Assignment 2: Recommendation System

By,

Ch Vishal 2015B5A70605H

Tanmay Kulkarni 2015B3A70647H

Phani Shankar Ede 2015B3A70420H

Aditya Addepalli 2015B1A70719H

Movie Recommendation System

The aim of the project is to obtain Movie recommendations for users given their initial preferences in the form of rating. We do this with the help of three methodologies, the first being Collaborative Filtering. In addition, we have made use of SVD and CUR as an additional way to extrapolate unknown ratings.

Methodology:

1. Collaborative Filtering

This can be implemented as either user-user CF or item-item CF. We use item-item CF for our purpose. It's methodology can be summarized by the following:

i. Similarity Calculation: First, we calculate similarities scores between each item using cosine similarity.
ii. We take the top k similar items and use their ratings for various movies to obtain a weighted average. This is used to replace movie ratings for users where ratings did not exist before.
iii. In the Baseline case, we first find the global average for every movie. This is followed by the average rating for that movie for all users and the average rating for a user.
iv. We get the base score by adding the values of the global scores to the average score of the user and the average score of the movie. In addition to this, we add the ratings given by different users for the same move adjusted for their rating bias and using their cosine similarity as the weight for the weighted sum.

2. SVD

This methodology tries to map the ratings data into concept space by decomposing it into 3 different matrices:
U: This matrix contains the strength of association between user and concept.
Sigma: This matrix is a diagonal matrix which contains the strength of each concept that can be obtained from the dataset
V^T: This matrix contains the strength of association between concept and movie.

On multiplying these 3 matrices together, we obtain reconstructed version of original matrix with imputed values for missing ratings.

For the 90% energy case, we keep removing concepts along with their corresponding rows and columns in U and V^T matrices until 90% of original Standard Deviation can be explained.

3. CUR

This methodology tries to decompose the original rating matrix into 3 different matrices C, U, R.
To obtain our C matrix, we randomly sample crows from the rating matrix.
To obtain our R matrix, we randomly sampler rows from the rating matrix.
The intersection of these 2 matrices gives us W. This can be decomposed using SVD into X, Sig, Y^T matrices. The final U matrix can be computed using the following:

$U= Y(psuedo\_inverse(Sig))^2X^T$

Similar to SVD, on multiplying these 3 matrices together, we obtain the reconstructed version of the original matrix with imputed values for missing ratings.

For the 90% energy case, we keep removing concepts from Sig along with their corresponding rows and columns in X and Y matrices until 90% of original Standard Deviation can be explained.

Running Time:

Technique	RMSE	Precision on Top K	Spearman Rank Correlation	Time taken to predict
Collaborative	1.3006	0.65	-0.006	251.43 seconds
Collaborative with Baseline	1.9342	0.45	0.017	378.58 seconds
SVD	1.9134	0.55	-0.043	320.85 seconds
SVD with 90% retained energy	1.9128	0.57	-0.074	419.60 seconds
CUR	1.9474	0.64	0.004	65.18 seconds
CUR with 90% retained energy	1.9239	0.67	-0.003	75.26 seconds

Flow of Control

1. We first parse the rating file by using Python's inbuilt OS routine. The end product obtained would be a dictionary containing the rating corresponding to every (movie, user) pair. The number of similar neighbors to look at is taken as 20.
2. Next, we apply the following pipeline for each of our methodologies :
    a. Divide the data into Training and Testing Data with 70:30 mix
    b. Separate Training data into another Ratings Matrix
    c. Following this, we impute the missing values in Ratings Matrix using Collaborative Filtering or SVD,CUR approach.
3. Now, we use our obtained matrix along with generated test data to compute our evaluation metrics: RMSE, Spearman Rank Correlation and Precision at Top K.