Collaborative Filtering
Assignment 1

Submitted by : Himanshu Aggarwal
Roll No. : MT17015
Branch: CSE

Approach

Dataset
- A simple and straight forward approach has been followed to do the given task.
- Given dataset was read from CSV file and converted to array of data.
- Later the data has been used to construct the User-Item rating matrix.
- The dataset taken is divided into training and testing datasets. Five such sets have been used.

Similarity
- Cosine similarity between users , as well items, has been used to act as weight in the weighted mean prediction formula for predicting user rating for a particular item.

Prediction
- The ratings, similarities and the number of nearest neighbours(optional), are used to calculate the predictions, that is, predicted rating for an item for a user.
- Formula used for prediction is weighting mean.

Other Tasks
- Variance Weighing and Significant weighing is applied with cosine similarity.

Results

100K Movielense Dataset

- User-User Based

User-User Similarity based / Neighbourhood Selection (lambda = 0.6, threshold = 0.6)

	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5
21	0.5175125	0.5333375	0.5307875	0.5232875	0.5236625
41	0.5351125	0.553775	0.544275	0.5411125	0.539725
61	0.5513875	0.5666625	0.55635	0.5526	0.5521625
81	0.5644125	0.578575	0.5673625	0.563925	0.5629375

- Item-Item based

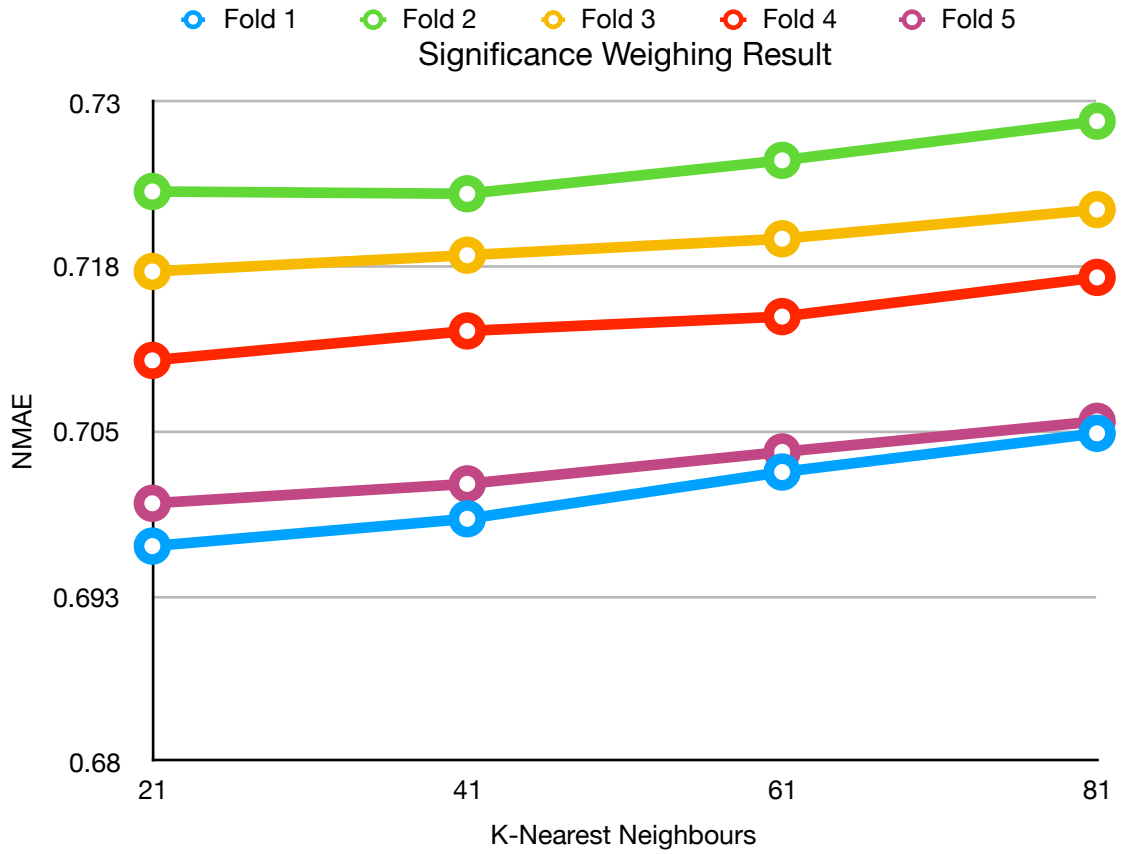
Item-Item based

	Fold1	Fold2	Fold3	Fold4	Fold5
Item-Item NMAE	0.8689625	0.8585125	0.8480125	0.85695	0.85845

- Significance Weighing with varying neighbourhood

Significance Weighing Result

	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5
21	0.6963125	0.7232375	0.7171625	0.7104125	0.6995625
41	0.6983875	0.7230625	0.7184	0.71265	0.7010375
61	0.7019375	0.7256	0.71965	0.7137375	0.7034625
81	0.704875	0.728575	0.7218625	0.7167125	0.7057875



- Variance Weighing

Variance Weighing

	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5
21	0.8998875	0.902125	0.8964375	0.8952125	0.8968375
41	0.9014875	0.9022	0.897	0.896175	0.898025
61	0.900825	0.900825	0.8957875	0.895575	0.897325
81	0.8993625	0.899125	0.8943625	0.8938875	0.896

- For User-Item combined approach, the result of the five folds are:

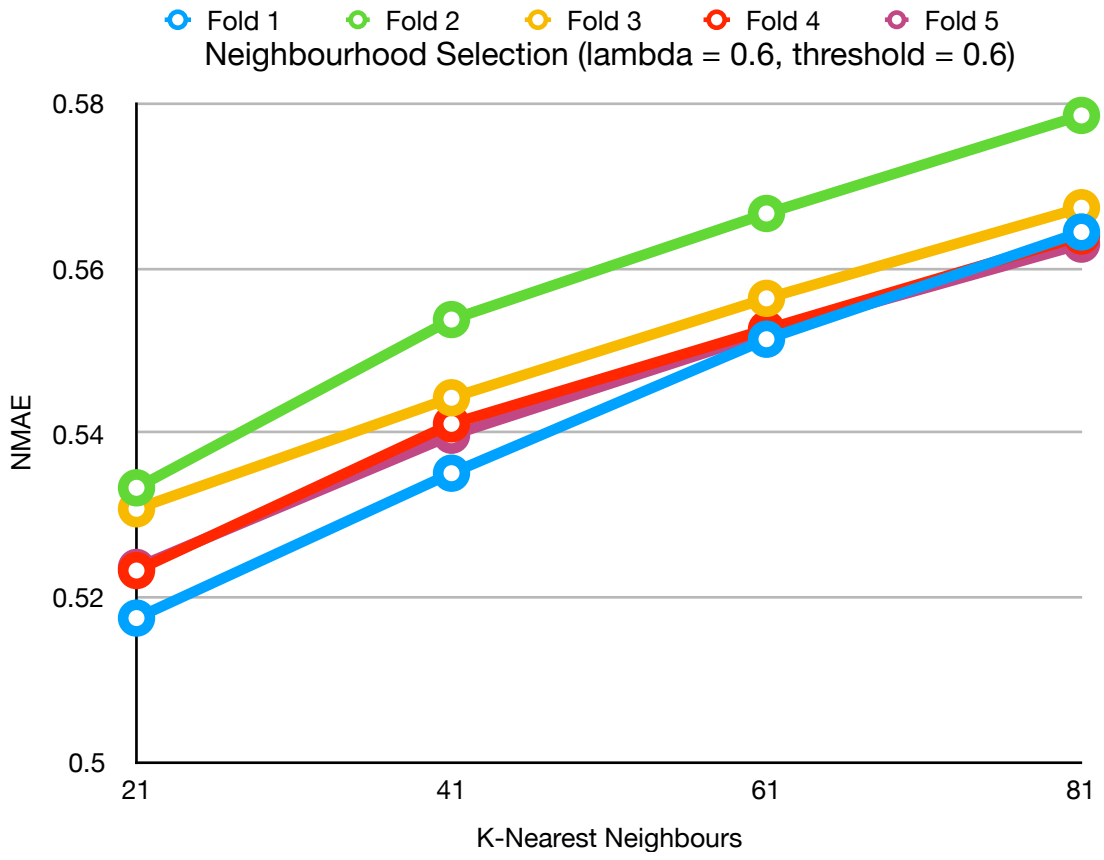
Results for Fold 1 :
User-Item based CF NMAE: 0.6508025

Results for Fold 2 :
User-Item based CF NMAE: 0.6574975

Results for Fold 3 :
User-Item based CF NMAE: 0.6516575

Results for Fold 4 :
User-Item based CF NMAE: 0.6507125

Results for Fold 5 :
User-Item based CF NMAE: 0.6507425



Misc:

- 1M movilense dataset has also been used, but is not providing correct results at the moment.
- Also, it is taking more than desired time to get split.
- **Significance Weighing** has been applied and is producing results as expected, but is taking a very long time for computation.
- **K-nearest neighbours** have been used in the prediction function.