MATH 640 Final Project Proposal

CinCin Fang & Michael Leibert

April 2019

In the war for limited screen time, it is natural for media companies to wonder if they can make their content more likable and sharable. One database to help answer that question is the Online News Popularity dataset available at the UCI Machine Learning Repository. It is a list of just under 40,000 articles published on Mashable from January 2013 to December 2014 with about 45 attributes ranging from category (business, entertainment, lifestyle, etc) to the subjectivity and polarity to the average length words. The goal is to find the attributes that can predict how sharable an article is.

Mashable (`www.mashable.com`) describes itself as "a global, multi-platform media and entertainment company. Mashable is the go-to source for tech, digital culture and entertainment content for its dedicated and influential audience around the globe," (as found on `https://mashable.com/about/`). Currently it has 45 million unique monthly views, 28 million social media followers, and 7.5 million shares per month.

Each row of the dataset consists of one article, the explanatory variables, and the response variable. We will split the dataset into a training and test set, with n = 30000 in the training set and n = 9644 in the test. This will be done by a simple stratified sample based on year. The response variable is the article's number of shares. We hope to classify the articles in the test dataset into five categories: viral, popular, mediocre, unpopular, and obscure via a multinomial logit model. In addition to this, we will also apply a cumulative logit model to find an article's cumulative odds for falling into each one of the five categories.

Because there are so many variables, many of which are varieties of one another, we plan to use different techniques to select variables and form a more robust model, such as penalization and LASSO with L1 and L2 norms. We would also do a thorough analysis of sensitivity of our results to different prior distributions.

In our exploratory analysis thus far, we find the dataset to be well formed and easy to work with, with only a few data quality concerns that should be easy to address before our analysis. The main issue is some articles have been online longer than others. An article published in early 2013 has more time to accumulate shares than an article from late 2014. We may have to normalize the shares variable; however an assumption can be made that an article has reaches close to its maximum amount of shares within a week or month. Thus all articles have reached close to their maximum amount of shares. This will be something we need to look at in a more thorough data exploratory analysis phase.