EECS 349
Project Proposal: Yelp Restaurant Recommender

Robert Luo

May 5, 2018

Background

Yelp is a highly used online platform where users can submit reviews about restaurants and businesses. A business' Yelp rating and its written reviews can attract or repel visitors, and consequently determine the success of the business. Many consumers depend on Yelp as well. In today's day and age, access to crowd-sourced reviews can highly influence consumer decisions on what stores to shop at and restaurants to dine in. Yelpers are, presumably, more likely to visit higher rated establishments. But is there way to predict how much a customer will enjoy their first visit to a new restaurant?

A fortunate result of Yelp's large user base is a publicly available dataset consisting of restaurant descriptions and user reviews. Using this dataset, I will attempt to create a model to predict users' restaurant ratings. A well-functioning predictor of an individual's restaurant rating can serve as a very useful recommender for Yelp, where the platform can suggest restaurants to users which are predicted to receive a high rating. This might not only serve as a means for restaurants to attract new customers, but could also provide Yelp another value proposition to attract new users and retain current ones. Recommenders can also extend beyond the scope of Yelp to any consumer-facing platform, such as Netflix to recommend movies, or Airbnb to suggest locations to visit.

Data Acquisition

Yelp provides a comprehensive dataset at www.yelp.com/dataset. I will be using a subset of the data which includes brief user profiles, user ratings of businesses, and business details (type of business, subcategories such as 'mexican food', location, etc). I will limit my project to data pertaining to restaurants. Furthermore, for the sake of being able to test predictions, I will only consider users that have more than 1 review.

Feature Selection

I hypothesize that some of the most important features to predict a user's preference of a restaurant will be: type of cuisine, price, whether or not alcohol is served, wait time, atmosphere/ambiance (music, noise level, dress style), kid-friendly, wifi, and hours of operation. The dataset offers several other attributes that may be helpful and will be taken into consideration.

Initial Approach

After sub-setting the data as mentioned in the Data Acquisition section, I will aggregate a matrix of users vs. restaurants which includes all user ratings of their visited restaurants. From there, I will divide this matrix into a train and test dataset. I will initially try Item-Based, and User-Based, kNN Collaborative Filtering, since there are many potential attributes that can be considered, and the outputs are categorical (integers 1-5 stars, or even classifications of "low", "ambivalent", and "high" ratings). In Item-Based kNN CF, I will need to define a similarity measure (likely using one hot representation) between restaurants to help kNN find the most similar restaurants that a user has previously rated. I will also need to define a related similarity measure to assist the User-Based kNN CF model, to help find the most similar users as the one considered. Since outputs are categorical, I plan on evaluating the accuracy of each model based on cross-validation accuracy. I will also evaluate my modes by comparing cross validation accuracies to the accuracy achieved by simply using the mode rating of the restaurant as the predicted class.