Data 621 Final Project Research Proposal:

Predicting the quality of white wine using data analysis techniques

**Critical Thinking Group One**

Organized thoughts of group consisting of

Joby John, Zachary 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 3 to 9 from a set of associated attributes.

**Model Generation**

With the response variable ‘quality’ (scored between 3 to 9) having multiple factor levels of ordinal ranking, we apply the following two advanced modeling techniques to generate predictive value, as described below. To test both algorithms, we divide the data into a test and training sets. We partition 898 data-points as the ultimate test set and the rest of the data points (4,000) as the training set. We elect not to 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.

**K-Nearest Neighbor Classiﬁcation**

K-Nearest Neighbor Classiﬁcation is a commonly-used model for 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 obtaining the class for the test data through majority voting. k-NN uses the distance between two points; as such it can be applied for a model without linear relationship. During the process, k-NN normalizes all the attributes between 0 to 1, alleviating the concern brought by outliers and collinearity. However, as k-NN analysis typically uses for a database with a few hundred observations, the white wine database is quite large, leading to time-consuming data processing. For instance, in most literature, we see scientists using k value around 7 to 9 in accordance to k-NN theory, which recommends k value approximate to the square root of number of observations. Give the sample size of our dataset, the k value should be around 69, that is our model is required to calculate the distance between 69 points or otherwise accuracy will be sacrificed.

**Ordinal Logistic Regression**

The second model built is the ordinal logistic proportional odds model first described in Walker and Duncan and later called the (PO) model by McCullagh. 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. 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. Ordinal regression can be performed using a general linear model that fits both a coefficient vector and a set of thresholds to a dataset.

**Discussion**

In addition to comparing the fitted results of the k-NN and Ordinal Logistic Regression Models, we test the following two research findings:

1. **There will be little difference between results obtained with ordinal regression and OLS regression approaches.** According to research by Kromrey & Rendina-Gobioff (2002) and Taylor, West, & Aiken (2006), if there are 5 or more categories and particularly with larger sample sizes and fairly normally distributed variables, the gains by performing an ordinal logistic regression are minimal. We investigate further the finding, considering ‘quality’ ratings as a discrete quantitative variable for comparison.
2. **It is better to use all of the ordinal values rather than collapsing into fewer categories or dichotomizing variables.** With sparse numbers of inferior (quality rating of 3) and superior (rating of 9) wines in the dataset, one is tempted to collapse these categories into the subpar and above par ones. While some analysts feel that combining categories improves the performance of test statistics when fitting PO models when sample sizes are small and cells are sparse, Murad et al. rebuke this notion, demonstrating that this actually causes more problems, resulting in overly conservative Wald tests. We seek to verify this claim that "for outcomes that can be considered ordinal, even with a sparse number of responses in some categories. Collapsing categories has been shown to reduce statistical power" (Ananth & Kleinbaum 1997; Manor, Mathews, & Power, 2000) and increase Type I error rates (Murad, Fleischman, Sadetzki, Geyer, & Freedman, 2003).

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 matrices, accuracy, and Wald tests also will be used to evaluate the models. Other measures are uses determine the best model for predicting the results on test data set.

References:

[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.

Frank E. Harrell , Jr. (auth.) (2015). *Regression Modeling Strategies - With Applications to Linear Models, Logistic and Ordinal Regression, and Survival Analysis.* Springer Series in Statistics

H. Murad, A. Fleischman, S. Sadetzki, O. Geyer, and L. S. Freedman. Small samples and ordered logistic regression: Does it help to collapse categories of outcome? *Am Statistician*, 57:155–160, 2003. 324

# UCLA Institute for Digital Research & Education. Ordinal Logistic Regression | R Data Analysis Examples https://stats.idre.ucla.edu/r/dae/ordinal-logistic-regression/

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/>

S. H. Walker and D. B. Duncan. Estimation of the probability of an event as a

function of several independent variables. Biometrika, 54:167–178, 1967. \_14,

220, 311, 313

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.