Wine Evaluation

Kyungjin Sohn, Seong Jin Lee

1 Introduction

Wine is one of the most popular alcoholic drink worldwide. There are many discussion about how to determine the quality of wines. Wine is generally certified through physiochemical tests and sensory tests. Physiochemical tests includes measurement of acidity, pH or alcohol, while the sensory tests are based on human experts. Cortez et al.(2009)[1] tries to predict the sensory aspects, or the "quality" with the physiochemical aspects based on mathematical models.

While the authors of Cortez et al.(2009) [1] use linear models, support vector machine and neural network models to make predictions, we focus on modifying the linear model to make better prediction on the quality of wine. Through these models we can suggest which aspects contribute to the quality of wine, and propose ways to improve wine production.

2 Data Description

The dataset "Wine Quality" was created by Cortez et al., 2009[1]. The two datasets were made using red and wine variants of the Portuguese "Vinho Verde" wine. Eleven input attributes were measured objectively, and one output attribute was subjectively scored by wine experts(median of at least three evaluations). Experts gave a minimum of 0(very bad) to a maximum

of 10(very excellent) points depending on the quality of the wine. Below is the outline of each columns of the data:

- Input variables (based on physicochemical tests):
    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
- Output variable
    1. Quality (score between 0 and 10)

3 Methods

First we analyze the data through building a linear regression model that uses all 11 factors to predict the quality. Note that the relationship between covariate and response variable might not be linear. So we apply different transformation to verify linear relationships.

Next we use variable selection methods including step-wise regression, LASSO regression to select valid variables. Through this process we can determine which aspects contribute to the quality of wine.

Also, note that the response variable is discrete. So the traditional linear model which takes value on the real line might not be appropriate. Therefore, we use different methods of evaluating the model other than simple sum of square errors. Also note that the response variable takes value in a bounded interval. So we might want to make transformations on the response variable as well.

References

[1] Paulo Cortez et al. "Modeling wine preferences by data mining from physicochemical properties". In: *Decision support systems* 47.4 (2009), pp. 547–553.