CS229 Final Project: Proposal

Yue Li (SUID: yulelee)
Haomiao Song (SUID: hmsong)

Proposed Title:

Predicting Yelp User's Rating Based on Previous Reviews

Data

The dataset is from Yelp Dataset Challenge (www.yelp.com/dataset_challenge), which contains about 3.7 million reviews from 687,000 users for 86,000 businesses. Within the dataset, we can find basic information of the businesses (opening hours, address, categories, average rating, parking, alcohol, etc), users (name, review counts, average rating, etc), and reviews (which user, wrote to which restaurant, the content, and the corresponding rating).

Goal

Currently, the interactions between the user and the Yelp application is mainly initiated by the user searching for some keywords, and then go through a list of the matches, potentially ranked by ratings, number of reviews, etc. Since the personalized recommendations are also crucial to better user experience, we want to build a model to recommend new places to users, or to predict whether a user would like a certain restaurant. Moreover, we decide to focus on the most interesting part of the dataset: reviews. Not only because reviews are a fairly complex system, which could potentially contain lots of information, but also because reviews are highly personalized, we can find out a lot about a user just by reading the reviews written by him/her.

Roadmap

We want to first try out the widely used collaborative filtering method to implement a control model, which would utilize the historical ratings of users to restaurants, trying to figure out which users are similar and/or which restaurants are similar to each other. After that, we could then utilize the reviews to build a separate model (or to enhance the first one), and compare the results. For the second model, our current plan is, for each user

and restaurant, extract a real-valued vector according to the reviews associated with it, and define a distance function based on those vectors, so then we can run various clustering and classification algorithms on them (k-means, perceptron, nearest neighbor, SVM, etc). If our model is robust enough, we can even use other review-like data (for example previous Tweets or Facebook status) to predict the restaurant preferences of a certain user.

Input Output and Evaluation

Because of the nature of our dataset, we choose to use the off-line evaluation: using both a user and a restaurant as input, and the output is the user's rating on this particular restaurant, or, under a less restricted setting, to predict whether the user would like this restaurant or not (binary classification).

Current Results

To familiarize us with the dataset and also with the scikit-learn library of Python, we've processed the data of businesses, extracted 6 features (take-out, wi-fi, price range, alcohol, tv, parking), and implemented two basic classification algorithms to predict its successfulness (which is defined by whether the average rating of this restaurant is higher than the median of all the restaurants). The testing error for Naive Bayes is 0.392, and for logistic regression is 0.433. It turns out that, it's hard to predict whether a restaurant would be successful just based on its basic information, that's why we need to build more sophisticated systems.