Predict movie rating based on User Based Collaborative Filtering(UBCF)

Author: Manas Ranjan Goth

This is a problem of predicting user ratings for various movies not rated by the user. There is a data file 'ratings.txt' which has the movie reviews from movielens. There are 1,00,000 lines of data. The first few lines are shown below:

userId	movieId	rating	timestamp
1	1061	3	1260759182
1	1129	2	1260759185
1	1172	4	1260759205
1	1263	2	1260759151
1	1287	2	1260759187

The idea is to split the data into training and test data set by means of stratified sampling. We will predict the ratings for the test data set and result would look like:

userId	movieId	actual rating	predicted rating
1	1029	3	2
1	1293	2	2
1	1343	2	2
1	1953	4	2
2	52	3	4
2	150	5	4

To achieve predicted rating, we will apply user based collaborative technique (UBCF). The assumption is that users with similar preferences will rate items similarly. Thus missing ratings for a user can be predicted by first finding a neighborhood of similar users and then aggregate the ratings of these users to form a prediction.

The neighborhood is defined in terms of similarity between users. We will apply the popular similarity measures Jaccard similarity, Pearson correlation coefficient and the Cosine similarity. We will use the R-package: recommenderlab.

A highly recommended read on the recommendation algorithms can be found here .

On the training data we will apply the prediction algorithm, which will predict a rating for every movie that a user has not rated.

Once the prediction is done, then we will parse the testing data row wise. For every user ID and movie ID combination we will extract the predicted rating.

We will perform the above 2 steps for all the 3-similarity metric, one at a time. Post this we will calculate the Normalized Mean Absolute Error (NMAE) for all the three models and perform comparison.

NMAE=Σ(|predicted rating – real rating|) / n(max rate – min rate)

Execution report:

Package used	Recommender
Sampling method	Stratified
Number of records in data set	1,00,000
Number of unique users	671
Number of unique movies	9066
Training set – % of records	80 %
Testing set – % of records	20%

NMAE scores for models generated by applying different similarity measure:

Similarity measure	NMAE
Cosine	0.184677
Jaccard	0.184433
Pearson	0.184611

Analysis of different models:

For the given data set the NMAE scores for different models are very similar and hence a best model cannot be decided. On an average the NMAE is 18% for all three models. This translates to an absolute error value of around **0.8 rating on an average**.

Details of attached files:

File name	Objective
Manas_RS.R	File containing R code
ratings.csv	Data set
submitfile_cosine.csv	Output based on cosine similarity
submitfile_jaccard.csv	Output based on jaccard similarity
submitfile_pearson.csv	Output based on pearson similarity
Github link	Github link