CZ4032 Data Analytics and Mining

Tutorial: Recommendation, Data

1. To make recommendations for user x, user-based collaborative filter method performs the following two steps:

   1) Find set N of other users whose ratings are most "similar" to x's ratings

   2) Estimate x's ratings based on ratings of users in N

Step 1: "Finding Similar Users". We can use measures such as cosine similarity measure and Pearson correlation coefficient to find the most similar N users.

Let $\boldsymbol{r_x}$ be the vector of user $\boldsymbol{x}$'s ratings,

- **Cosine similarity measure**
  - $\text{sim}(\boldsymbol{x}, \boldsymbol{y}) = \cos(\boldsymbol{r_x}, \boldsymbol{r_y}) = \frac{r_x \cdot r_y}{||r_x|| \cdot ||r_y||}$
- **Pearson correlation coefficient**
  - $\boldsymbol{S_{xy}}$ = items rated by both users $\boldsymbol{x}$ and $\boldsymbol{y}$
  - $\boldsymbol{r_{xs}}$ ***: rating of user x on item s***
  - $\boldsymbol{r_{ys}}$ ***: rating of user y on item s***
  - $\overline{r_x}$: mean of the values of the r_x-variable (average rating of user x)
  - $\overline{r_y}$: mean of the values of the r_y-variable (average rating of user y)

$$\boldsymbol{sim(x,y)} = \frac{\sum_{s \in S_{xy}} (r_{xs} - \overline{r_x})(r_{ys} - \overline{r_y})}{\sqrt{\sum_{s \in S_{xy}} (r_{xs} - \overline{r_x})^2} \sqrt{\sum_{s \in S_{xy}} (r_{ys} - \overline{r_y})^2}}$$

Step 2: **From similarity metric to recommendations:**

- Let $\boldsymbol{r_x}$ be the vector of user $\boldsymbol{x}$'s ratings
- Let $\boldsymbol{N}$ be the set of $\boldsymbol{k}$ users most similar to $\boldsymbol{x}$ who have rated item $\boldsymbol{i}$

you can use the following formula to predict the rating of user x on item i

- $r_{xi} = \frac{\sum_{y \in N} sim(x,y) \cdot r_{yi}}{\sum_{y \in N} sim(x,y)}$

Now given the following matrix

Movie \ user	1	2	3	4	5	6	7	8	9	10	11	12
1	1		3			5			5		4	
2			5	4		5	5			4	4	5
3	2	4		1	2		3		4	3	5	
4		2	4		5			4			2	
5			4	3	4	2						5
6	1		3		3			2				
7			3		3			4	4			
8			4			4		1	2		3	
9	4		1	2		5			2			
10	2	2		2					3	3	3	
11		4	3		3				1	2		
12			3		3			2			4	

CosineSim(u1, u2) = (2*4 + 2*2)/ ()

1) Compute the rating of movie 1 by user 2 using user-based Collaborative Filtering of 3 most similar users.

2) Compute the rating of movie 1 by user 2 using item-based Collaborative Filtering of 3 most similar movies.

If we use the following to **Prediction for item *i* of *user x*:**

- $r_{xi} = \frac{1}{k} \sum_{y \in N} r_{yi}$

What will be results for the last two questions?

2. Can you use Graph Neural Networks for recommendation?

3 In order to perform data mining, what important steps do you need to perform in the following cases:

a. Trying to detect oil slicks from satellite images to give early warning of ecological disasters and deter illegal dumping
b. A bank is trying to reduce customer attrition
c. A supermarket is planning store layouts

4 Suppose a group of 12 sales price records has been sorted as follows:

5, 10, 11, 13, 15, 35, 50, 55, 72, 92, 204, 215

Partition them into 3 bins by each of the following methods:

- equal-frequency (equidepth) partitioning
- equal-width partitioning

Use some appropriate method to smooth.

5 Given two sets of restaurant objects: one set is collected from Google Maps, and the other set is collected from Yelp. Each restaurant object is associated with a set of attributes, such as name, address, coordinates, phone numbers, etc. Please design an algorithm for entity matching to find all the pairs of the same restaurant objects from the two sets. Analyze the algorithm complexity.