IT-562 Recommendation Systems and Engines

<u>Assignment-5</u>

Collaborative Filtering: Estimating SVD through Stochastic Gradient Descent

Name: Rishab Arora
ID: 201501070
Group Name: 404

Dataset:

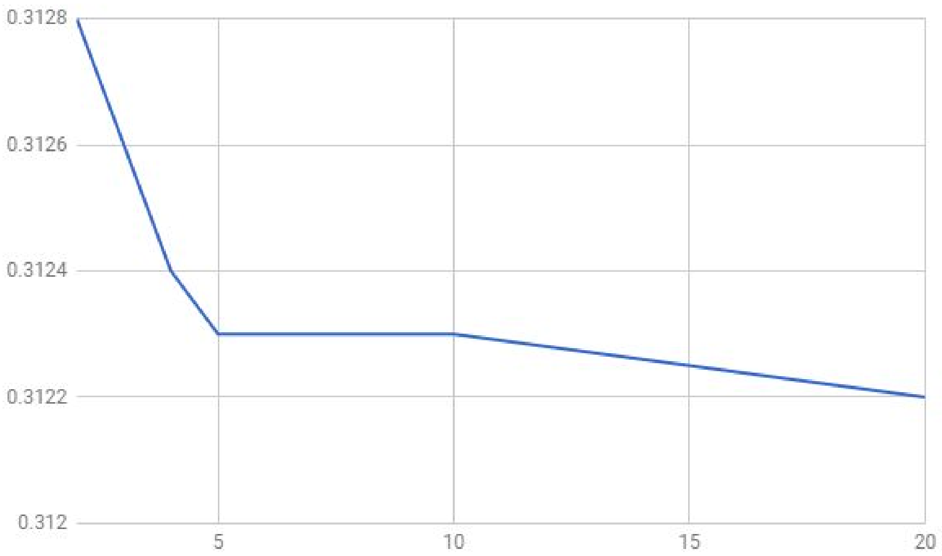
The dataset that we've used is the Goodbooks-10K dataset with ratings of around 10,000 books by different users. The ratings have been normalized between 0-1 before putting the dataset to use for different algorithms below.

1. **<u>Finding the Optimum Number of Folds</u>:**
   (For number of epochs 10, Learning Rate: 0.03)

Table:

No of Folds	**Mean RMSE**
2	0.3124
4	0.3123
5	0.3123
10	0.3123
20	0.3122

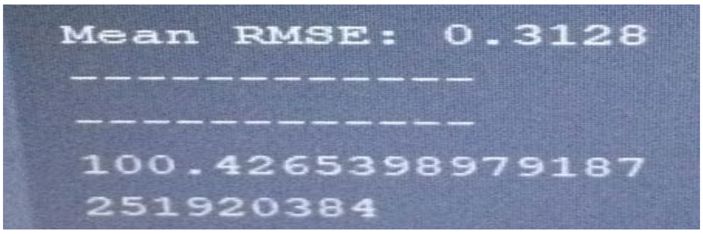
Graph: (Folds vs Mean RMSE)



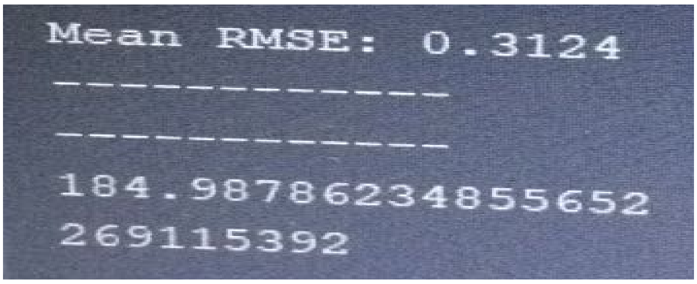
2. Finding the Optimum Number of Epochs:

There was no such significant difference in the RMSE on changing the number of epochs from 10 to 20 and then to 30. Also increasing the number of epochs, increased the time significantly. Thus, we have used the Number of Epochs as **10** everywhere as a trade-off between time and minute accuracy.

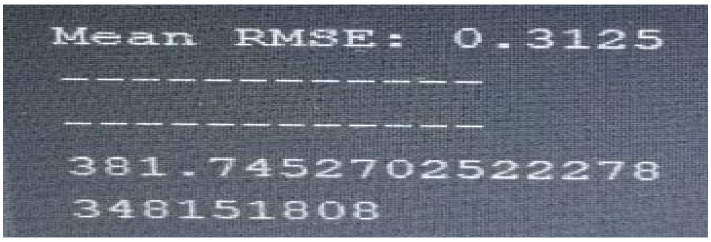
a. **No_of_Folds =2, Epochs: 10**



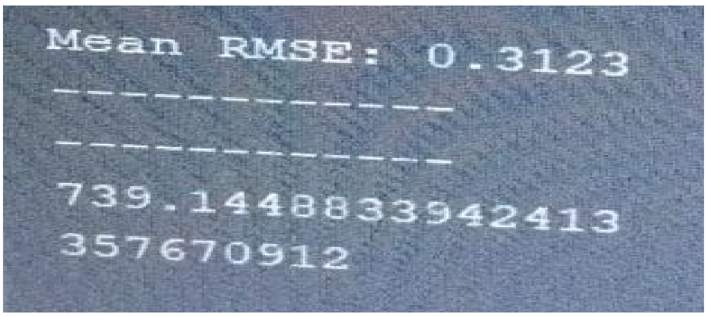
b. **No_of_Folds= 2, Epochs: 20**



c. **No_of_Folds=5, Epochs: 10**



d. **No_of_Folds=5, Epochs=20**



3. <u>Finding optimal learning rate</u>:

(Epoch=10, Splits=2)

Learning Rate	**Mean RMSE**	**Time**	**Memory**
0.005	0.3132	102.53956	254988288
0.01	0.3126	101.0542	251649841
0.02	0.3125	102.22452	253849600
0.03	**0.3124**	**101.06726**	**252669952**
0.04	0.3124	101.6551	253456384
0.05	0.3125	101.7446	254861312
0.08	0.3133	106.3357	254758912

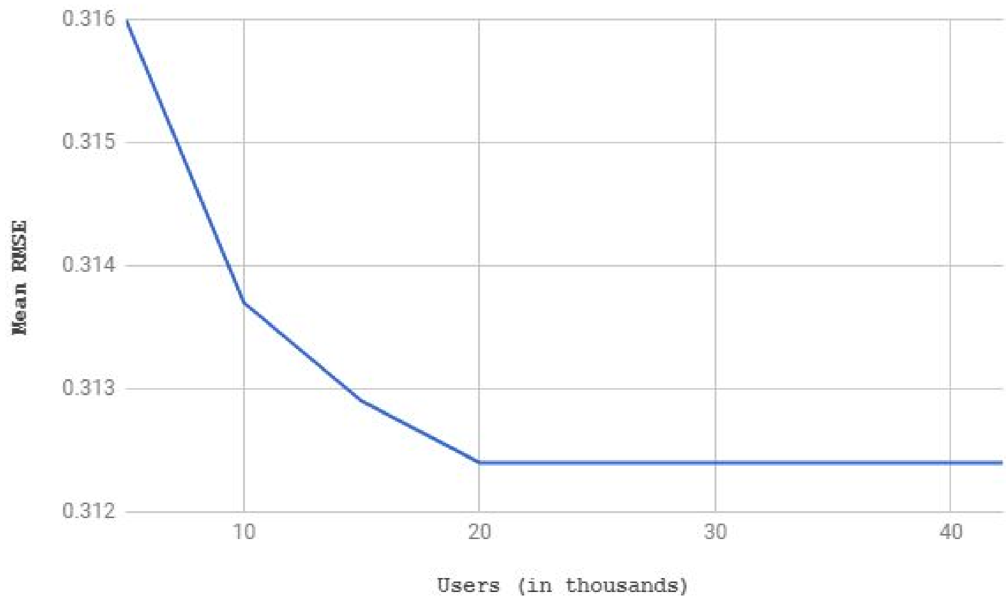
Thus, Optimum Learning Rate is 0.03.

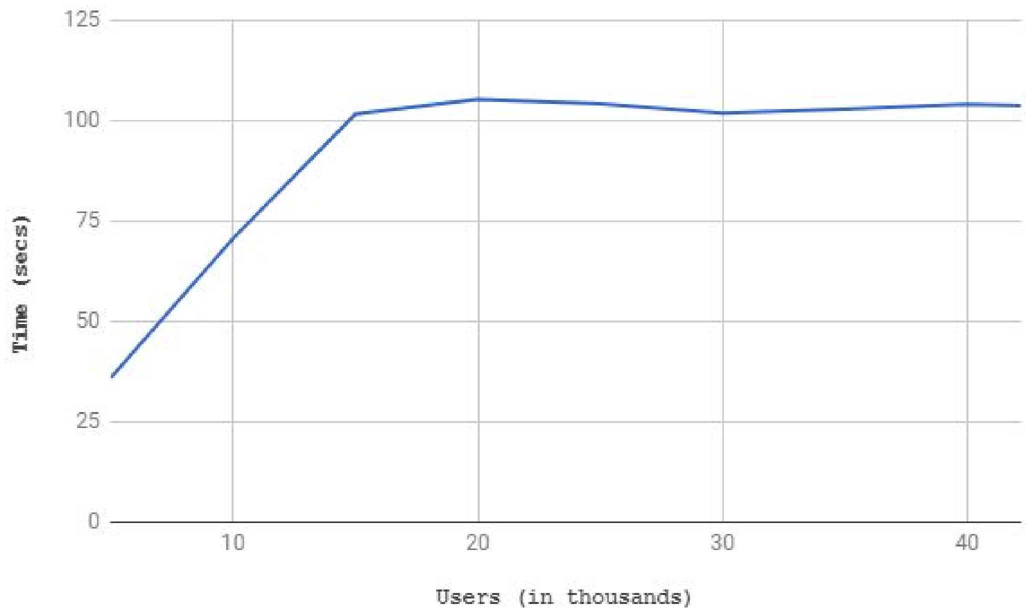
4. Increasing no. of users

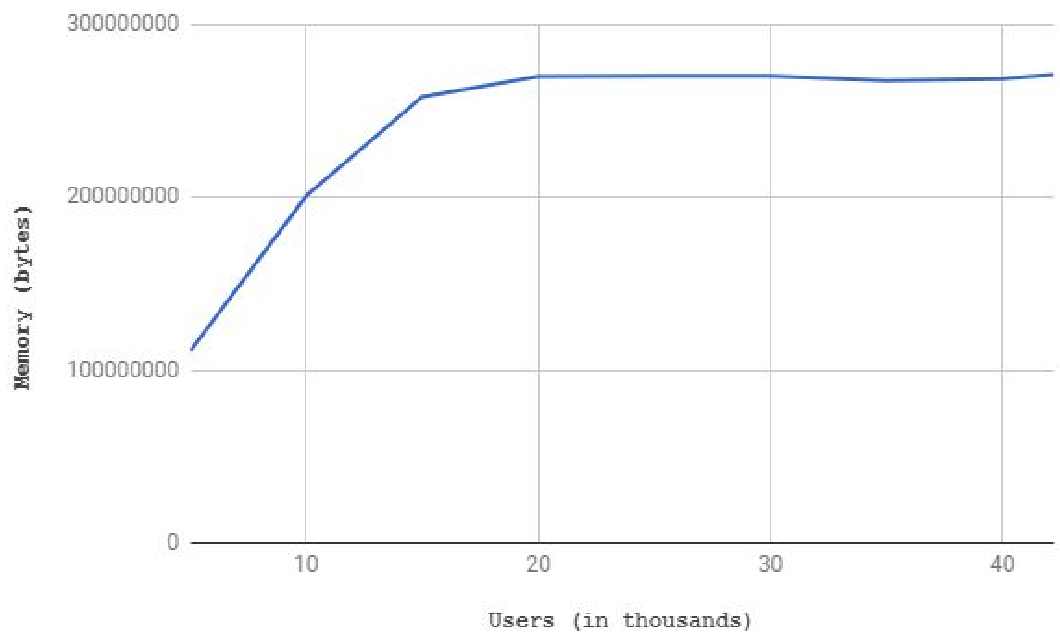
Total items = 10K

No. of Users	Mean RMSE	Time (secs)	Memory
5K	0.3160	35.945	111009792
10K	0.3137	70.716	200847360
15K	0.3129	101.718	258334720
20K	0.3124	105.333	269963264
25K	0.3124	104.266	270221312
30K	0.3124	101.932	270340096
35K	0.3124	102.866	267522048
40K	0.3124	104.163	268574720
All (42208)	0.3124	103.794	271097856

Note: Even though the total number of users is 42.2K, the ratings data available for users having id > 16K is very less. On analysing the dataset, it turns out that the effective number of users for which we have substantial number of ratings data available is around 13K. This characteristic of the dataset is the reason we get almost constant mean RMSE after the first 20K user ids.







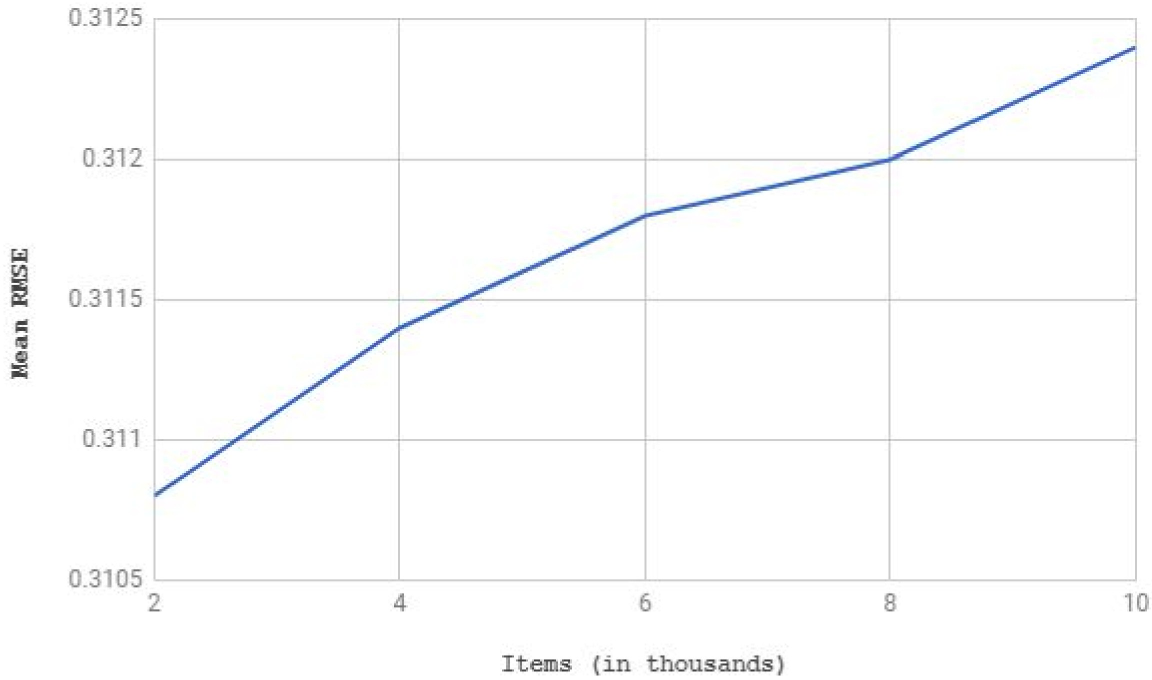
Observation:

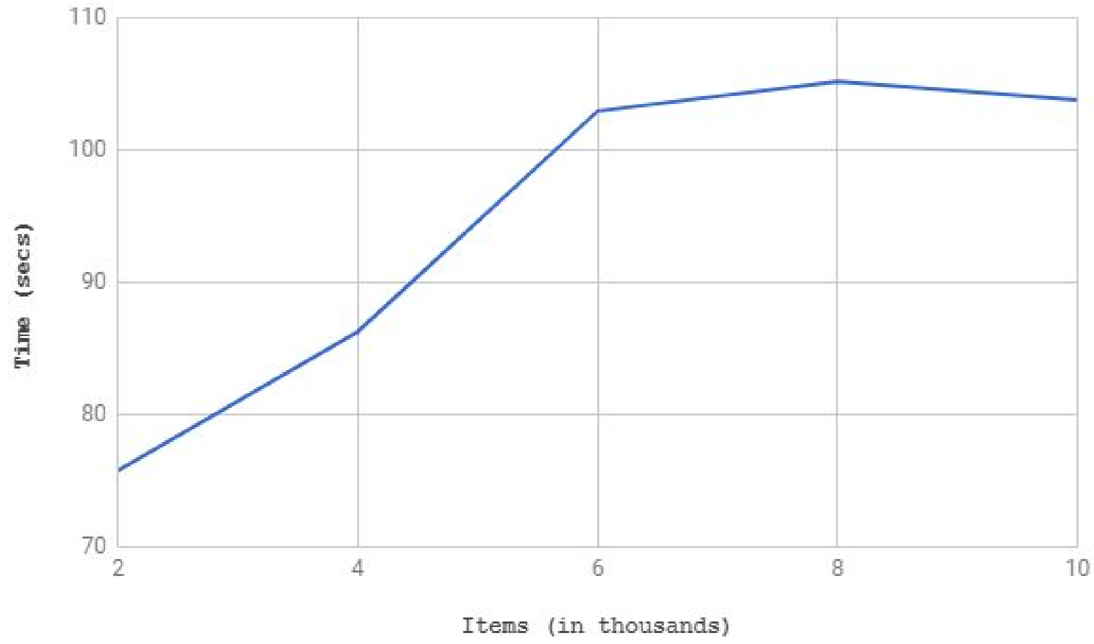
- With the increase in number of users, the system gets more data to work with, and hence the results we get are more accurate (as seen by the decrease in mean RMSE).
- However, there is a cost that the system bears: that of time and memory (as seen in the graphs above).
- It is interesting to note that the rate of change of mean RMSE decreases with increase in the number of users, and becomes almost stagnant after a point, hence giving us hints to the optimum number of users that we might want to consider while training the model.

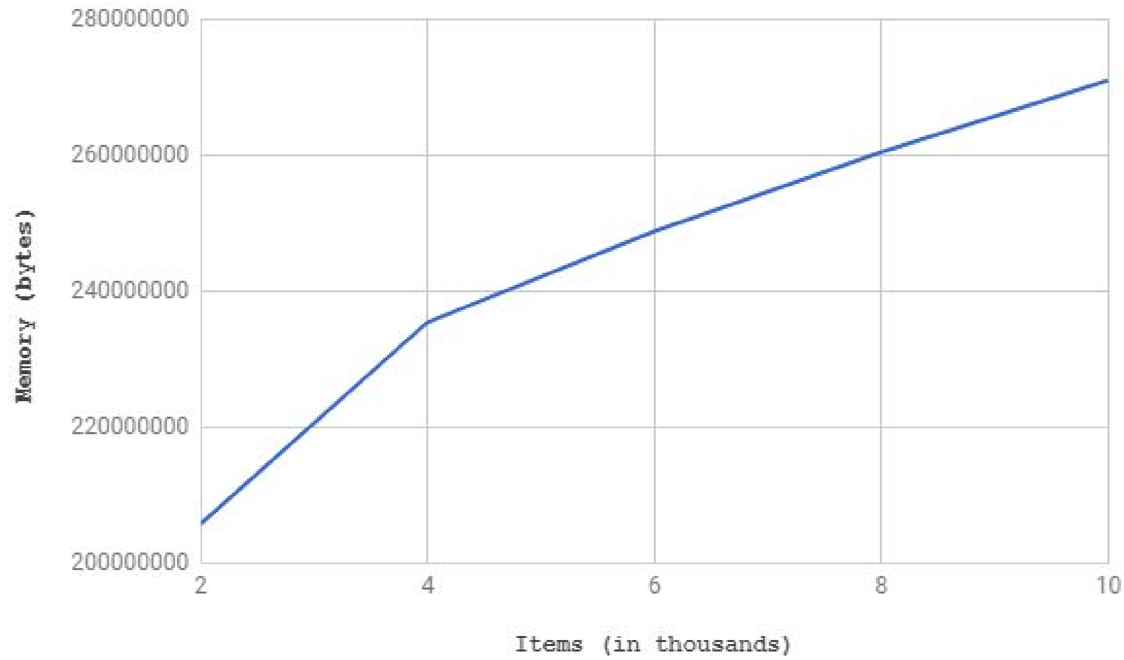
5. Increasing no. of items

Total users = 42208

No. of items	Mean RMSE	Time	Memory
2K	0.3108	75.747	205795328
4K	0.3114	86.273	235483136
6K	0.3118	102.955	248844288
8K	0.3120	105.180	260509696
All (10K)	0.3124	103.794	271097856







Observations:

- With the increase in the number of items, the trend in the mean RMSE shows that it increases and thus, accuracy decreases. One plausible reason for this trend could be that as the number of items increases, the no. of question marks that the algorithm needs to predict increases and with limited data available, the overall mean error in the prediction may increase.
- As expected, the execution time and memory consumption increases as we increase the items in the dataset provided to the algorithm.

6. Comparison between different Methods:

(Using all data, No of Folds: 2, epoch 10)

Method	Fold-1 RMSE	Fold-2 RMSE	Mean RMSE
1. SVD	0.3120	0.3126	0.3123
2. SVD++	0.3119	0.3124	0.3122
3. NMF	0.3118	0.3124	0.3121