Dorina Ismaili & Deniz Gaiserdenizgaiser42@gmail.com
04.10.2024

Mining Massive Datasets
Homework 2

Exercise 1

Consider a web shop that sells furniture and uses a recommendation system. When a new user creates an account and likes one product, he will be presented with similar products on his next visit.
How can a competitor - in principle - try to steal the valuable data for recommendation from this website? Does this work better when the web shop implemented a content-based or a collaborative filtering system? What data would the competitor be able to infer? Would this technique have an impact on the recommendation system, i.e., would this attack create a bias on the data? Why is this attack probably not viable in any case?

Solution

- A competitor could attempt to reverse-engineer the recommendation system by simulating user behavior on the website. They could follow pattern analysis by analyzing the results of multiple interactions, the competitor could infer the similarity matrix. This could reveal groups of products that are frequently recommended together or associated with similar attributes. If the competitor interacts with different categories, they could identify category-level recommendation strategies. Additionally, the competitor has the potential to gather extensive data on product attributes from the website, including descriptions, styles, prices, and categories, which they could then utilize to create their own recommendation model that replicates the original.
- This method is more effective in content-based systems, which are closely tied to product features, than against collaborative filtering systems, which rely on aggregated user behavior and would need a substantial amount of data for accurate simulation. The competitor could infer product association, which items are frequently recommended together; attribute similarities; behavior patterns: general patterns in recommendations following user action.
- This technique could create a bias in the recommendation system by artificially inflating the popularity or connection of specific items or categories. This attack is not viable as it could cause legal and ethical risks, as manipulating datas from a recommendation system causes expose to legal action. Additionally it is not feasible because, for an attack to induce a meaningful bias, a competitor would have to replicate or manage a considerable number of accounts, each demonstrating substantial activity, which would be both time-consuming and expensive.

Exercise 2

The following table shows a utility matrix with ratings on a 1-5 star scale of eight items, a through h, by three users A, B, and C.

-	a	b	c	d	e	f	g	h
A	4	5		5	1		3	2
B		3	4	3	1	2	1	
C	2		1	3		4	5	3

a) Treating each blank entry in the utility matrix as 0, compute cosine distance between each pair of users.

b) Treating ratings of 3, 4, and 5 as 1 and 1, 2, and blank as 0, compute the cosine distance between each pair of users. Compare these results to those obtained in Part a).

c) Normalize the matrix by subtracting from each non-blank entry the average value for its user. Then compute the cosine distance between each pair of users (blank entries are treated as 0).

d) Compute the Pearson correlation coefficient between each pair of users as defined in Lecture 1 (slide From Cosine to Pearson). Compare these results to those obtained in Part c) and state your conclusions.

Solution

Cosine Similarity $=\frac{\vec{u}\cdot\vec{v}}{||\vec{u}||\cdot||\vec{v}||}$
User A: 4,5,0,5,1,0,3,2
User B: 0,3,4,3,1,2,1,0
User C: 2,0,1,3,0,4,5,3

a) **User A and B**

1. Dot product: 4*0+5*3+0*4+5*3+1*1+0*2+3*1+2*0 = 34
2. $||A|| = \sqrt{80} = 8.9$
3. $||B|| = \sqrt{40} = 6.3$
4. Cosine Similarity $= \frac{34}{\sqrt{80}\sqrt{40}} = 0.6$
5. Cosine Distance = 1 - Cosine Similarity = 0.4

User B and C
1. Dot product: $B \cdot C = 26$
2. $||B|| = \sqrt{40} = 6.3$
3. $||C|| = \sqrt{64} = 8$
4. Cosine similarity = 26/50.4 = 0.51
5. Cosine distance = 1 -0.51 = 0.49

User A and C
1. Dot product: $A \cdot C = 44$
2. $||A|| = \sqrt{80} = 8.9$
3. $||C|| = \sqrt{64} = 8$
4. Cosine similarity = 44/71.2 = 0.61
5. Cosine distance = 1 -0.61 = 0.39

b) User A:1,1,0,1,1,0,1,2
User B:0,1,1,1,1,2,1,0
User C:2,0,1,1,0,1,1,1

User A and B
1. Dot product: 4
2. $||A|| = 3$
3. $||B|| = 3$
4. Cosine Similarity $= \frac{4}{9} = 0.44$
5. Cosine Distance = 1 - Cosine Similarity = 0.55

User A and C
1. Dot product: 6
2. $||A|| = 3$
3. $||C|| = 3$
4. Cosine Similarity $= \frac{6}{9} = 0.66$
5. Cosine Distance = 1 - Cosine Similarity = 0.33

User B and C
1. Dot product: 5
2. $||B|| = 3$
3. $||C|| = 3$
4. Cosine Similarity $= \frac{5}{9} = 0.55$
5. Cosine Distance = 1 - Cosine Similarity = 0.44

The transformation increases the difference for A,B meaning that they are not less similar. While for B and C, and A and C it decreases meaning that they are now more similar.

c) User A:0.6, 1.7 ,0 ,1.7 ,-2.3 ,0 ,-0.3 ,-1.3
User B:0 ,0.5, 1.5, 0.5 ,-1.5 ,-0.5 ,-1.5 ,0
User C: -1, 0 ,-2 ,0 ,0 ,1 ,2 ,0

User A and B
1. Dot product: 0.85 + 0.85 + 3.45 + 0.45 = 5.6
2. $||A|| = 3.63$
3. $||B|| = 2.7$
4. Cosine Similarity $= \frac{5.6}{9.8} = 0.57$
5. Cosine Distance = 1 - Cosine Similarity = 0.42

User A and C
1. Dot product: -0.6 - 0.6 = -1.2
2. $||A|| = 3.63$
3. $||C|| = 3.24$
4. Cosine Similarity $= \frac{-1.2}{11.76} = -0.10$
5. Cosine Distance = 1 - Cosine Similarity = 1.11

User B and C
1. Dot product: -3 -0.5 -3 = -6.5
2. $||B|| = 2.7$
3. $||C|| = 3.24$
4. Cosine Similarity $= \frac{-6.5}{8.7} = -0.74$
5. Cosine Distance = 1 - Cosine Similarity = 1.74

The normalization step increases the distance between two users. This reflects a reduction in similarity.

d) Compute the Pearson correlation coefficient.
User A: 0.6, 1.7, 0, 1.7,-2.3,0,-0.3,-1.3
User B: 0, 0.5, 1.5, 0.5, -1.5, -0.5, -1.5, 0
User C: -1, 0, -2, 0, 0, 1, 2, 0

User A and B
1. Dot product: 0.85 + 0.85 + +3.45 +0.45 = 5.6
2. $||A|| = 3.63$
3. $||B|| = 2.7$
4. Pearson Correlation $= \frac{5.6}{3.63*2.7} = 0.57$

User A and C
1. Dot product: -0.6 - 0.6 = -1.2
2. $||A|| = 3.63$
3. $||C|| = 3.24$
4. Pearson Correlation $= \frac{-1.2}{11.76} = -0.10$

User B and C
1. Dot product = -6.5
2. $||B|| = 2.7$
3. $||C|| = 3.24$
4. Pearson Correlation = $\frac{-6.5}{8.7} = -0.74$.

The Pearson correlation results give an indication of the linear relationship between the users' normalized ratings, with positive values showing similarity and negative values showing dissimilarity.

Exercise 5

a) Research how you could inspect the memory usage of a JAX program. If you need to use some tools, try to install them in Google Colab. Then attempt to find out whether such large intermittent matrix is indeed allocated. Report your findings about methods for memory inspection, whether you could confirm the memory problem (or not), and submit relevant tool outputs (do not include any large data sets). Hint: you might need to split the function mse_loss_one_batch into smaller parts.

b) Assuming that the original implementation is indeed not memory-scalable, design an approach which uses significantly less memory for loss computation. Hint: an intermittent matrix of dimensions at most B × B is acceptable in terms of memory usage. Give your solution as pseudo-code with all relevant details. In addition, provide a small example of a hypothetical execution for B = 4 with imagined but concrete values for rows, columns, ratings and the relevant entries of mat_u and mat_v.

c)

Solution

a) We discovered the package memory_profiler to be useful. To use it, we can tag methods with ”@profile”, to log the memory usage per line of the method. We then wrote a profiled method, and defined mocked data to prevent a long execution time. The profiling showed us, that in our example, around 400MB were used, with a memory loss of 100 MB after termination. This confirms the given assumption. An example output log has been copied and placed into a separate file ”e02-t05-a.txt”.

b) The idea is, to only compute the required data entries of the whole matrix. For this, we define batches of data, with size B, i.e., we consider B entries total. For the matrix, this then creates a result of size $B \times B$.

Algorithm 1 function mse_loss_one_batch(mat_u, mat_v, record)

```
total_loss = 0.0
batch_size = length(record[”rows”])
for i from 0 to batch_size - 1 do do
    user_index = record[”rows”][i]
    item_index = record[”columns”][i]
    actual_rating = record[”ratings”][i]
    predicted_rating = mat_u[user_index] × mat_v[:, item_index]
    error = (predicted_rating - actual_rating)$^2$
    total_loss += error
end for
mean_squared_error = total_loss / batch_size
return mean_squared_error
```

c)

Exercise 6

What data can you use to populate a utility matrix in the following cases? Think about what are users and what constitutes the items. What relations other than like or a numerical rating can be exploited to express a positive or negative connection?

a) A website where students can rate their professors. Users are students that can share their experiences of lectures, seminars and exams. If a rating is given for a so far unknown professor, there will be created a fresh profile for this professor.

b) An online community for sharing artwork. Each user can upload pictures of his or her artwork, and also rate those of others.

c) A dating platform. Every user has an online profile and can visit and like those of others. Users can describe a hidden profile of their dream partner. Users can send messages to users and block users. What is special about this scenario?

Solution

a) In this scenario, we distinguish between users and items. The **users** in this matrix are students who provide feedback on the professors. The **items** are professors who are being evaluated. Data types to populate the matrix could be numerical ratings, text reviews,comments can be analyzed for sentiment, yielding a score that indicates positive, neutral, or negative sentiment toward the professor, the frequency of interaction: the number of times a student has attended the course.

b) The **users** are individuals who upload and rate artwork. The **items** are individual pieces of artwork, each associated with a user who uploaded it. Each artwork can have associated attributes such as style, medium, and tags. Data types to populate the matrix could be numerical ratings, likes or favorites, frequency of interaction, tags and comments.

c) The **users** are both users and **items**. Each user interacts with other users, making the matrix symmetric. When two users like each other, it forms a mutual relationship, which is different from typical one-way ratings. Matches, or two-way likes, are particularly important as they directly indicate compatibility. Users can define an ideal partner profile, which introduces a layer of explicit preference. Matching users' profiles with these hidden preferences can suggest compatibility. The frequency, length, and sentiment of messages exchanged indicate levels of interest and compatibility between users etc. There are different ways to express a connection.