BIOS 735 Group 3 Project Proposal

Julian Sim, Li Jiang, Qikai Jiang, Hongyu Yu, Qinghua Li

1 Introduction

One natural consequence of graduate school is an increased tendency for students to start drinking wine (n = 1) and many new wine drinkers may want to fit in and understand what qualities of a wine make it a “good wine”. To answer this question, we utilize the Vinho Verde Wine Quality dataset on the UCI ML Repository which contains various physiochemical attributes of a wine alongside its quality, a continuous variable, and a binary variable indicating whether the wine is a white or a red. We seek to model quality to understand which attributes are the strongest and most accurate predictors as well as to understand whether these attributes may differ by color.

The dataset contains 6497 different wines with thirteen different variables. Two are modeled as outcomes, namely color and quality, whereas the remaining eleven are factored in as features: fixed acidity, volatile acidity, citric acid, residual sugar levels, chloride levels, free sulfur dioxide, total sulfur dioxide, density, pH, sulphate level, and alcohol content.

For our analysis, we consider three different guiding questions: (1) what chemical properties are predictive of wine quality (considering whites and reds in separate analyses), (2) what chemical properties are predictive of wine color, and finally combining these two, (3) determining whether different chemical properties are valued in different color wines. For aim one, we proceed via modeling wine quality through binomial regression, and then using Newton Raphson to update our estimates for the distribution. Then, we can further refine our model via using the Wald test on each covariate to test for significance.

For aim two, we propose three different methods. First, we consider fitting a Bayesian logistic regression model with a prior. We could also utilize a Support Vector Classifier using a “rbf” kernel to increase the dimension of the predictors. Finally, a Multi-layer perceptrons (MLP) with a ReLU activation and softmax loss could be used with a stochastic gradient descent optimization model.

For the final aim, we mimic the analysis conducted in aim one, however, we include interaction terms into the model. These can then each be tested for significance via Wald test.

2 Aims

1. Find key chemical properties that most strongly influence wine quality (alcohol content, residual sugar, PH etc.)

2. Build a classifier to differentiate between red and white or (good quality and bad quality) wine based on chemical composition.

3. Are there clusters within wines based on chemical composition, and do these clusters correspond to quality levels/wine categories?

3 Dataset

Wine quality dataset can be downloaded from here. Wine quality data include white wine quality dataset and red wine quality dataset. There are 12 columns in each dataset and the variables are the same. The description of the variables is shown in Table 1.

The dataset has 6,497 observations in total. We created a binary outcome ‘color’, which includes red and white. There are 1,599 (24.6%) observations for red wine and 4,898 (75.4%) for white wine.

We also plan to calculate the correlation between different variables. If two variables have a high correlation, they may function similarly and thus lead to unstable estimates. Therefore, for variable pairs with correlation higher than 0.8, we remove the second one and keep the first one.

4 Methods

Since the outcome, wine quality, is an integer ranging from 3 to 9, we first subtract all quality by 3 so that the range is between 0 and 6. Then we fit the following binomial regression model:

$$\eta(\mu_i) = \mathbf{X}\boldsymbol{\beta}, \quad \text{where } \eta(x) = \log \frac{x}{6-x} \text{ and } \mu_i = \mathrm{E}[Y_i]. \tag{1}$$

Here Y_i is the quality of the ith wine, $Y_i \sim \text{Bin}(6, \mu_i)$, and $\mathbf{X} \in \mathbb{R}^{n\times(p+1)}$ are the p covariates and the intercept. To fit the model, we first write down the joint likelihood:

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{n} \mu_i^{Y_i}(1-\mu_i)^{6-Y_i} \quad \Rightarrow \quad \ell(\boldsymbol{\beta}) = \sum_{i=1}^{n} Y_i \log \mu_i + (6-Y_i)\log(1-\mu_i). \tag{2}$$

To obtain the maximum likelihood estimates (MLEs), we derive the first and second derivatives. Then we use the Newton-Raphson algorithm to iteratively update the estimates:

$$\hat{\boldsymbol{\beta}}^{(t+1)} \leftarrow \hat{\boldsymbol{\beta}}^{(t)} + I_n(\hat{\boldsymbol{\beta}}^{(t)})^{-1}\dot{\ell}(\hat{\boldsymbol{\beta}}^{(t)}), \quad \text{where } I_n(\hat{\boldsymbol{\beta}}^{(t)}) = -\text{E}_{\boldsymbol{Y}}[\ddot{\ell}(\hat{\boldsymbol{\beta}}^{(t)})]. \tag{3}$$

The model can be used to determine the key properties that most strongly influence wine equality. To test whether the kth covariate is significant, we use Wald test with statistic:

$$W_n = R(\hat{\boldsymbol{\beta}})^\top (H(\hat{\boldsymbol{\beta}})^\top I_n(\hat{\boldsymbol{\beta}})^{-1} H(\hat{\boldsymbol{\beta}}))^{-1} R(\hat{\boldsymbol{\beta}}) \to_d \chi_1^2, \tag{4}$$

where $R(\hat{\boldsymbol{\beta}}) = \hat{\beta}_k$ and $H(\hat{\boldsymbol{\beta}})$ is the Jacobian of $R(\hat{\boldsymbol{\beta}})$. An advantage of using the Wald test is that the MLE is the same regardless of H_0, so only one MLE is required to test different sets of parameters. This is in contrast to likelihood ratio test or score test.

We next investigate the predictive power of covariates on wine type. We consider the following models:

1. Logistic regression: treat the wine type as a zero-one variable and use all other features left as the covariate. We consider a Bayesian logistic regression model of the form:

   $$Y_i \mid \beta \sim_{\text{i.i.d.}} \text{Ber}(\pi_i) \text{ with } \log\left(\frac{\pi_i}{1-\pi_i}\right) = \beta\mathbf{X}$$

   where π_i is the probability of the wine i being red, β represents the fitted values of the parameters and X is our data. However, we expand this to utilize a Bayesian framework and establish priors on the parameters β. To establish informative priors, we plan to rely on existing literature via fitting their estimated parameters for various parameters to a normal distribution.

2. Support vector classifier (SVC): treat the wine type as the response and use all other features as predictors. Use “rbf” kernel to increase the dimension of the predictors.

3. Multi-layer perceptrons (MLP): ReLU activation function and softmax loss function. Optimized with stochastic gradient descent.

Accuracy, recall, and precision will be calculated for each model to evaluate their performance. ROC curve will also be used and the area under the curve will serve as an additional performance evaluation metric.

For the third aim, we will include additional interaction terms in the binomial regression model. Similar as before, the significance of interaction terms can be tested via Wald test, by nesting the reduced model into the full model.

5 Reference

1. 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. ISSN: 0167-9236.

2. Azza Ali, Mervat Abu-Elkheir, Ahmed Atwan, Mohammed Elmogy, Missing values imputation using Fuzzy K-Top Matching Value, Journal of King Saud University - Computer and Information Sciences, Volume 35, Issue 1, 2023, Pages 426-437, ISSN 1319-1578 Sikka, H. (2022, June 1). Combinatorial Analysis of Dimensionality Reduction and Augmentation Approaches on Neural Network Performance

3. Patryk Kudła, Tomasz P. Pawlak, One-class synthesis of constraints for Mixed-Integer Linear Programming with C4.5 decision trees, Applied Soft Computing, Volume 68, 2018, Pages 1-12, ISSN 1568-4946

4. Kudła, Patryk, and Tomasz P. Pawlak. "One-class synthesis of constraints for Mixed-Integer Linear Programming with C4. 5 decision trees." Applied Soft Computing 68 (2018): 1-12.

5. Nazanin Alipourfard and Keith Burghardt and Kristina Lerman. DoGR: Disaggregated Gaussian Regression for Reproducible Analysis of Heterogeneous Data, 2021, 2108.13581, arXiv

6. Ekambaram, Rajmadhan, et al. "Active cleaning of label noise." Pattern Recognition 51 (2016): 463-480.

SVM, SVC:

1. Zeng, Ming. "The classification of red wine quality based on machine learning techniques."

2. Second International Conference on Statistics, Applied Mathematics, and Computing Science (CSAMCS 2022). Vol. 12597. SPIE, 2023.

3. Maita, Ana Rocío Cárdenas, et al. “Process mining through artificial neural networks and support vector machines: A systematic literature review.” Business Process Management Journal 21.6 (2015): 1391-1415.

4. Wang, Jie, Peter Wonka, and Jieping Ye. “Scaling SVM and least absolute deviations via exact data reduction.” International conference on machine learning. PMLR, 2014.

5. Gnip, Peter, Liberios Vokorokos, and Peter Drotár. “Selective oversampling approach for strongly imbalanced data.” PeerJ Computer Science 7 (2021): e604.

Variable name	**Role**	**Type**	**Mean (SD)**	**Median (min, max)**	**Missing**
Fixed acidity	Feature	Continuous	7.22 (1.30)	7.00 (3.80, 15.9)	no
Volatile acidity	Feature	Continuous	0.34 (0.17)	0.29 (0.08, 1.58)	no
Citric acid	Feature	Continuous	0.32 (0.15)	0.31 (0, 1.66)	no
Residual sugar	Feature	Continuous	5.44 (4.76)	3 (0.6, 65.8)	no
Chlorides	Feature	Continuous	0.06 (0.04)	0.05 (0.01, 0.61)	no
Free sulfur dioxide	Feature	Continuous	30.5 (17.7)	29 (1, 289)	no
Total sulfur dioxide	Feature	Continuous	116 (56.5)	118 (6, 440)	no
Density	Feature	Continuous	1.00 (0.003)	1.00 (0.99, 1.04)	no
pH	Feature	Continuous	3.22 (0.16)	3.21 (2.72, 4.01)	no
Sulphates	Feature	Continuous	0.53 (0.15)	0.51 (0.22, 2)	no
Alcohol	Feature	Continuous	10.5 (1.19)	10.3 (8, 14.9)	no
Quality	Outcome	Continuous	5.82 (0.87)	6 (3, 9)	no
Color	Outcome	Binary	Red 1,599 (24.6%) / White 4,898 (75.4%)		no

Note: n (%) was used for binary variables.

Table 1: Summary of variables in the data set