Machine Learning Project: **PROPOSAL**

Yixi Chen: yca3160, William Ehrich: wde220, Chungyuen Li: clt2503 April 19th, 2017
EECS 349: Machine Learning, Spring 2017 Prof. Downey

Project Proposal

Task

In this project we seek to create a model to predict the number of comments on a Facebook post based on other Facebook information including (1) the characteristic of the page on which the post was made, (2) information about the post itself, and (3) non-commenting behavior of users in response to the post.

Social networking platforms are ripe for investigation because of the rapid growth of data maintained by these services that has taken place since the turn of the century. In particular, Facebook, founded in 2004, has grown to the extent that today it has more than 100 petabytes of disk space in one of its Hadoop clusters and witnesses more than 500 terabytes of data uploads and 2.5 billion content shares daily. This expansion is also seen in Twitter, which handled 5,000 tweets daily in 2007 and 500,000,000 daily in 2013 and is visible in Flickr, which takes in 3000-5000 images a minute.

Given the plentiful supply of new information, Facebook is particularly valuable because Facebook constructs correspond to real life people and things. For example, there are public Facebook profiles for a variety of different use cases, including businesses, brands, celebrity, and others. Therefore, developing a more sophisticated understanding of behaviors on Facebook such as commenting can correspond to a greater understanding of people's views and behavior outside of social media.

Data

Source: http://archive.ics.uci.edu/ml/datasets/Facebook+Comment+Volume+Dataset#
Description: The data is found in the University of California, Irvine Center for Machine Learning and Intelligent Systems and can be accessed via the UCI Machine Learning Repository. With respect to the specifics of the data set, there is information for 40949 Facebook posts, each of which contains 54 attributes with no missing values. The Facebook data was compiled and donated to the repository on March 11th, 2016.

Features

The data set includes a large number of features, including but not limited to:

- Number of likes on the page.
- Number of individuals who have visited the place (if the page corresponds to an institution or place)
- Daily interest of page (measured by number of other comments, likes, posts, shares in a given day)
- Page category (Place, brand, institution)
- Total number of comments before a given date/time
- The number of comments in the preceding 24 hours, and in the preceding 48 to 24 hours
- Number of characters in the post, and number of post shares
- Whether or not the post has been promoted and the day of the week in which the post was made

Initial Approach

Our initial approach will be to implement the C4.5 algorithm for a decision tree with pruning. With respect to data preprocessing, the data set contains no missing values, and already features derived attributes such as the minimum, maximum, average, median, and standard deviation of essential attributes. We will evaluate the accuracy of our prediction model using the Hits@10 metric, which compares the predicted top 10 posts with the actual top 10 posts, measured by comment volume in the several hours following a post.