

Answer – This is a question of collaborative filtering. It is an application of machine learning which is generally used to give shopping recommendations to users. Like if you buy a thing on amazon, then amazon would suggest few other things that you might be interest in.

So, there are two ways of doing collaborative filtering. The first way is item-based approach. This approach finds items that are similar. Like let’s say, Calzone and Pizza are similar items. So, if someone likes a calzone, we can say with some confidence that they will also like pizza.

The second way of doing this is by using a ‘user-based’ approach. In this approach, we find two people who are like each other. So, if we know that a 18 year old – upper-middle class, English speaking guy likes to go in concerts of Ed Sheeran, we can tell with some confidence that some other 18 year old -upper-middle class, English speaking guy would like to go in concerts of Ed Sheeran too.

Here in the question, it is clearly mentioned that we must use an item-based approach. There is a rating missing for product 6. So we need to find a product that is most similar to product 6 and based on that product, we can predict the missing rating.

Now the question reduces to, how can we find that which item is most like product 6?

Well, we can do that by calculating the cosine of product 6 with other products and the product that would have the highest cosine value with respect to product 6 will be the most similar product to product 6.

Cosine value is an indicator of the angle between two vectors. Therefore, if two vectors are similar, they will lie on each-other and the angle between them will be zero. And cos(0) = 1 (which is the maximum possible value). So, the higher the cosine value, higher is the similarity.

So the algorithm has following steps –

1. Find the mean rating of all the participants
2. Centralize ratings for all participants
3. Find the L2-norm for all the products.
4. Calculate cosine between L2-norm of product 6 and all other products.
5. Select the product with highest cosine value.

While doing all these steps, remember that the rating is missing for participant 1, so we will exclude participant 1 from the calculations. (data of participant 1 is the test data here. We do not include test data in training).

So, step 1-

Find mean rating

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | Product 1 | Product 2 | Product 3 | Product 4 | Product 5 | Product 6 | Mean |
| Participant 2 | 3 | 4 | 2 | 3 | 4 | 2 | 3 |
| Participant 3 | 2 | 3 | 4 | 4 | 3 | 5 | 3.5 |
| Participant 4 | 5 | 4 | 5 | 4 | 1 | 2 | 3.5 |

Now, we will centralize ratings. (Subtract ratings from the mean). So, for participant 2, we will subtract 3 from all ratings. Similarly, for participant 3 and participant 4, we will subtract 3.5 from each rating given by them.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | Product 1 | Product 2 | Product 3 | Product 4 | Product 5 | Product 6 |
| Participant 2 | 0 | 1 | -1 | 0 | 1 | -1 |
| Participant 3 | -1.5 | -0.5 | 0.5 | 0.5 | -0.5 | 1.5 |
| Participant 4 | 1.5 | 0.5 | 1.5 | 0.5 | -2.5 | -1.5 |

Now, we will calculate L2 norm for all the products. To calculate L2 norm, we square all the ratings, add them up and then take their square-root. So for product 1, it will be, sqrt(0^2 + (-1.5)^2 + 1.5^2) = 2.12

Similarly,

For product 2, sqrt(1^2 + (-0.5)^2 + 0.5^2) = 1.22

For product 3, sqrt((-1)^2 + 0.5^2 + 1.5^2) = 1.871

For product 4, sqrt(0^2 + 0.5^2 + 0.5^2) = 0.707

For product 5, sqrt(1^2 + (-0.5)^2 + (-2.5)^2) = 2.739

For product 6, sqrt((-1)^2 + 1.5^2 + (-1.5)^2) = 2.345

Now, we have all the L2 norms. We can now calculate cosines. For which, the formula is,

Graphical user interface, text, application

Description automatically generated

Now look at the denominator. Here is where our L2-norm calculations would come in handy. Denominator is nothing but L2-norms multiplied. So effectively, we only need to calculate numerators.

For that, we gotta multiply all the centralized ratings for product 6 with the product that we are comparing it with.

Lets start with Product 1.

For product 1, Ratings are -> (0, -1.5 and 1.5). For product 6, ratings are (-1, 1.5, -1.5)

So, Cosine = (0 x -1 + (-1.5) x 1.5 + 1.5 x (-1.5)) / (2.12 x 2.345) = -0.90

For product 2, Ratings are -> (1, -0.5, 0.5)

Cosine = (1 x -1 + (-0.5 x 1.5) + 0.5 x -1.5)/ (1.22 x 2. 345) = -0.87

For product 3, Ratings are -> (-1, 0.5, 1.5)

Cosine = (-1 x -1 + 0.5 x 1.5+ 1.5 x -1.5)/( 1.871 x 2.345) = -0.11

For Product 4, Ratings are -> (0, 0.5, 0.5)

Cosine = (0 x -1 + 0.5 x 1.5+ 0.5 x -1.5)/(0.707x 2.345) = 0

For product 5, Ratings are -> (1, -0.5, -2.5)

Cosine = (1 x -1 + (- 0.5 x 1.5) + (-2.5 x -1.5))/( 2.739x2.345) = 0.311

So, the cosine value is highest for Product 5. We will select it as the most similar product.

So, participant 1, assigned a rating of 5 to Product 5. Hence, it is more likely that he would assign a rating of **5 to the product 6**