Wine Samples by Vapour-Generation Inductively Coupled Plasma–Optical-Emission Spectrometry." *Analytical and Bioanalytical Chemistry* 383 (3): 483–88.

Cortez, Paulo, António Cerdeira, Fernando Almeida, Telmo Matos, and José Reis. 2009. "Modeling Wine Preferences by Data Mining from Physicochemical Properties." *Decision Support Systems* 47 (4): 547–53.

Lustig, William P, Soumya Mukherjee, Nathan D Rudd, Aamod V Desai, Jing Li, and Sujit K Ghosh. 2017. "Metal–Organic Frameworks: Functional Luminescent and Photonic Materials for Sensing Applications." *Chemical Society Reviews* 46 (11): 3242–85.

Monro, Tanya M, Rachel L Moore, Mai-Chi Nguyen, Heike Ebendorff-Heidepriem, George K Skouroumounis, Gordon M Elsey, and Dennis K Taylor. 2012.

"Sensing Free Sulfur Dioxide in Wine." *Sensors* 12 (8): 10759–73.

Nebot, Àngela, Francisco Mugica, and Antoni Escobet. 2015. "Modeling Wine Preferences from Physicochemical Properties Using Fuzzy Techniques." In *SIMULTECH*, 501–7.

Soccol, Carlos R, Luciana PS Vandenberghe, Cristine Rodrigues, and Ashok Pandey. 2006. "New Perspectives for Citric Acid Production and Application." *Food Technology & Biotechnology* 44 (2).