Final Project: **STATUS REPORT**
Yixi Chen: yca3160, William Ehrich: wde220, Chungyuen Li: clt2503
May 17th, 2017
EECS 349: Machine Learning, Spring 2017
Prof. Doug Downey

Facebook Comment Volume Prediction - Project Status Report

Task

The goal of our project is to predict the number of comments a Facebook post will receive based on characteristics of the page and information about the post.

This pursuit is of great potential value because the number of comments on a Facebook post can be used as a proxy for interest in the subject of the page and content of the post. Therefore, by formulating a model to predict the number of Facebook comments based page and post information, one can gain insight into the thoughts and feelings of people active on social media. This is insight, in turn, can be used by advertisers, marketers, as well scientists studying social media.

Data set

Source:
http://archive.ics.uci.edu/ml/datasets/Facebook+Comment+Volume+Dataset#
Description:
The data source for this project is a collection of information regarding a set of Facebook posts compiled by Kamaljot Singh, who obtained the data using Facebook's API for data crawling and donated the resulting data set to the University of California, Irvine Machine Learning Repository.

The data set includes, for its training set, the number of comments, as well as 54 attributes, of which 29 are base attributes and 25 are derived attributes, for 40949 example Facebook posts. As mentioned in the proposal, the list of attributes includes, but is not limited to the following features:

- Number of likes on the page.
- Number of individuals who have visited the place (if the page corresponds to a physical location)
- 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
- Number of comments in the preceding 24 hours
- Number of comments in the preceding 48 to 24 hours
- Number of characters in the post
- Number of post shares
- Whether or not the post has been promoted
- Day of the week in which the post was made

In addition, the repository includes 10 sets of 100 test examples for measuring the accuracy of the machine learning models. Currently, our team is using all 40949 examples for our training set and all 10 sets of 100 examples for our test sets.

Preliminary results

Our team is taking three-pronged approach to machine learning, in which we are creatings models using (1) regression trees, (2) neural networks, and (3) naïve Bayes. To encode our solution, we are writing Python scripts that utilize the scikit-learn machine learning package, as well as the matplotlib, numpy libraries. To measure accuracy we are using the Hits@10 metric, which take a test set of 100 Facebook posts, ranks the posts by the predicted number of comments and counts how many of the posts in the predicted top ten are in the actual top ten. For the 10 tests, with the regression tree, we obtained Hits@10 scores ranging from 4 to 9, with an average score of 6.4, as can be seen in Figure 1.

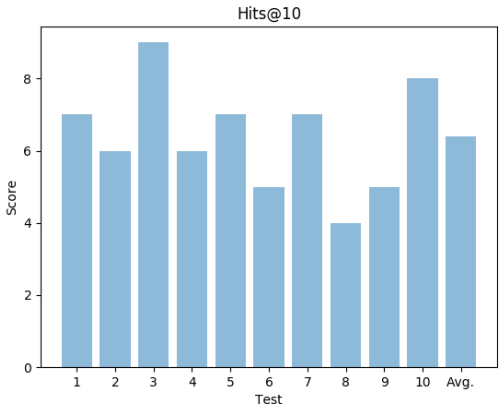

Figure 1: Regression tree Hits@10 Scores

With the use of neural networks, we were able to obtain a Hits@ 10 score of 6 on the first test set. Currently we have not been able to exceed the random baseline accuracy through the use of naïve Bayes.

Future plans

Single Feature Baseline

We will try to determine which particular attribute is the greatest indicator of comment volume using either Weka or a custom Python script. From there, we can measure the accuracy of the regressors when only that attribute is used for training. Alternatively, we can perform an ablation study in which we record the reduction in accuracy when best indicator is removed from the complete set of input attributes prior to training.

Derived Data Set

Once the most important feature is identified, derived values from that attribute will be used to train the repressors. Because the data often varies by magnitudes of 10, the logs of that attribute will likely be used as an input.

Additional Metrics

It appears that Hits@10 is the most appropriate metric for our investigation, but we will consider adding mean-squared error and the ratio of predicted to actual values as additional accuracy metrics.

Analysis

Through our investigation, we will attempt to gather a deeper understanding of the data and try to reason why some regression models feature higher accuracies than others.