Benjamin Arancibia

12/20/2015

CUNY IS 622 – Machine Learning and Big Data

Building a Recommendation System for Restaurant Reviews

**Introduction**

The objective of this project is to build a recommendation system for restaurants using collaborative filtering. The process of identifying similar users and recommending what similar users like is called collaborative filtering. This project is structured in the following manner:

1. Exploratory data analysis
2. Neighborhood-based CF recommender
3. Predict a rating that user gives a restaurant that the user has not encountered before
4. Error analysis

**Background**

Collaborative filtering (CF) is a technique that is used by recommender systems. CF is the process of filtering for information or patterns using techniques involving collaboration among multiple agents, viewpoints, and data sources. CF is a method of making automatic predictions (filtering) about the interests of a user by collecting preferences or taste information from many users (collaborating). The underlying assumption of the collaborative filtering approach is that if a person A has the same opinion as a person B on an issue, A is more likely to have B's opinion on a different issue x than to have the opinion on x of a person chosen randomly. These predictions are specific to the user, but use information gleaned from many users. This project uses the Yhat example of building a beer recommendation system found [here](http://nbviewer.ipython.org/gist/glamp/20a18d52c539b87de2af).

**Methodology**

*Exploratory Data Analysis*

The first step in the analysis is collect the data from Yelp on restaurant reviews and then joined information about users and businesses. Fields in this data are:

['user\_id',

'business\_id',

'date',

'review\_id',

'stars',

'usefulvotes\_review',

'user\_name',

'categories',

'biz\_name',

'latitude',

'longitude',

'business\_avg',

'business\_review\_count',

'user\_avg',

'user\_review\_count']

After preparing the data, an exploratory analysis was performed. The exploratory analysis that was performed were to plot histograms of the review count grouped by the user\_id and business\_id. There are around 34,789 users and 4,503 business that with review counts. Afterwards then the average rating of reviews in the data set and a histogram of all the ratings in the dataset. The next step in the exploratory data analysis is to plot histograms of the average user rating in the smaller data set, and the average business rating in the data set. The last part of the exploratory data analysis is to look at the common user support (the number of common reviewers) of each pair of restaurants on the data set.

*Similarity*

The next step in the analysis is to calculate similarity and then create a database of similarities. By calculating similarities, the k-nearest restaurants to a given restaurant based on the database of similarities are calculated. The issue with this methodology is that there might be a small number or large number of common reviewers. This is resolved by normalizing the the common reviewers by:

After normalizing the data, the recommendation was made based on the idea of finding restaurants that users might also like based on the nearest neighbor to the original restaurant. After performing the nearest neighbor recommendation, the top recommendations for a user were then calculated.

*User Based Recommendation*

The next step in the analysis and creation of the recommendation project was to create a user based recommender with predicted ratings. The first part of this analysis define the predicted rating and then predict the rating for a user and an item. Then compute the predicted rating and compare it with the average rating over all users available.

*Error Analysis*

The next step in the project is to take a set of actual ratings, and a set of predicted ratings, and plots the latter against the former. This graph allows for a user to see how accurate the predictions are for the recommended restaurants. For selecting the nearest neighbor parameter of k, users are selecting a k that avoids outliers and this can be achieved by plotting graphs to see which k gives the most structured predictions with minimum outliers. Also by plotting the graphics, the user can see how close the mean lies to the slope of the predictions.

**Results**

**Conclusion**