Michael Mason
ECE/ME 532 Course Project Proposal
Due: 10/22/2020

Overview

For this project, I propose to utilize machine learning to analyze a relatively large dataset provided by Fernandes, et al. through the UCI Machine Learning Repository. The dataset contains a range of 61 identifying attributes for nearly 40,000 articles posted to the entertainment news website *Mashable* between 2013 and 2014. The output metric that will be used to assess the articles is the total number of shares each article produced on social media, with the assumption that "success" of an article is directly tied to this number. The attribute selection is diverse, including Booleans and count values, and can be broken into several subsets, including day of the week of publication, quantifications of the text and article content, and pre-calculated metrics regarding positivity/negativity and subjectivity of each article.

The analysis will employ and compare/contrast three types of algorithms: a simple least squares implementation, a k-nearest neighbors model, and a more advanced neural network approach. Given that several groups of attributes can be identified from the 61 total, for example the several metrics quantifying the tone of the articles, my goal is to perform several levels of analysis using each of these models. Each subgroup of attributes will be tested individually in addition to a more comprehensive, computationally extensive analysis encompassing all attributes, with the lessons of the subgroup analyses used to refine and improve the overall analysis. The final deliverable will be a set of models that use some or all of the given attributes to predict the number of shares a new article will generate, with each model based on one of the three listed approaches. Analysis of the viability and success of each of the approaches will also be developed and presented in the final report.

Dataset

The dataset that I will be using for this project is titled *Online News Popularity Data Set*, and the citation and link are provided in the subsections below. The data incorporates measured attributes of 39,797 articles posted on the website *Mashable* between January 7, 2013 and December 27, 2014. The 61 attributes include basic statistics like word count in the title and in the article, link and tag statistics, more complex grammar-related metrics including unique wordcount and positivity/negativity of the title and article contents, image and video content, day of the week of publication, and genre (with day and genre in Boolean form).

I chose this dataset over others for several reasons:

1. The high number of samples made it more appealing over datasets on the order of 1000 or fewer, mainly to have an abundance for the validation steps.
2. Relative to datasets with a set of potentially disconnected features, the subgroups of features (day of week, word count metrics, etc) will make it possible to examine models using individual subgroups of the overall features, as well as fully-encompassing approaches using all of the attributes.
3. The share counts encompass a wide range of values, with extremes of 1 to ~800,000, but generally ranging from a 200 to 100,000, as seen in **Fig. 1**. Some other promising datasets, with large sample size and attractive features, had a very narrow, low-resolution range of actual result values (i.e. wine ratings that were almost exclusively 4, 5, 6, or 7), so the possibility for a complex, meaningful model seems higher with this set.

The goal of the project will be to find a model that predicts social media shares for an article, which for this project are presumed to be the "success" metric, using the attributes available in the dataset. Some pre-examination will be performed to account for obvious trends in the dataset and normalize the data to account for them, such as if articles posted on weekend days average five times more shares than weekday articles. The reason for this will be to account for these clear biases in the parts of the project that examines unrelated subgroups, such as word count and grammar. As shown in **Fig. 1**, there does not appear to be a clear trend upwards or downwards in the shares over the two-year span of the collection, but these pre-filtering efforts will be aimed at identifying other easily-identified linear trends before the actual models are applied.

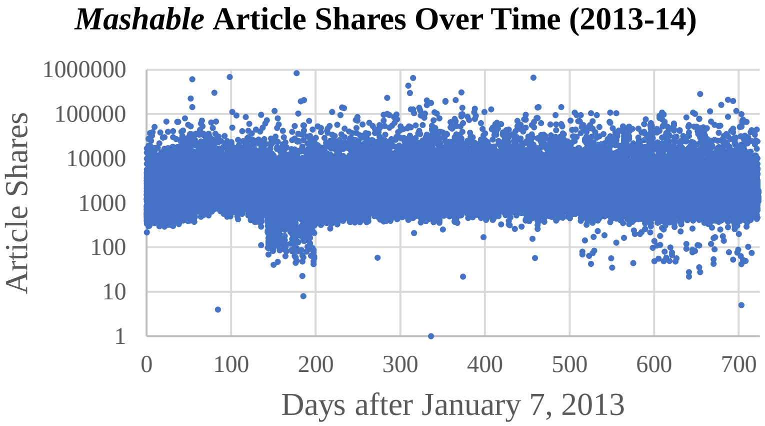

Figure 1: Shares for each of the 39,797 articles in the dataset as a function of the date of publication.

Proposed Project Algorithms

The project will employ three machine learning approaches: a simple least squares implementation, a k-nearest neighbors model, and a more advanced neural network approach. My experience in the course so far is limited to the first two, and so my commentary and predictions for them are more developed at this time than for the neural networking. For the least squares approximations, I believe that some simple trends will be identified especially when applied to certain subgroups of the features, in particular towards the more "obvious" features such as genre. On a related note, the nearest neighbors approach may also reveal clusters of "types" of articles that perform well or poorly.

For the final model incorporating neural networks, I plan on allotting the most time to development and implementation. My first usage of it will be with partial and full sets of all of the available features, but I am also interested in pre-treating the data using potential lessons from the simpler models to create a multi-part model. As this last, all-encompassing step is highly hypothetical and dependent on meaningful trends being developed from the simpler models, I don't want to make a strong commitment to a multi-model implementation, but my timeline (below) will be organized to provide time to develop such an approach if possible at the end.

Validation of all approaches, both those encompassing the full range of features and those using subgroups, will be performed via holdout sets used as testing data for models developed from the rest of the data (training sets). This process will be repeated for holdout sets encompassing different selections from the data as taught in the course, with the average errors from these many holdout tests used to validate each approach.

Timeline

October 22: Submit proposal
November 8: Pretreatment and implementation of least squares and nearest-neighbors approaches complete
November 17: Validation of first two models complete, major effort underway for neural networking, ***update due***
November 27: Neural networking complete
December 1: Validation of all methods complete or near-complete, ***update due***
December 7: Analysis of results complete, report written (*complete early to potentially allow all-encompassing method/model to be attempted, as well as a buffer for unforeseen delays in project*)
December 12: ***Project submitted***

Dataset URL & Citation

https://archive.ics.uci.edu/ml/datasets/Online+News+Popularity

K. Fernandes, P. Vinagre and P. Cortez. *A Proactive Intelligent Decision Support System for Predicting the Popularity of Online News*. Proceedings of the 17th EPIA, 2015.

Link to GitHub Repository:

https://github.com/mamason2/ECE532Project_Mason