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

Version 4

05/18/2019

Abstract

It is important to develop mathematical model and algorithm to predict wine quality. A successful model with high accuracy will be useful to support the oenologist wine tasting evaluations and improve wine production. Furthermore, similar techniques can help in target marketing by modeling consumer tastes from niche markets.

In this study, we developed ordinary logistic regression model and knn models to predict human wine taste preferences that is based on easily available analytical physicochemical tests at the certification step. A large dataset with 4,898 observers of white wine from Portugal was used from UCI machine learning repository. Two regression techniques were applied under a computationally efficient procedure that performs simultaneous variable and model selection. Cross validation, confusion matrix, sensitivity and specificity in each class were used to evaluate the predict accuracy and overall performance of the models.

The knn model achieved promising results, outperforming the ordinal regression methods due to less affected by outliers, collinearity and nonlinear relationship between the response variable and independent variables. However, knn model are limited to be further investigate by its expensive computational effort.

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 holds 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 an 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, 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 used for predicting an 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 (Winship and Mare, 1984). Ordinal regression turns up often in the 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. In 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 was 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 left over 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) 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 (rstatistics.co) and correlation variables were evaluated.

The response variable, score of quality, was converted to factors. The white wine dataset was 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 neighbors uses distance to classify the response variable. Hence, we normalized 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 neighbors, 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 removing 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).

For Ordinal Logistic Regression, we choose the polr method of MASS package (rdocument.org), with method = "logistic". We could have also used the vglm () function from the VGAM package, lrm() from the rms package, and clm() from the ordinal package.

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 matrix, 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 (figure 1). 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, figure 2). Data were further explored by density plot for each variable by quality of scores (figure 3). 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 (table 3). 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 4). 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 5), 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. In order to avoid collinearity, we would like to drop the density variable for further analysis.

Before analysis, responsible variable was converted to 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 14). 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” outperformed the alternatives. The overall accuracy is 60.5%

and kappa value of 0.4223 (figure 15). The highest sensitivity is 68.31 (quality score is 5). The highest specificity is 98.60 (quality score is 4) (figure 15). For reduced model, we can see that k equals to 13, distance 2, kernel “rectangular” was outperform than the alternatives (figure 16). The overall accuracy is 59.98%. The kappa value is 40.47. The highest sensitivity is 64.62 (quality score is 4) (figure 17). 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 (Table 18A and 18B).

Ordinal logistic regression (all variables) output coefficient table includes the value of each coefficient, standard errors, and t value, which is simply the ratio of the coefficient to its standard error. One notes the estimates for the six intercepts (for each one-point step in quality improvement, from 3 to 9), which are sometimes called cutpoints. The intercepts indicate where the latent variable is cut to make the seven groups that we observe in our data (figure 6). By using ANOVA method to compare the t-value against the standard normal distribution, we have found that it lacked statistical significance in “citric acid”, “chlorides” and “total.sulfur.dioxide” associated with quality (figure 7). After dropped the above variables, we analyzed the dataset with ordinal logistic regression model (significant variables) again. The residual deviance and AIC values of the significant variable model are only slightly less favorable compared with the all variable model (figure 8). All the variables in the significant variable model were double checked by ANOVA. All the variables were showing $p < 0.001$ (figure 9). After removed variables of “citric acid”, “chlorides” and “total.sulfur.dioxide”, it showed that “density” is positively correlated with “residual.sugar” and inversely correlated with “alcohol” (figure 10). Due to the collinearity, we developed third model by adding interaction (model 3, figure 11). The residual deviance and AIC has gone down. Compared all three ordinal logistic regression model, we believed that the model 3 was the best. However, the confusion matrix showed the overall accuracy was 51.8% (figure 13) which was much lower than knn model. It predicted only 2 out of 163 actual "4"s and only 13 out of 175 actual "8"s. In this model specificity surpassed sensitivity for all levels except average ("6"), suggesting that there were many false results for above-average wines, this was the model fails to have much predictive value for qualities other than "5" and "6". Far too many superior wines were classified as average under the model. The red and blue areas of this mosaic plot (figure X) revealed the "blind spots" of the model, extreme values of Pearson residuals, indicating a large discrepancy between observed and expected values. If red, the cell's observed frequency was less than the expected frequency, meaning false "5"s when the actual quality rating was "6" and false "6"s when the quality rating was "5". Collapsing the category to only 3 instead of 7 improved the accuracy of prediction in our study (figure 20). The accuracy for prediction was about 55%.

Discussion and

In statistics, collinearity is a phenomenon in which one feature variable in a regression model is highly linearly correlated with another feature variable (Farrar and Glauber, 1967). It is a special case when two or more variables are exactly correlated. This means the regression coefficients are not uniquely determined. In turn it hurts the interpretability of the model as then the regression coefficients are not unique and have influences from other features. It is interesting to note the apparent inverse relationship between residual sugar and alcohol. Rough rules of thumb say if a wine's alcohol content is 10% or less it will have sweet characteristics. Wines that are even lower (especially down around 8 or 9 percent) will be sweet. One would suspect the more residual sugar, the lower the potential alcohol that wine could have, since sugar is converted into alcohol in the fermentation process. Sugar's role in dictating the final alcohol content of the wine (and such its resulting body and "mouth-feel") sometimes encourages winemakers to add sugar (usually sucrose) during winemaking in a process known as chaptalization solely to boost the alcohol content - chaptalization does not increase the sweetness of a wine. Grape juice is denser than water. Thus, before we fermented the grape juice the specific gravity was over 1.0. As the yeast converted the sugar into alcohol and carbon dioxide during fermentation, the density of the wine has been decreasing. A specific gravity less than 0.990 tells us that the primary fermentation has slowed down enough that racking is necessary. Due to important interplay between residual.sugar, density and alcohol, we refine our model by adding an interaction between these interrelated components. After correcting the collinearity, residual deviance and AIC values have been decreased, the accuracy of prediction has been improved.

Our best model of ordinary logistic regression has an accuracy of 51.8%. It predicts only 2 out of 163 actual "4"s and only 13 out of 175 actual "8"s. In this model specificity surpasses sensitivity for all levels except average ("6"), suggesting that there are many false results for above-average wines, this is the model fails to have much predictive value for qualities other than "5" and "6". Far too many superior wines were classified as average under the model. One interpretation is that our model is still not sensitive enough to density and pH and alcohol, which are monotonous decreasing, increasing and increasing respectively as quality increases. That is as density goes down and pH and alcohol go up, quality improves at each threshold when removing the highly inferior "3" and "4" quality levels. The mosaic plots reveal the “blind spots” of the model, extreme values of Pearson residual. That is the reason of a large discrepancy between observed and expected values.

During analysis, we believe that it is better to use all 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. Although, reduce categories can improve the accuracy of prediction, it causes more problem later. 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 causes more problems, resulting in overly conservative Wald tests. 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). Publications on this database shows that most people use all the ordinal values (Uniyal et al, 2017; Lemionet et al, 2015).

Ordinal logistic regression or (ordinal regression) is used to predict an ordinal dependent variable given one or more independent variables (Ananth and Kleinbaum, 1997). Ordinal regression enables us to determine which of our independent variables (if any) have a statistically significant effect on our dependent variable. To run the ordinal logistic regression, it is important to deal with missing data. If any are, we may have difficulty to use this model. It is fortunate that we do not have any missing data in this database.

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 (knn, Wikipedia). During the process, k-NN normalizes all the attributes between 0 to 1, alleviating the concern brought by outliers and collinearity. So, it is not necessary to deal with outliers and collinearity in our case which is an important feature of using this model.

In our study, 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. 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). We noticed that most of the independent variables, are not linear related to our response variable. So, using knn model to compare orange to orange, apple to apple has much better accuracy than ordinal logistic regression. In our study, all knn model show much better accuracy for prediction than all the ordinal logistic regression models.

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. In terms of computational effort, the knn is the most expensive method, particularly for the larger white dataset. 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 conclusion, quality of wine can be predicted by the following variables: “alcohol”, “residual.sugar”, “pH”, “fixed.acidity”, “volatile.acidity” and “free.sulfur.dioxide”. Outliers and collinearity are needed to be justified for ordinal logistic regression models but not knn models. K-nn models performs better than ordinal logistic regression models because most of the independent variables are not linear related to the response variable. All the ordinal regression models and knn models failed to predict wine quality scores of “3” and “9”, due to lack of cases, fit into that two categories. Although collapsed categories can improve prediction accuracy, it loses prediction power. It is not wise to collapse the categories in this study. A large discrepancy between observed and expected values is due to the blind spots of models. There are still room for improve the knn models. But the computational cost will be increased significantly.

Review of Literature

Review other study using the same database, Lemionet used knn, weighted linear regression, additive logistic regression, they found additive logistic regression had least test error. They believed that additive logistic regression does better at leveraging the ordinal structure of the data and hence produces better results. As for weighted linear regression, they noted that weighted linear regression performed well when the number of predictors was small. In the case of 10 variables, the predictor space may be too sparse to generate good results. (This can also be explained by the curse of dimensionality). Because they did not use the same methods such as overall accuracy, sensitivity and specificity to evaluate their model, we could not make comparison with the models.

Uniyal et al reported using machine learning algorithm to build a linear regression model based on this database. We feel that they used the wrong model for their study. How could they use a linear regression model for an ordinal response variable? Also, we could not find how they had treated the outliers and collinearity between the lines of their paper. They neither provided ROC curve, nor provided any detailed predictor between each quality scores. We felt that it was a poorly written manuscript.

Cortez et al had spent significant amount effort to use neural network and small vector model to build a couple of prediction models on wine database. The overall accuracy was slightly higher than our knn models. However, they combined 8/9 of wine quality score together might be the reason outperform accuracy than our models. We tried some other model such as random forest, which got similar results as our knn. Because those models were quite time consuming, we did not dig deeper this time. We were failed to perform SVM because our computer stopped running after a couple hours computation. In his later part of his report, he claimed that he had improved accuracy around 90%. But he was failed to provide detailed information for us to repeat his model.

Other studies such as using machine learning algorithm to predict red wine and white wine. We believed it did not make too much sense to do so. Because we can use our naked eye to make a judgment of red wine or white wine based on their color, it is not necessary to build a complicated model to calculate it although the overall accuracy was around 80%.

Based on our preliminary analysis of the white wine dataset and review of the literature, we feel: (1) small vector model, random forest, knn models are better predict the wine quality because most the variables are not linear correlated; (2) the computer expense are enormous because the nature of model; (3) those three models may be more practical on red wine dataset due to the number of observers in the dataset.

References:

Altman, N. S. (1992). "An introduction to kernel and nearest-neighbor nonparametric regression". The American Statistician. 46 (3): 175–185.

Ananth CV, Kleinbaum. Regression models for ordinal responses: a review of methods and applications. International Journal of Epidemiology. 1997. 26(6):1323-1333.

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

Cortez P, Cerdeira A, Almeida A, Matos T and Reis J. Modeling wine preferences by data mining from physicochemical properties. In Decision Support Systems, Elsevier, 47(4):547-553, 2009.

Farrar, Donald E.; Glauber, Robert R. (1967). "Multicollinearity in Regression Analysis: The Problem Revisited". Review of Economics and Statistics. **49** (1): 92–107.

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. Small samples and ordered logistic regression: Does it help to collapse categories of outcome? Am Statistician, 57:155–160, 2003. 324

Lemionet A, Liu Y and Zhou Z. Predicting quality of wine based on chemical attributes. CS 229 Project, 2015, Stanford University.

Rdocumentation.org. https://www.rdocumentation.org/packages/MASS/versions/7.3-51.4/topics/polr

Rstatistics.co http://r-statistics.co/Outlier-Treatment-With-R.html

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.

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

Uniyal1 X, Barthwal P, Joshi A. Wine Quality Evaluation Using Machine Learning Algorithms. Asia-pacific Journal of Convergent Research Interchange 2017: 3(4): 1-9.

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.