STAT1361: Project Proposals

James Hahn

Proposal 1

Title: Predicting the Popularity of Online News Articles

Big Idea: The general idea is to predict the popularity of a news article from a popular digital media news source, Mashable, which is determined by predicting the number of shares the article will achieve.

Data Description:
The data can be found at UCI's Machine Learning Repository: `https://archive.ics.uci.edu/ml/datasets/Online+News+Popularity` . The data comes as a 24MB csv file. It has 39,797 samples across 61 features, with 58 being predictors, 2 being descriptive (i.e. url and days between article publication and dataset collection), and 1 being a response. The response variable is number of shares of the news article. It is important to note all news articles in this dataset are taken from Mashable, a digital media website founded in 2005. Many of them focus on number of links, images, and videos, news category, worst, best, and average keyword shares, number of referenced articles, day published, sentiment polarity, positive words, negative words, text subjectivity, and title subjectivity. As such, it is a well-rounded, in-depth dataset capable of providing a fantastic regression problem for the class. Thankfully, none of the data is missing.

The best part of this dataset is the ratio of number of samples to dimensionality. This can cause a real pain with other datasets that either have too many features for the number of samples, or not enough features to find hard-cut differences in the data. As such, it is the perfect dataset for this classroom setting since it allows students to jump in immediately, rather than disposing of some data. With that being said, that does not mean there is not preprocessing work. Most of the features are one-hot encoded if need be, but the numerical features need to be normalized or fit to some distribution, such as a sigmoid. Also, some of these features seem useless, such as is_weekend, even though weekday_is_saturday and weekday_is_sunday are both features. Dimensionality reduction can be performed to see if it is worth reducing the number of features.

One part of the dataset that makes it especially useful for a statistical learning term project is its mix of binary, or categorical, and numeric features. Many times, if all the data is numeric, it allows for easy curve fitting. Categorical predictors make it more interesting, as more sophisticated learning approaches need to be used or extensive statistical analysis must be performed. Finally, he most interesting features are most likely going to be title polarity, weekday published, and the news category since many people filter news by category, then sift through the interesting stories based on title.

Questions of Interest:
The main goal is predicting overall popularity of a given news article. Thankfully, since the data is high-dimensional, there are a plethora of useful features available for prediction. Additionally, one side question to investigate is popularity within the type of media. For example, each news article has its own category: lifestyle, entertainment, business, social media, tech, and world. One category may be inherently more popular than the other, so it may be worth investigating popularity within each of those six categories to see if specific details for a category can determine popularity, rather than giving a general cure-all formula for any news story, which may not always work. There are other interesting ideas, but the scope of the project should really be limited to avoid extensive time commitment and failure to find results for all the problems.

Interested Parties:
Predicting news popularity is obviously applicable to nearly every media source. More specifically, since this is a Mashable dataset, it is applicable to English-speaking, Western pop news sources. A larger agency, such as the New York Times may be able to gain from this project, but less so since their news sources are not as click-baity. With that being said, with applicable news sources, this can save time and money since it would be provided to an editor to check popularity before its release, forcing the journalist to re-write the story or modify some portions. Because of this, inference is a key portion of this regression problem in addition to the actual prediction.

Proposal 2

Title: Detecting Phishing Websites

Big Idea: We want to predict, given a website's metadata, whether it is a phishing source before a user visits the website and involuntarily forfeits their information.

Data Description:
This dataset is found on a website called Mendeley, located at the following url: `https://data.mendeley.com/datasets/h3cgnj8hft/1`. The important thing to note is the file comes in a 1MB .arff format, but further investigation shows you can open it with any text editor and the data itself can be easily converted into a csv. The dataset was collected during the periods of January-May 2015 and May-June 2017 using Selenium, a popular web automation scripting framework, so it is up-to-date and accurate. Also, the data for phishing sources are collected from PhishTank and OpenPhish, while the legitimate sources are from Alexa and Common Crawl, so the data is reliable. Finally, there are 10,000 samples, 5000 of which are phishing websites and 5000 legitimate sources, across 48 features, therefore not requiring any class balancing. So, the task becomes slightly easier, with a significant number of samples for the number of features. It is important to note there is a different phishing dataset on UCI, but it is not connected to this dataset in the slighest. The UCI dataset merely consists of ternary predictors, making the problem boring and trivial. There are no missing features or samples in the dataset.

All of the data is metadata, such as length of the url, IP address, domain, certificate, and more. I believe the UrlLength, HttpsInHostname, and IpAddress features will be the most useful as they indicate whether the host is legitimate, secure, and the url is easily memorized, which indicates people should be able to travel there with easy access. Also, all features are numeric, some being binary and others continuous. Finally, the NumHash, NumAmbersand, NumUnderscore, and NumPercent features might be able to be reduced to NumSpecialCharacters, reducing dimensionality. I suspect we can do this with other sets of features.

Questions of Interest:
The main goal is predicting whether a website is phishing, which is obviously a binary classification task. A side project to this is predicting confidence, or probability, of a website being a phishing website. However, that is an extra added benefit to the underlying classification task if everything goes correctly and is timely. I do not think there are any other pressing questions of interest, as the dataset is very tailored toward one goal, for a niche area of technology, in a refined industry.

Interested Parties:
Interested parties mostly concern search engines. For example, Google or Bing present tens of links to consumers on every search, so users are vulnerable to unexpected attacks from websites that appear normal. This may lead to loss of user credit card, social security, or bank account information. In this case, the users are fumed at the company for providing access to these links. Therefore, it is in the search engines' best interests to minimize, or at least warn about, the amount of phishing websites.