PROJECT 5

stat4squirrels

Xiaoqian Zhang

Xinyi Liu

Shan He

Rachneet Kaur

1. Introduction

Our goal is to build a movie recommender system based on the MovieLens 1M Dataset. In the train.dat, it contains about 60% rows of the ratings.dat from the MovieLens 1M dataset (of the same format). And in the test.csv, it contains about 20% of the user-movie pairs from the ratings.dat from the MovieLens 1M dataset. In our data preprocessing procedure, we get the rating matrix and the movie feature matrix at first. Then we apply content-based method and collaborative filtering method to predict the ratings.

2. Preprocessing

2.1 Building Rating Matrix

Firstly, we replace the separator as comma and reread the data so it can be recognized correctly. Then reshape the dataset into a matrix so that its row names are userID, column names are movieID and values are ratings. In this way we get the sparse rating matrix.

2.2 Building Movie Feature Matrix

We also need to change the delimiter for this dataset and read it in a correct way. Next, we create a feature matrix based on the movie genres. If the movie belongs to certain genre, its value under this column is 1, otherwise 0.

2.3 Building Similarity Matrix

Based on the movie feature matrix we created in last step, we further calculate the correlation between every two movies and get a correlation matrix. The row and column names are movieID.

2.4

3. Model Selection

3.1 Content-Based Method

By the content-based method, we can only predict the ratings of old users instead of those of new users. Thus, we will use content-based method to predict the ratings of old users and use the average ratings of the user group to predict the ratings of new users.

Using the content-based method, firstly we need to normalize the rating matrix for each user. Then for each old user and for each movie, we can get all the ratings of the user from the rating matrix and the similarity among this movie and other movies from the similarity matrix. Then we can calculate the dot product of the specific row from rating matrix and the specific column from the similarity matrix, divided by the sum of the similarity weights. At last, we can get our predicted ratings by adding back the normalization terms.

3.2 Collaborative Filtering Method

Using item-based CF and user-based CF to build the recommendation system. Also, we use different kinds of method to measure similarity, including Jaccard similarity and Cosine similarity. After that, we compare the RMSE between predicted ratings and true ratings and select the best performed model.

Since the collaborative filtering model can only predict ratings in the rating matrix, which only includes old movies and old users. Thus, for new movies, we use results of the content-based model to predict their ratings, while for new users, we use the classification model to predict their ratings.

The table below shows models' performance. We can see that the Popular method has the smallest RMSE and the shortest running time. Thus, we choose it as our final model 2.

Model	Similarity Measurement	RMSE	Time(min)
UBCF	Cosine	1.018613	30.81
UBCF	Jaccard	1.013895	25.17
IBCF	Jaccard	1.235863	13.29
Popular	----	0.938186	6.90

3.3 Grouping and Prediction for New Users

Methods we mentioned above can only deal with prediction for old users. But for new users, we only have personal information like gender and age. So we group all users into several groups based on the personal information. And calculate the mean score of ratings for each movie in each group. The prediction of certain new user for a movie is the average rating he/she belongs to for that movie.

4. Result

Model	RMSE	Time (Minute)
Content-based	1.141890	2.84688
Popularity-based	0.938186	2.55765

In the first model, we get the ratings of the old users by the content-based method and the ratings of the new users by the average ratings of the specific user group.

In the second model, we get the ratings of the old movies and old users by the popularity-based method, the ratings of the new movies and old users by the content-based method, the ratings of the new users by the average ratings of the specific user group.