Applied Spatial and Temporal Data Analysis

Vijay K. Shah

I. Preliminary Details

For this assignment, I have created three different folders, namely

1) **PythonFiles:** This folder contains all the python files.

2) **OutputFiles:** This folder contains all the output files corresponding to each python file. For example, for python file $svd.py$ (in folder *PythonFiles*, I have saved its output in a file named $svd.txt$ in this folder.

3) **images:** This folder contains all the plots that have been shown in this report.

Besides these, I have a file named *recommendation_command.txt* which basically contains all the **GNUPLOT** scripts to generate the plots. To run this file, you may simple run <**gnuplot recommendation_command.txt** >in terminal and all the plots would be generated in folder **images**.

II. Comparative Performance Analysis

Q 9. Compare the performances of User-based collaborative filtering, item-based collaborative filtering, SVD, PMF, NMF on fold-1 with respect to RMSE and MAE. Since data.split(n-folds=3) randomly split the data into 3 folds, please make sure you test the five algorithms on the same fold-1 so the results are comparable.

Q 10. Compare the performances of User-based collaborative filtering, item-based collaborative filtering, SVD, PMF, NMF on fold-2 with respect to RMSE and MAE. Please make sure you test the five algorithms on the same fold-2 so the results are comparable.

Q 11. Compare the performances of User-based collaborative filtering, item-based collaborative filtering, SVD, PMF, NMF on fold-3 with respect to RMSE and MAE. Please make sure you test the five algorithms on the same fold-3 so the results are comparable.

Q 12: Compare the average (mean) performances of User-based collaborative filtering, item-based collaborative filtering, SVD, PMF, NMF with respect to RMSE and MAE. Please make sure you test the five algorithms on the same 3-fold data split plan so the results are comparable.

In this section, we first define the two famous metrics (i) MAE and (ii) RMSE for evaluating the performances of a recommender system. Then, we discuss the comparative performance analysis of two classes of recommendation algorithms – (i) Matrix Factorization (such as SVD, PMF and NMF) and (ii) Collaborative Filtering (such as UCF and ICF).

A. Definitions

- **Mean Absolute Error (MAE)** measures how close the predictions are to the true (actual) values. MAE is calculated as follows:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|f_i - y_i| \tag{1}$$

  where n is the total number of observations, f_i is the prediction and y_i is the true value.
- **Root Mean Square Deviation (RMSD)** measures the differences between the predicted values by a model and the values actually observed. RMSD is calculated as follows:

$$RMSD = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(f_i - y_i)^2} \tag{2}$$

  where n is the total number of observations, f_i is the predicted value and y_i is the actual value.

B. Fold-wise Performance Analysis for Different Recommendation Algorithms

As shown in Fig. 1 (Fold 1), Fig. 2 (Fold 2), Fig. 3 (Fold 3) and Fig. 4 (Average), for both RMSE and MAE, matrix factorization recommendation algorithms viz. SVD, PMF and NMF perform better than collaborative filtering ones i.e. UCF and ICF. In particular, SVD yields least error whereas User (UCF) and Item-based collaborative filtering (ICF) perform the worst. This is because SVD uses the k-largest singular values of matrix A (where the rows are users, columns are items and value of each cell is the rating) to construct a matrix A_k to approximate A. Whereas, UCF simply estimates each user's preferences by referring to his/her similar user's tastes and ICF recommends items which are similar to the target user's selected items. These two collaborative filtering approaches are prone to noise and local optimal recommendations, hence yield higher error than the SVD constructs a provably-best k-rank approximation, and hence does generalized globally optimized recommendations.

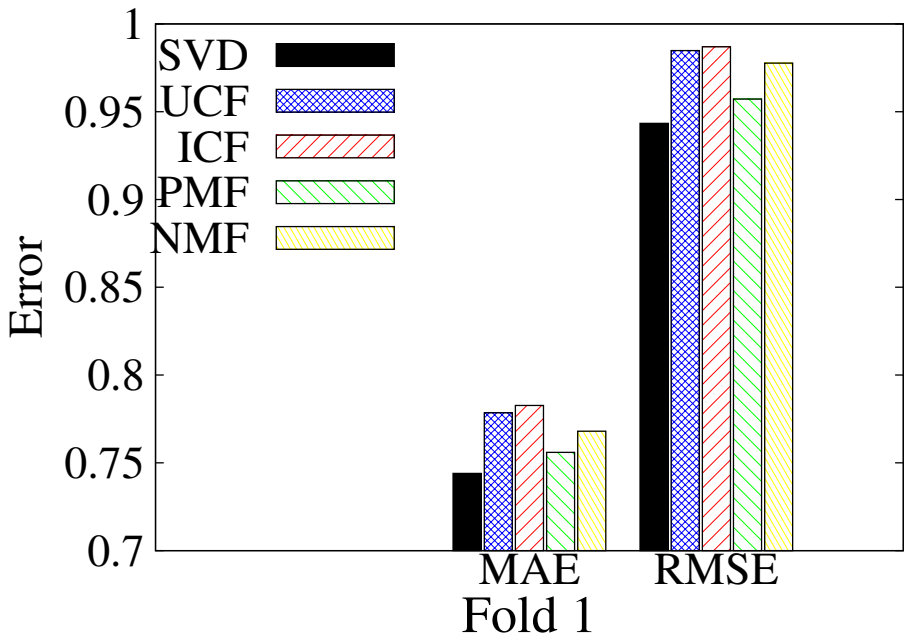


Fig. 1: Fold 1 - Error against different algorithms

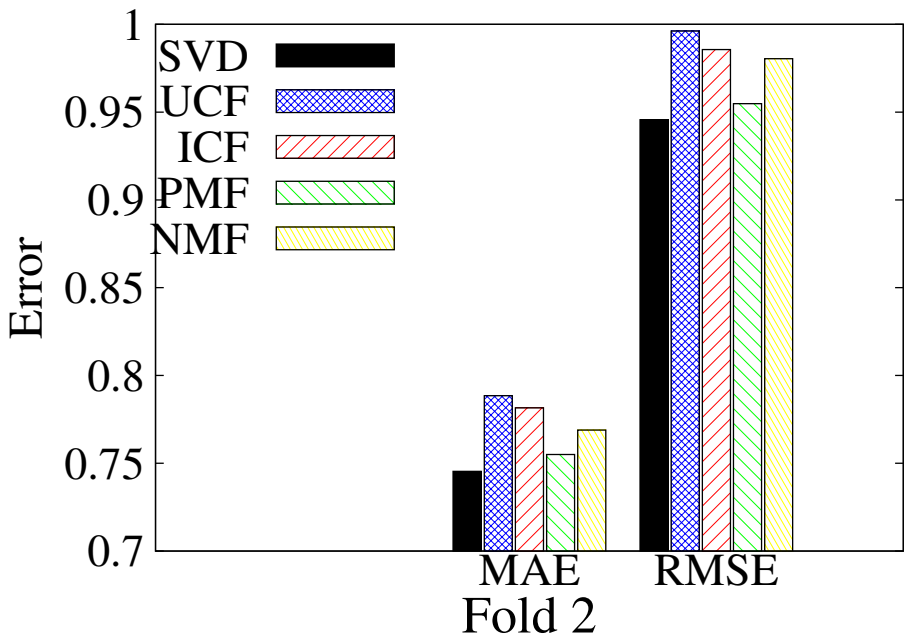


Fig. 2: Fold 2 - Error against different algorithms

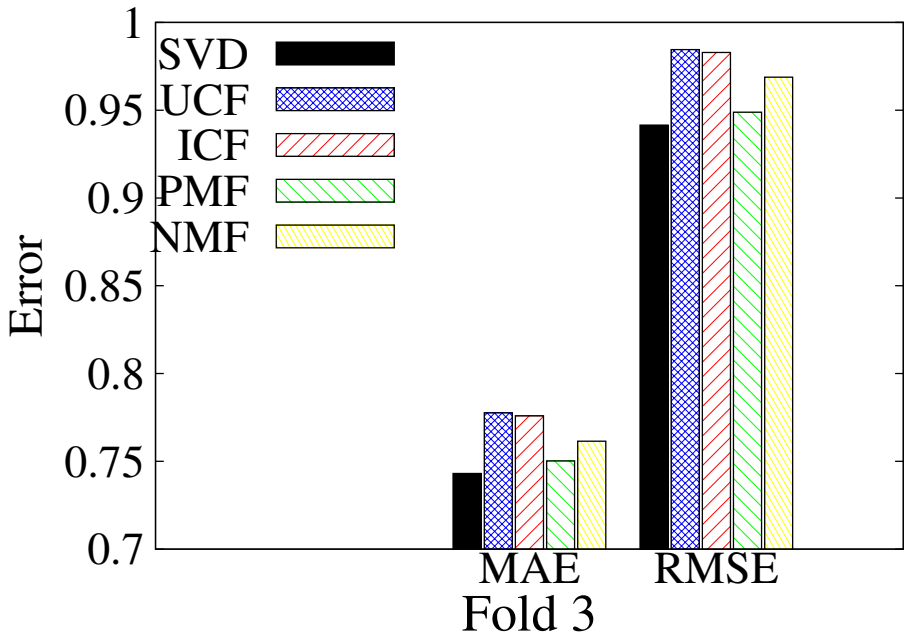


Fig. 3: Fold 3 - Error against different algorithms

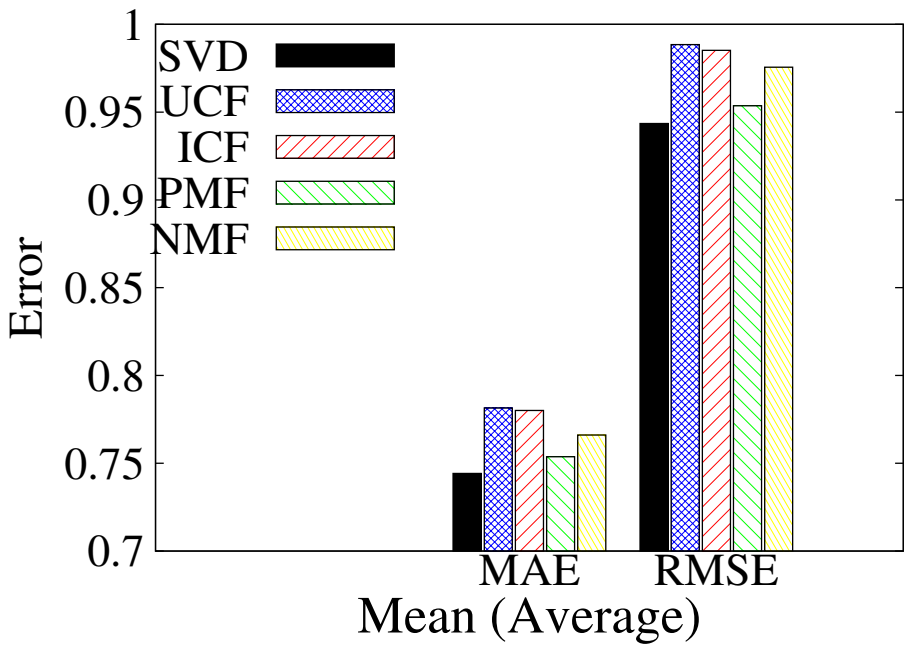


Fig. 4: Average - Error against different algorithms

Q 13: Examine how the cosine, MSD (Mean Squared Difference), and Pearson similarities impact the performances of User based Collaborative Filtering and Item based Collaborative Filtering. Finally, 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? Plot your results.

In this section, we first formally introduce cosine, MSD and Pearson similarity metrics followed by detailed analysis of these similarity metrics on the performances of UCF and ICF.

C. Definitions

- **Cosine**: The cosine similarity is defined as follows:

$$cosine - sim(u,v) = \frac{\sum_{i=1}^{m} r_{ui} r_{vi}}{\sqrt{\sum_{i=1}^{m} r_{ui}^2}\sqrt{\sum_{i=1}^{m} r_{vi}^2}} \tag{3}$$

  where u and v are two users (or items) and m is the total number of users (or items). It is important to note that only common users (or items) are taken into account.
- **Mean Squared Difference (MSD)** The Mean Squared Difference is defined as:

$$msd(u,v) = \frac{1}{m}\sum_{i=1}^{m}(r_{ui} - r_{vi})^2 \tag{4}$$

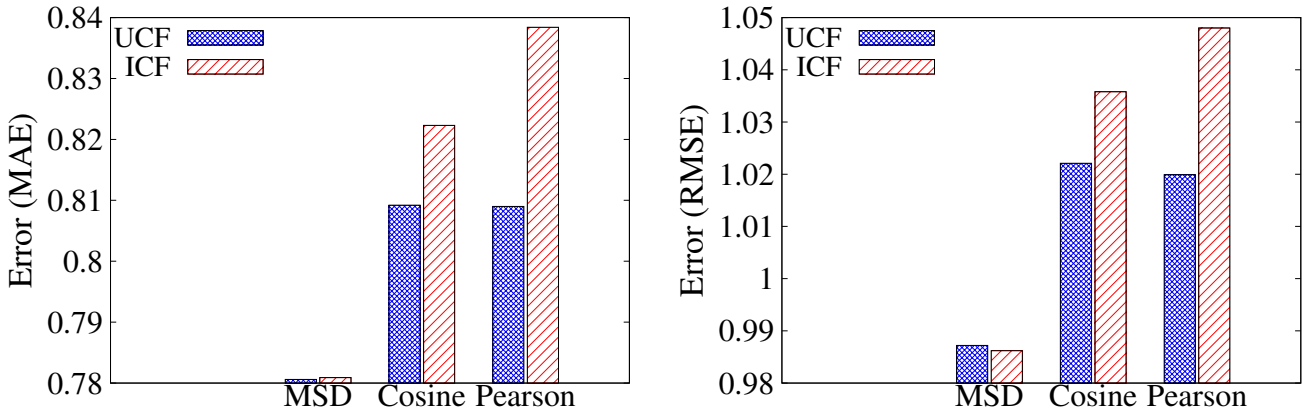
  where u and v are two users (or items) and m is the total number of users (or items). It is important to note that only common users (or items) are taken into account.
- **Pearson correlation coefficient**: The Pearson correlation coefficient is defined as:

$$pearson - sim(u,v) = \frac{\sum_{i=1}^{m}(r_{ui} - \mu_i).(r_{vi} - \mu_i)}{\sqrt{\sum_{i=1}^{m}(r_{ui} - \mu_i)^2}\sqrt{\sum_{j=1}^{m}(r_{uj} - \mu_j)^2}} \tag{5}$$

  where $\mu_i = \frac{1}{m}\sum_{i=1}^{m} r_{ui}$ is the sample mean value and analogously for μ_j.

D. Performance Similarity Metrics Analysis

As shown in Fig. 5(a) and 5(b), the UCF and ICF are consistent in terms of Mean Squared Difference (MSD). Whereas, in case of Cosine and Pearson, ICF suffers a significantly higher error value (both MAE and RMSE) compared to that of UCF.



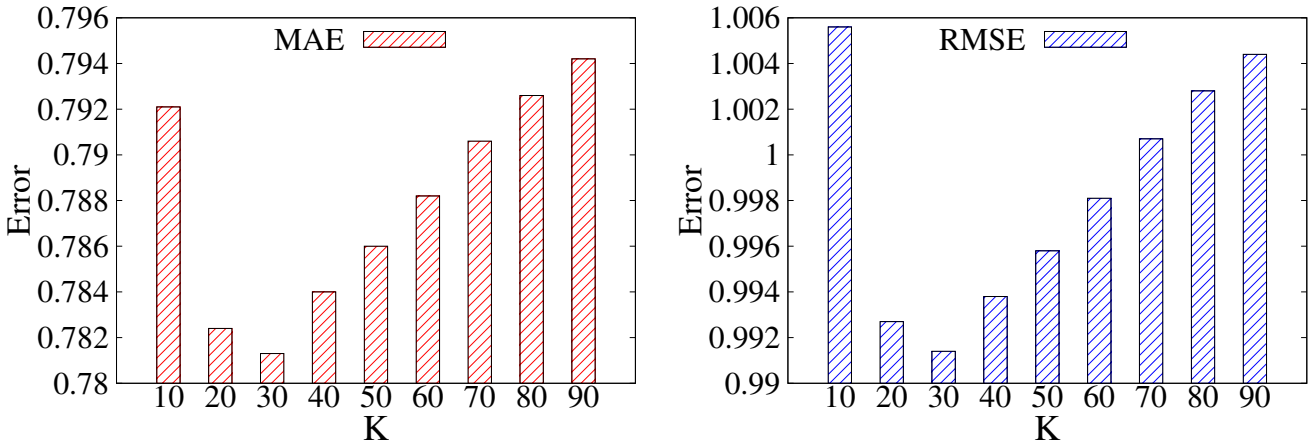
(a) (b)

Fig. 5: UCF vs ICF (a. MAE b. RMSE)

Q 14: Examine how the number of neighbors impacts the performances of User based Collaborative Filtering or Item based Collaborative Filtering? Plot your results.

As shown in Fig. 6(a) (for MAE) and 6(b) (for RMSE), with increase in K (number of neighbors) the error value for UCF recommendation algorithm first decreases until $K = 30$. Also refer Fig. 7. However, after that, the error value increases. Hence, the best K is 30 for the given data set.

In case of ICF recommendation algorithm also, the error value first decreases until $K = 30$, then, it again start peaking up as shown Fig. 8(a) (for MAE) and 8(b) (for RMSE) and 9. Here, also, the best K is 30.



(a) (b)

Fig. 6: UCF: a. Error (MAE) vs K b. Error (RMSE) vs K

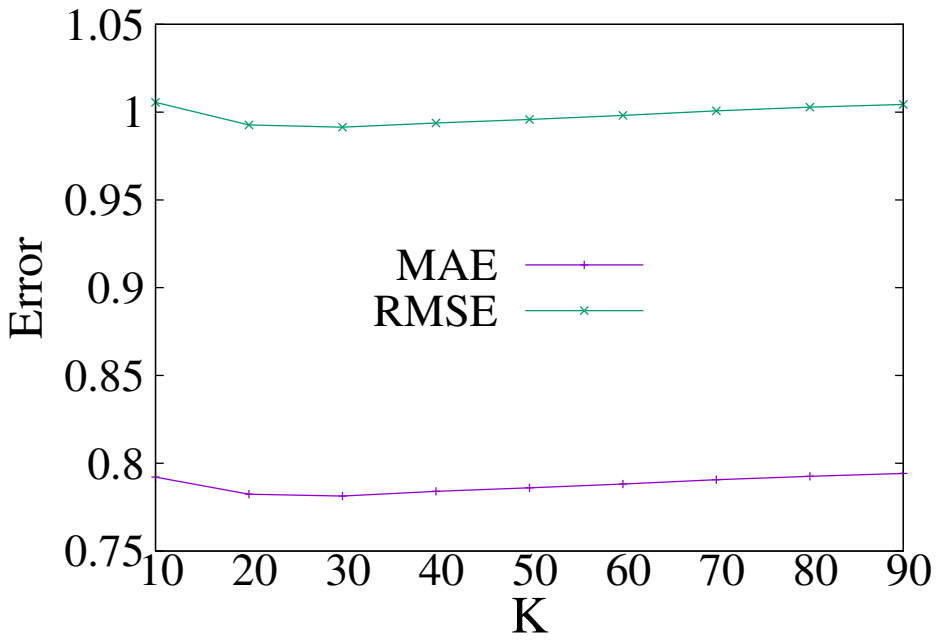
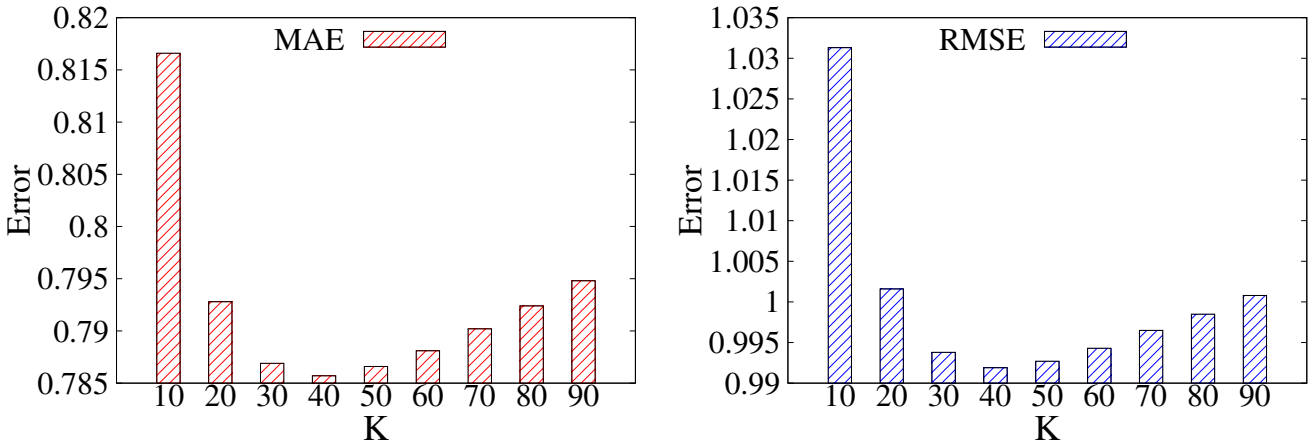


Fig. 7: UCF: Error vs K



(a) (b)

Fig. 8: ICF: a. Error (MAE) vs K b. Error (RMSE) vs K

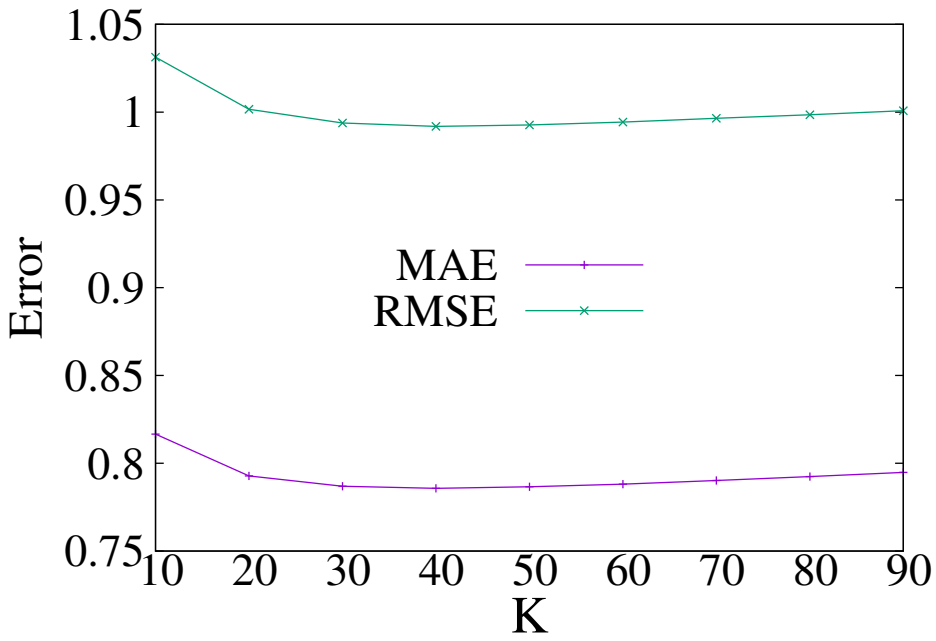


Fig. 9: ICF: Error vs K

III. Conclusion

From the aforementioned analysis, we can conclude that the matrix factorization recommendation algorithms such as SVD, PMF and NMF predict better than the counterpart Collaborative filtering (CF) algorithms viz. UCF and ICF. Also, among CF recommendation algorithms, the MSD exhibits lower error value (for both MAE and RMSE) compared to that of Cosine and Pearson similarity metrics. Finally, it is also evident that MAE consistently exhibits low error value compared to that of RMSE irrespective of recommendation algorithms and similarity metrics employed.