EECS 349: Project Status Report

Team Name: kbbz
Group Members: Kevin Chen, Basil Huang, Brittany Lee

Task

Our task is to predict the popularity of a Facebook status based on that user's personal Facebook data. Popularity, measured by the number of "likes" a post gets, is important to users because it eases the anxiety that users feel when sharing on social networking sites. Facebook is one of the core ways people express themselves, but users often struggle to decide whether their thoughts are worth sharing. Our task ensures that users feel comfortable that their post will be approved.

Data Set

Our dataset includes 85,957 instances of Facebook statuses including six features. The data was gathered using each of our group member's Facebook tokens and we collected from them and all of their friends. The features that we collected include the number of friends, age, and gender of the user who posted the status, the time of the status (month and hours), the time since the last status (seconds), and the "score" of the Facebook status itself. To find the score of a Facebook status, a dictionary was built with individual scores of keywords from statuses in our entire dataset. The individual score was calculated by averaging the number of likes that a status with that word receives. Once the individual scores for each word is calculated and stored in a dictionary, the status is scored by averaging those values for each word in the status. The feature that we are predicting is number of likes on each status, which is included in the training and validation set. We created three sets: training, testing and validation. Each set has roughly the same amount of examples and contains a random subset of the entire dataset. The training set includes 29,252 instances, the validation set includes 28,564 instances and the testing set includes 28,414 instances.

Preliminary Results

We tried two nearest neighbor/distance classifiers on our dataset. We used Weka and built our models using the training set and then tested our models using the validation set. The first classifier we tried was IBk and we changed the kNN parameter to 3. This resulted in a root-mean-square error (RMSE) of 12.749 and a correlation of 0.0859. The second classifier we tried was KStar with no parameter changes. This resulted in a RMSE of 11.2465 and a correlation of 0.234.

Next Steps

After collecting our initial dataset we came across a few issues that we hope to fix during the remainder of the quarter. These issues include changing the status time from month and hours to month and the time of day (ex. morning, afternoon, and evening). Currently, the time since last status is recorded in seconds and we hope to change this to hours. Also, we want to collect the user's location as a feature and want our prediction to be a proportion of likes to friends

rather than a straight number of likes value. We also want to figure out exactly which statuses we are pulling right now. There are user who are shown to have no statuses, but when we check Facebook they clearly have posted statuses. We expect that this is a privacy setting issue. The final issue we want tackle in regards to data is that we want to create the word dictionary across all of our status instead of creating a separate word dictionary for each access token. We predict that this will help our status scores be more accurate.

Once we make these corrections to our dataset, we plan to try out more classifiers. We want to try different algorithms and compare the results to make sure that we chose the optimal classifier. After finalizing our model, we will run the model on our test set and visualize the results. We will create the final web page where our report and results will be posted and create the final class video.

Questions

- Do you have any suggestions on ways for us to improve the features that we are currently using (either adding more, or refining the ones we already have)?
- Do you have any suggestions for classifiers that would work better?
- What is an acceptable RMSE? Also, is correlation a useful metric for determining whether our classifier is good?