Collaborative Filtering
End-Semester Exam

Mohammed Kashif
MT15035

Approach Used :

We had to apply matrix factorization on 10M movielens dataset . The Divide and Conquer approach used in the paper can be divided into 3 parts :

1. **Divide :** First the 10M dataset had to be divided into smaller parts since it could not be loaded into the memory .
2. **Factorize :** Then matrix factorization is performed on these individual parts .
3. **Conquer :** Finally these individual results are combined to give an approximate result for the complete dataset .

Divide Step :

Total number of users in 10M dataset : 71567
Total number of items in 10M dataset : 10681

Strategy Used : Random Sampling of Rows

As mentioned in the paper , one way of dividing the dataset is to partition it into T L-column sub matrices , given that mod(no_of_columns , T) = 0. Now the factors of 10681 are 1 , 11 , 971 and 10681 . So I chose T as 11 , hence L is 971 .

The steps for Divide strategy are as follows :

1. Create an array containing integers 1 to 10681.
2. Randomly shuffle the array using numpy shuffle method .
3. Partition the shuffled array into 11 parts of equal size .
4. Create a hashmap named cluster as follows , suppose if the number 5 falls into the 2nd bin , the cluster[5] = 2 .
5. Create an instance of dataframe to read the dataset . **NOTE : THIS DOES NOT LOAD THE WHOLE DATASET BUT CREATES AN INSTANCE**
6. Read the data LINE BY LINE only and store the line in variable ‘line’
    a. Now line has the data of the form “ UserID::ItemID::Rating::Timestamp”
    b. Find the cluster for the itemID using the hashmap created in step4 .
    c. Cluster_ID = cluster [ItemID in the line]
    d. Store the line in file named train_part_cluster_ID.csv .
    e. Read the next line
7. Once the file is read line by line , 11 parts will be created having the name train_part_1.csv , train_part_2.csv , etc. and so on , each containing matrices of size no_of_users X L , i.e. 71567 X 971

Factor Step :

I use the factorization as discussed in Assignment 2 for each of the individual chunks .

Combine Step :

Input at this step : The Factorized user and Item latent vectors are stored in the individual files named as C_hat_1.txt , C_hat_2.txt and so on . Each file contains the corresponding U and V matrix obtained after matrix factorization . These will be denoted by U_1 and V_1 and so on .

Approach Used :I used the equation given in the paper

$$\mathbf{L}^{proj} = \mathbf{C}\mathbf{C}^{+}\mathbf{M}$$

As mentioned in the paper , I project [C_hat_1 C_hat_2 … C_hat_11] into the column space of C_hat_1 by using the following equations :

X . U = C_hat_1 . U

X . V = [C_hat_1.V (C_hat_1.V * pseudoInverse(C_hat_1)*C_hat_2.U)*C_hat_2.V
(C_hat_1.V * pseudoInverse(C_hat_1)*C_hat_2.U)*C_hat_3.V
…….
(C_hat_1.V * pseudoInverse(C_hat_1)*C_hat_2.U)*C_hat_11.V]

Step for combine step are as follows :

1. Read the first cluster ,i.e. C_hat_1
2. STore the final result in X
3. Now X.U = C_hat_1.U
4. Calculate X.V using the equation shown above for the rest of the clusters .
5. Now X.U and X.V contain the user and Item latent vectors respectively .
6. Once the latent factors for users and items are obtained I predict the ratings and compare the ratings with the test ratings and calculate the NMAE .

Number of Factors	**NMAE**
40	0.29987