**Build Models to Predict White Wine Quality by Data Mining**

**from Physicochemical Properties of Wine Quality Data**

Joby John, Zachery Herold, Jun Pan

Graduate Students of Master of Science in Data Science Program,

School of Professional Study, City University of New York

119 W 31st St, New York, NY 10001

**Abstract**

**Key words:** wine quality, k-NN, ordinal logistic regression, physicochemical

**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. 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. Wine certification is generally assessed by physicochemical and sensory tests (Cortez 2009). 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.

Advances in information technologies have made it possible to analysis and process massive complicated datasets. All this data hold valuable information which can be used to improve decision making and optimize chances of success (Turban 2007). Data mining (DM) techniques aim at extracting high-level knowledge from raw data (Witten and Frank 2005). The response variable “quality” in wine quality dataset is on a scale of 1 to 10. So it is a ordinal variable. For ordinary variables, there are several DM algorithms such as k-NN, random forest, neural network, support vector machines, multinomial regression, ordinal logistic regression. We are going to use k-NN and ordinal logistic regression methods in our study. 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. Indeed, k- NN has hyperparameters that need to be adjusted (Hastie 2001), such as the number of NN hidden nodes or the SVM kernel parameter, in order to get good predictive accuracy. The second model built is the ordinal logistic proportional odds model first described in Walker and Duncanand 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.

**Methods**

Wine quality data were collected from May/2004 to February/2007. 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). This database were downloaded from uci machine learning repository (<https://archive.ics.uci.edu/ml/datasets/Wine+Quality>).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). The data set contains eleven explanatory variables that measure wine attributes and one response variable: “wine quality”. In our study, we just used the white wine dataset.

Here is 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).

Data exploration and analysis were conducted using the most updated version R studio ([www.rstudio.com](http://www.rstudio.com)) equipped with following packages: kknn, caret, corrplot, ggplot2, kernlab. To visualize the data, plots for each predictor variable were displayed. Mean, quartiles, median and standard deviation were calculated. Missing data, outliers and correlation variables were evaluated.

The response variable, score of quality, was converted to factors. The white wine dataset were split into training (80%) and test sets (20%). Training was performed with the help of the caret package’s train function. The cross-validation method was 5 fold, repeated 5 times. K-nearest neighbours uses distance to classify the response variable. Hence, we normlized all the predictor variables’ values in the range of 0 and 1. Using the above method, we prevented predictors with larger values from being over-emphasized by the algorithm. The “preprocess” argument in the train function was used to center and scale the predictors for standardization. Firstly, a full model was built using all the independent variables. For k-nearest neighbours, 5 kmax, 2 distance, and 3 kernel values were used. For the distance value, 1 is the Manhattan distance, and 2 is the Euclidian distance. K-max were tried to use a series values (3, 5, 7, 9, 11) to find the best k-max value. Kernel models were evaluated using “rectangular”, “Gaussian”, “cos”. For the distance value, 1 is for Manhattan distance, and 2 for Euclidean distance. Secondly, we built a reduced model by remove the variables of citric.acid, free.sulfur.dioxide, and sulphates. The rest of process was the same as the full model except the k-max value using (7,9,11,13,15).

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 ordinal regression model might be better than the logistic regression. So we are going to use k-NN model. The classification algorithms will be evaluated by 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 will be used to evaluate the models. The best model will be used to predict the results on test data set.

For each classification model, we analyzed how the results vary whenever test mode was changed. The study included the analysis of classifiers on each model. The results are described in percentage of correctly classified instances, precision, recall, F measure, and ROC after applying the cross-validation or percentage split mode. Different classifiers like k-nearest-neighborhood evaluated on dataset.

**Results**

After visualization the white wine data set, we found that 4898 samples and 12 variables in the data set. There is no missing information in the database. All predictors are continuous variables while the response is a categorical variable which takes wine quality scores from 1 to 10. All the variables are summarized in table (table 1). Number of wines is not evenly distributed along with the scores, for example, there were 20 samples with score of 3 and 5 samples with score of 9 (table 2). Data were further explored by density plot for each variable by quality of scores (figure 1). Compared to the mean of each variable by quality scores, we can see many variables are not linear related to quality scores, such as citric acid, residual sugar, free sulfur dioxide, total sulfur dioxide, density, pH and sulphates. So k-NN model is more appropriate for building prediction model. Box plots were used to see the correlation of explainable variable to the responsible variables (figure 2). Box plots further confirmed the non-linear relationship of between the explainable variables to the response variable (score of quality).

The first thing that stands out in the plots is the presence of outliers for most of the predictor variables. The UCI wine dataset was cleaned prior to its posting, so we did not treat it as errors. For residual sugar, the outlier had a residual sugar level of 65.8. The next highest sugar level in the dataset was 31.6. In wine industry, a wine with more than 45g/l of sugar is considered a sweet wine. So we removed the highest value of residual sugar. Also, the sample wine had the density outlier (sugar contribute to the high density of wine), we removed it as well. “Free sulfur dioxide” had an outlying sample 2 times greater than the next largest one. But this wine had a score of quality of 3 which is the lowest quality in the dataset. The high value for “free sulfur dioxide” may be linked to the sample’s poor quality rating. So, we just kept it.

From the correlation plot (figure 3), we found weak relationships between quality and citric.acid, free.sulfur.dioxide, and sulphates. It might be a good idea to remove those predictor variables to reduce the dimensionality of the data. Feature selections were applied after the data is split into training and test sets. Density had a 0.83 correlation with residual.sugar and a -0.80 correlation with alcohol.

Before analysis, responsible variable were convert to be factor variable. Data were split into train data set and test dataset with the ration of 8:2. Training was conducted using caret package’s train function. In order to prevent predictors with larger ranges being over-emphasized by the algorithms, the preprocess argument in the train function was used to center and scale the predictors for standardization. The first step in adjusting the k-nearest neighbors model was to fix the number of neighbors k. We used 10-fold cross validation and chose k such that the CV residual mean squared error (RMSE) is minimized. This yielded to a result of k = 11 (figure 4). To adapt the mean k-nearest neighbor regression to our ordinal data, we rounded the resulting value to obtain a integer number. Among them, we can see that k equals to 11, distance 1, kernel “cos” was outperform than the alternatives. The overall accuracy is 60.5% and kappa value of 0.4223 (figure 5). The highest sensitivity is 68.31 (quality score is 5). The highest specificity is 98.60 (quality score is 4) (figure 5). For reduced modle, we can see that k equals to 13, distance 2, kernel “rectangular” was outperform than the alternatives (figure 6). The overall accuracy is 59.98%. The kappa value is 40.47. The highest sensitivity is 64.62 (quality score is 4) (figure 7). The highest specificity is 99.7 of (quality score of 3) which only has 20 wines in total. After compared the accuracy, kappa value, and overall performance of sensitivity and specificity, we consider full mode is a better model.

Reviewing the density plots of all the variables, we feel that the low quality wine and high quality wine did not have enough data to play with. Due to lack of sufficient data in those categories, it was difficult to conclude which model was better than the others.

**Discussion**

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 means our model is required to calculate the distance between 69 points or otherwise accuracy will be sacrificed.

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 teststatistics when fitting PO models when sample sizes are small and cells aresparse, 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 Walk testsalso will be used to evaluate the models. Other measures are uses determine the best model for predicting the results on test data set.

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).

**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

Hastie T., Tibshirani R.,and Friedman J.. The Elements of Statistical Learning:Data Mining, Inference and Prediction. Springer-Verlag, NY, USA, 2001.

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

# UCLA Institute for Digital Research & Education. Ordinal Logistic Regression | R Data Analysis Exampleshttps://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/>

Turban E., Sharda R., Aronson J., and King D. Business Intelligence, A Managerial Approach. Prentice-Hall, 2007.

Walker S. H. and Duncan D. B. 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 ,AytenAtasoy. 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.

Witten I.H. and Frank E.. Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations. Morgan Kaufmann, San Francisco, CA, 2nd edition, 2005.