James Simmons
Machine Learning
Fall 2021
Homework 5: Recommender Systems

Note about data configuration

For this assignment, I used the "ratings_small.csv" file from the movies dataset, rather than the full dataset. This smaller dataset contain ~100,000 ratings, and is a subset of the full ~26,000,000 ratings dataset. The reasoning for this is that the full dataset was so large, that the user-item matrix would have required over 100 gigabytes to store, and as such was computationally infeasible given the resources I had available. The professor indicated in class on November 8th that using the smaller dataset was acceptable.

Note about implementation

To solve the problems of this homework, I used the Surprise library. For PMF, I used the SVD with biases=False, because the official documentation explicitly stated that this configuration was identical to the PMF algorithm. For CF, I used KNNBasic. Surprise supports multiple algorithms listed as KNN-based, and all of them also mention that this is the way the library performs CF. I opted for KNNBasic because it is the simplest and I had no prior knowledge with which to decide to use a more complex KNN instead. In either case (PMF and CF), I used the builtin cross_validate function, which automatically produces all the necessary training and testing computations, and outputs the RMSE and MAE.

For this assignment, I performed CF on all 900 combinations of {user-based, item-based} X {cosine, msd, pearson} X {k | 1 <= k <= 150}.

Overall Cross-Validation Performance Error (3c, 3d)

Below is a table featuring the mean validation performances of PMF, user-based CF, and item-based CF with respect to RMSE and MAE. For CF, the data listed below is selected from the BEST combination of similarity metric and number of neighbors. As RMSE and MAE are error measures, lower is better.

	PMF	**User-Based CF (MSD)**	**Item-Based CF (MSD)**
RMSE	1.00745	0.96132 (14 neighbors)	0.93091 (68 neighbors)
MAE	0.77881	0.73560 (12 neighbors)	0.71742 (69 neighbors)

As you can see, by both RMSE and MAE **Item-Based CF is the best,** user-based is second-best, and PMF is worst.

Effect of Similarity Metric and Number of Neighbors (3e, 3f, 3g)

Similarity Metrics (3e)

User-based: MSD is best, then cosine, then pearson is worst
Item-based: MSD is best, then pearson, then cosine is worst
The above orderings are true for both RMSE and MAE
For both user-based and item-based CF, cosine and pearson were very close in accuracy, especially as the number of neighbors increased.

The impact of similarity metric for user-based CF and item-based CF are NOT THE SAME

Number of Neighbors (3g)

User-based: MSD RMSE performed best with 14 neighbors. MSD MAE performed best with 12 neighbors.
Item-based: MSD RMSE performed best with 68 neighbors. MSD MAE performed best with 69 neighbors.

The best number of neighbors for user-based CF and item-based CF are NOT THE SAME

The table below gives the best number of neighbors for each combination of {user, item} X {cosine, msd, pearson}. The values that resulted in the best overall performance are highlighted.

	User-Based			**Item-Based**		
	Cosine	MSD	Pearson	Cosine	MSD	Pearson
RMSE	61	14	95	144*	68	150*
MAE	32	12	95	144*	69	148*

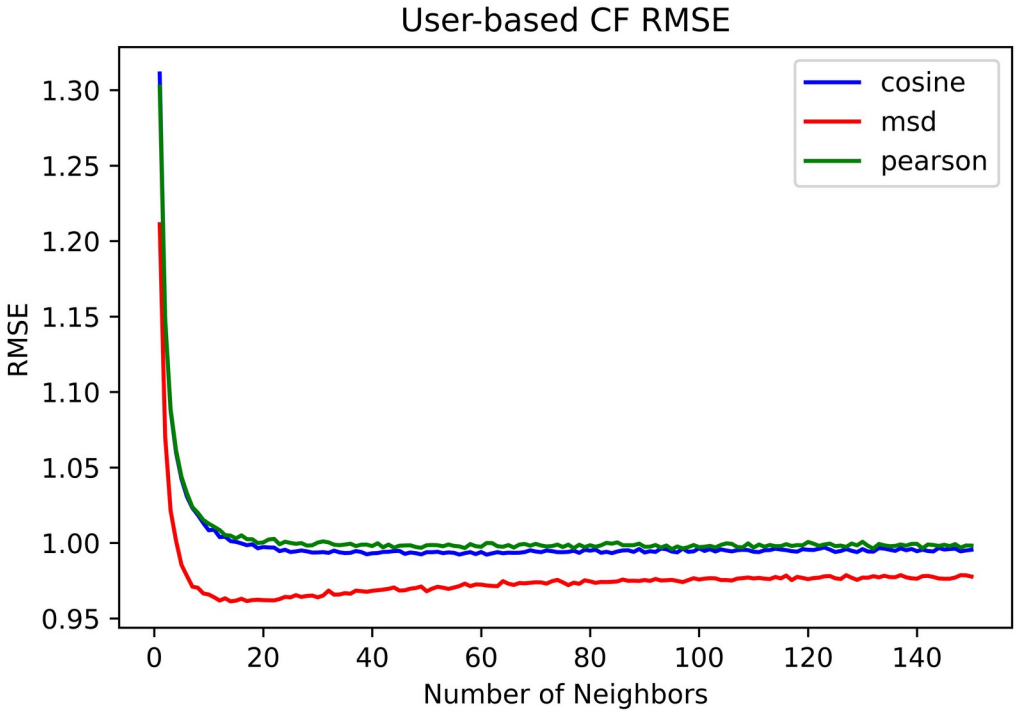
* For item-based CF, cosine and pearson appear to still be decreasing even with 150 neighbors. Their true ideal number of neighbors may be greater than 150.

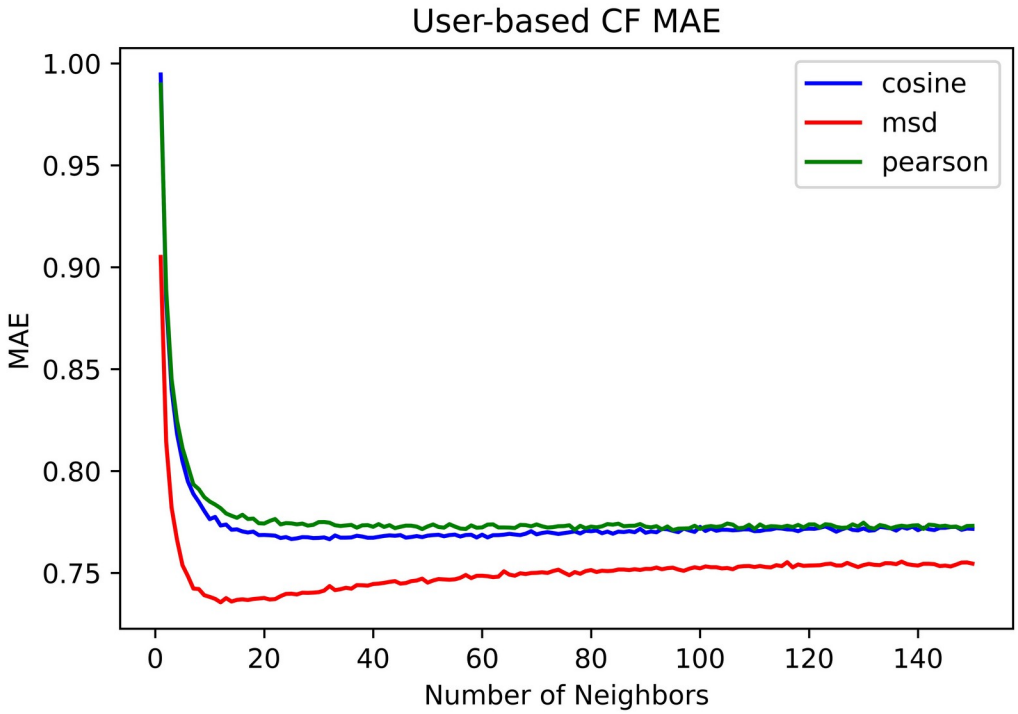
Plots

See following pages for plots.

User-Based CF (3e, 3f)

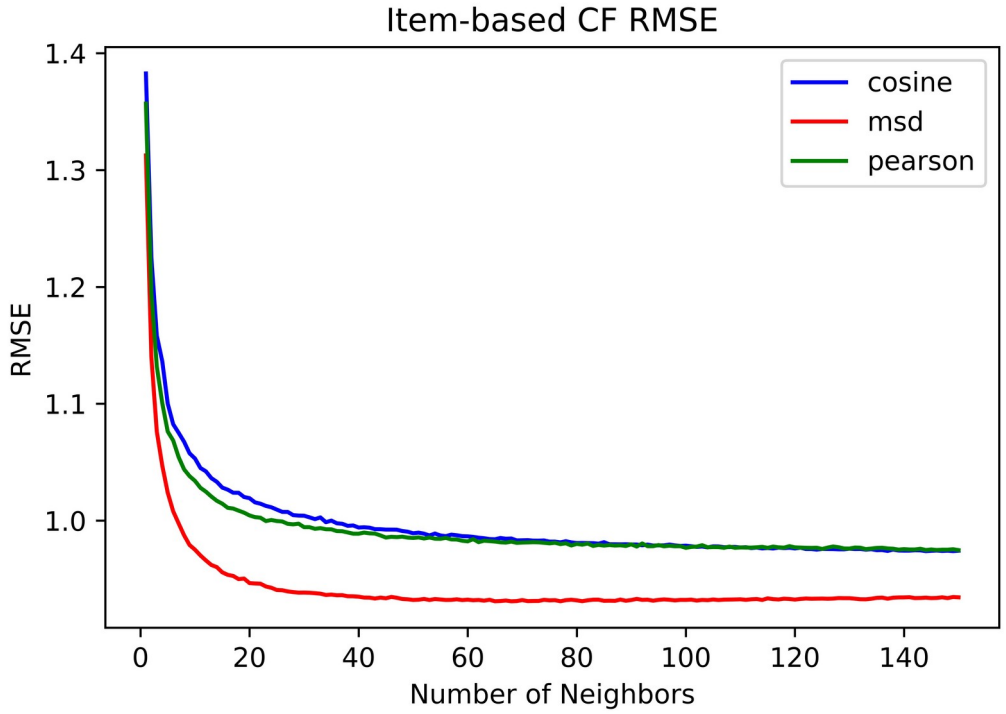
MSD performed the best, with cosine and pearson performing almost identically (cosine slightly better). Cosine and pearson mostly just leveled off once they reached their effective optimum, but MSD got noticably worse as the number of neighbors increased beyond the ideal of 12-14.

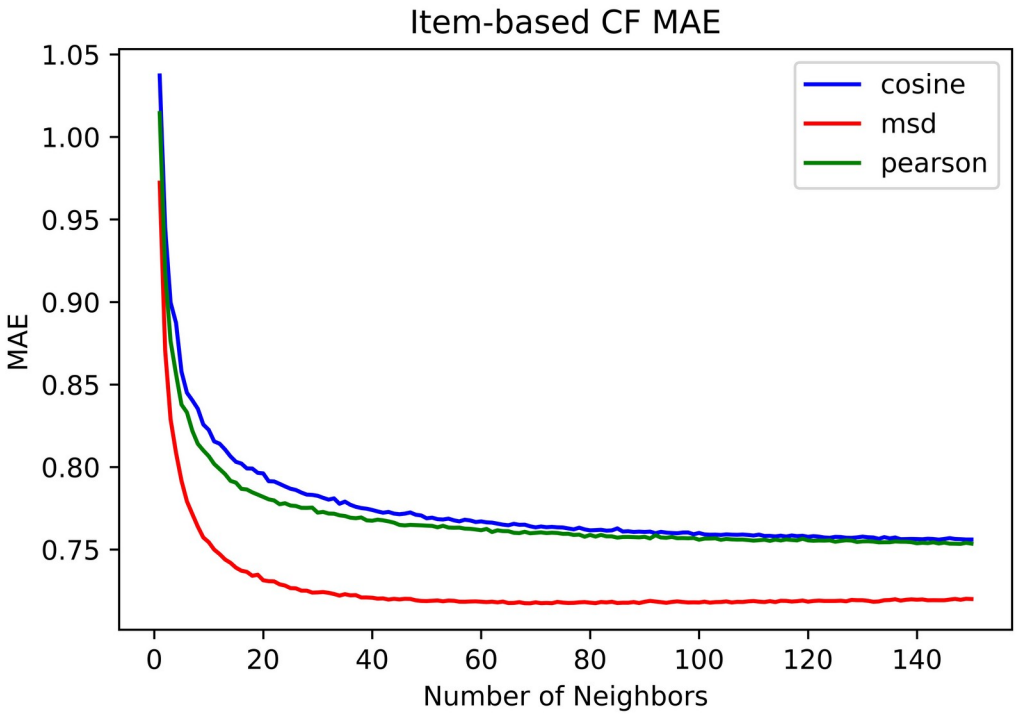




Item-Based CF (3e, 3f)

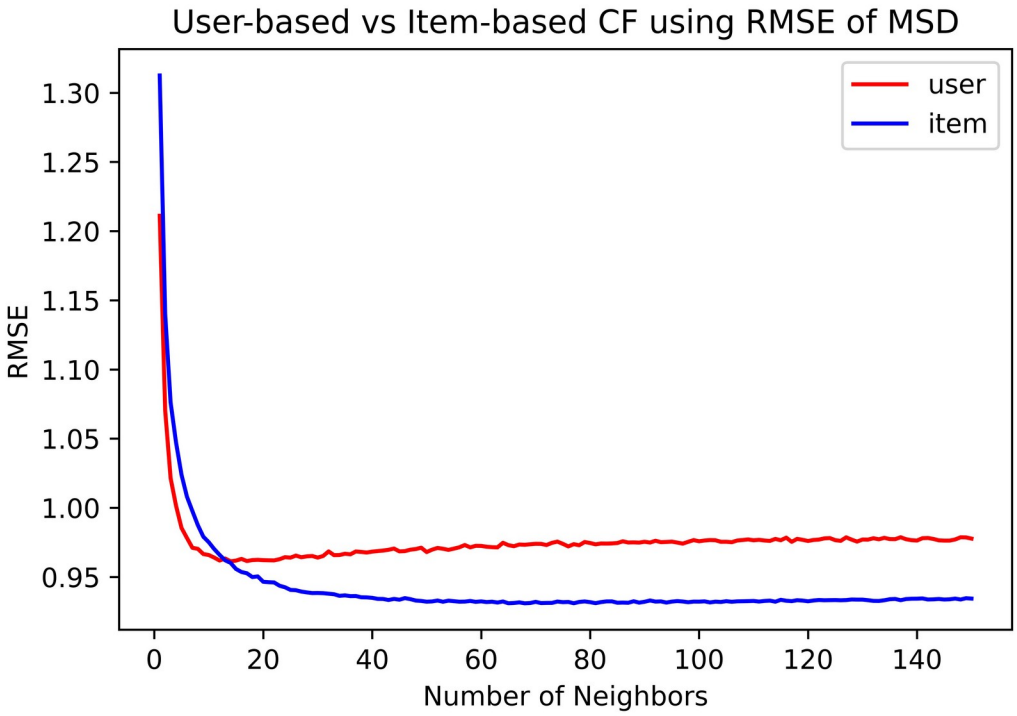
MSD performed the best, with cosine and pearson performing almost identically (pearson slightly better). MSD reached an ideal number of nieghbors at 68-69, but doesn't get substantially worse thereafter. Cosine and pearson appear to still be improving even at 150 neighbors.

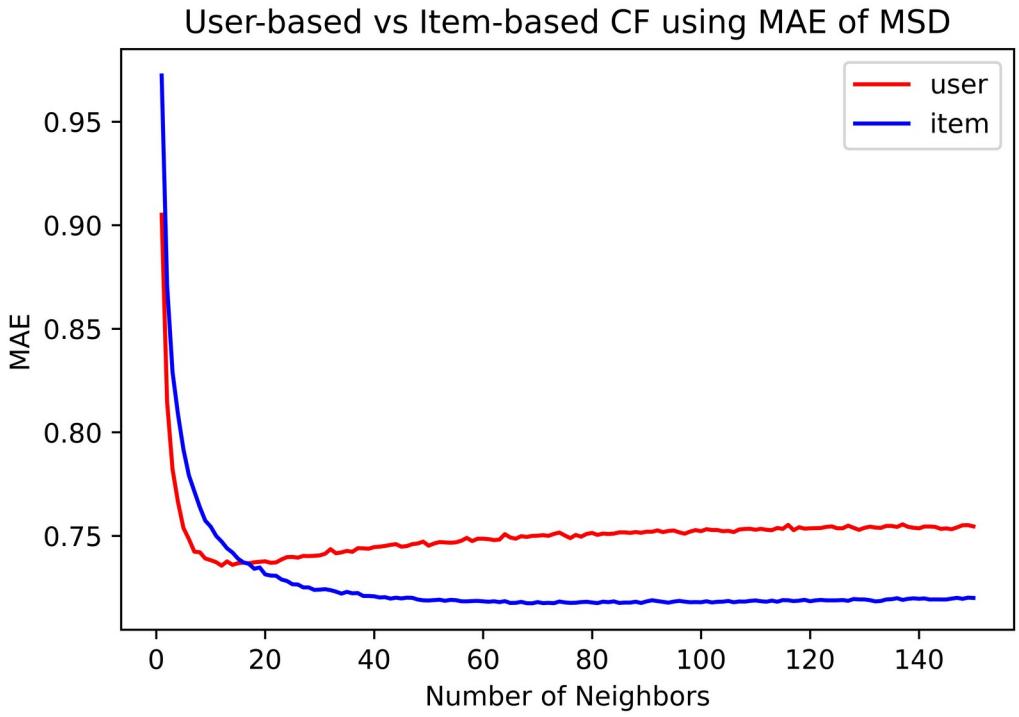




User-Based vs Item-Based MSD (3d, 3f)

MSD was used to compare user-based and item-based CF directly because both methods performed better with it than with any other similarity metric. User-based takes an early lead as it approaches its optimal number of neighbors (12-14) while item-based is a bit slower. Overall, however, item-based is clearly superior for any number of neighbors greater than 20.





Code:
All code and output files (CSV and pyplot JPG) are available at:
https://github.com/Mesaj2000/Machine_Learning_5610/tree/master/hw5