Collaborative Filtering Assignment 1

Name: Shaney Waris Roll no.: 2018308

My User-Based Approach & Assumptions:

- Pivot the training dataframe such that the User ID are in the rows and the Item IDs are in the columns.
- Replaced NA values (No ratings are available) with 0
- Find the cosine similarities between all the users and saved in a dataframe.
- Now in my users similarity dataframe, the similarities between the same user is 1. And I am not supposed to consider these similarities while predicting the rating. Hence, I replaced the similarity between the same user from 1 to 0 so that I will ignore these values in my upcoming steps.
- Now during the rating prediction (Given userId and itemId),
    - It might be possible that this itemId is a completely new movie in my testing dataset which was not present in training dataset. So I have ignored this case. This is my [**Assumption**].
    - Fetched the cosine similarities of user userId with all other users.
    - Fetched the ratings of all the users for item itemId.
    - Ignored all the cases when user similarity is lesser than the given 'tau' value.
    - Ignored all those cases when no user gave the rating to that movie.
    - After the above two cases, if the number of similar users becomes 0. Then this is a coverage problem. And in this case, I have taken the average rating of that movie as my predicted rating. This is my [**Assumption**].
    - After the above two cases, if the number of similar users are greater than 0. Then I just calculate the weighted ratings by multiplying the ratings with their corresponding similarities. And Normalized it by the sum of all the similarities.

Table 1. MAE values User-based (5 Marks)

Fold #	**τ = 0.4**	**τ = 0.5**	**τ = 0.6**	**τ = 0.7**
1	0.8475006447582307	0.836024471030855	0.8256316310412225	0.8261415793915304
2	0.8478761926090771	0.8345119280063997	0.8185844766963448	0.8183431468757628
3	0.8397190856313358	0.8266754201144628	0.8095313805227538	0.8103010226084848
4	0.8203078252670899	0.8105526572396692	0.7918141995985887	0.7928346899239228
5	0.8186155541110084	0.8096246972385351	0.7953889430830614	0.7944856462231144
Average	0.8348038604753484	0.8234778347259842	0.8081901261883943	0.808421217004563

My Item-Based Approach & Assumptions:

- Pivot the training dataframe such that the Item ID are in the rows and the User IDs are in the columns. (Just took the transpose of the same pivot table from the user-based.)
- Replaced NA values (No ratings are available) with 0
- Find the cosine similarities between all the items and saved in a dataframe.
- Now in my items similarity dataframe, the similarity between the same item is 1. And I am not supposed to consider these similarities while predicting the rating. Hence, I replaced the similarity between the same item from 1 to 0 so that I will ignore these values in my upcoming steps.
- Now during the rating prediction (Given userId and itemId),
    - It might be possible that this itemId is a completely new movie in my testing dataset which was not present in training dataset. So I have ignored this case. This is my [**Assumption**].
    - Fetched the cosine similarities of item itemId with all other items.
    - Fetched the ratings of all the items for user userId.
    - Ignored all the cases when item similarity is not in the top K values of item similarity.
    - Ignored all the cases when the rating to any item in the above fetched item ratings is not available.
    - After the above two cases, if the number of similar items becomes 0. Then this is a coverage problem. And in this case, I have taken the average rating of that movie as my predicted rating. This is my [**Assumption**].
    - After the above two cases, if the number of similar items are greater than 0. Then I just normalized all the similarities and then linearly interpolate it with their corresponding ratings.

Table 2. MAE values Item-based (5 Marks)

Fold #	K = 10	K = 20	K = 30	K = 40
1	0.8065361949096769	0.7884116848570262	0.7831470866249407	0.7791024069436101
2	0.7939964674859856	0.7699057629281751	0.7618879943222987	0.7614343040790187
3	0.7797270876387902	0.7681496007174117	0.7619395582626856	0.7581520186952564
4	0.7696899678196355	0.760143479878848	0.7578787448650866	0.757736183303115
5	0.7666360317959755	0.7529098809777055	0.7482164840226724	0.7494023824822783
Average	0.7833171499300129	0.7679040818718332	0.7626139736195369	0.7611654591006556

********** **Assignment 1 Ended** **********