CS449 - Project Milestone 1

Saoud Akram - SCIPER : 273661

Remark : In this report, the term "movie" and "item" are used interchangeably.

Part 3.1

1. Do ratings, on average coincide with the middle of the rating scale ? If not, are they higher or lower on average ? By how much ?

The global average we obtained using the provided data is approximately 3.53, which is slightly higher from the middle of the rating scale (by about 0.53), but not by a substantial amount.

2. Do all users rate, on average, close to the global average? Do most users rate, on average, close to the global average?

No, all users do not rate, on average, close to the global average. Some users have an average rating that can go as low as 1.5 or as high as 4.8. However, this concerns only a handful of users as we have found that around 75% of users tend to rate, on average, close to the global average (i.e. their average rating score does not deviate by more than 0.5 to the global average for those users). So yes, most users do rate, on average, close to the global average.

3. Are all items rated, on average, close to the global average? Are most items rated, on average, close to the global average?

No, all items are not rated close to average. Some movies have an average rating of 1.0 (i.e. they are universally and highly disliked) and others have an average rating of 5.0 (i.e. they are universally praised and highly liked).

As for the ratio of movies that have average ratings that are close to the global average, we estimate it to be around 49%, which means that most movies, are, in fact, not rated, on average, close to the global average.

Remark : When assuming that a movie is universally liked (or disliked), the word "universal" should be understood in the context of the set of users that are in the dataset.

4. Compare the prediction accuracy (of the previous methods to the proposed baseline. Report the results you obtained in a table. Discuss the difference(s) you observed and why you think they occur.

The results we obtained are the following:

Global Average MAE	User Average MAE	Item Average MAE	Baseline MAE
0.968	0.850	0.827	0.779

First, we can observe that the accuracy obtained using the global average method is the worst. As mentioned earlier, the global average is close to the middle of the rating scale. As such, this method predicts the same rating for every movie (i.e. $p_{u,i} = \overline{r}_{\bullet,\bullet}$ for all $u \in U$ and for all $i \in I$). As such, this method does not take into account two facts :

- All users do not vote close to the global average. We have seen that around 25% of them do not do so.
- Most movies do not have a score close to the global average (which is the case for around 51% of movies).

The user average method has a better accuracy, as it takes into account a given user's average rating (i.e. for two users u_1 and u_2, it not always the case that $p_{u_1,i} = p_{u_2,i}$). As such, contrary to the global average method, it takes advantage of the fact that 25% of users do not vote close to the global average to make predictions. As such for those 25% of users, the predictions were, on average, much better than using the global average method. For the remaining 75% of users, predictions were, on average, either slightly better than using the global average method or equally as good (since $\overline{r}_{u,\bullet} - \overline{r}_{\bullet,\bullet} \leq 0.5$ for these 75% of users). However, this method does not take into account that some movies have a higher score than others (i.e. using this method, it is always the case that for a user u, $p_{u,i_1} = p_{u,i_2}$ for all $i_1 \neq i_2$).

Similarly, the item average method takes into account a given movie's average rating (i.e. for two different movies i_1 and i_2, it is not always the case that $p_{u,i_1} = p_{u,i_2}$). As such, contrary to both the global average method, it takes advantage of the fact that most movies' average rating does not sit close to the global average rating. However, this method does not take into account the fact that different users do not rate the same way (i.e. using this method, it is always the case that for a given movie i, $p_{u_1,i} = p_{u_2,i}$ for all $u_1 \neq u_2$).

This begs the question : Why does the item average method have a better accuracy than the user average method? The information that is used by the item average method (i.e. the fact that most movies are not rated close to the global average) is more valuable than the information used by the user average rating (i.e. the fact that users do not necessarily rate close to the global average). This is due to the fact that only 25% of users do not rate, on average, close to the global average while 51% of movies are not rated, on average, close to the global average.

Finally, the baseline method is the one that performs best as it takes advantage of both facts (i.e. a movie's average rating as well as a user's average rating). On top of that, this method also uses how a user's average rating deviates from the rating itself, in order to get more accurate results (these deviations are normalized in order to make sure that the resulting prediction is between 1.0 and 5.0).

5. Measure the time required for computing 5 predictions for all ratings in the test set with all four methods.

Technical specifications:

- **Model :** Surface Pro 4
- **CPU Type and Speed :** Intel(R) Core(TM) i5-6300U CPU @ 2.40Ghz - 2.50Ghz
- **RAM :** 8.00 GB
- **OS :** Windows 10
- **System type :** x64 based system (64 bit Operating System)
- **Scala Version :** 2.12.13

Here are the results we got:

	Global Method	User Method	Item Method	Baseline Method
min	0.155	0.9691	0.79	3.8171
max	0.2215	1.2068	0.9687	4.3071
mean	0.1763	1.1018	0.8467	4.0134
std	0.0213	0.0665	0.0213	0.1305

(Please note that the timings were converted into seconds and rounded to the 4th decimal, as it allows for better readability. For more precise answers, please consult the appropriate .json file which contains the exact timings in μs)

We can observe that the Baseline Method is by far the most expensive (time-wise) to compute. Compared to the Global Average method, the Baseline Method is about 22x slower (i.e. the ratio between the Baseline Method and the Global Average is approximately 22). As such, the Baseline Method is almost 4 seconds more expensive than the Global Average.

This is very understandable, as there are very few transformations and actions required to obtain the global average. The user method and the item method take relatively the same time to complete, but are more expensive than the global average by a little less than 1 second. The fact that the Baseline Method takes more time than the User Method also makes sense since the Baseline method also computes the user's average rating (on top of all the rest) before computing the predictions.

Part 4.1

1. Report your personal top 5 recommendations using the baseline predictor, including the movie id, the movie title, and the prediction score. Are these movies you have actually liked (but did not rate) or would like to see in the future?

Movie id	Title	Prediction score
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

I have seen none of these movies. I have heard about the last two, but haven't watched them yet. *"Star Kid (1997)"* is in my "to watch" list.

2. (Bonus) How could you modify the predictions to favour more popular movies, e.g. by smoothly decreasing the prediction score of movies with few ratings while keeping the prediction score of those with many ratings almost identical?

In order to give more weight to movies with a higher number of ratings, we used IMDB's weighted rating formula which is the following :

$$w_i = \frac{v_i}{v_i + k} R_i + \frac{k}{v_i + k} C$$

where:

w_i is the weighted rating for a given movie i

v_i is the number of ratings the movie i got

k is the minimum of votes required for a movie to be deemed "worthy" (we set $k := 15$ through trial and error, as it yielded the best predictions)

R_i is the average rating of the movie

C is the mean across the whole dataset (a.k.a. the global average rating)

The idea of this formula is that, if there are very few votes for a given movie i, it's rating is decreased more than a movie with a high number of ratings. This favours movies with a high number of ratings while simultaneously disfavouring movies with few ratings

Since IMDB rates its movies with a score that is in the range $[1.0, 10.0]$, we had to scale our ratings to this interval. To do this, we defined the following :

$\text{shift\_range}_{a,b,c,d}(x) = c + \frac{d-c}{b-a}(x - a)$ where the number x is mapped from an interval $[a, b]$ to an interval $[c, d]$. This allowed us to compute the following :

$$R_i = \text{shift\_range}_{-1,1,1,10}(\overline{\hat{r}}_{\bullet,i})$$

$$C = \text{shift\_range}_{1,5,1,10}(\overline{r}_{\bullet,\bullet})$$

Remark 1: Why use the global average deviation for R_i and not the actual movie average rating? We determined through trial and error that this leads to better predictions

Remark 2: Obviously, we also used shift_range on the result to map the weighted rating back to $[-1, 1]$: $w_i := \text{shift\_range}_{1,10,-1,1}(w_i)$

Remark 3: To compute $p_{944,i}$ we simply used :

$$p_{944,i} = \overline{r}_{944,\bullet} + w_i \cdot scale((\overline{r}_{944,\bullet} + w_i), \overline{r}_{944,\bullet})$$

The results we got are the following :

Movie id	Title	Prediction Score
318	Schindler's List (1993)	4.497
483	Casablanca (1942)	4.47
408	Close Shave	4.458
169	Wrong Trousers	4.443
12	Usual Suspects	4.412

I have already watched (and liked) all these movies, except for "Close Shave" and "Wrong Trousers". I've heard about these two but I haven't watched them nor am I planning to.