Milestone II

20 April 2021

CS-449

Systems for Data Science

Submitted By:
Vignoud Julien

Sciper
282 142

2. Similarity-based Predictions

1. The prediction MAE using adjusted cosine similarity is 0.7478. Given a baseline MAE of 0.7669, the difference is -0.0191. Thus, the prediction accuracy is better using the cosine similarity.

2. $$Jaccard\_similarity(u,v) = \frac{|I(u) \cap I(v)|}{|I(u) \cup I(v)|} = \frac{|I(u) \cap I(v)|}{|I(u)| + |I(v)| - |I(u) \cap I(v)|}$$

   The prediction MAE using the Jaccard similarity is 0.7623. The difference with the prediction using the cosine similarity is 0.0145. Thus, using the Jaccard similarity yields a worse prediction accuracy than the cosine similarity.

3. The number of possible similarities is the number of unordered pair of users. Indeed we don't need to compute both similarities for (u,v) and (v, u) since they are equal. Moreover we don't need to compute similarity for (u,u) as we know that it is 1. That is equal to $\frac{U(U-1)}{2}$ with U the number of users. That is 444'153 possible similarities for the 'ml-100k' dataset.

4. The minimum number of multiplication is 1, the maximum is 332, the mean is 13.0889 and the standard deviation is 18.1720.

5. Given that each value is stored as a 64-bit double, each similarity needs $64/8 = 8$ bytes of storage. The total number of bytes needed is therefore $8 * \frac{U(U-1)}{2}$. The number bytes required to store all the non-zero similarities is 3'553'224.

6. The minimum time required to compute the predictions is 60342385 microseconds, the maximum is 64062768 μs, the average is 62591166 μs and the standard deviation is 1093576 μs. The average is higher than the previous methods due to the similarity computations. Indeed, we are using the same prediction method overall but using similarity measures on top of it, therefore it is coherent that the average run time is higher due to the overhead of computing and using the similarities.

7. The minimum time required to compute the similarities is 18020584 μs, the maximum is 18593943 μs, the average is 18195402 μs and the standard deviation is 153754 μs. The average time spent per $s_{u.v}$ is 40.9665 μs. On average, the ratio between the computation of similarities and the total time required to make predictions is 0.2907. The time to compute the similarities accounts for 30% of the total time needed to make predictions, it is therefore significant.

1 3. Neighbourhood-Based Predictions

1. The prediction accuracy for different values of k is reported in the Table 2. We can see that for low values of k the MAE is really high and decreases when k increases. At this point the model overfits the training data and doesn't generalize well to the testing set. Increasing k improves the performance by reducing the overfitting. The first k for which the MAE is lower than the baseline MAE is 100, with MAE equal to 0.7561. The difference between the baseline and the knn MAE's is -0.0108. The lowest MAE is reached at k = 300, with value 0.7469. After that the MAE increases, it it now underfitting because the number of neighbours is too high.

k	MAE
10	0.8407036862423933
30	0.7914221792247494
50	0.7749407796360591
100	0.7561353222065881
200	0.7484528977469224
300	0.7469140388149909
400	0.7471389103638708
800	0.7475383223779415
943	0.7477281438504022

Table 1: Prediction MAE for different number of nearest neighbours.

2. The number of bytes need to store the similarities as a function of U and k is $8 * k * U$.

k	Number of bytes
10	75440
30	226320
50	377200
100	754400
200	1508800
300	2263200
400	3017600
800	6035200
943	7113992

Table 2: Number of bytes required to store the similarities w.r.t. different number of nearest neighbours.

3. Given a RAM of 8GB, the maximum number of users that we can fit in memory is $\left\lfloor \frac{RAM}{3*8*k} \right\rfloor = \left\lfloor \frac{8.10^9}{3*8*100} \right\rfloor = 3333333$.
4. For any k, we have to compute all the similarities in order to find the top k similarities.
5. Here are the top 5 recommendations for k = 30:
   (a) (8, "Babe (1995)", 5.0)
   (b) (86, "Remains of the Day", 5.0)
   (c) (133, "Gone with the Wind (1939)", 5.0)
   (d) (165, "Jean de Florette (1986)", 5.0)
   (e) (166, "Manon of the Spring (Manon des sources) (1986)", 5.0)

   The top 5 recommendations for k = 300 are:
   (a) (850, "Perfect Candidate", 5.0)
   (b) (884, "Year of the Horse (1997)", 5.0)
   (c) (889, "Tango Lesson", 5.0)
   (d) (1144, "Quiet Room", 5.0)
   (e) (1189, "Prefontaine (1997)", 5.0)

   We can see that the recommendations are completely different for different values of k. The recommendations are also different from the baseline, except for Prefontaine that has been recommended too. The baseline recommended top rated movies while I have never heard of any movies in the knn recommendations.