Final Project Proposal: Statistical Learning (UNL STAT 983) Imbalanced Classification and Prediction of Wine Quality

by Sarah Aurit, Alison Kleffner, Emily Robinson

March 20, 2021

Needed in Project Proposal (1-3 Pages)

- Project Description
- How and where we obtained our data (Maybe a quick description of it as well? Link to Kaggle)
- Scientific Research Questions you may want to address and corresponding machine learning and statistical learning methods

1 Introduction

- Industries use quality certifications to promote their products. This takes time and must be evaluated by human experts which can be expensive. [Gupta, 2018]
- Machine learning techniques perform quality assurance process with the help of available characteristics of product and automate the process by minimizing human interfere. The work also identifies the important features to predict the values of dependent variables.
- Wine has lots of characteristics...
- Wine quality assessment:
  - Physiochemcial: determined by lab tests, no human expert required
  - Sensory test: Expert required (side note: taste is the sense in which the least is understood about)
  - Relationship between physiochemical and sensory analysis not understood.
  - These factors include intrinsic characteristics (visual, taste, smell), environmental (climate, region, site) and management practices (viticultural practice), as well as physicochemical ingredients (acid, pH, etc.). [Hu et al., 2016]
- Methods of assessment: research in the food industry & machine learning techniques.

2 Data

- Wine Quality Data Set found at `https://archive.ics.uci.edu/ml/datasets/Wine+Quality`. [UCI, Cortez et al., 2009]
- Collection of white (4898 samples) and red (1599 samples) wines (Portuguese "Vinho Verde" wine)
- 12 physiochemical variables: fixed acidity, volatile acidity, citric acid, residual sugar, chlorides, free sulfur dioxide, total sulfur dioxide, density, pH, sulphates, alcohol, and quality rating.
- Quality rating is based on a sensory test carried out by at least three sommeliers and scaled in 11 quality classes from 0 - very bad to 10 - very excellent.
- Separated wine in to 3 classes: (Low Quality: 3, 4; Normal: 5, 6, 7; High Quality: 8, 9)
- Ordered and imblanced (there are many more normal wines than excellent or poor ones)
- **Do we need to talk about any preprocessing of data? (e.g. differing magnitudes of some of the predictor variables)**

3 Classification Methods

In this paper we are going to be looking at two main research questions:

- Developing a model that would work well to classify new wines into three categories of poor quality, normal quality and high quality.It is desirable to classify wines using physicochemical properties since this does not involve human bias that would come into play with human tasters. Methods we are considering to be Compared:
  - Random Forests: In prior research, this seems to work well on just the white wine data set, so would look at how this would work using red and white wine.
  - XGBoost: gradient boosting framework that can be done in R. Previous papers on this data set have used different versions of gradient boosting like adaptive boost
  - KNN
- As described above, an issue with our data is the imbalanced nature of the data set, with most of the wines being classified into the normal quality category. So we are interested in looking at different resampling techniques in order to determine. The different methods we are going to compare are:
  - No resampling
  - SMOTE - an algorithm where oversample the minority classes (so in this case low quality and high quality). Running the risk of overfitting our model
  - Random under sampler: randomly remove cases of the majority class from the data set. Running the risk of losing valuable infomation in our model.

Comment: Need to still clean this up and write into paragraphs but want to wait until earlier stuff writen in order to have it to reference.

References

Uci machine learning repository: Wine quality data set. URL `https://archive.ics.uci.edu/ml/datasets/Wine+Quality`.

P. Cortez, J. Teixeira, A. Cerdeira, F. Almeida, T. Matos, and J. Reis. Using data mining for wine quality assessment. In *International Conference on Discovery Science*, pages 66–79. Springer, 2009.

Y. Gupta. Selection of important features and predicting wine quality using machine learning techniques. *Procedia Computer Science*, 125:305–312, 2018.

G. Hu, T. Xi, F. Mohammed, and H. Miao. Classification of wine quality with imbalanced data. In *2016 IEEE International Conference on Industrial Technology (ICIT)*, pages 1712–1217. IEEE, 2016.