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

Critical thinking group 1

Organized thoughts of group by Joby John, Zachery Herold, Jun Pan

April 15, 2019

Introduction

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. Wine quality 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). This database is available for downloading from uci machine learning repository (<https://archive.ics.uci.edu/ml/datasets/Wine+Quality>). Our analyses focus in a Portuguese white wine database consisting of 4,898 observations. The data set contains eleven explanatory variables that measure wine attributes and one response variable: “wine quality”. Here are the information regarding variables in the dataset: (1) Fixed acidity: a measurement of the total concentration of titratable acids and free hydrogen ions present in the wine. (2) Volatile acidity: a measure of steam distillable acids present in a wine. (3) Citric acid: one of the many acids that are measured to obtained ﬁxed acidity. (4) Residual sugar: measurement of any natural grape sugars that are leftover after fermentation ceases. (5) Chlorides: the amount of salt in the wine. (6) Free sulfuric dioxide: the free form of SO2 exists in equilibrium between molecular SO2 (as a dissolved gas) and bisulﬁte ion; (7) Total sulfuric dioxide: amount of free and bound forms of SO2; (8) Density: measure of density of wine. (9) pH: value for pH. (10) Sulfates: a wine additive which can contribute to sulfur dioxide gas (S02) levels, which acts as an antimicrobial and antioxidant. (11) Alcohol: the percentage of alcohol present in the wine. (12) Quality: subjective measurement ranging from 1 to 10 (although the observed data ranges from 3 to 8).

Our study is to build a model assessing the white 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.

Build Models

The response variable quality (score 1 to 10) has multiple levels of factors, we feel that ordinal regression model might be more fit than the logistic regression. K-Nearest Neighbor Classiﬁcation is common used model in ordinal classiﬁcation in the industry. An ordinary k-nearest neighbors involves ﬁnding the k nearest neighbors of the test data in the variable space and obtain the class for the test data through majority of votes.

Another model we are going to use is multinomial regression. Multinomial regression is much similar to logistic regression but is applicable when the response variable is a nominal categorical variable with more than 2 levels (r-statistics.co). It is sometimes considered an extension of [binomial logistic regression](https://statistics.laerd.com/spss-tutorials/binomial-logistic-regression-using-spss-statistics.php) to allow for a dependent variable with more than two categories. As with other types of regression, multinomial logistic regression can have nominal and/or continuous independent variables and can have interactions between independent variables to predict the dependent variable.

Ordinal regression (also called "ordinal classification") is a type of [regression analysis](https://en.wikipedia.org/wiki/Regression_analysis) used for predicting an [ordinal variable](https://en.wikipedia.org/wiki/Ordinal_variable), i.e. a variable whose value exists on an arbitrary scale where only the relative ordering between different values is significant. With a binary variable, the ordinal logistic model is the same as logistic regression. It can be considered an intermediate problem between regression and [classification](https://en.wikipedia.org/wiki/Statistical_classification) (Winship and Mare, 1984). Ordinal regression turns up often in the [social sciences](https://en.wikipedia.org/wiki/Social_sciences), for example in the modeling of human levels of preference (on a scale from, say, 1–5 for "very poor" through "excellent"), as well as in [information retrieval](https://en.wikipedia.org/wiki/Information_retrieval). In [machine learning](https://en.wikipedia.org/wiki/Machine_learning), ordinal regression may also be called ranking learning. So it is a potential model we would like to try.

To test the algorithm, we will divide the data into a test set and a training set. We will set aside 898 data-points as the ultimate test set and the rest of the data points (4,000) as the training set. Because 4,000 data points does not consider a giant data set, we will use cross-validation to ﬁnd the respective optimal tuning parameters for k-nearest neighbors and multinomial regression. We will not normalize the data because most of the data are concentrations measured at the same level and hence the scale is important to the data set.

To evaluate different models, some of the standard performance measures (statistics) are instituted. The standard performance measures are recall, precision, F measure, and ROC values. Confusion metrix, accuracy also will be used to evaluate the models. The best model will be used to predict the results on test data set.

Our hypothesis is multinomial regression model overall is better than ordinal regression and k-NN model.

Discussion

Ordinal regression arises frequently in social sciences and information retrieval where human preferences play a major role (Wei et al). Ordinal responses are comprised of both a ranking structure of real numbers and a discrete structure of classiﬁcation. This therefore makes the response more difficult to model than real numbers or categorical variables alone. There are altogether eleven chemical attributes serving as potential predictors. All predictors are continuous while the response is a categorical variable which takes values from 1 to 10. One method used in ordinal classiﬁcation in the industry is k-nearest neighbors. An ordinary k-nearest neighbors involves ﬁnding the k nearest neighbors of the test data in the variable space and obtain the class for the test data through majority of votes (knn Wikipedia). k-NN using the distance between 2 points, so it can be used for a model with no linear relationship. During the process, k-NN normalizes all the attributes between 0 to 1, it is not necessary to worry about outliers and collinearity. However, k-NN analysis usually uses for a database with a few hundred observers. Our data base is quite big. To process the analysis is quite time consuming. According to k-NN theory, k value should be square root of number of observation. In our dataset, k value should be around 69. That means we need to calculate the distance between 69 points. That is a huge amount of work. After review the literature, most scientist using k value around 7 to 9 for similar size database because the computer memory issue. It will sacrifice the accuracy of the model. Ordinal regression can be performed using a general linear model that fits both a coefficient vector and a set of thresholds to a dataset. After review the data, we feel the predictors do not have a typical linear relationship with the response variable. So we think the ordinal regression definitely has disadvantage in this case. In our study, we still choose multinomial regression over the other methods. First, Multinomial logistic regression does necessitate careful consideration of the sample size and examination for outlying cases (Schwab 2002). Second, Multinomial logistic regression is often considered an attractive analysis because; it does not assume normality, linearity, or homoscedasticity. This makes multinomial regression a better choice compared to k-NN and ordinal regression models.

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. There will be 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>)

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.

r-statistics.co (<http://r-statistics.co/Multinomial-Regression-With-R.html>).

Schwab, J. A. (2002). Multinomial logistic regression: Basic relationships and complete problems. http://www.utexas.edu/courses/schwab/sw388r7/SolvingProblems/

Wei Chu, S. Sathiya Keerthi, Support Vector Ordinal Regression http: //www.gatsby.ucl.ac.uk/∼chuwei/paper/svor.pdf

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

Winship, Christopher; Mare, Robert D. (1984). “Regression Models with Ordinal Variables”. American Sociological Review. 49(4):512-525.