Systems for Data Science

CS-449

Movie Recommendation with k-NN - Milestone 1

Author:
Kamil Seghrouchni

Professors:
Erick Lavoie
Athanasios Xygkis

Spring Semester 2021

Contents

3 Prediction based on global average deviation

3.1 Questions

3.1.1 Reporting global average $\bar{r}_{\bullet,\bullet}$:

The global average was found to be equal to :

$$\bar{r}_{\bullet,\bullet} = 3.5299$$

Ratings are therefore on average higher than the middle of the rating scale by approximately 0.53 (rounded two decimals).

3.1.2 Reporting average users rating $\bar{r}_{u,\bullet}$:

Table 1: user average rating statistics

	min	max	average
UsersAverageRating	1.491954	4.869565	3.588191

The average rating for each user was computed and one has found that all the users do not rate on average close to the global average. This result was obtained by assessing whether or not for all users, the difference between the minimum of users average rating and the global average as well as the difference between the maximum of users average rating and the global average are lesser than a given threshold of 0.5. The ratio of users with average rating close to global average was computed and one has founded it to be equal to :

$$ratio_u = 0.7466$$

3.1.3 Reporting average item rating $\bar{r}_{i,\bullet}$:

Table 2: item average rating statistics

	min	max	average
ItemsAverageRating	1.0	5.0	3.076045

The average rating for each item was computed and one has found that all the items do not rate on average close to the global average. This result was obtained by assessing whether or not for all items, the difference between the minimum of items average rating and the global average as well as the difference between the maximum of items average rating and the global average are lesser than a given threshold of 0.5. The ratio of items with average rating close to global average was computed and one has founded it to be equal to :

$$ratio_i = 0.4899$$

3.1.4 Comparing prediction accuracy :

Table 3: Prediction errors

Predictions	Mean absolute errors
$\bar{r}_{\bullet,\bullet}$	0.968049
$\bar{r}_{u,\bullet}$	0.850191
$\bar{r}_{i,\bullet}$	0.827568
$p_{u,i}$	0.768055

It can be observed that the best prediction results are achieved with the baseline method, followed by the item average than the user average. It also observed that using the global average gives the worst prediction accuracy. As expected, the baseline predictions are significantly better as this technique is able to capture both user bias (in the form average user) and user "item" bias (in the form of a global average item deviation) making it all the more specific for a given user, item pair. Also expected, the global average gives the worst results as it does not capture any specificity for a user and item pair. Then the item average gives significantly better results than the user average. This can be explained by the fact that there are less ratings per items than ratings per user, respectively on average 59.45 versus 106.044 ratings. Therefore, item average can be considered as more specific than user average. This added to the fact that there are at least 20 ratings per user when there is only one rating at least per item.

3.1.5 Comparing computing prediction times:

Table 4: Computing times in micro seconds

	min(ms)	max(ms)	average(ms)	stddev(ms)
DurationInMicrosecForGlobalMethod	55237.447	151261.975	8.021010e+04	27888.381493
DurationInMicrosecForPerItemMethod	267706.293	927580.914	3.807835e+05	190350.431321
DurationInMicrosecForPerUserMethod	478819.894	902244.572	5.738828e+05	130482.172800
DurationInMicrosecForBaselineMethod	1175765.907	1252437.931	1.206466e+06	25642.552669

Technical specifications :

- Machine : MacBook Pro (13-inch, 2018, Four Thunderbolt 3 Ports)
- Processor : 2.3 GHz Intel Core i5 four cores

The Baseline method appears to be the more time consuming. In order to have better idea of the time consumption, computing the ratio between the computing time for the baseline model and the global average model gives out :

$$ratio_{computing} = 15.0413$$

Meaning that the baseline prediction takes approximately 15 more time to be computed.

4 Recommendation

4.1 Questions

4.1.1 Top 5 recommendations

Table 5: Top 5 recommendations

item number	movie title	rate
814	Great Day in Harlem	5.0
1122	They Made Me a Criminal (1939)	5.0
1189	Prefontaine (1997)	5.0
1201	Marlene Dietrich: Shadow and Light (1996)	5.0
1293	Star Kid (1997)	5.0

To be honest, i have never heard of any of the movies suggested and therefore have no particular opinion about them. After a quick look up on the internet, i think both Star kid and Prefontaine are movies i might actually really like and want to see in the future. For the other ones, i will have to watch to be sure of my opinion.

4.1.2 Bonus

According to the "Recommender Systems" Text Book by Charu C. Aggrwal, page 84, equation 3.8 it is possible to smooth the prediction using a "Laplacian smoothing" technique with a parameter α. In essence, it implies adding to the numerator a factor α and in the denominator $\alpha * l$ where l would be the total number of user.Yet, the formula is not applicable directly to our case and some thoughts had to be given on parameter α as well as to what formula should this method be applied to. Therefore, the following smoothed global average deviation is proposed :

$$\bar{\hat{r}}_{\bullet} i = \frac{\sum_{u \in U_i} \hat{r}_{u,i} + \alpha}{U_i + l * \alpha}$$

With l = 943 and alpha = 0.0013, meaning adding to the denominator approximately $\beta = \alpha * l =$ 1.223. Analyzing this formula, it can be seen that for that for popular movies with large number of users, β will have very little influence. But intuitively, for less popular movies, with a number of user having rated of the order of 10, because of β, the denominator is significantly increased, leading to a smaller global average deviation and decreased predictions.Where as for more popular movies, the added factor will have little to no influence compared to the large value of U_i. The small magnitude of α has on the other hand very little influence on the numerator. Thus, after a small grid search the proposed α yields to a new mean absolute error of 0.76753 which improves the baseline model.

Table 6: Top 5 recommendations after smoothing

item index	movie title	rate
1189	Prefontaine (1997)	4.679291
1293	Star Kid (1997)	4.679291
1449	Pather Panchali (1955)	4.652425
318	Schindler's List (1993)	4.630964
408	Close Shave	4.627302

Interestingly, the two movies i was really interested in turned out to be the top predictions when smoothing.