**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**

The goal of our project is to predict the blind taster quality score of a wine based on chemical tests, using the “Wine Quality” data from UCI Machine Learning Repository, <https://archive.ics.uci.edu/ml/datasets.html>. The response variable is the taster quality scale with eleven explanatory variables from various phytochemicals in wine. There are two datasets, 4898 white and 1599 red vinho verde wine samples from Northern Portugal, we concentrated on the white wine data.

The taster quality is a discrete scale ranging from 0 to 10, with 0 indicating ‘very bad’ and 10 indicating ‘very excellent’. The median of taster quality in white wine is 6 (n = 2198) with no graded as 0, 1, 2, or 10.

**Description of the machine learning method**

Training and testing sets were done with stratified sampling.

*K*-Nearest Neighbor Regression

*K*-nearest neighbor (KNN) is “a supervised learning algorithm where the result of new instance query is classified based on majority of *k*-nearest neighbor category” (Teknomo, n.s.). The KNN regression is using some measure of distance to find the nearest neighbors in the dataset. Entries are ordered by increasing distance, and an “optimal” number, i, of nearest neighbors. An inverse distance weighted average is calculated with the k-nearest multivariate neighbors. In order to obtain the result, we used the fit function from rminer package in R, which offers many regression types.

Ordinal Regression

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. Proportional odds were used that coefficients stay the same, and the intercept value changes. All explanatory variables have the same weight for all categories. We picked the one with the highest probability after put them in possible categories. Model selection using backward and forward was conducted to find the good-fit model. “ordinal” package installed in R.

Multiclass Classification

The classification approach is multiclass classification algorithm called One vs. All.  The algorithm trains logistic regression parameters for each class--it computes the probability of the class. To predict the class for a new observation, the algorithm picks the class with highest probability. We used three data sets such as training set, cross-validation set, and test set given the amount of data in training set. For preventing high bias, fitted 4th-degree polynomial regression. This gives a model with 44 predictors. Downside of this is over fitting (high variance). To prevent over fitting, we applied a shrink (penalty) parameter *lambda* to reduce the effect of each predictor. Model selection is conducted to find the best model with using *lambda* values and computing training set error along with cross validation error.

**Summary findings**

K-Nearest Neighbor Regression had a 59.6% success rate. Nothing was allocated to category 3 or 9.

Ordinal Regression had a 53.3% success rate. Nothing was allocated to 3, 8 or 9.

Classification had a 54.4% success rate. Nothing was allocated to category 3 or 9.

**Discussion including assumptions/limitations**

Assumptions:

* Regular logistic regression: individual logistic regression is independent meaning that the probability of all categories does not sum to one. Also, the category with higher probability is more likely of occur that other categories.
* Multicolinearity is not an issue with prediction.
* Ordinal regression:‘ordering of categories’ has some repercussions.
* K-Neighbors: options available for the “search method” for KNN algorithm were not explored. This changes how the hyper-parameters of the algorithm are tuned.
* Cross validation was not explored.

Limitations:

* K-Neighbors: if category distribution is skewed, larger categories can dominates, which is what we see in our result.
* Regression does not always scale well, adding covariates can bog down the nuber of comparisons, especially with model selection.
* Random or stratified sampling of data to get a reasonable set size could help.
* When two or more categories have the same probability of success, then the approach will just pick one. Aiddiotionaly, the algorithm is computationally expensive but run in about 3mns scalability is an issue for the algorithm.

**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>

Teknomo, K. (n.s.). What is K nearest Neighbors Algorithm? Retrieved May 30, 2014 from

http://people.revoledu.com/kardi/tutorial/KNN/What-is-K-Nearest-Neighbor-Algorithm.html

**Prediction Results for K Nearest Neighbors and Classification**

**Mac HD:Users:choiso:wine-st599:images:KNNRegression_Results.pdfMac HD:Users:choiso:wine-st599:images:Classification_Results.pdf**