CS 412: Intro to Machine Learning
Project Proposal

Group Members

Hongwei Jin
Krutarth Joshi
Aayush Kataria
Natawut Monaikul
Ashwin Sattiraju
Zhan Shi
Dan Zhao

Dataset

We will be using the dataset provided directly from Yelp, which can be found at

`http://www.yelp.com/dataset_challenge`.

The data consist of JSON files which contain reviews and tips about businesses, information about businesses such as hours and location, and information about users giving the reviews.

Machine Learning Tasks

Our main task to predict the rating a user will assign to a business. Ratings are a whole-valued number between 1 and 5 inclusive.

Techniques

There are two directions in which we intend on going for this task. One direction is to predict the user's rating based only on the text in that user's review. In this approach, we would take into account the following features:

- Ratio of capital letters

- Positive and negative words
- Frequency of common versus arcane words
- Length of text
- Amount of punctuation
- Occurrences of consecutive repeated letters

The other direction is to predict the user's rating based on the user's profile, i.e., information about the user as well as his/her past reviews of other businesses. In this approach, we would take into account the following features:

- Length of time of elite status
- Number of votes (funny, useful, or cool)
- Average star ratings
- "Yelping since"
- Number of compliments
- Number of friends
- Number of reviews given
- Ratings of past reviews of other businesses, potentially separated by the type of business to find businesses that are closest to the one in question

In both approaches, we intend to create a classifier using these features that will predict to which of five classes (whole-valued star ratings from 1 to 5) a review belongs. We would like to test different methods of classification, including Naive Bayes and decision trees. We plan to see which approach produces better results or if a combination of features from both approaches can produce even better results. Our target programming language is Python.