Project 3 : Collaborative Filtering

Aadithya Venkatanarayanan 404946465
Narendran Raghavan 404945767
Mina Shahi 604717571
Srividhya Balasubramanian 205023825

21st February 2018

1 Question 1

Sparsity refers to the behavior of occurring or growing or settling in widely placed intervals; it is not thick or dense but thinly dispersed or scattered. Similarly, the data set that we consider is sparse because the users providing the ratings will not essentially rate all the available movies but only a fraction of it and as a result most of the ratings are not specified. We calculate sparsity by the following formula :
Sparsity = Total number of available ratings / Total number of possible ratings
Output: Sparsity of the dataset is 0.016332850172508883

2 Question 2

We are plotting a histogram showing the frequency of the rating values. We have binned the rating values with width=0.5 in horizontal axis. According to the ratings matrix,we understand that out of 6122825 movies, only 10004 have got user ratings. **Output**: The plot is uneven and non-uniform. This is because the data set is sparse and hence frequency of rating values is wavering for specific ratings.

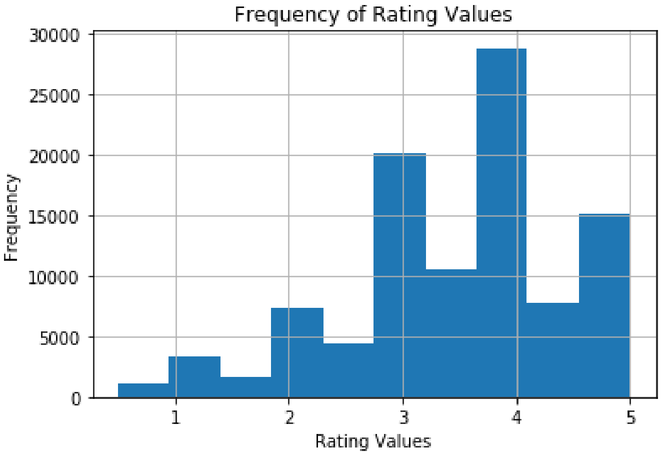


Figure 1: Frequency of rating values

3 Question 3

Here, we plot how many ratings a movie has received.We take movie ID ordered by decreasing frequency in x axis and number of ratings for each movie in y axis. **Output**:

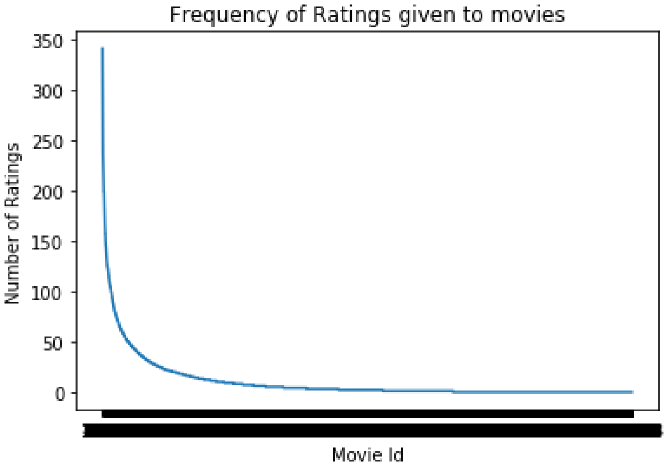


Figure 2: Frequency of ratings given to various movies

4 Question 4

Here, we plot how many movies the user has rated.We take user index ordered by decreasing frequency in x axis and number of movies the user has rated in y axis. **Output**:

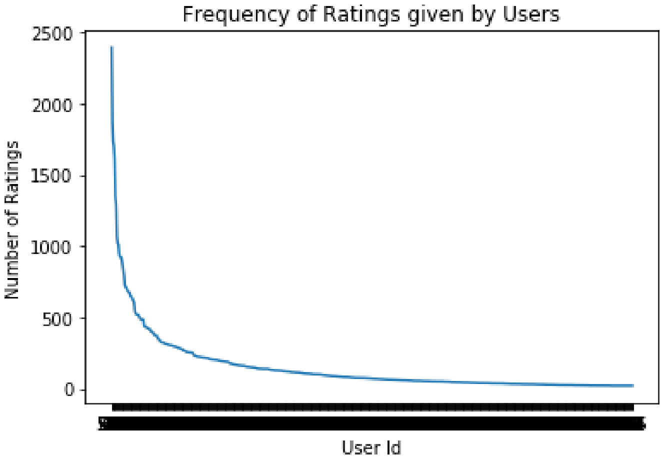


Figure 3: Frequency of ratings given by various users

5 Question 5

We analyze the histogram plotted for question 3. The x axis labels are not displayed because the number of movies are in the range of 10,000. We can see that the graph is a decreasing function curve as the x axis is aligned in the order of decreasing frequency. By seeing the graph, one can highly recommend the first set of movies in the x axis, by identifying them according to their respective movie index. This is because they have higher rating than the movies that occur further in the x axis line. This kind of plot gives the user an easy, graphical way to quickly find out about a movie according to the ratings.

6 Question 6

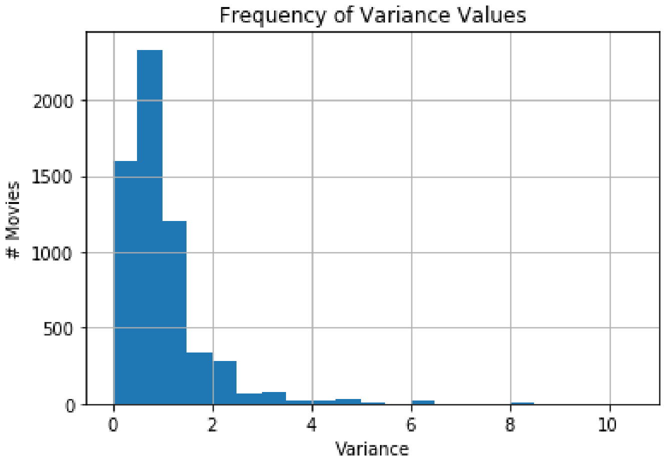


Figure 4: Frequency of Variance Values of movie ratings

The shape of the histogram is exponentially decaying.

7 Question 7

μ_u: Mean rating for user u computed using her specified ratings
r_{uk} : Rating of user u for item k
μ_u = (r_{uk} x n.k) / n where n is the number of items for which user u specifies rating.

8 Question 8

$I_u \cap I_v$ refers to set of item indices that are common in the list specified by both users u and v (I_u refers to Set of item indices for which ratings have been specified by user u, I_v : Set of item indices for which ratings have been specified by user v).
Yes, $I_u \cap I_v$ can be zero or a null set.This is because the two users u and v can have unique and distinct item index with a specific rating individually. They are called mutually exclusive if their intersection results in a null set. Since the ratings matrix is sparse, there will be cases where the above stated condition can occur-each user having dissimilar elements.

9 Question 9

In some cases, certain users will rate all the set of movies high while there will also be other set of users who might rate all set of movies poorly. This is an extreme case on both ends scenario. In this situation, the prediction function will be affected. The centering of raw ratings aid us to avoid negative impact on prediction function. The centering makes all samples equally similar to the data centroid. This is done by shifting the origin to the data centroid. This process is expected to reduce hubs. When the data set is presented in a high dimensional feature space, hubness phenomenon occur. They tend to affect the k-nearest neighbor(kNN).A sample which is similar to the data centroid tends to become a hub. Centering similarity measures help to eliminate this problem.

10 Question 10

k Nearest Neighbor Algorithm

In this algorithm, a distance measure is needed to determine the "closeness" of instances. We classify an instance by finding its nearest neighbors and picking the most popular class among the neighbors. Advantages:

- Simple to implement and use
- Easy to explain prediction
- Robust to noisy data by averaging k-nearest neighbors

Root-mean-square error (RMSE) is a frequently used measure of the differences between values (sample and population values) predicted by a model and the values actually observed. The Mean Absolute Error(MAE) is the average of all absolute errors. n = the number of errors,; $\sum$ = summation symbol; $|yj{-}y|$ = the absolute errors In 10-fold cross-validation, the original sample is randomly partitioned into 10 equal size subsamples.

$$\mathrm{RMSE} = \sqrt{\sum \frac{(y_{pred} - y_{ref})^2}{N}}$$

(a) RMSE formula

$$\mathrm{MAE} = \frac{1}{n}\sum_{j=1}^{n} |y_j - \hat{y}_j|$$

(b) MAE formula

Of the 10 subsamples, a single subsample is retained as the validation data for testing the model, and the remaining 9 subsamples are used as training data. The cross-validation process is then repeated 10 times (the folds), with each of the 10 subsamples used exactly once as the validation data. The 10 results from the folds can then be averaged to produce a single estimation. The advantage of this method is that all observations are used for both training and validation, and each observation is used for validation exactly once.In this question, we design a k-NN collaborative filter and perform 10 fold cross validation. We sweep k(number of neighbors) from 2 to 100, compute average RMSE and MAE for each k and plot it. Here we design the filter to predict the ratings of the movies in the MovieLens dataset.

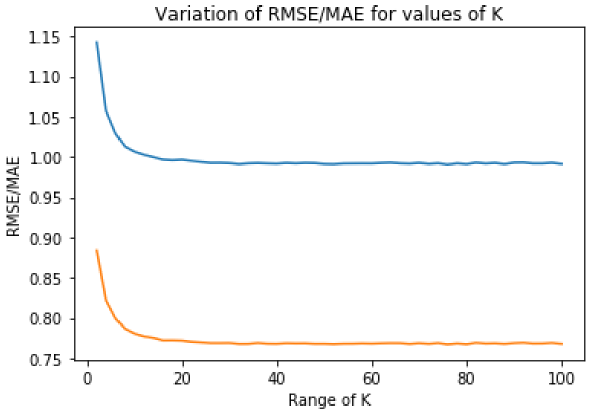


Figure 5: Variation of RMSE/MAE for values of K for kNN collaborative filtering

11 Question 11

The minimum k value is 20 for the steady state RMSE value of 1 and MAE value of 0.77. The 'minimum k' would correspond to the k value for which average RMSE and average MAE converges to a steady-state value.

12 Question 12

We analyze the performance of the movies in the trimmed data set. The types of trimmed data set are: Popular movie trimming- data set containing movies having rating more than 2, Unpopular movie trimming- data set containing movies having rating less than 2, High variance movie trimming- data set containing movies that has variance at least 2 and 5 ratings in the entire data set.Here we design the filter to predict the ratings of the movies in the popular movie trimmed dataset.We design a k-NN collaborative filter and perform 10 fold cross validation. We sweep k(number of neighbors) from 2 to 100, compute average RMSE for each k and plot it. Minimum average RMSE = 0.9853

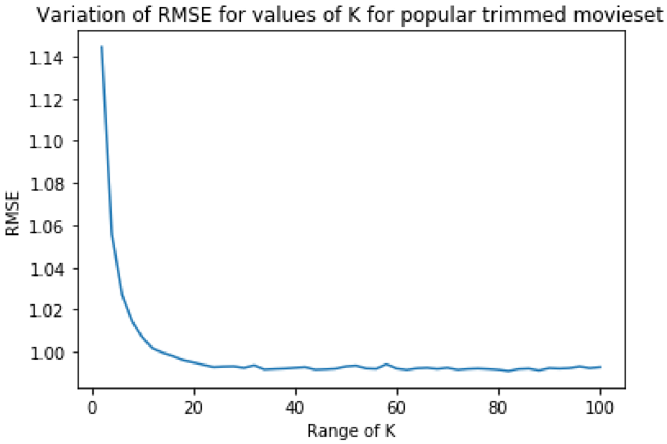


Figure 6: Variation of RMSE for values of K for popular trimmed movieset for kNN collaborative filtering

13 Question 13

Here we design the filter to predict the ratings of the movies in the unpopular trimmed movie dataset.We design a k-NN collaborative filter and perform 10 fold cross validation. We sweep k(number of neighbors) from 2 to 100, compute average RMSE for each k and plot it as shown in Figure 7.
Minimum average RMSE = 1.0702

14 Question 14

Here we design the filter to predict the ratings of the movies in the high variance movie trimmed dataset.We design a k-NN collaborative filter and perform 10 fold cross validation. We sweep k(number of neighbors) from 2 to 100, compute average RMSE for each k and plot it as shown in Figure 8.
Minimum average RMSE = 1.4602

15 Question 15

ROC curve is a measure of the relevance of the items recommended to the user in context to recommendation systems. By applying threshold to the observed ratings, we convert them to binary scale. Observed ratings

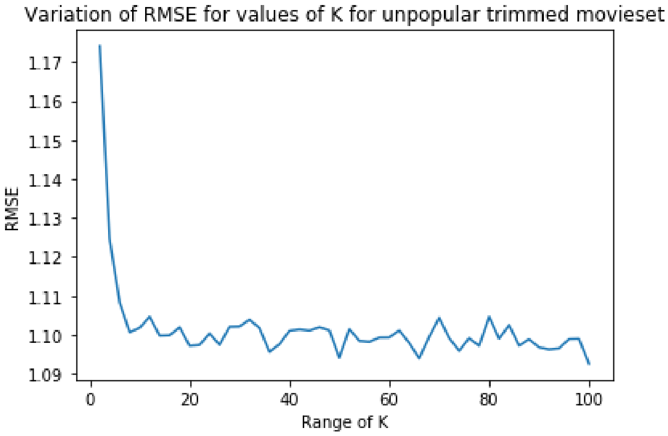


Figure 7: Variation of RMSE for values of K for unpopular movie trimmed movieset for kNN collaborative filtering

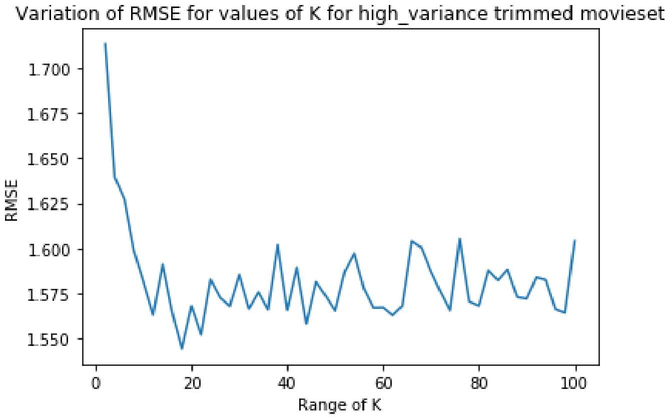
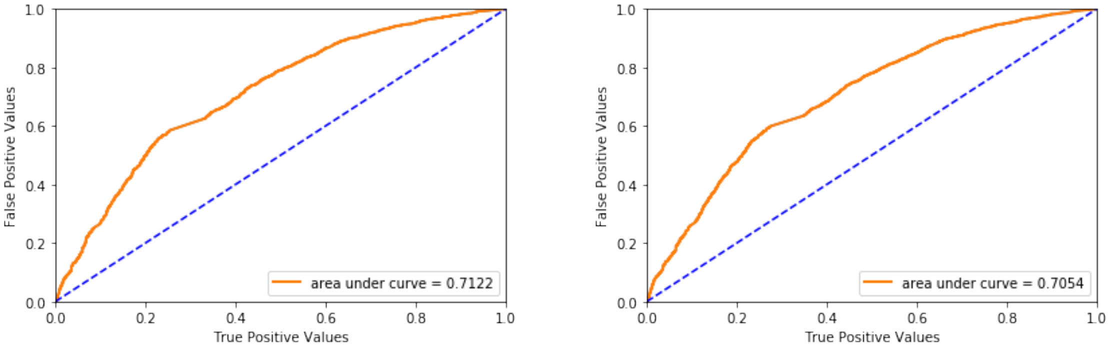


Figure 8: Variation of RMSE for values of K for high variance movie trimmed movieset for kNN collaborative filtering

greater than the set threshold is assigned value 1 and those below the threshold are assigned value 0. In this question, we set different threshold values- 2.5,3,3.5,4. We take k value solved from question 11 as 20. Area under the curve (AUC) value is also plotted for them all as shown in Figure 9 and Figure 10.

16 Question 16

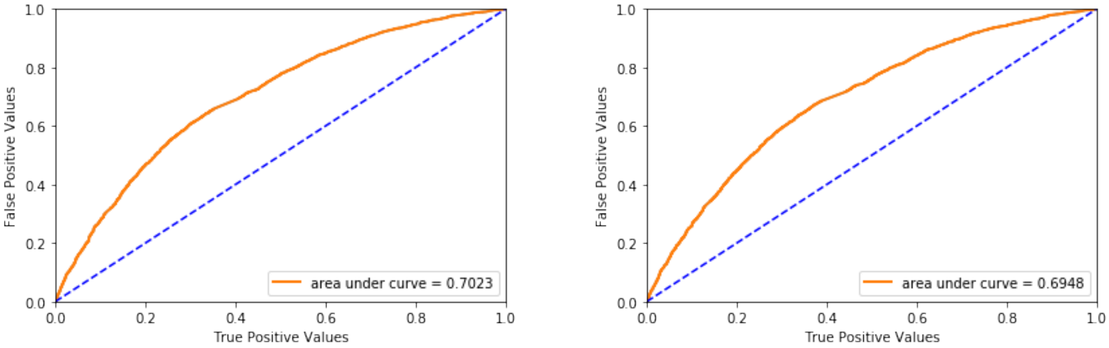
The equation 5 is a non-convex equation because both U and V are unknown. In order to change it to a least square problem, we assume one of the values of U and V. So by fixing either one of these values, the optimal value of the other parameter can be determined by minimizing the least square problem which is of the form shown in figure 11.



(a) For threshold = 2.5

(b) For threshold = 3.5

Figure 9: ROC Curve for threshold=2.5/3 in KNN collaborative filtering



(a) For threshold = 4

(b) For threshold = 4.5

Figure 10: ROC Curve for threshold=3.5/4 in KNN collaborative filtering

17 Question 17

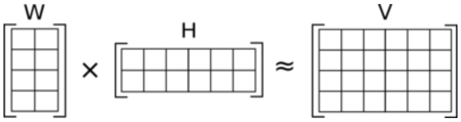
Non-negative matrix factorization is a group of algorithms in multivariate analysis and linear algebra where a matrix V is factorized into (usually) two matrices W and H, with the property that all three matrices have no negative elements. This non-negativity makes the resulting matrices easier to inspect. In this question, we design a NNMF-based collaborative filter and perform 10 fold cross validation. We sweep k(number of neighbors) from 2 to 50, compute average RMSE and MAE for each k and plot it. Here we design the filter to predict the ratings of the movies in the MovieLens dataset as shown in Figure 12.

18 Question 18

Optimal number of latent factors is the value of k that gives the minimum average RMSE or the minimum average MAE. According to our results,
number of latent factors = 15.
Minimum average RMSE= 0.95
Minimum average MAE= 0.72
No, the number of movie genres are 18(without counting no genres specified) which are not same as number of latent factors we obtained (12).

$$| \sum_{i:(i,j)\in S} W_{ij} \; (r_{ij} - \sum_{s=1}^{k} u_{is} \; v_{js})^2$$

Figure 11: Question 16 formula



(a) NNMF

$$\begin{aligned} &\underset{U,V}{\text{minimize}} && \sum_{i=1}^{m}\sum_{j=1}^{n} W_{ij}(r_{ij} - (UV^T)_{ij})^2 + \lambda \|U\|_F^2 + \lambda \|V\|_F^2 \\ &\text{subject to} && U \geq 0, V \geq 0 \end{aligned}$$

(b) Optimization Formulation

19 Question 19

Here we design the filter to predict the ratings of the movies in the popular trimmed movie dataset.We design a NNMF collaborative filter and perform 10 fold cross validation. We sweep k(number of neighbors) from 2 to 50, compute average RMSE for each k and plot it as shown in Figure 13. Minimum average RMSE = 0.88

20 Question 20

Here we design the filter to predict the ratings of the movies in the unpopular trimmed movie dataset.We design a NNMF collaborative filter and perform 10 fold cross validation. We sweep k(number of neighbors) from 2 to 50, compute average RMSE for each k and plot it as shown in Figure 14.
Minimum average RMSE = 1.025

21 Question 21

Here we design the filter to predict the ratings of the movies in the high variance trimmed movie dataset.We design a NNMF collaborative filter and perform 10 fold cross validation. We sweep k(number of neighbors) from 2 to 50, compute average RMSE for each k and plot it as shown in Figure 15.
Minimum average RMSE = 1.38

22 Question 22

Here, we evaluate the performance of NNMF collaborative filter for threshold values 2.5,3.5,4,4.5 using ROC curve.We take the number of latent factors to be 12 as we found in question 18. The area under the curve is also reported as shown in Figures 16(a)(b) and 17(a)(b).

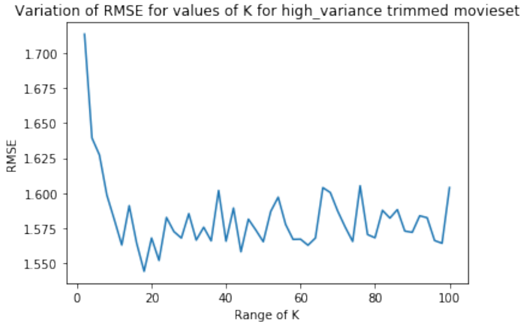


Figure 12: Variation of RMSE/MAE for values of K for NMF

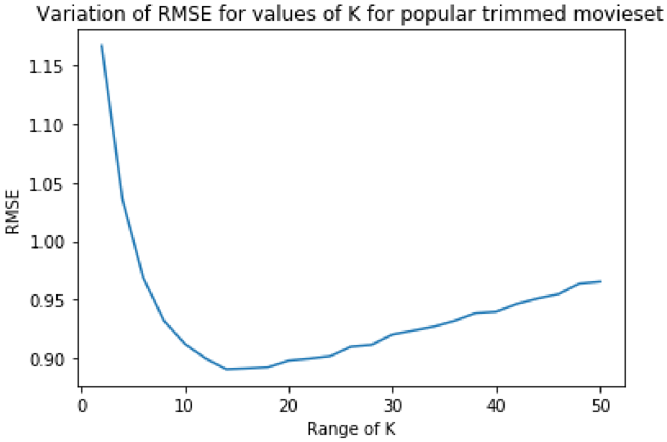


Figure 13: Variation of RMSE for values of K for popular movie trimmed movieset for NMF

23 Question 23

We explore the interpretability of NNMF. We perform NNMF on ratings matrix R to get matrices U and V. The movies are sorted in descending order and the genres of top 10 movies are reported.
For column 1:
8039 Action—Sci-Fi—Thriller—IMAX
4519 Comedy—Crime—Drama
7296 Comedy
4029 Comedy
1710 Children—Comedy
2455 Adventure—Animation—Children—Fantasy—Sci-Fi
3003 Adventure—Animation—Children—Comedy—Fantasy
5881 Drama
2406 Drama
4814 Children—Comedy

For column 2:
7485 Comedy
3536 Comedy
279 Animation—Children
354 Adventure—Animation
1443 Documentary
2384 Drama—War

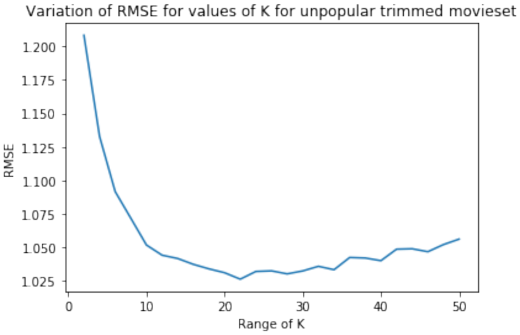


Figure 14: Variation of RMSE for values of K for unpopular movie trimmed movieset for NMF

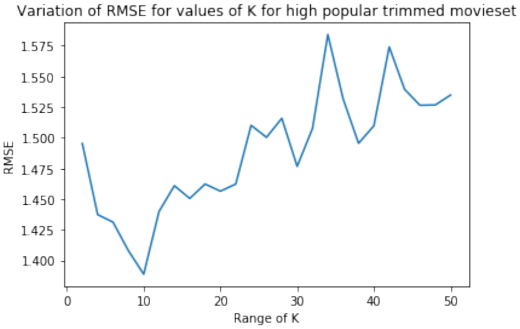


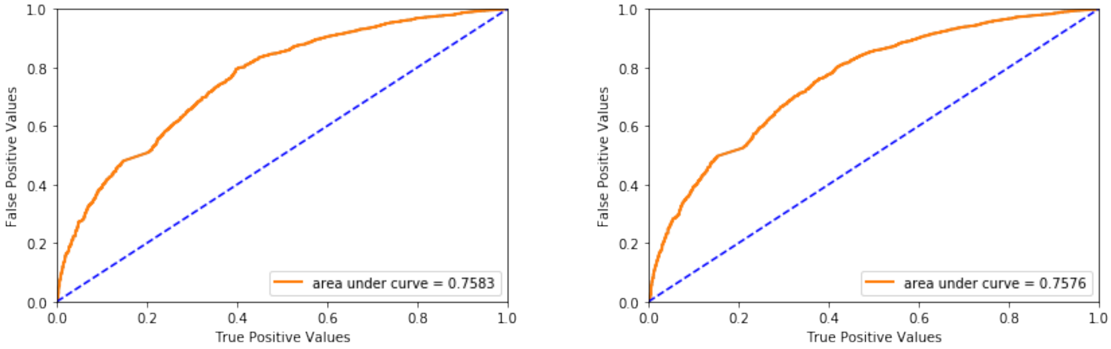
Figure 15: Variation of RMSE for values of K for high variance trimmed movie dataset for NMF

809 Children—Comedy
3516 Documentary—IMAX
430 Drama
7313 Crime—Drama—Mystery—Romance—Thriller

For column 3:
987 Drama—Mystery—Sci-Fi
7199 Comedy—Romance
4822 Action—Adventure—War
3028 Drama—Mystery
809 Children—Comedy
1652 Adventure—Animation—Children—Fantasy—Musical—Romance 5185 Comedy—Fantasy—Musical—Romance
4131 Comedy
285 Comedy—Drama—Thriller
6534 Comedy—Documentary
The genre ”Comedy” seems to appear the most in the few columns (3) reported above. Other than that, the displayed result imply that the top 10 movies do not belong to a specific genre. The number of genres are 19 and the number of latent factors are 20.

24 Question 24

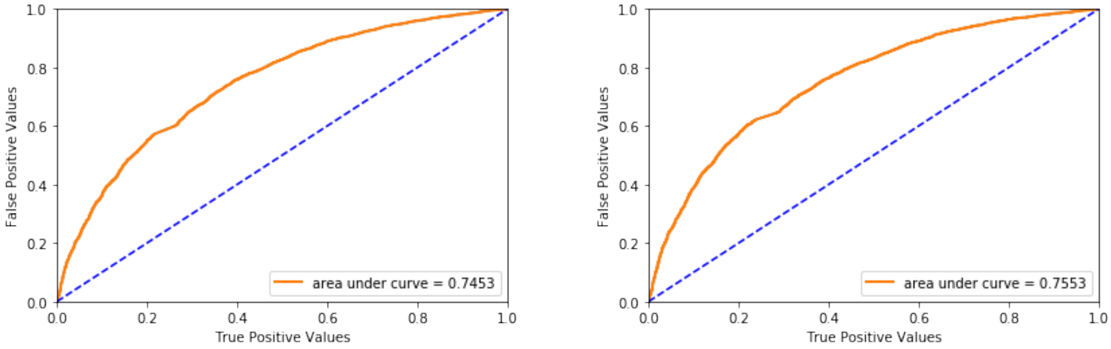
We design a Matrix Factorization with bias collaborative filter and test it's performance via 10-fold cross validation. We sweep k(number of neighbors) from 2 to 50, compute average RMSE and MAE for each k and plot it. Here we design the filter to predict the ratings of the movies in the MovieLens dataset as shown in Figures 18(a)(b).



(a) For threshold = 2.5

(b) For threshold = 3.5

Figure 16: ROC Curve for threshold=2.5/3 in NMF



(a) For threshold = 4

(b) For threshold = 4.5

Figure 17: ROC Curve for threshold=3.5/4 in NMF

25 Question 25

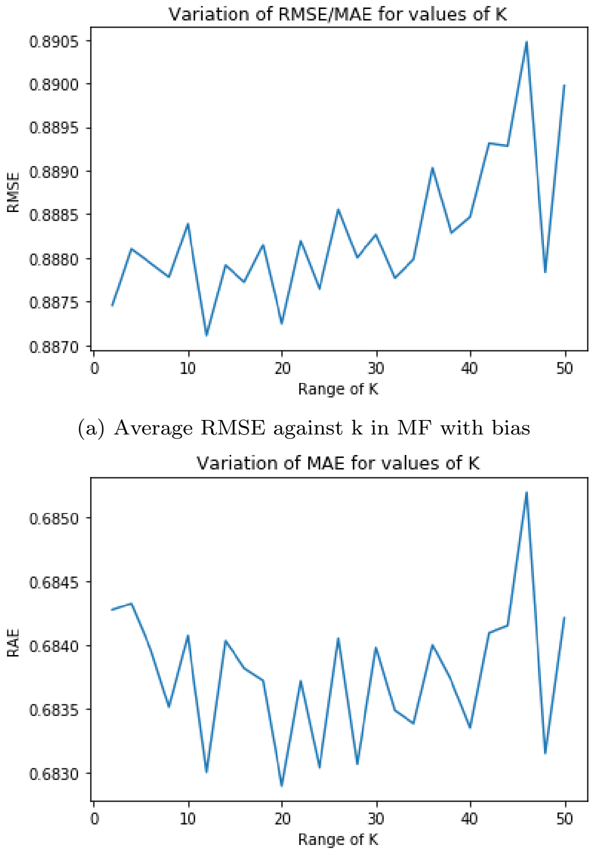
Optimal number of latent factors is the value of k that gives the minimum average RMSE or the minimum average MAE.
Number of latent factors=20
Minimum average RMSE = 0.8872
Minimum average MAE = 0.6828

26 Question 26

Here we design the filter to predict the ratings of the movies in the popular trimmed movie dataset.We design a MF with bias collaborative filter and perform 10 fold cross validation. We sweep k(number of neighbors) from 2 to 50, compute average RMSE for each k and plot it as shown in Figure 19. Minimum average RMSE = 0.86342

27 Question 27

Here we design the filter to predict the ratings of the movies in the unpopular trimmed movie dataset.We design a MF with bias collaborative filter and perform 10 fold cross validation. We sweep k(number of neighbors) from 2 to 50, compute average RMSE for each k and plot it as in Figure 20. Minimum average RMSE = 0.935



(a) Average RMSE against k in MF with bias

(b) Average MAE against k in MF with bias

Figure 18: Average RMSE and MAE against k in MF with bias

28 Question 28

Here we design the filter to predict the ratings of the movies in the high variance trimmed movie dataset.We design a MF with bias collaborative filter and perform 10 fold cross validation. We sweep k(number of neighbors) from 2 to 50, compute average RMSE for each k and plot it as in Figure 21. Minimum average RMSE = 1.38

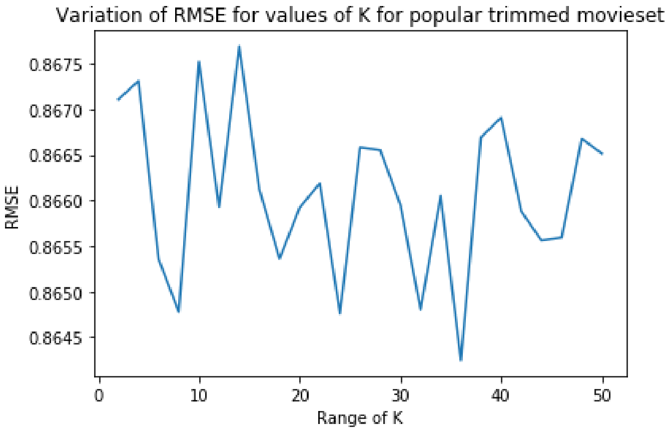


Figure 19: Average RMSE against k in MF with bias for popular movie set

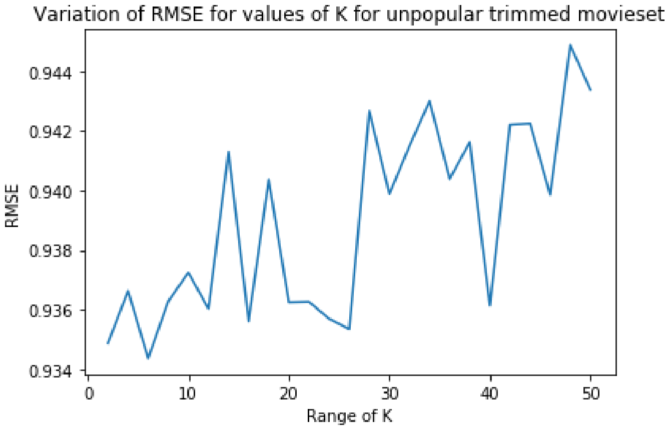


Figure 20: Average RMSE against k in MF with bias for unpopular movieset

29 Question 29

Here, we evaluate the performance of MF with bias collaborative filter for threshold values 2.5,3,3.5,4 using ROC curve.We take the number of latent factors to be as we found in question 25. The area under the curve is also reported as shown in Figure 22.

30 Question 30

We design a Naive collaborative filter and test it's performance via 10-fold cross validation. We compute average RMSE and MAE for across all 10 folds and plot it. Here we design the filter to predict the ratings of the movies in the MovieLens dataset. The average RMSE (naive collaborative filter) is 0.913

31 Question 31

Here we design the filter to predict the ratings of the movies in the popular trimmed movie dataset.We design a naive collaborative filter and perform 10 fold cross validation. We compute the average RMSE by averaging the RMSE across all 10 folds. The average RMSE (naive collaborative filter, Popular Movies) is 0.883.

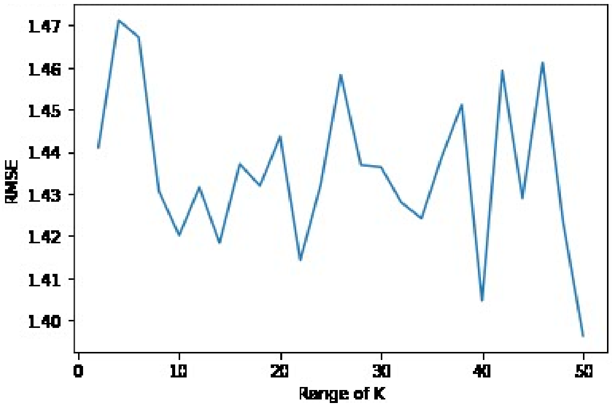


Figure 21: Average RMSE against k in MF with bias for high variance movieset

32 Question 32

Here we design the filter to predict the ratings of the movies in the unpopular trimmed movie dataset.We design a naive collaborative filter and perform 10 fold cross validation. We compute the average RMSE by averaging the RMSE across all 10 folds. The average RMSE (naive collaborative filter, Unpopular Movies) is 0.982.

33 Question 33

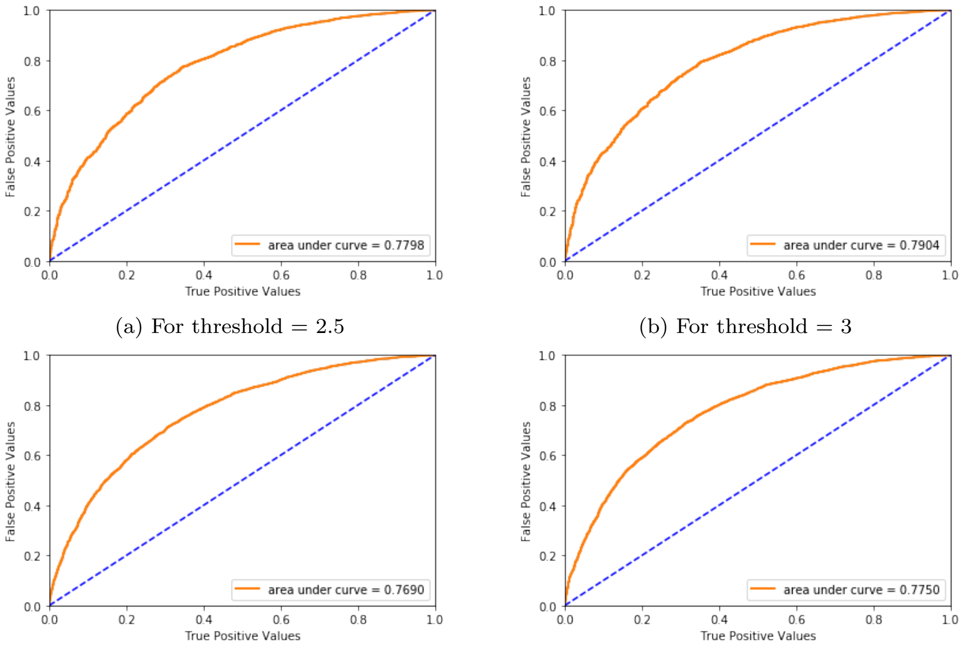
Here we design the filter to predict the ratings of the movies in the high variance trimmed movie dataset.We design a naive collaborative filter and perform 10 fold cross validation. We compute the average RMSE by averaging the RMSE across all 10 folds. The average RMSE (naive collaborative filter, High Variance) is 2.135.

34 Question 34

In this question, we will compare the performance of the various collaborative filters in predicting the ratings of the movies in the MovieLens dataset. For this, we plot the ROC curves (threshold = 3) for the k-NN, NNMF, and MF with bias based collaborative filters as shown in Figure 23. From the figure, we conclude that MF with bias performs the best than the rest. It is followed by NMF and kNN in the decreasing order of performance.

35 Question 35

Precision is defined as Precision(t)=$|S(t) \cap G|/|S(t)|$ according to the formula where S(t) refers to set of items recommended to the user and G is the set of items liked by the user. This means that it refers the fraction of the set of items recommended to the user that are common in both the recommended list and the liked set.It can also be evaluated at a given cut-off rank, considering only the topmost results returned by the system. This measure is called precision at n or P@n.
Recall on the other hand is defined as Recall(t) = $|S(t) \cap G|/|G|$ according to the formula. This means that it refers to the fraction of the ground truth positives (the relevant items in the given data set list) that belong to both recommended list and user-liked list.



(a) For threshold = 2.5

(b) For threshold = 3

(c) For threshold = 3.5

(d) For threshold = 4

Figure 22: ROC Curve for threshold=2.5/3/3.5/4 in MF with Bias

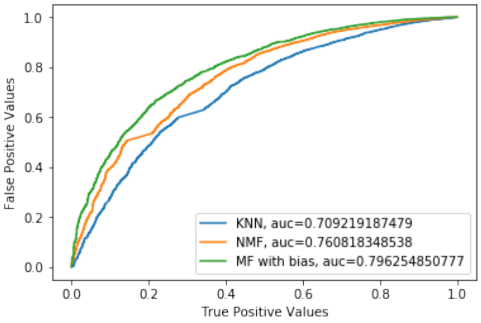


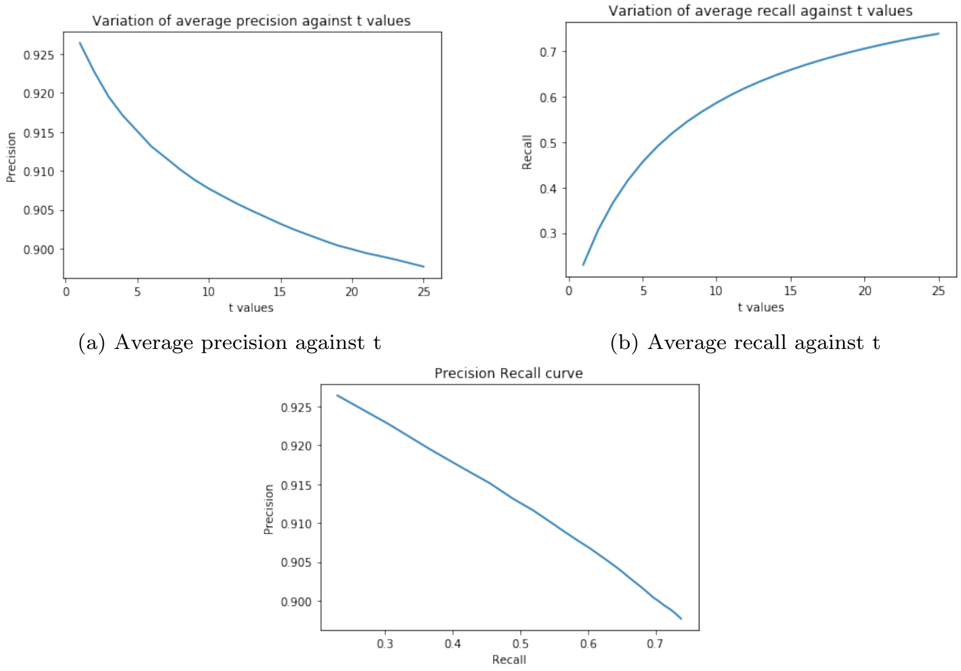
Figure 23: ROC curves (threshold = 3) for the k-NN, NNMF and MF

36 Question 36

Here we plot using k-NN collaborative filter predictions, 1. average precision against t, 2. the average recall against t and 3. average precision against average recall. We use the k=40 found in question 11 and sweep t from 1 to 25 in step sizes of 1 as shown in Figure 24. The shape of the average precision against t plot is decreasing.
The shape of the average recall against t increasing.
The shape of the average precision against average recall: as recall increases, precision decreases and this shows the efficiency of the method in predicting the precision and recall. It is a decaying curve.

37 Question 37

Here we plot using NNMF-based collaborative filter predictions, 1. average precision against t, 2. the average recall against t and 3. average precision against average recall. We use the optimal number of latent factors=12 found in question 18 and sweep t from 1 to 25 in step sizes of 1 as shown in Figure 25. The shape



(a) Average precision against t

(b) Average recall against t

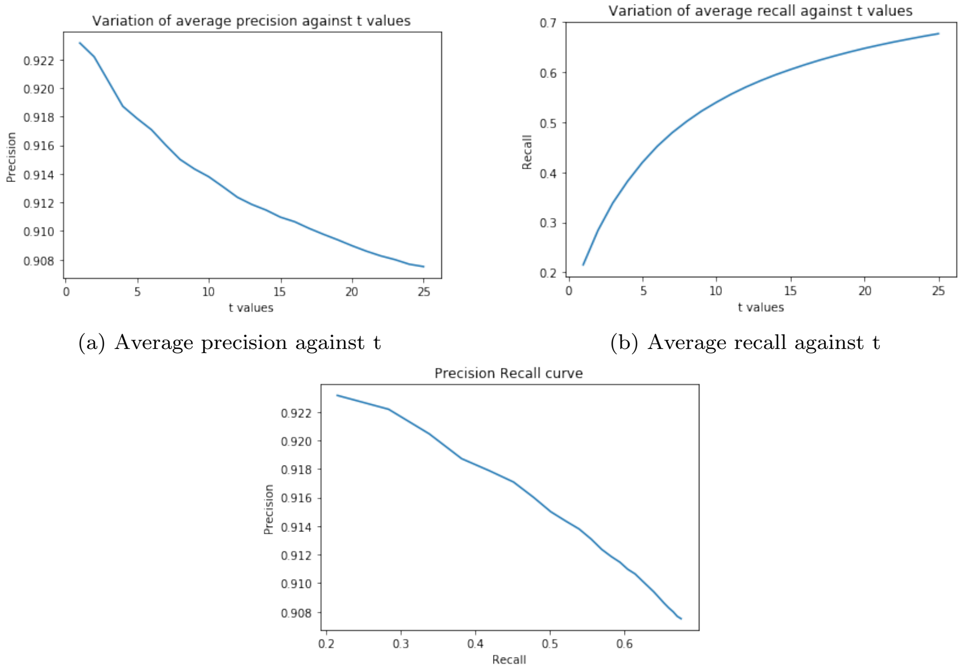
(c) average precision against average recall

Figure 24: PR curve for k-NN collaborative filter

of the average precision against t plot is decreasing.
The shape of the average recall against t exponentially increasing.
The shape of the average precision against average recall: as recall increases, precision decreases and this shows the efficiency of the method in predicting the precision and recall. It is a decaying curve.

38 Question 38

Here we plot using MF with bias-based collaborative filter predictions, 1. the average precision against t, 2. the average recall against t and 3. average precision against average recall. We use the optimal number of latent factors= found in question 25 and sweep t from 1 to 25 in step sizes of 1 as shown in Figure 26. The shape of the average precision against t plot is decreasing.
The shape of the average recall against t increasing.
The shape of the average precision against average recall: as recall increases, precision decreases and this shows the efficiency of the method in predicting the precision and recall. It is a decaying curve.



(a) Average precision against t

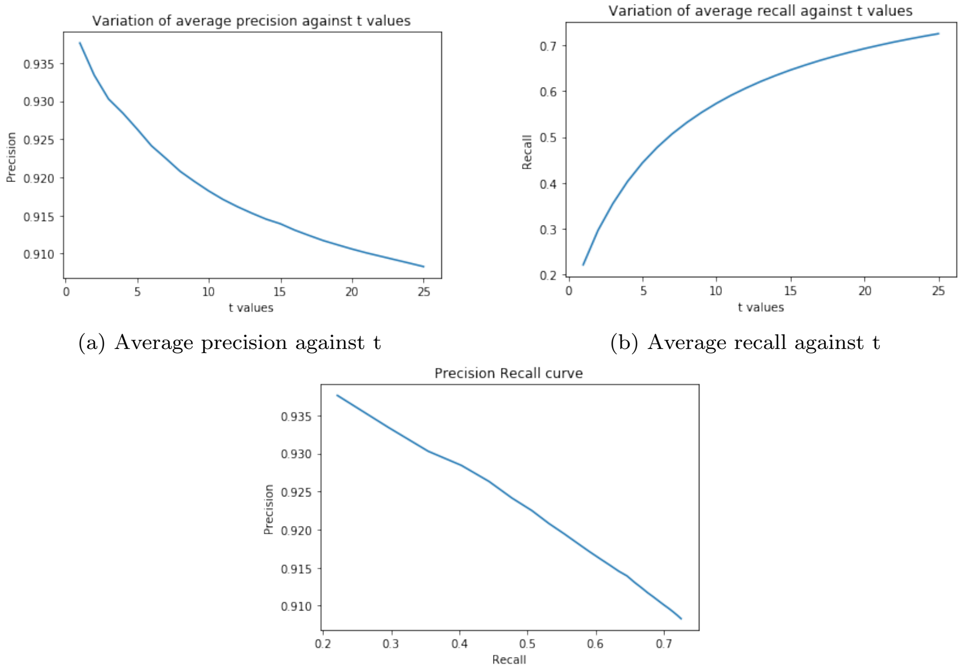
(b) Average recall against t

(c) average precision against average recall

Figure 25: PR curve for NNMF based collaborative filter

39 Question 39

In this question, we will compare the relevance of the recommendation list generated using k-NN, NNMF, and MF with bias predictions. For this, precision-recall curve obtained in questions 36,37, and 38 in the same Figure 27.



(a) Average precision against t

(b) Average recall against t

(c) average precision against average recall

Figure 26: PR curve for MF with bias based collaborative filter

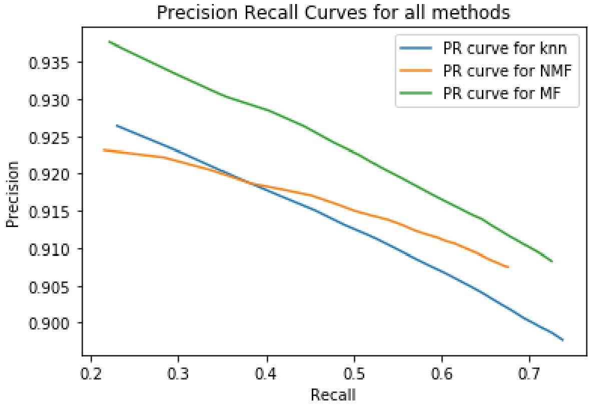


Figure 27: Precision-recall curve obtained for k-NN, NNMF, and MF