ECE219 — Large Scale Data Mining

Name: Lu Ren UID: 704706181
Name: Tianxue Chen UID: 004943548
Name: Yuting Tang UID: 304881011
Name: Xiao Peng UID: 005033608
Due Date: Feb 22 2018, 11:59pm Assignment: HW 3

Question (1)

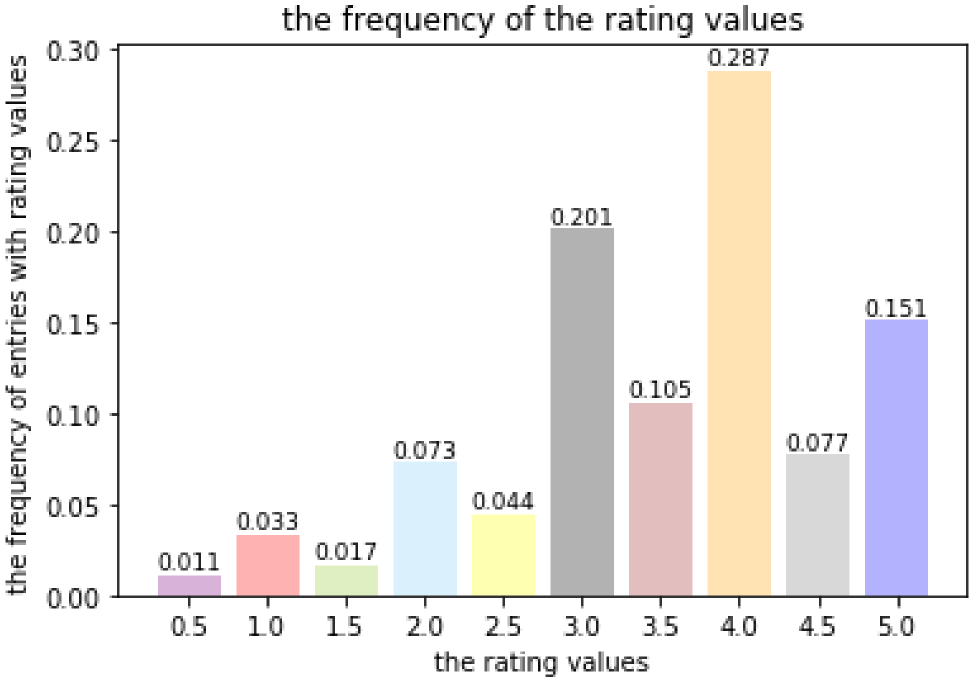
Description: Compute the **sparsity** of the movie rating dataset.

Results: The sparsity of the movie rating dataset: 0.016.

Question (2)

Description: Plot a histogram showing the frequency of the rating values. Briefly comment on the shape of the histogram.

Plot:



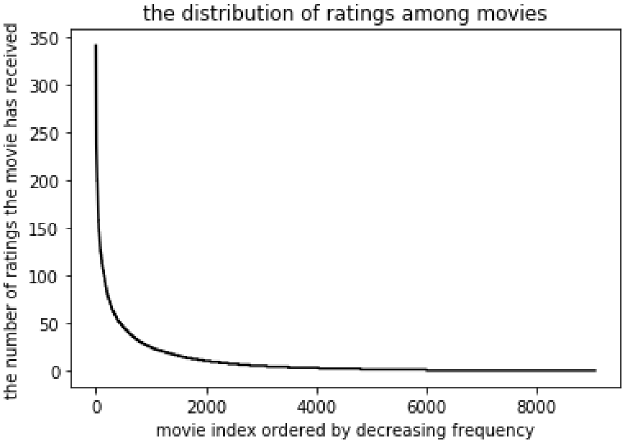
Discussion:

The histogram above shows the frequency of the rating values, according to which, we can find that the most common ratings concentrate on range from 3.0 to 5.0. And comparing to them, other ratings are negligible since their low showing frequency in the whole dataset.

Question (3)

Description: Plot the distribution of ratings among movies.

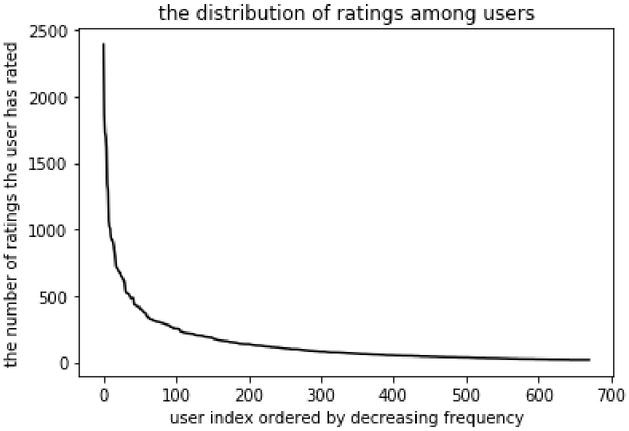
Plot:



Question (4)

Description: Plot the distribution of ratings among users.

Plot:



Question (5)

Description: Explain the **salient features** of the distribution found in question 3 and their implications for the recommendation process.

Explanation:

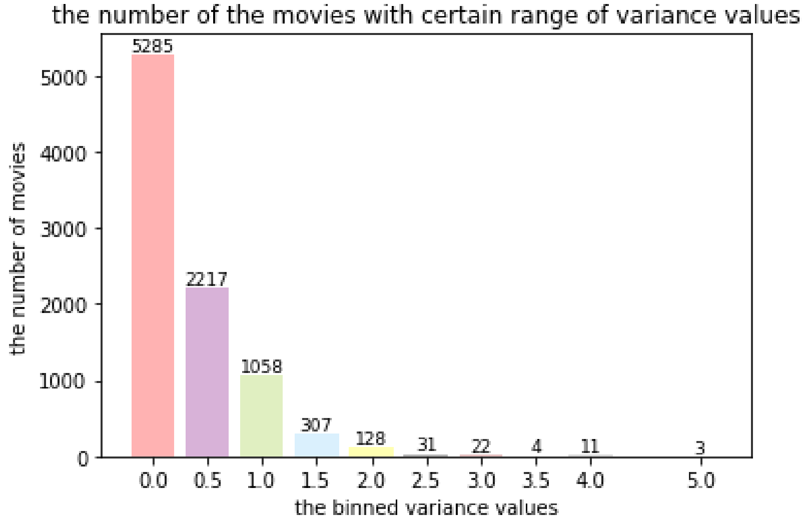
According to the figure in question 3, it’s evident that movies with a lot of ratings only account for a very small part of all the movies. In other words, more ratings concentrate on a small number of movies, resulting in the fact that the distribution of ratings is uneven.

As for the processing of the recommendation system, movies with more ratings have more user preference information, thus the recommendation for this part of the movies predicted by the system is based on more rating data, leading to more accurate recommendation result. On the contrary, the predictions and recommendations of the other part movies that have less ratings, some may only possessing one or two ratings, may not be accurate enough due to the information inadequacy.

Question (6)

Description: Compute the variance of the rating values received by each movie. Then, bin the variance values into intervals of width 0.5 and use the binned **variance** values as the horizontal axis. Briefly comment on the shape of the histogram.

Plot:



Discussion:

The histogram shows the amount of movies with certain rating variance value. According to this figure, we can intuitively gain the observation that the overall trend of the histogram is exponential decline. The variance of the most movie ratings are at the range from 0 to 1.5, whose amounts are dominant to those of the others.

Question (7)

Description: Write down the **formula** for μ_u in terms of I_u and r_{uk} .

Discussion:

The formula for μ_u in terms of I_u and r_{uk} is like,

$$m_u = \frac{1}{N} \mathring{a}_{k\hat{I}\, I_u} r_{uk}$$

where N is the amount of the ratings have been specified by user u.

Question (8)

Description: In plain words, explain the meaning of $I_u \cap I_v$. Can $I_u \cap I_v = Ø$?

Discussion:

$I_u \cap I_v$ means the movie that has been rated both by user u and user v. In this sense, of course $I_u \cap I_v$ can be equal to 0, which means user u and user v never rate the same movie.

Question (9)

Description: Can you explain the reason behind **mean-centering** the raw ratings ($r_{vj} - \mu_v$) in the prediction function?

Discussion:

In fact, when it comes to anything concerning of personal taste, such as the movie rating, we must take the diversity of different people's standards into consideration, which means we may have different evaluation criteria for average movie.

The prediction function is like,

$$\hat{r}_{uj} = m_u + \frac{\underset{v\hat{I}\;P_u}{\text{å}} Pearson(u,v)\left(r_{vj} - m_v\right)}{\underset{v\hat{I}\;P_u}{\text{å}} \left|Pearson(u,v)\right|}$$

By mean-centering the raw ratings with ($r_{vj} - \mu_v$), we can eliminate the differences owing to different people having different evaluation criteria. For example, we have two users, one of which, user u is prone to be critical of the movie with the mean rating of 2, while the other one of which, user v, is looser about the movie with the mean rating of 4. If we choose not to mean-centering the raw ratings with,

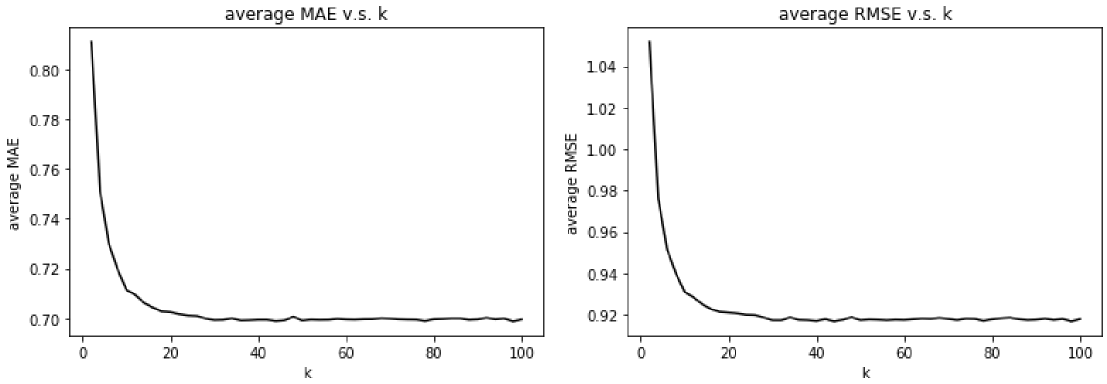
$$\tilde{r}_{uj} = \frac{\sum_{v \in P_u} Pearson(u,v) \cdot r_{vj}}{\sum_{v \in P_u} \left|Pearson(u,v)\right|}$$

The predicted rating $\overline{r}_{uj}$ gives more weights to the user v than to the user u, since the former has μ_v of 4, obviously higher than that of the latter of 2.

Question (10)

Description: Design a **k-NN collaborative filter** to predict the ratings of the movies in the MovieLens dataset and evaluate its performance using **10-fold cross validation**. Sweep k from 2 to 100 in step sizes of 2, and for each k compute the average RMSE and average MAE obtained by averaging the RMSE and MAE across all 10 folds. Plot average RMSE (Y-axis) against k (X-axis) and average MAE (Y-axis) against k (X-axis).

Plots:



Question (11)

Description: Use the plot from question 10, to find a **'minimum k'**. Please report the **steady state values** of average RMSE and average MAE.

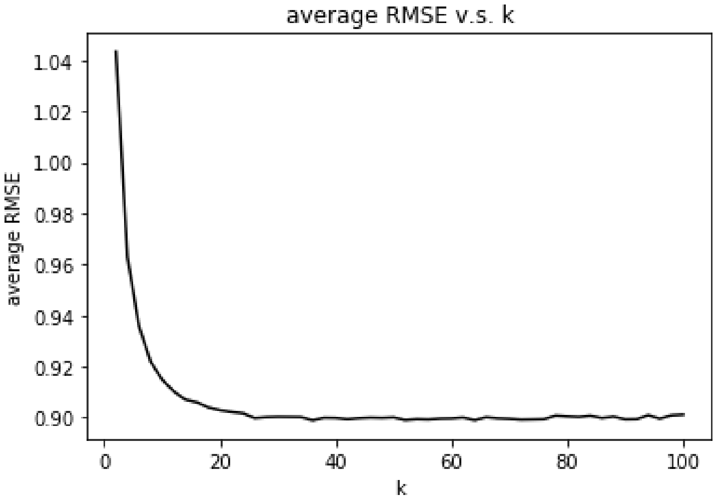
Result:

The 'minimum k' is 20. Accordingly the steady state values of average RMSE is 0.914 and average MAE is 0.698.

Question (12)

Description: Design a **k-NN collaborative filter** to predict the ratings of the movies in **the popular movie trimmed test set** and evaluate its performance using **10-fold cross validation**. Sweep k from 2 to 100 in step sizes of 2, and for each k compute the average RMSE obtained by averaging the RMSE across all 10 folds. Plot average RMSE (Y-axis) against k (X-axis). Also, report the minimum average RMSE.

Plots and results:

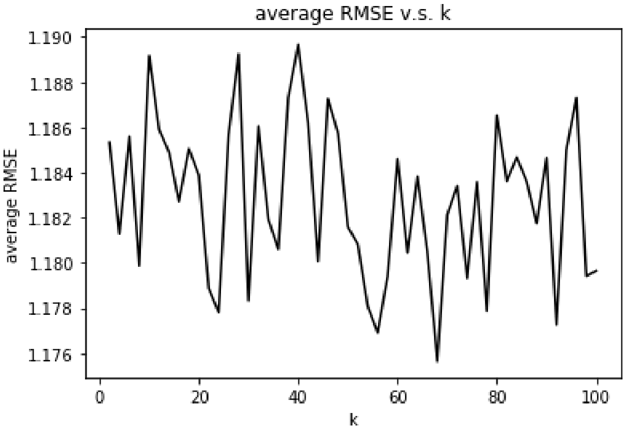


The minimum average RMSE is 0.8988

Question (13)

Description: Design a **k-NN collaborative filter** to predict the ratings of the movies in **the unpopular movie trimmed test set** and evaluate it's performance using **10-fold cross validation**. Sweep k from 2 to 100 in step sizes of 2, and for each k compute the average RMSE obtained by averaging the RMSE across all 10 folds. Plot average RMSE (Y-axis) against k (X-axis). Also, report the minimum average RMSE.

Plot and result:

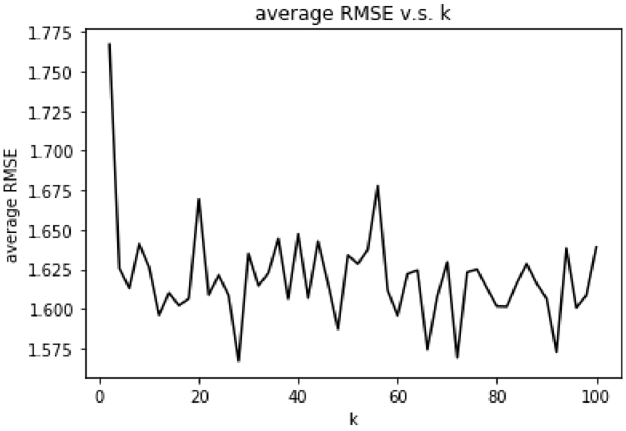


The minimum average RMSE is 1.1737

Question (14)

Description: Design a **k-NN collaborative filter** to predict the ratings of the movies in **the high variance movie trimmed test set** and evaluate its performance using **10-fold cross validation**. Sweep k from 2 to 100 in step sizes of 2, and for each k compute the average RMSE obtained by averaging the RMSE across all 10 folds. Plot average RMSE (Y-axis) against k (X-axis). Also, report the minimum average RMSE.

Plot and result:

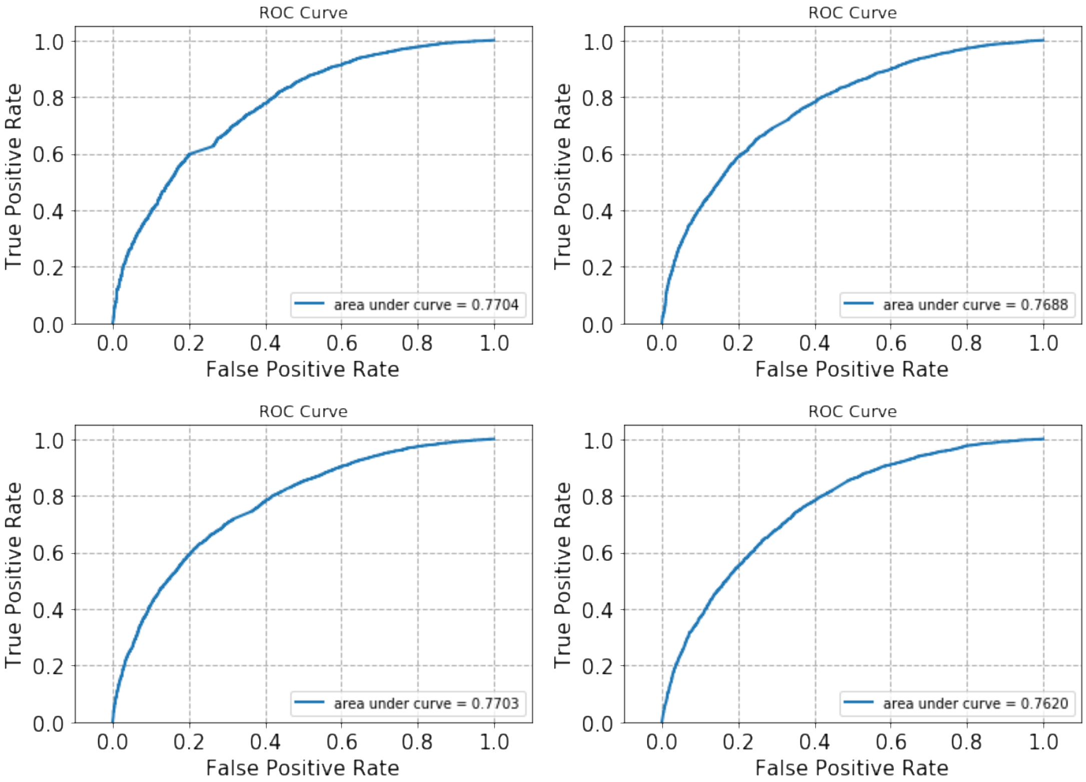


The minimum average RMSE is 1.5761

Question (15)

Description: Plot the **ROC curves** for the **k-NN collaborative filter** designed in question 10 for **threshold values** [2.5, 3, 3.5, 4]. For the ROC plotting use the k found in question 11. For each of the plots, also report the area under the curve value.

Plots and results:



Question (16)

Description: Is the optimization problem given by equation 5 convex? For U fixed, formulate it as a least-squares problem.

Solution: Yes, it is convex.

$$\frac{\partial}{\partial \mathrm{V}}\sum_{i=1}^{m}\sum_{j=1}^{n} W_{ij}\big(r_{ij}-(UV^T)_{ij}\big)^2$$

$$=\frac{\partial}{\partial \mathrm{V}}W(R-UV^T)^T(R-UV^T)$$

$$=\mathrm{W}\frac{\partial}{\partial \mathrm{V}}(R^TR-VU^TR-R^TUV^T-VU^TUV^T)$$

$$\because \frac{\partial R^TR}{\partial V}=0 \qquad \frac{\partial VU^TR}{\partial V}=\frac{\partial R^TUV^T}{\partial V}=R^TU \qquad \frac{\partial VU^TUV^T}{\partial V}=2U^TUV$$

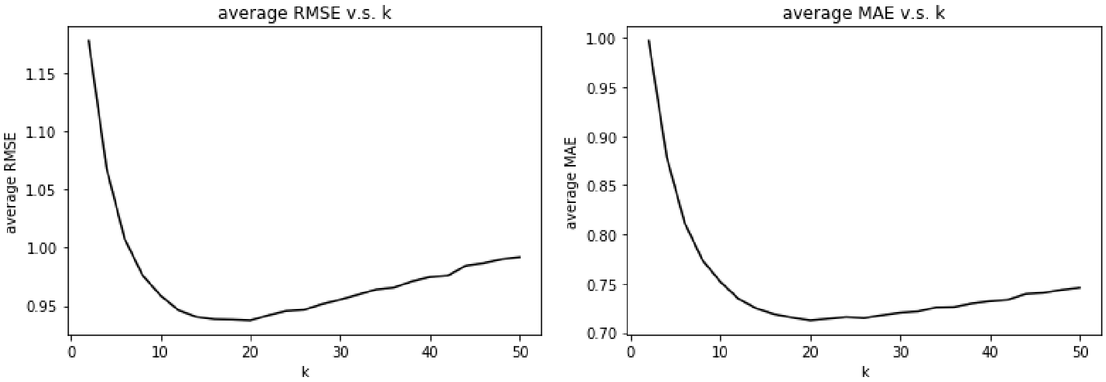
$$\therefore \mathrm{W}\frac{\partial}{\partial \mathrm{V}}(R^T R - VU^T R - R^T UV^T - VU^T UV^T)$$
$$= \mathrm{W}(2R^T U - 2U^T UV)$$

Since the second order derivative exists, the optimization problem given by equation 5 is convex.

Question (17)

Description: Design a **NNMF-based collaborative filter** to predict the ratings of the movies in the Movie Lens dataset and evaluate its performance using **10-fold cross-validation**. The number of latent factors range from 2 to 50 in step sizes of 2 and the plot of average RMSE and MAE over different k value.

Plots:



Question (18)

Description: Use the plot from question 17, to find the optimal number of latent factors

Results:

Optimal number of latent factors is 20.
Minimum average RMSE is 0.9374
Minimum average MAE is 0.7125

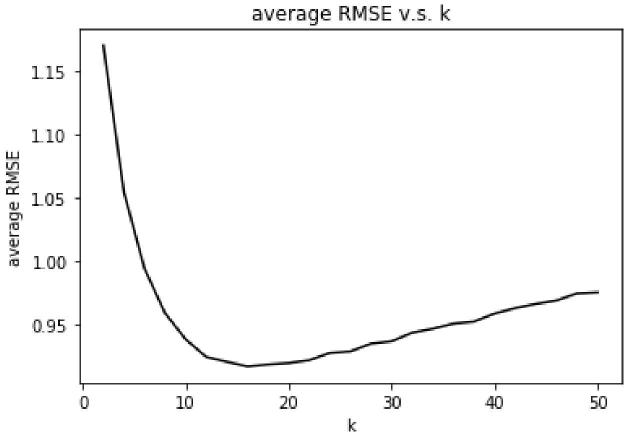
Q: Is the optimal number of latent factors same as the number of movie genres?

A: Yes, the optimal number of latent factors is 20 and the number of movie genres is 20, they are the same.

Question (19)

Description: Design a **NNMF collaborative filter** to predict the ratings of the movies in the **popular movie trimmed test set** and evaluate its performance using **10-fold cross-validation**. The number of latent factors range from 2 to 50 in step sizes of 2 and **the plot of average RMSE** obtained by averaging the RMSE across all 10 folds **over different k value**. And report the **minimum average RMSE**.

Plots and Results:

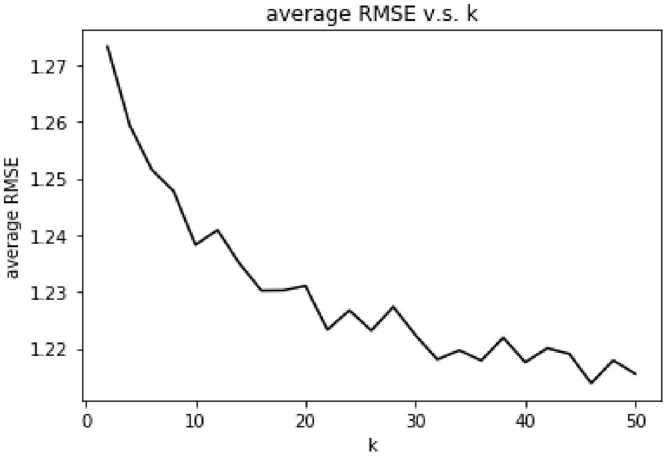


Minimum RMSE is 0.9215

Question (20)

Description: Design a **NNMF collaborative filter** to predict the ratings of the movies in the **unpopular movie trimmed test set** and evaluate its performance using **10-fold cross-validation**. The number of latent factors range from 2 to 50 in step sizes of 2 and **the plot of average RMSE** obtained by averaging the RMSE across all 10 folds **over different k value**. And report the **minimum average RMSE**.

Plots and Results:

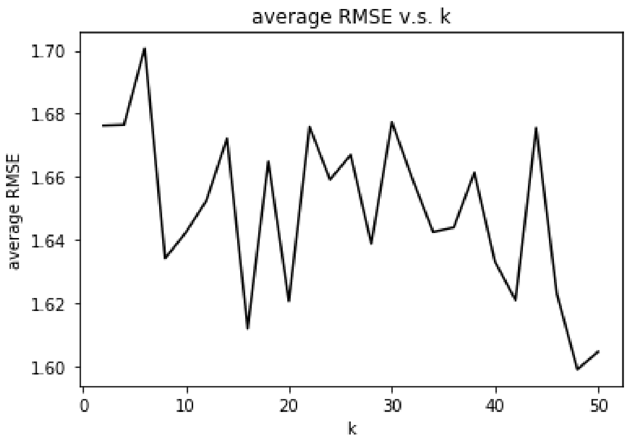


Minimum RMSE is 1.2139

Question (21)

Description: Design a **NNMF collaborative filter** to predict the ratings of the movies in the **high variance movie trimmed test set** and evaluate its performance using **10-fold cross-validation**. The number of latent factors range from 2 to 50 in step sizes of 2 and **the plot of average RMSE** obtained by averaging the RMSE across all 10 folds **over different k value**. And report the **minimum average RMSE**.

Plots and Results:

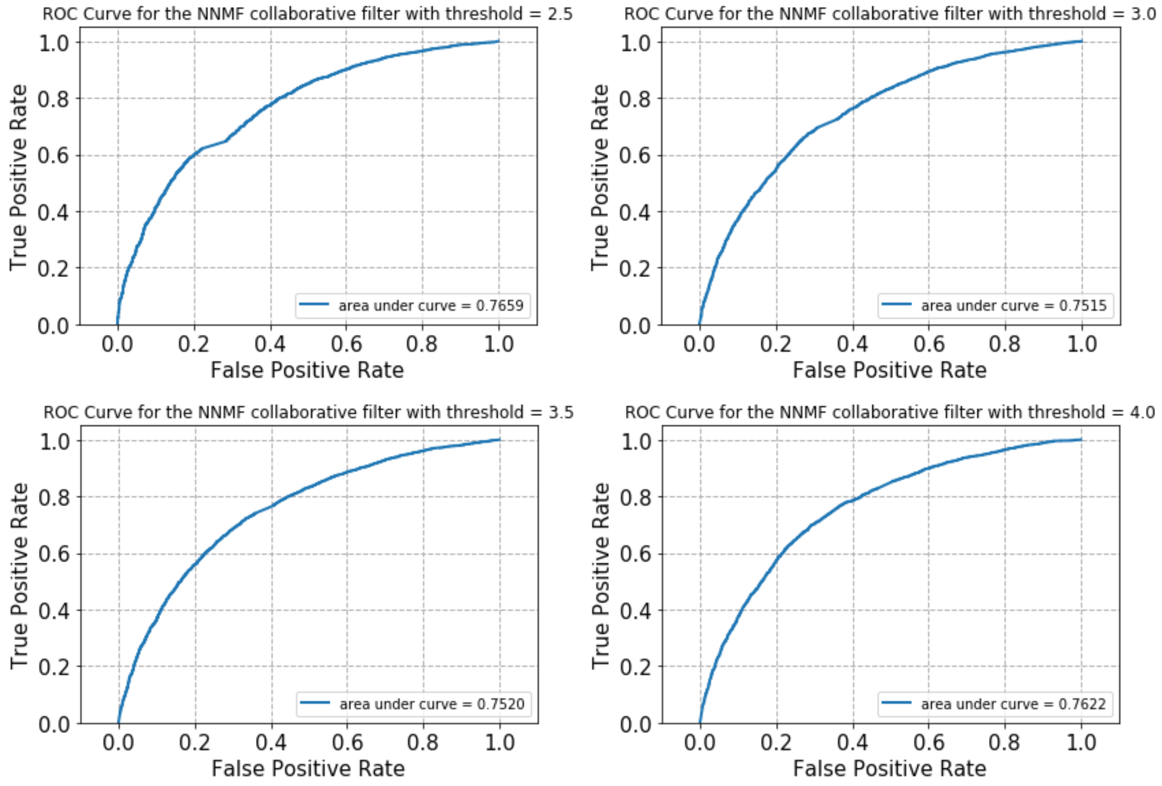


Minimum RMSE is 1.5989

Question (22)

Description: Plot the **ROC curve** for the NNMF-based collaborative filter designed in question 17. Use different **threshold as 2.5, 3, 3.5 and 4**. For ROC plotting use the **optimal number of latent factors 20**. Also calculate the area under the curve (**AUC**) score.

Plots:



Question (23)

Description: Perform **Non-negative matrix factorization** on the ratings matrix R to obtain the **factor matrices U and V**, where U represents the user-latent factors interaction and V represents the movie-latent factors interaction. For the first 5 columns of V, sort the movies in descending order and report the genres of the top 10 movies.

Solution:

Column #1:

Movie ID	genre
4630	Action
78105	Action|Adventure|Fantasy|Romance|IMAX
482	Drama
2888	Comedy|Romance
138036	Action|Adventure|Comedy
7155	Comedy
5334	Comedy|Crime
6345	Comedy|Drama|Musical
122902	Action|Adventure|Fantasy|Sci-Fi
94018	Action|Sci-Fi|Thriller|IMAX

Column #2:

Movie ID	genre
1735	Drama|Romance
39446	Horror|Thriller
2460	Horror
3865	Comedy|Documentary
4835	Drama
4520	Comedy
62081	Action|Crime|Thriller|IMAX
5172	Comedy|Drama|Romance
4343	Comedy|Sci-Fi
3910	Drama|Musical

Column #3:

Movie ID	genre
5597	Comedy|Sci-Fi
87522	Comedy|Drama|Romance
7577	Comedy|Fantasy|Musical|Romance
109187	Drama|Fantasy|Sci-Fi
429	Comedy

1150	Drama
5034	Drama|Romance
37731	Crime|Drama
6413	Comedy|Western
171	Comedy|Drama

Column #4:

Movie ID	genre
3910	Drama|Musical
6993	Comedy|Drama|Romance
58047	Comedy|Drama|Romance
1173	Comedy|Drama
3415	Drama
3865	Comedy|Documentary
4248	Comedy
3520	Comedy
80551	Drama|Romance
1005	Children|Comedy

Column #5:

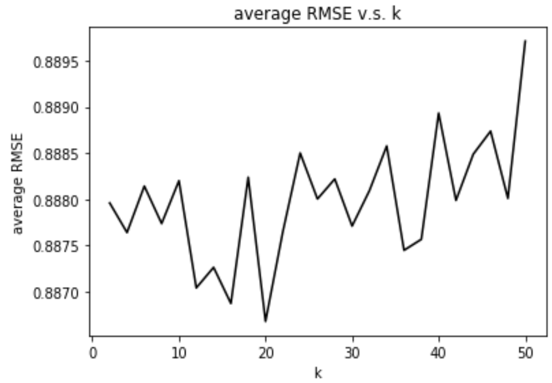
Movie ID	genre
6219	Action|Drama|Thriller
67255	Crime|Drama|Mystery|Thriller
65585	Comedy|Romance
67788	Comedy|Romance
3503	Drama|Mystery|Sci-Fi
63853	Adventure|Drama|War|Western
102993	Comedy|Drama
2570	Drama|Romance
6223	Comedy|Crime|Drama
1180	Comedy

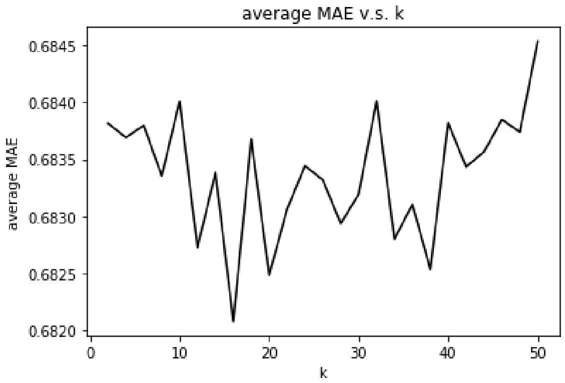
Discussion: From the above results, we randomly choose 5 columns from the total 20 columns and can be seen that the top 10 movies can belong to a small collection of genres, which is comedy, drama, Action and Horror. The movie genre and latent factor is actually a one to one mapping. Actually, each latent factor represents a specific movie genre, that is to say, for a unique movie ID, is belongs to the same genre for the same column (latent factor k). That can also help to explain why the top movies can belong to a small collection of genres.

Question (24)

Description: Design a **MF with bias collaborative filter** to predict the ratings of the movies in the MovieLens dataset and evaluate its performance using **10-fold cross-validation**. The number of latent factors range from 2 to 50 in step sizes of 2 and plot the average RMSE and MAE against different k value.

Plots:





Question (25)

Description: Use the plot from question 24, to find the optimal number of latent factors

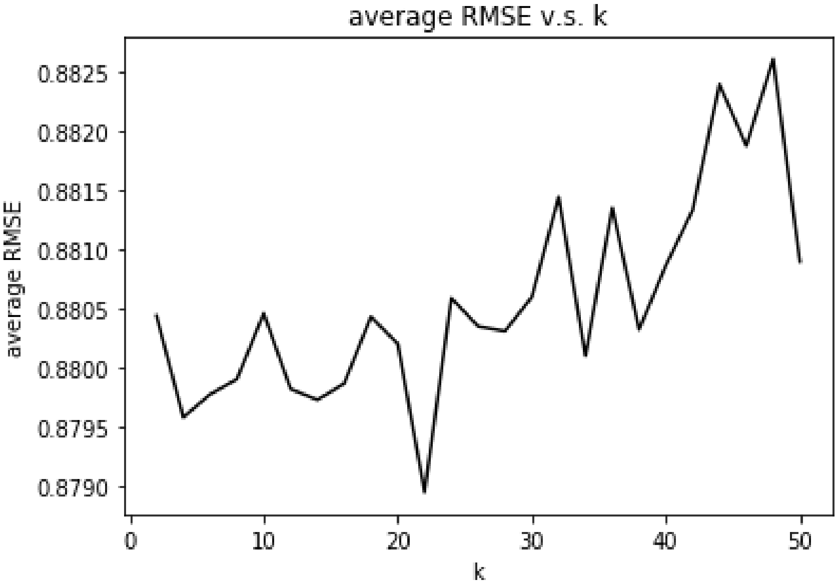
Results:

Optimal number of latent factors is 20.
Minimum average RMSE is 0.8867
Minimum average MAE is 0.6821

Question (26)

Description: Design a **MF with bias collaborative filter** to predict the ratings of the movies in the **popular movie trimmed test set** and evaluate its performance using **10-fold cross-validation**. The number of latent factors range from 2 to 50 in step sizes of 2 and **the plot of average RMSE** obtained by averaging the RMSE across all 10 folds **over different k value**. And report the **minimum average RMSE**.

Plots and Results:

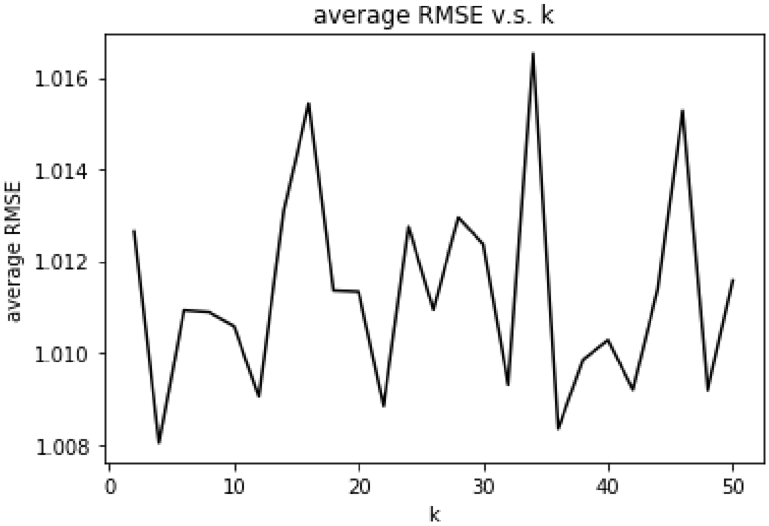


Minimum RMSE is 0.8789

Question (27)

Description: Design a **MF with bias collaborative filter** to predict the ratings of the movies in the **unpopular movie trimmed test set** and evaluate its performance using **10-fold cross-validation**. The number of latent factors range from 2 to 50 in step sizes of 2 and **the plot of average RMSE** obtained by averaging the RMSE across all 10 folds **over different k value**. And report the **minimum average RMSE**.

Plots and Results:

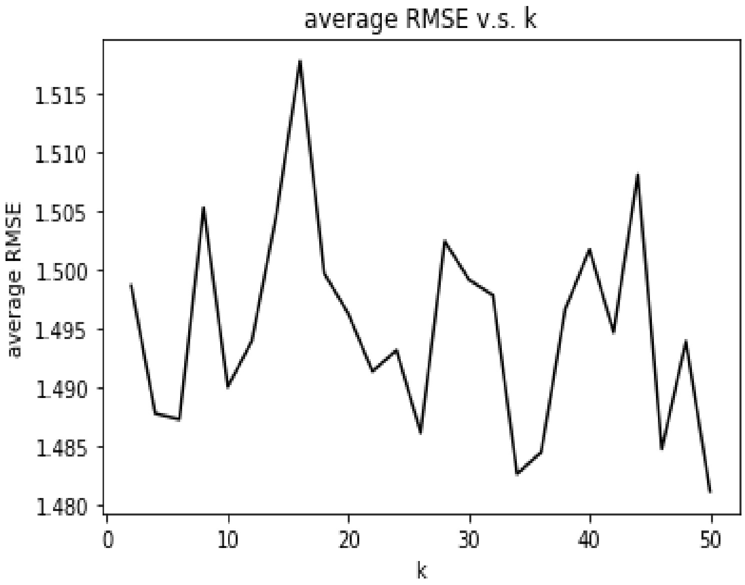


Minimum RMSE is 1.0080

Question (28)

Description: Design a **MF with bias collaborative filter** to predict the ratings of the movies in the **high variance movie trimmed test set** and evaluate its performance using **10-fold cross-validation**. The number of latent factors range from 2 to 50 in step sizes of 2 and **the plot of average RMSE** obtained by averaging the RMSE across all 10 folds **over different k value**. And report the **minimum average RMSE**.

Plots and Results:

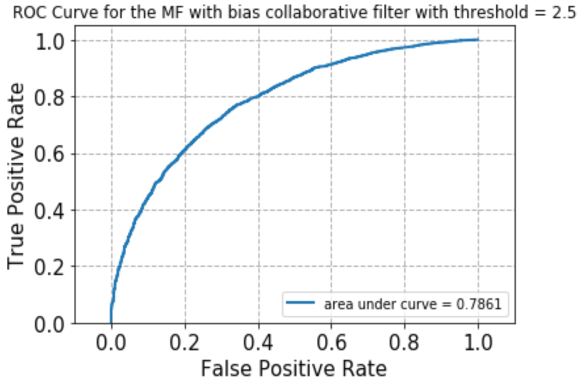


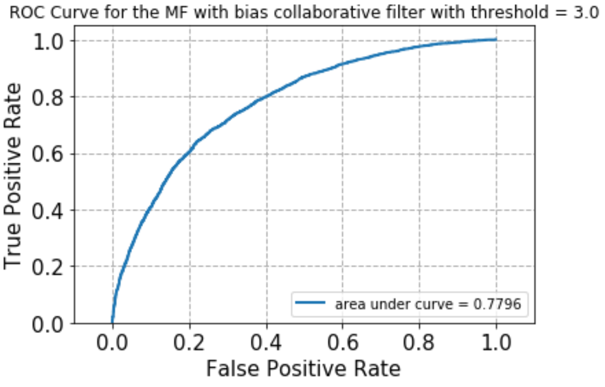
Minimum RMSE is 1.4819

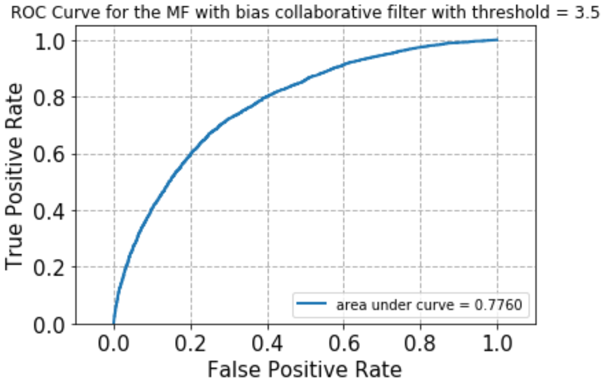
Question (29)

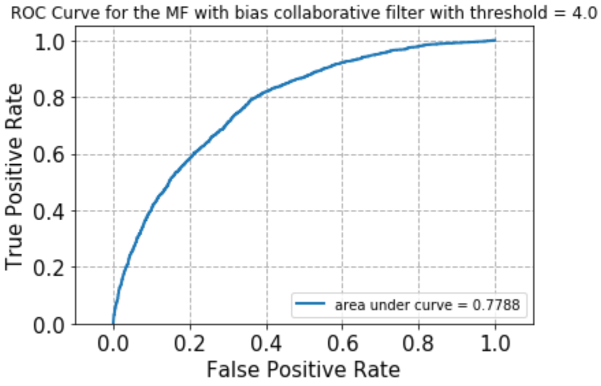
Description: Plot the **ROC curve** for the **MF with bias** collaborative filter designed in question 24. Use different **threshold as 2.5, 3, 3.5 and 4**. For ROC plotting use the **optimal number of latent factors 20**. Also calculate the area under the curve (**AUC**) score.

Plots:






ROC Curve for the MF with bias collaborative filter with threshold = 3.5


ROC Curve for the MF with bias collaborative filter with threshold = 4.0

Question (30-33)

Description: Design a **naive collaborative filter** to predict the ratings of the movies in the Movie Lens dataset and evaluate its performance using **10-fold cross validation**. Compute the **average RMSE** by averaging the RMSE across all 10 folds. Report the average RMSE.

Results:

Since the predicted rating of naïve collaborative filter is given in equation:

$$\hat{r}_{ij} = \mu_i$$

There is no difference whether using 10-fold cross validation or not. Thus, using random split to get ten different test set and compute the RMSE of ground-truth rating and predict rating.

The average RMSE results of different test set is given in Table 1 :

Table 1 Average RMSE results on different test set

Test Set	Average RMSE
All movie	9.554274120864244e-05
Popular Movie	0.00010072640595083424
Unpopular Movie	0.0018475033777616094
High Variance Movie	0.03645813964784499

Discussion:

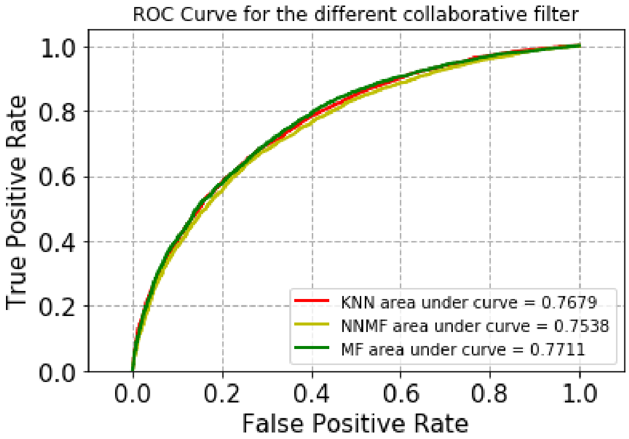
This result intuitively make sense because:

1. Since naïve collaborative filter provides predict results based on mean rating from all ratings, when dealing with all movie in the dataset, the ground truth will have smaller difference with predict results.
2. When dealing with popular movies/unpopular movies/high variance movies, predict results will have higher RMSE since mean of all ratings is lower/higher/lower than ground truth, respectively.

Question (34)

Description: Plot the **ROC curve** (threshold = 3) for the **k-NN**, **NNMF** and **MF with bias** based collaborative filter in the same figure. Use the figure to **compare** the performance of the filters in prediction the ratings of the movies.

Plots:



Discussion:

MF with bias performs the best, while NNMF has the least AUC score among the three. By adding bias term for each user and item, MF with bias model improves the accuracy of predicting ratings. The idea behind such models is that attitudes or preferences of a user can be determined by a small number of hidden factors. Matrix decomposition can be reformulated as an optimization problem with loss function and constraints and the constraints are chosen based on property of the model. As for Non negative matrix decomposition, it requires non negative elements in resultant matrices. The idea of clustering based algorithm (KNN) is same as that of memory-based recommendation systems. In memory-based algorithms, we use the similarities between users and/or items and use them as weights to predict a rating for a user and an item. The difference is that the similarities in this approach are calculated based on an unsupervised learning model, rather than Pearson correlation of cosine similarity. In this approach, we also limit the number of similar users, which makes system more scalable.

Question (35)

Description: Precision and Recall are defined by the mathematical expressions given by equations 12 and 13 respectively. Please explain the meaning of precision and recall in your own words.

Explanation:

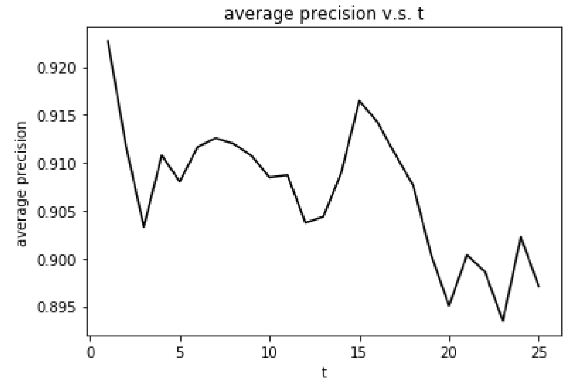
As is given by these two equations below, the precision in this case is the ratio between the number of correct recommendation given by the system and the number of all the recommendation given by the system. And "correct" here means the movie given by the system is truly liked by the user. Similarly, the recall in this case is the ratio between the number of the correct recommendation given by the system and the number of all the items liked by the user, where "correct" has the same meaning as precision.

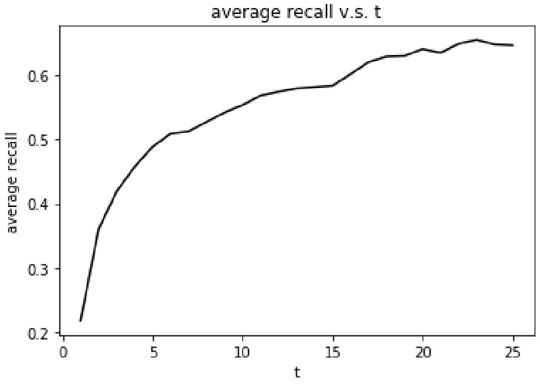
$$Precision(t) = \frac{|S(t) \cap G|}{|S(t)|}$$
$$Recall(t) = \frac{|S(t) \cap G|}{|G|}$$

Question (36)

Description: Plot average precision (Y-axis) against t (X-axis) for the ranking obtained using k-NN collaborative filter predictions. Also, plot the average recall (Y-axis) against t (X-axis) and average precision (Y-axis) against average recall (X-axis). Use the k found in question 11 and sweep t from 1 to 25 in step sizes of 1. For each plot, briefly comment on the shape of the plot.

Plots:





Discussion:

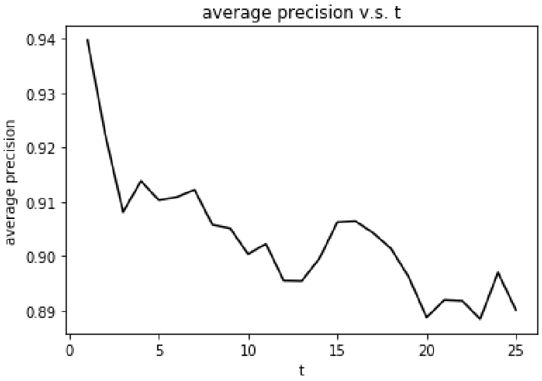
The average precision decreases with an increase in t, showing a decline with fluctuation. It is easy to understand that the probability of dislike movies increases as the recommended range expands with t increases which would lead to the result that the precision decreases.

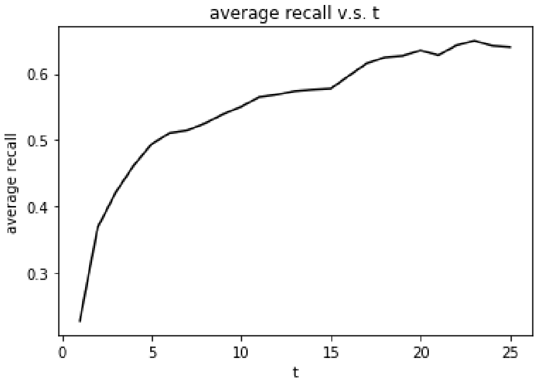
The average recall increases with an increase in t. As explained in Question 35, the recall in this case is the ratio between the number of the correct recommendation given by the system and the number of all the items liked by the user, where "correct" means the movie given by the system is truly liked by the user. Therefore, as t increases, the number of the movies recommended increases which would result in the increase of the probability of the movie liked by the user to be recommended.

Question (37)

Description: Plot average precision (Y-axis) against t (X-axis) for the ranking obtained using NNMF-based collaborative filter predictions. Also, plot the average recall (Y-axis) against t (X-axis) and average precision (Y-axis) against average recall (X-axis). Use optimal number of latent factors found in question 18 and sweep t from 1 to 25 in step sizes of 1. For each plot, briefly comment on the shape of the plot.

Plots:





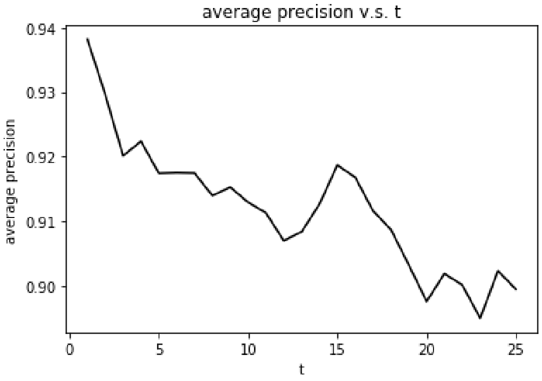
Discussion:

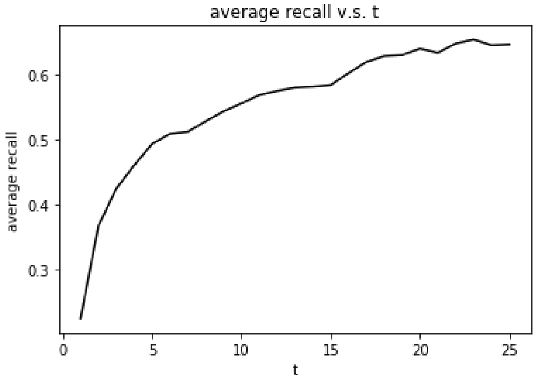
The shape of the plot is similar to that of Question 36 which uses k-NN collaborative filter. The average precision decreases with an increase in t, showing a decline with fluctuation, while the average recall increases with an increase in t. But the fluctuation is more severe than the result of k-NN collaborative filter.

Question (38)

Description: Plot average precision (Y-axis) against t (X-axis) for the ranking obtained using MF with bias-based collaborative filter predictions. Also, plot the average recall (Y-axis) against t (X-axis) and average precision (Y-axis) against average recall (X-axis). Use optimal number of latent factors found in question 25 and sweep t from 1 to 25 in step sizes of 1. For each plot, briefly comment on the shape of the plot.

Plots:





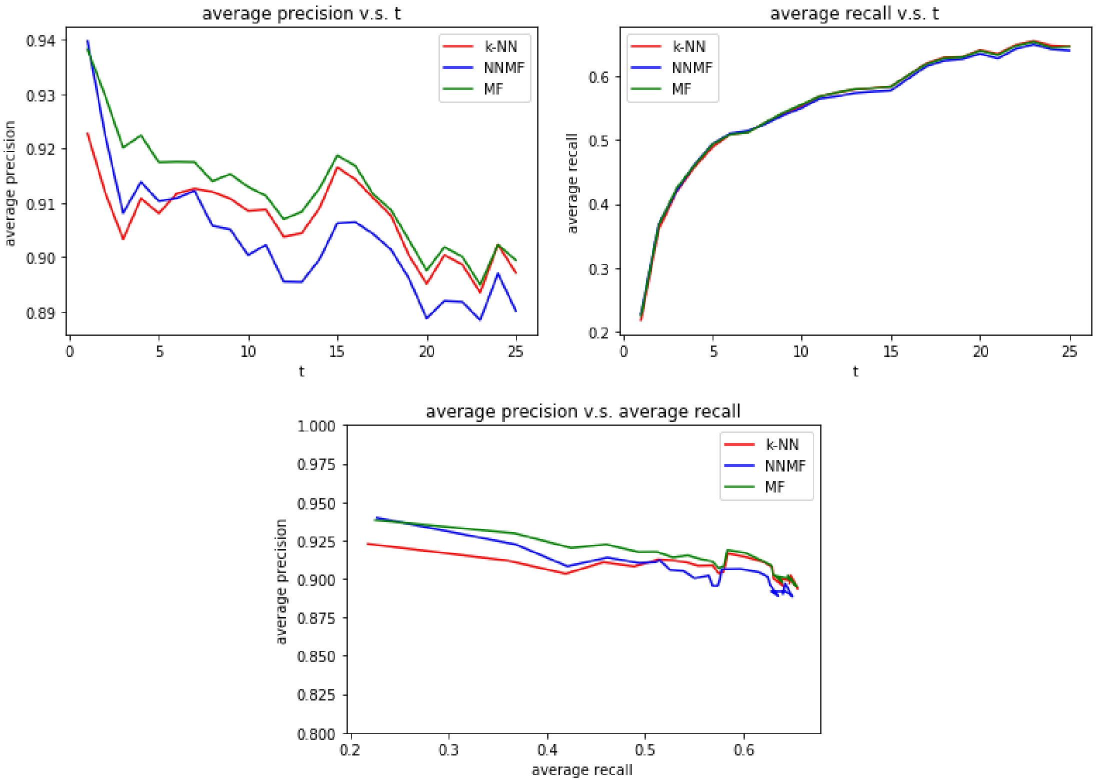
Discussion:

The shape of the plot is similar to that of Question 36 which uses k-NN collaborative filter and Question 37 which uses NNMF-based collaborative filter. The average precision decreases with an increase in t, showing a decline with fluctuation, while the average recall increases with an increase in t.

Question (39)

Description: Plot the precision-recall curve obtained in questions 36,37, and 38 in the same figure. Use this figure to compare the relevance of the recommendation list generated using k-NN, NNMF, and MF with bias predictions.

Plots:



Discussion:

The shapes of the three plots similar. The average precision decreases with an increase in t, showing a decline with fluctuation, while the average recall increases with an increase in t. And it is obvious from the plot of the average precision that the result of MF filter is the best whose fluctuation is slighter and the precision is relatively high. And the three lines of the average recall v.s. t are almost the same. According to the plot average precision v.s. average recall, we could learn that the average precision would decrease as the average recall increases, which is consistent with the definitions of the two features. And the MF performs the best with the slightest fluctuation and the highest value.