Project Proposal

Team Michael Coughlin Aaron Holt Michael Hutchison
CSCI 5622
April 3, 2015

Project

We will be working on the default Kaggle project, “When they Buzz”, predicting how long it will take a user to answer a quiz question.

Baseline for first deliverable:

A submission was generated as a baseline by setting the estimate for all users in the test data to the overall mean of all *position* values from the training data, assuming all guesses would be correct. That is, the value 39.30 was used for all users’ estimated *position*. This results in a root mean squared error of 84.45 when submitted to the Kaggle competition.

Approach

We plan to complete this project in two stages. In the first stage, we plan to use logistic regression in combination with an initial set of features in order to determine how well the data can be classified by a simple algorithm. Our initial set of features is listed in the next section.

Based on how well logistic regression performs on the data using our initial feature set, we will select one of several options to further classify the data. If logistic regression performs well, we may simply choose to spend our time engineering more features. However, if it does not perform well, we may need to use additional algorithms, such as SVMs, or even boosting to combine the output of several different algorithms. This will be determined after analyzing the results of our initial logistic regression attempts.

As there are a limited number of submissions to Kaggle on any given day, cross-validation will be used for development. Training will be done on 80% of the data and testing will be done on the remaining 20%.

POTENTIAL FEATURES:

Predicting time to answer (position)

Average time for the user to answer any question
Average time for the user to answer questions in the same category
Average time for all users to answer questions in the same category
Average time for all users to answer the same question
Length of question

Predicting correctness

Average success rate for the same user for all questions
Average success rate for the same user for question in the same category
Average success rate for all users for questions in the same category

Average success rate for all users for the same question
Question vocabulary (long vs short words, vocab diversity)
Number of names/proper nouns in question
Answer type(place, thing, person etc.)
Answer information(look up on wiki/freebase to see if the answer is “well known” via amount of info available)

Timeline	
Date	Item
4-2-15	Project Proposal
4-9-15	First submission
4-9-15 to 5-3-15	Feature engineering, error analysis, alternate classification methods
5-3-15	Final submission, further analysis based on official results
5-6-15	Final Presentation