**MovieLens Datasets Study**

Zhuangye Chen(Michael), Alexander Trent Dewey, Xingxing Tong(David)

Summary

…details needed…

Part 1 Introduction

**) Background of the recommender systems**

Recommender systems apply statistical and knowledge discovery techniques to the problem of making product recommendations.[2] Good recommender systems can help customers easily find the products they want, and create a value-added relationship between buyers and sellers.

There are generally two types of recommender systems: Content-based filtering & Collaborative filtering (CF). Two types of systems share one “consistency” assumption which is those who agreed in the past tend to agree again in the further. Content-based filtering is based on the attributes of products. Consumers preference’s is generated by the attributes of products their have bought before. However, on the other hand, collaborative filtering is based on the performance of similar users. In other words, defining groups of similar users is the key process in the type of recommender systems.

We focus on the application of CF on movies recommendations. So much emphasis will be put on this particular area in the rest of introduction part. Based on CF we can predict the rating of an unrated movies or create a top-N list of unrated movies for a specific user. Memory-based CF focus on finding K nearest neighbors for the user in the user-item matrix. Then recommendation will be given based on the items liked by the K nearest neighbors. Model-based CF tries to build a model from the rating data (clustering, latent semantic structure, etc.) and then us this model to predict missing ratings.

**) Background of the MovieLens datasets**

MovieLens datasets was first released in 1998. People’s preferences for movies were collected in this dataset. Nowadays, these datasets are widely used in education, research, and industry. They are downloaded hundreds of thousands of times each year. This popularity is, to a certain degree, a reflation of the incredible rate of growth of personalization and recommendation research, in with datasets such as these have substantial value in exploring and validating ideas. [1]

Part 2 General Process and Key Question:

Our project focus on the MovieLens 100K dataset which contains 100,000 ratings (1-5) from 943 users on 1682 movies. Three crucial subsets we focused are “u.data”, “u.genre” and “u.user”. These three sub datasets contains information of movie rating scores, user’s demographic information and movie’s genre. Emphasis is put on digging the raw datasets in the first stage. Interesting stories are found via visualization and basic data summary. Along the way, we decided to focus on dataset “u.user” and building an “affinity” matrix to describe the similarity of users which is also the first step of CF recommender system. The closeness is defined by a distance mainly based four demographic variables in the sub dataset. In the fourth part, we use this “affinity” matrix to make rating prediction of users.

…details needed…

Part 3 Initial Exploration:

**Different behavior between movie lovers and enthusiasts**

For each user we counted the number of movies they rated and plot the result.



Figure1. Histogram of the number of movies users rated (Blue dashed line indicates median)

The number of movies users rated ranges from 20 to 737. The median is 65. So we divided our user group into two groups based on this median. Users who rated less or equal to 65 movies are named movie lovers and those who rated more than 65 movies are named movie enthusiasts. Then we combined the demographic information in “u.user” dataset with the division which gives us two graphs in the following.



Figure2. Distribution of average rating scores for two groups given each occupation



Figure3. Distribution of occupation and gender for two groups

…details needed…

**Genre**

A similar method is applied to “u.genre”. Instead of user, here we focus on the movie. The figure below shows the average rating for each genre of film. The figure below shows the number of genres associated with each film film. The exact distribution is (833, 569, 215, 51, 11, 3) for i = 1, 2,..., 6 labels. There is about a 50% chance that a given film will have multiple genres.



Figure4. Graphical summary for genre

In the figure above, note that comedies and dramas dominate the genre (reflected by the size of the points). Comedies are notably rated around 0.3 stars lower than dramas. Given that most of the variation in ratings by genre is primarily a result of having very few films, this sticks out as the most obvious trend--"fun" genres like action, children's, horror, and family are very low, while "serious" genres like Drama, Documentary, Crime, Mystery, Romance, and War are very high rated.

Part 4 Ratings Adjustment

A general assumption we made is that the big problem of recommendation is not quality level, but the general problem of search within a niche or genre of interest. Different users have different tastes and may even have different aversions to risk, but it's unlikely that they will completely disagree on most movies' quality level--what's more important is that users who log in knowing very little about a movie will tend to be presented with movies they and users like them will tend to want to watch.

Given the above discussion of genres, it was important to adjust for rating variations by genre (and that that films may have many genres). So we ran a simple regression which gave us a coefficient for each genre describing its effect size on star rating. This allowed us to fit an expected score for each film based only on its genre.

(NOTE: In order to prepare the design matrix for this regression, we simply divided each indicator variable by the number of genres for each film. Therefore, if Jaws is listed as Action/Adventure, the predictor row for Jaws would be all 0s except for Action and Adventure, which would get 0.5. And the response variable was the average rating over all users. We should probably weight this somewhat by number of ratings or exclude films with few ratings.)

Based on this regression and its fitted values, we created a "genre-adjusted rating" by subtracting the fitted value from the film's actual rating (with a prior of 20 ratings at the expected genre rating). This helped us isolate films that were better than expected for their genre. For example, while Jaws was 321st among all 1682 movies (or around the 80th percentile) with an average rating of just 3.775, it was 8th using this adjustment at 1.123 stars above its expected genre rating.

It doesn't seem that that this algorithm actually isolates quality "independent of genre" any better than just using the prior--but it does seem to isolate quality relative to genre expectation--someone watching Jaws or The Birds, knowing only its genre, wouldn't necessarily be expecting an exceptional film.

Similarly, someone deciding what to watch may not have the choice of genre, as they may, for example, have to choose among children's movies--knowing whether a movie is good "for a kids' movie" is often more important than knowing whether it's better than David Lynch's "Lost Highway". Finally, someone might only be looking for comedies, and "Jean de Florette"'s rating is probably somewhat irrelevant to them.

Part 5 Affinity Matrix

A basic goal of this project is to build a recommendation engine, and one natural place to start is figuring out how users and/or movies are clustered together somehow.

For users we decided to use a demographic similarity score for each ordered pair of users. The math is a little bit complex for this, but in essence we quantified age, gender, and salary (with the proxy of occupation) and tested if two users were close to one another by the correlation of these demographic variables. Since we computed the score for each pair, the end result is a "similarity matrix" or "affinity matrix", with the diagonal being 1 and each entry varying from -1 to 1.

Demographically, we used a simple proxy for salary from occupation by informally checking recent government salary statistics, with "students, homemakers, and none" being assigned to $20,000 and "other" being assigned to the U.S. median of $50,000. Gender was recorded as a binary predictor and age was continuous, from 0 (youngest) to 1 (oldest).

From here we noted that it would be possible to build an affinity matrix for movies in a similar way, by noting if there were movies which had both been rated more or less than expected assuming statistical independence. For example, if Jaws has been rated by 50% of users and the Godfather has been rated by 30% of users, we'd expect 15% of the users to have rated both if they were independent--the proportion more or less than 15% could serve as a measure of correlation. We haven't built this yet, but it could allow us to weight the genre-adjusted ratings above for a user who has rated a certain subset of movies--by recommending films that are substantially correlated with the films they've seen. Two problems that make this difficult are handling films with lower proportions and incorporating a measure of similarity by genre.

**Reference:**

[1] F. Maxwell Harper and Joseph A. Konstan. 2015.TheMovieLens datasets: History and context. ACM Trans.Interact. Intell. Syst. 5, 4, Article 19 (December 2015), 19 pages.

[2] (Sarwar et al., 2000).