**Predicting Wine Quality: A Conundrum**

Would you like some cheese with that?

ST 599 Statistical Computing and Big Data-Project 3

June 2, 2014

Matt Edwards, Song Hoa Choi, Gina Shellhammer, Kalbi Zongo

**Introduction**

To predict the blind taster quality score of a wine based on chemical tests, we collected “Wine Quality” data from UCI Machine Learning Repository, <https://archive.ics.uci.edu/ml/datasets.html>. There are twelve variables such as fixed acidity, volatile acidity, citric acid, residual sugar, chlorides, free sulfur dioxide, total sulfur dioxide, density, ph, sulphates, alcohol, and taster quality scale 0 to 10 in two datasets, 4898 white and 1599 red vinho verde wine samples from Northern Portugal. The taster quality scale 0 indicates ‘very bad’ and 10 indicate ‘very excellent’. The response variable is the taster quality scale with eleven explanatory variables from various phytochemicals in wine. The median of taster quality in white (n = 2198) and red wine (n = 681) is between 5 and 6 (Figure1). We decided to use 5 and 6 from taster quality scale to predict.

**Description of the machine learning method**

There are steps to obtain the finding, 1) Stratified sampling method 2) Principal component analysis 3) Ordinal regression.

First, we constructed training and testing set by using stratified the quality variables. 37.5% of items in strata were randomly selected to be in the testing set and remaining 62.5% were the training set. Secondly, principal component analysis was conduct to find the scores of wine. Finally, ordinal regression is to predict the ordinal categorical dependent variable – scale with the explanatory variables.

Ordinal regression is one of the general linear models and its formula is similar to logistic binomial regression. This is to predict the ordinal categorical dependent variable – scale with the explanatory variables. Because taster quality scale is ordinal categorical variable, ordinal regression is reasonable to predict the taster quality from various phytochemicals. This is also to determine which the significant effect of various phytochemicals on the taster quality scale. Y ={r1, …rq}, (r1… rq ) with applied the stochastic ordering and the cumulative model, all pairwise different x1 and x2 is P(y ri|x1) P(y ri|x2) for all ri Y or P(y ri|x1) P(y ri|x2) for all ri Y (Herbrich, Graepel, & Obermayer, 1999). If our interest is the observing event, use the odds – j = prob(score j) / prob(score j) or j = prob(score j) / (1-prob(score j)). Thus, the ordinal logistic model is ln(j )= j – \*X1 + …+ \*Xj where j is the threshold values. Selecting model is important to decide whether phytochemical variables effects or not.

**Summary findings**

**----will do it after our meeting tomorrow..**

**Discussion including assumptions/limitations**

* multicollinearity
* High variation

**References**

Ordinal Regression. Retrieved May 27, 2014 from <http://www.norusis.com/pdf/ASPC_v13.pdf>

Herbrich,R., Graepel, T., & Obermayer, K. (1999). Regression models for ordinal data: A machine learning approach. Retrieved May 29, 2014 from

<http://research.microsoft.com/apps/pubs/default.aspx?id=65632>

**Appendix**

**Figure 1. White/red wine taste quality histogram**

**Macintosh HD:Users:choiso:Downloads:white_hist.pdf**

**-------------will add red wine histogram later**

**Figure 2. Random prediction**

**:images:RandomPrediction.pdf**