Machine Learning with Matrix Data for Recommender Systems

1. Recommender systems are a hot topic. Recommendation systems can be formulated as a task of matrix completion in machine learning. Recommender systems aim to predict the rating that a user will give for an item (e.g., a restaurant, a movie, a product).

2. Download the movie rating dataset from: https://www.kaggle.com/rounakbanik/themovies-dataset. These files contain metadata for all 45,000 movies listed in the Full MovieLens Dataset. The dataset consists of movies released on or before July 2017. Data points include cast, crew, plot keywords, budget, revenue, posters, release dates, languages, production companies, countries, TMDB vote counts and vote averages. This dataset also has files containing 26 million ratings from 270,000 users for all 45,000 movies. Ratings are on a scale of 1-5 and have been obtained from the official GroupLens website.

3. Building a small recommender system with the matrix data: "ratings.csv". You can use the recommender system library: Surprise (http://surpriselib.com), use other recommender system libraries, or implement from scratches.

a. Read data from "ratings.csv" with line format: 'userID movieID rating timestamp'.

b. MAE and RMSE are two famous metrics for evaluating the performances of a recommender system. The definition of MAE can be found via: https://en.wikipedia.org/wiki/Mean_absolute_error. The definition of RMSE can be found via: https://en.wikipedia.org/wiki/Root-mean-square_deviation.

c. Compute the average MAE and RMSE of the Probabilistic Matrix Factorization (PMF), User based Collaborative Filtering, Item based Collaborative Filtering, under the 5-folds cross-validation

Probabilistic Matrix Factorization (PMF)

Running recommendation algorithm PMF with 5 fold cross validation results in :

```
Evaluating RMSE, MAE of algorithm SVD on 5 split(s).

                  Fold 1  Fold 2  Fold 3  Fold 4  Fold 5  Mean    Std
RMSE (testset)    1.0190  0.9982  1.0069  1.0066  1.0006  1.0063  0.0072
MAE (testset)     0.7856  0.7720  0.7798  0.7792  0.7726  0.7778  0.0050
Fit time          6.27    6.74    7.92    7.24    6.10    6.85    0.66
Test time         0.14    0.17    0.12    0.28    0.13    0.17    0.06
```

Mean of RMSE across 5 folds cv = 1.0063

Mean of MAE across 5 folds cv = 0.7778

User based Collaborative Filtering

Running recommendation algorithm User Based Collaborative Filtering with 5 fold cross validation results in :

	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	Mean	Std
RMSE (testset)	0.9282	0.9127	0.9160	0.9176	0.9183	0.9185	0.0052
MAE (testset)	0.7093	0.6979	0.7008	0.7048	0.7036	0.7033	0.0038
Fit time	0.32	0.38	0.30	0.33	0.38	0.34	0.03
Test time	3.07	2.30	2.53	2.40	2.81	2.62	0.28

Mean of RMSE across 5 folds cv = 0.9185

Mean of MAE across 5 folds cv = 0.7033

Item based Collaborative Filtering

Running recommendation algorithm Item Based Collaborative Filtering with 5 fold cross validation results in :

	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	Mean	Std
RMSE (testset)	0.9198	0.9233	0.9165	0.9192	0.9180	0.9193	0.0023
MAE (testset)	0.7049	0.7041	0.7007	0.7046	0.7053	0.7039	0.0017
Fit time	0.22	0.26	0.27	0.33	0.27	0.27	0.03
Test time	2.04	1.70	2.05	1.89	1.79	1.89	0.14

Mean of RMSE across 5 folds cv = 0.9193

Mean of MAE across 5 folds cv = 0.7039

d. Compare the average (mean) performances of User-based collaborative filtering, item-based collaborative filtering, PMF with respect to RMSE and MAE. Which ML model is the best in the movie rating data?

```
pmf_mae     0.777836
ubcf_mae    0.703279
ibcf_mae    0.703930
```

Looking at the MAE comparison the User based collaborative filtering has the lowest MAE among all

```
pmf_rmse     1.006258
ubcf_rmse    0.918548
ibcf_rmse    0.919346
```

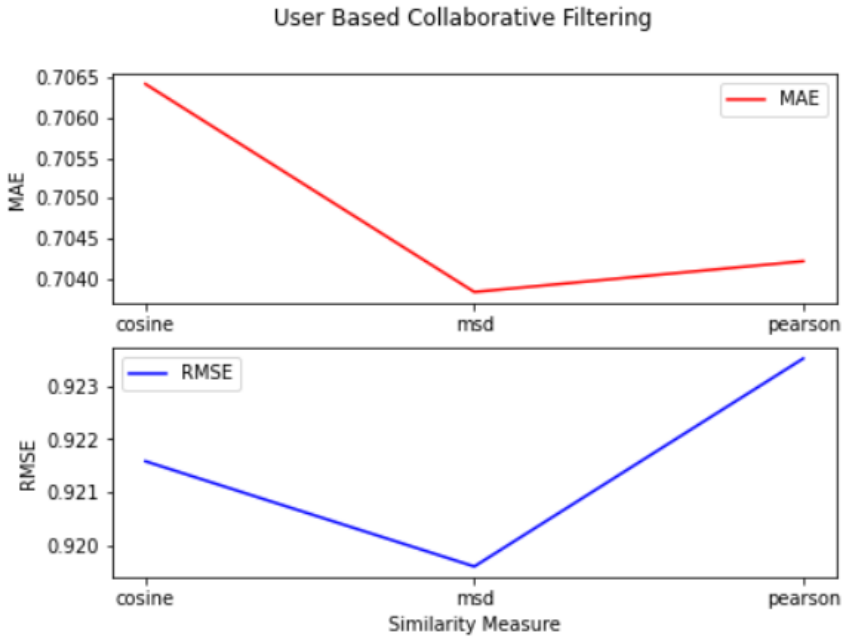
Looking at the RMSE comparison the User based collaborative filtering has the lowest RMSE among all

Considering both MAE and RMSE the User Based Collaborative Filtering with lowest error is relatively the best

e. Examine how the cosine, MSD (Mean Squared Difference), and Pearson similarities impact the performances of User based Collaborative Filtering and Item based Collaborative Filtering. Plot your results. Is the impact of the three metrics on User based Collaborative Filtering consistent with the impact of the three metrics on Item based Collaborative Filtering?

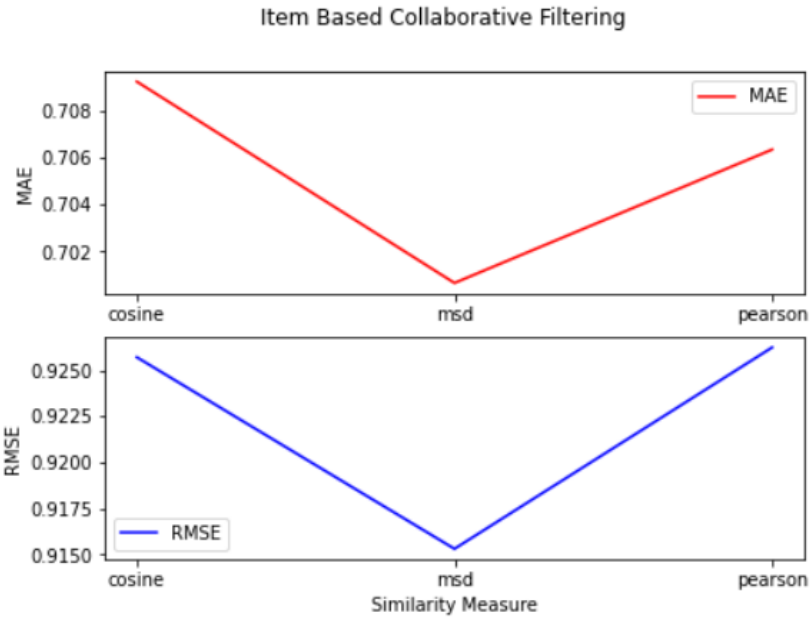
For the User Based Collaborative Filtering the performance measures using different similarity methods are as below:

	cosine	msd	pearson
MAE	0.706422	0.703830	0.704213
RMSE	0.921581	0.919601	0.923514



For the Item Based Collaborative Filtering the performance measures using different similarity methods are as below:

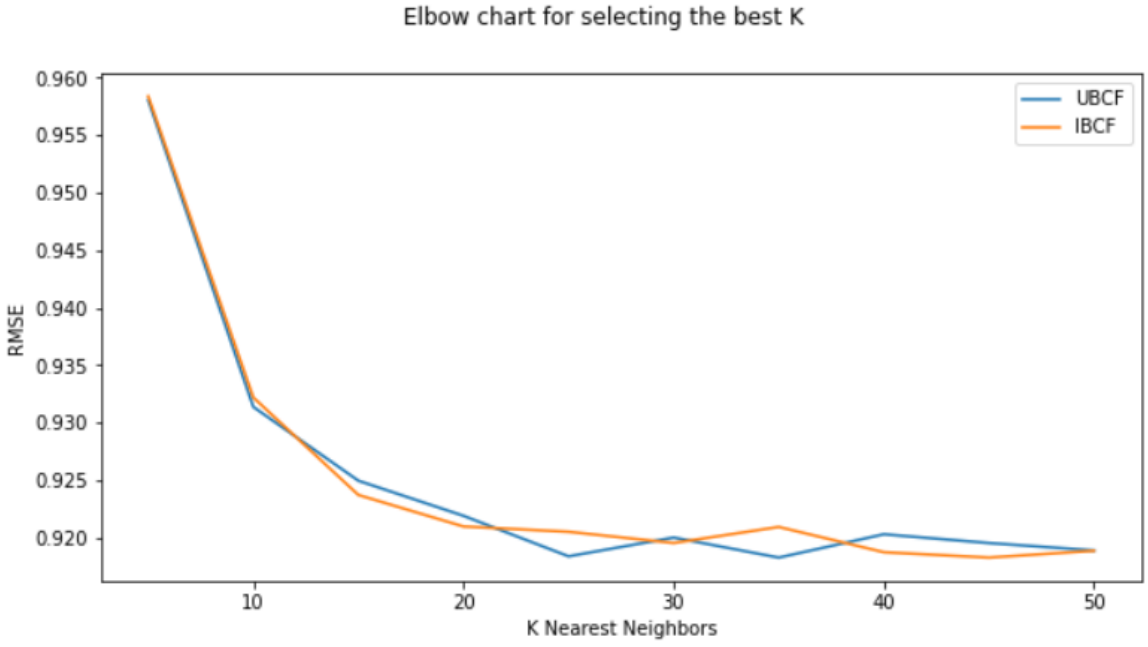
	cosine	msd	pearson
MAE	0.709236	0.700655	0.706341
RMSE	0.925703	0.915310	0.926235



For the Item Based Collaboration Filtering, the MSD was showing the lowest error among other similarity methods (cosine & Pearson) in both MAE and the RMSE. Pearson shows the highest RMSE and Cosine showing the highest MAE. These measures in Item Based Collaborative Filtering is consistent with the User Based Collaboration Filtering.

f. Examine how the number of neighbors impacts the performances of User based Collaborative Filtering and Item based Collaborative Filtering? Plot your results.

We ran the User based Collaborative Filtering and Item Based Collaborative Filtering on multiple K range d from 5 to 50 stepping 5 → [5, 10, 15, 20, 25, 30, 35, 40, 45, 50] and plotted the RMSE elbow chart to f ind the best value for K.



g. Identify the best number of neighbor (denoted by K) for User/Item based collaborative filtering in terms of RMSE. Is the best K of User based collaborative filtering the same with the best K of Item based collaborative filtering?

Looking at the elbow plot above we recognize the elbow at K=30 which shows relatively the best RMSE for both User Based Collaborative Filtering and Item Based Collaborative Filtering

Please find the code at the address below:

https://github.com/nimaarvin83/CAP5610/blob/main/HW5_Recom_Sys/HW5_RecSys.ipynb