Data 621 Final Project Research Proposal: Can we predict the quality of wine using data analysis techniques?

Critical thinking group 1

Organized thoughts of group by Jun Pan

April 9th, 2019

Introduction

Vinho Verde wine exclusively produced in the demarcated region of Vinho Verde in northwestern Portugal. It is only produced from the indigenous grape varieties of the region, preserving its typicity of aromas and flavors as unique in the world of wine. To support its growth, the wine industry is investing in new technologies for wine certification and quality assessment. Certification prevents the illegal adulteration of wines (to safeguard human health) and assures quality for the wine market. Wine certification is generally assessed by physicochemical and sensory tests (Cortez 2009).

Wine quality data were collected from May/2004 to February/2007 using only protected designation of origin samples that were tested at the official certification entity (CVRVV). The data were recorded by a computerized system (iLab), which automatically manages the process of wine sample testing from producer requests to laboratory and sensory analysis. Each entry denotes a given test (analytical or sensory) and the final database was exported into a single sheet (.csv). During the preprocessing stage, the database was transformed in order to include a distinct wine sample (with all tests) per row. To avoid discarding examples, only the most common physicochemical tests such as alcohol, chlorides, density, total sulfur dioxide, free sulfur dioxide, residual sugar, and pH were selected. Since the red and white tastes are quite different, the analysis will be performed separately, thus two datasets were built with 1599 red and 4898 white examples (Cortez 2009). This database is available for downloading from uci machine learning repository (<https://archive.ics.uci.edu/ml/datasets/Wine+Quality>).

Physicochemical laboratory tests routinely used to characterize wine include determination of density, alcohol or pH values, while sensory tests rely mainly on human experts. It should be stressed that taste is the least understood of the human senses, thus wine classification is a difficult task. Moreover, the relationships between the physicochemical and sensory analysis are complex and still not fully understood.

Our goal is to predict the wine quality based on various psychochemical tests using wine quality database.

Build Models

Globally, wine industry is nearly worth 300 billion dollars. Being able to predict the quality of wine would be very valuable addition to this industry. Our study is to build a model assessing the wine quality for a given sample based on a given set of attributes. This is modeled by predicting the quality on a scale of 1 to 10 from a set of associated attributes.

In this study, we are going to split the data to form a train set and a test set. By tradition, linear regression models will be used in our study. We may develop a couple of linear regression models by using different predictors on train data set. Confusion matrix, RMSE, sensitivity, specificity, ANOVA will be used for the best fit of the model. The best model will be used to predict the results on test data.

After we explored the dataset and conducted some preliminary analysis of the database, we feel that non linear regression model might be better than the linear regression. So we are going to use the following methods to build model: (1) k-NN, (2) random forest, (3) support vector machine. The classification results of the three classification algorithms are evaluated both test modes which are 10-fold cross-validation, and 80% percentage split. Also, some of the standard performance measures (statistics) are calculated to evaluate the performance of the algorithms. The standard performance measures are recall, precision, F measure, and ROC values. Confusion metrix, accuracy, kappa value will be used to evaluate the models. The best model will be used to predict the results on test data set. Reviewing the density plots of all the variables, we feel that the low quality wine and high quality wine do not have enough data to play with.

So our hypothesis is the non linear models overall are better than linear models. However, there is no model has better performance on high quality wine and low quality wine due to lack of sufficient of data.

Reference:

[Altman, N. S.](https://en.wikipedia.org/wiki/Naomi_Altman) (1992). "An introduction to kernel and nearest-neighbor nonparametric regression". The American Statistician. **46** (3): 175–185.

knn ( <https://en.wikipedia.org/wiki/K-nearest_neighbors_algorithm>)

Linear regression (<https://en.wikipedia.org/wiki/Linear_regression>)

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.

Wine quality dataset (<https://archive.ics.uci.edu/ml/datasets/Wine+Quality>)

Yesim Er , Ayten Atasoy. The Classification of White Wine and Red Wine According to Their Physicochemical Qualities. IJISAE, 2016, 4(Special Issue), 23–26 | 23