Milestone I

19 March 2021

CS-449

Systems for Data Science

Submitted By:
Vignoud Julien

Sciper
282 142

3. Baseline: Prediction based on Global Average Deviation

1. The global average rating is approximately 3.53. It is therefore higher by 0.53 than the middle of the rating scale.

2. The minimum average rating of a user is around 1.492 while the maximum is around 4.87. On average, users have an average rating of 3.588. Considering that a difference smaller than 0.5 is small, we can say that not all users are close to the global average. In fact 74.7% of users are close to the global average.

3. The minimum average rating of an item is 1 and the maximum is 5. On average, an item has an average rating of 3.08. Again assuming that a difference smaller than 0.5 is small, we can say that not all the items rated are close to the global average. Indeed, only 48.98% of the items are close to the global average.

4. The different MAE are reported in Table 4. The worse (greatest) MAE is the global average method, followed by the average per user method. It is coherent that the latter method has a higher accuracy since the average per user method takes more information of the user into account. The average per item method has a smaller MAE than the average per user method. Indeed it is more likely that all users have a similar rating for an item, rather than a user that have similar ratings for all of its items. In other words, it is more useful to know what other users thought of the item than to know how the user rated other of its items. Finally the baseline method has the best MAE, coherent with the fact that it uses both the average rating of the user and the average deviation of the item.

Global	Per user	Per item	Baseline
0.968	0.850	0.828	0.768

Table 1: MAE of the different methods

5. Here are my specs:

   - MacBook Pro (Retina, 13-inch, Early 2015)
   - 2.7 GHz Dual-Core Intel Core i5
   - 8 GB of RAM
   - MacOS Catalina
   - Scala version 2.12.13

The most expensive method is the baseline method, that takes 2.8 seconds. It is 2.5 seconds more than the global average method. The ration between the two is 12.75.

	Global	Per user	Per item	Baseline
Min	159554	87277	75332	1745566
Max	348043	211687	109589	6385189
Average	216948	128151	86574	2766171
Stddev	59288	36587	9566	1383905

Table 2: Time in microseconds for the different methods

4. Recommendation

1. My top 5 recommendations (in order) are the following:
   (a) (814, "Great Day in Harlem", 5.0)
   (b) (1122, "They Made Me a Criminal (1939)", 5.0)
   (c) (1189, "Prefontaine (1997)", 5.0)
   (d) (1201, "Marlene Dietrich: Shadow and Light (1996)", 5.0)
   (e) (1293, "Star Kid (1997)", 5.0)

   I don't know any of these movies but after looking up on internet, I would indeed be interesting in watching *Great Day in Harlem* and *Marlene Dietrich*.

2. As an improvement, to take into account the movies popularity I came up with this equation for the prediction:

$$p_{u,i} = \overline{r}_{u,\bullet} + \overline{\hat{r}}_{\bullet,i} * scale((\overline{r}_{u,\bullet} + \overline{\hat{r}}_{\bullet,i}), \overline{r}_{u,\bullet}) + popularity(i)$$

$$popularity(i) = -\frac{\log(\max_j(x_j)) - \log(x_i)}{\log(\max_j(x_j)} \qquad x_i = \sum_{(u,i)} 1$$

   In words, we penalize up to 1 the most obscure movies while we don't change the prediction of most popular movies. I decided to penalize instead of augmenting the ratings of popular movies due to the number of 5-star rating in the baseline predictions. The popularity function ranges from 0 to 1, 0 for the movie that has the most ratings and 1 for movies with only one rating. The scale is computed using the logarithm of the number of ratings due to the exponential distribution of ratings: most ratings are given to a few movies, while most movies have few ratings (e.g. one third of the movies have less than 10 ratings, 80% have less than 100 ratings while the maximum number of rating is 584)

Here is my new top 10 recommendations:

(a) (318, "Schindler's List (1993)", 4.352801588279078)

(b) (64, "Shawshank Redemption", 4.297062520404296)

(c) (483, "Casablanca (1942)", 4.294749165980445)

(d) (12, "Usual Suspects", 4.236552716938676)

(e) (127, "Godfather", 4.209743188755581)

(f) (408, "Close Shave", 4.201139888317453)

(g) (169, "Wrong Trousers", 4.186034489133709)

(h) (1189, "Prefontaine (1997)", 4.1724692882400705)

(i) (1293, "Star Kid (1997)", 4.1724692882400705)

(j) (603, "Rear Window (1954)", 4.172301939221159)

I already saw Schindler's List and (painfully) loved it. Moreover Shawshank Redemption was already on my watch list so I consider these new recommendations better than the previous ones. We can note that the predicted ratings are way lower than the previous perfect ratings yet we can still see the previously recommended movies in the lower ranks with Prefontaine and Star Kid in 8th and 9th positions. There is a trade-off between popular movies and personalized movies and that can be be found using a variable $\alpha * popularity(i)$ with $\alpha \in [0, 1]$.