Seoul National University

M1522.001400-001 Introduction to Data Mining

Homework 5: Advertising on the Web & Recommendation (Chapter 8 & 9)

2017-18538 Hwang SunYoung

Question1

Given a bipartite graph G with the following set of edges:

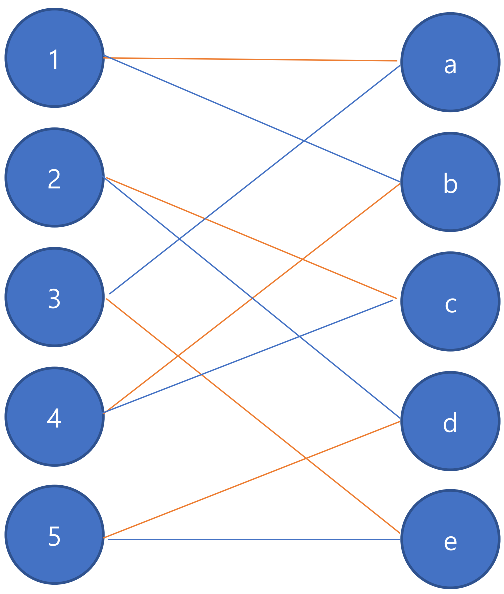
(1, a), (1, b), (2, c), (2, d), (3, a), (3, e), (4, b), (4, c), (5, d)

(a) consider a greedy match for the graph G. The first edge, (1, a), surely becomes part of the matching. The second edge, (1, b), cannot be chosen, because node 1 already appears in the matching. The third edge, (2, c), is selected, because neither node 2 nor node c appears in the matching so far. We can apply this process to all edges by turns. The matching produced by the greedy matching algorithm for this ordering of the edges is {(1, a), (2, c), (3, e), (4, b), (5, d)}.

(b) perfect matching for G is {(1, a), (2, c), (3, e), (4, b), (5, d)}.

(c) perfect matching for G′ is M_1 ={(1, a), (2, c), (3, e), (4, b), (5, d)} or M_2 ={(1, b), (2, d), (3, a), (4, c), (5, e)}. M_1 is same with perfect matching for G, and M_2 is different from perfect matching for G.

(d) Because each vertex has 2 edges, the value of k is 2. Graph G′ is partitioned into k edge-disjoint perfect matchings M_1 and M_2 from part (c).



M_1: orange lines, M_2: blue lines

Question2

(a) Because content-based recommendation recommend items based on the individual's rating history while collaborative filtering requires diverse users' ratings, content-based recommendation would be better.

(b)

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

(Ⅰ) Jaccard similarity

sim(A, B) = 2/5, sim(B, C) = 2/5, sim(A, C) = 2/4 = 1/2

(Ⅱ) normalized utility matrix

Average of row A: (1+4+3)/3=2.67

Average of row B: (2+5+4+1)/4=3

Average of row C: (5+5+3)/3=4.33

	a	b	c	d	e	f	g	H
A		-1.67		1.33			0.33	
B	-1	2	1	-2				
C	0.67			0.67			-1.33	

(Ⅲ) cosine similarity

sim(A, B) = $\frac{-1.67\times 2+1.33\times(-2)}{\sqrt{(-1.67)^2+1.33^2+0.33^2}\times\sqrt{(-1)^2+2^2+1^2+(-2)^2}}$ = -0.88

sim(B, C) = $\frac{-1\times 0.67+(-2)\times 0.67}{\sqrt{(-1)^2+2^2+1^2+(-2)^2}\times\sqrt{0.67^2+0.67^2+(-1.33)^2}}$ = -0.39

sim(A, C) = $\frac{1.33\times 0.67+0.33\times(-1.33)}{\sqrt{(-1.67)^2+1.33^2+0.33^2}\times\sqrt{0.67^2+0.67^2+(-1.33)^2}}$ = 0.13

(IV)

Jaccard similarity: sim(A, B) = 0.4, sim(B, C) = 0.4, sim(A, C) = 0.5

Cosine similarity: sim(A, B) = -0.88, sim(B, C) = -0.39, sim(A, C) = 0.13

Cosine similarity is more appropriate for the given scenario. Because Jaccard similarity ignore values in the matrix, it loses importance in matrix detailed ratings. On the other hand cosine similarity reflect ratings.

Question3

When the query sequence is xxyyzz, worst case greedy choice is CCBB_ _. In worst case, greedy algorithm assigns 4 of these 6 queries. So the greedy algorithm assigns at least 4 of these 6 queries.

Question4

$$M = \begin{array}{c} \\ A \\ B \\ C \\ D \end{array} \begin{array}{c} \begin{array}{cccc} I_1 & I_2 & I_3 & I_4 \end{array} \\ \begin{pmatrix} 2 & / & 4 & 1 \\ / & 1 & 3 & / \\ 2 & 4 & / & 3 \\ 1 & 5 & / & / \end{pmatrix} \end{array}$$

M is a utility matrix. / is a blank.

(a)

$$U_0 \times V_0 = \begin{bmatrix} 1 & 2 \\ 2 & 1 \\ 1 & 1 \\ 2 & 0 \end{bmatrix} \times \begin{bmatrix} 1 & 2 & 2 & 0 \\ 0 & 1 & 2 & 1 \end{bmatrix} = \begin{bmatrix} 1 & 4 & 6 & 2 \\ 2 & 5 & 6 & 1 \\ 1 & 3 & 4 & 1 \\ 2 & 4 & 4 & 0 \end{bmatrix}$$

We ignore blanks when computing the RMSE. We subtract each entry in U_0V_0 from the entries in M. In the first row, we get 1, -2, 1 and the sum of squares is 6. In the second row, we get -4, -3 and the sum of squares is 25. In the third row, we get 1, 1, 2 and the sum of squares is 6. In the fourth row, we get -1, 1 and the sum of squares is 2. 6+25+6+2=39. We divide by 10(the number of nonblank entries in M) and take the square root. In this case $\sqrt{39/10} = 1.975$ is the RMSE.

∴RMSE=1.975

(b)

$$U \times V = \begin{bmatrix} x_1 & 2 \\ x_2 & 1 \\ x_3 & 1 \\ x_4 & 0 \end{bmatrix} \times \begin{bmatrix} 1 & 2 & 2 & 0 \\ 0 & 1 & 2 & 1 \end{bmatrix} = \begin{bmatrix} x_1 & 2x_1 + 2 & 2x_1 + 4 & 2 \\ x_2 & 2x_2 + 1 & 2x_2 + 2 & 1 \\ x_3 & 2x_3 + 1 & 2x_3 + 2 & 1 \\ x_4 & 2x_4 & 2x_4 & 0 \end{bmatrix}$$

We only consider terms that include unknown.

$$(2 - x_1)^2 + (4 - (2x_1 + 4))^2 = (2 - x_1)^2 + (-2x_1)^2$$

$$(1 - (2x_2 + 1))^2 + (3 - (2x_2 + 2))^2 = (-2x_2)^2 + (1 - 2x_2)^2$$

$$(2 - x_3)^2 + (4 - (2x_3 + 1))^2 = (2 - x_3)^2 + (3 - 2x_3)^2$$

$$(1 - x_4)^2 + (5 - 2x_4)^2$$

We want the value of x that minimizes the sum, so we take the derivative and set that equel to 0, as:

$-2 \times \{(2 - x_1) + (-4x_1)\} = 0$ -> $x_1 = 0.4$

$-4 \times \{(-2x_2) + (1 - 2x_2)\} = 0$ -> $x_2 = 0.25$

$-2 \times \{(2 - x_3) + (6 - 4x_3)\} = 0$ -> $x_3 = 1.6$

$-2 \times \{(1 - x_4) + (10 - 4x_4)\} = 0$ -> $x_4 = 2.2$

$$\therefore x_1 = 0.4, \quad x_2 = 0.25, \quad x_3 = 1.6, \quad x_4 = 2.2$$

(c)

$$U \times V = \begin{bmatrix} 0.4 & 2 \\ 0.25 & 1 \\ 1.6 & 1 \\ 2.2 & 0 \end{bmatrix} \times \begin{bmatrix} 1 & 2 & 2 & 0 \\ y_1 & y_2 & y_3 & y_4 \end{bmatrix} = \begin{bmatrix} 0.4 + 2y_1 & 0.8 + 2y_2 & 0.8 + 2y_3 & 2y_4 \\ 0.25 + y_1 & 0.5 + y_2 & 0.5 + y_3 & y_4 \\ 1.6 + y_1 & 3.2 + y_2 & 3.2 + y_3 & y_4 \\ 2.2 & 4.4 & 4.4 & 0 \end{bmatrix}$$

We only consider terms that include unknown.

$$(2 - (0.4 + 2y_1))^2 + (2 - (1.6 + y_1))^2 = (1.6 - 2y_1)^2 + (0.4 - y_1)^2$$

$$(1 - (0.5 + y_2))^2 + (4 - (3.2 + y_2))^2 = (0.5 - y_2)^2 + (0.8 - y_2)^2$$

$$(4 - (0.8 + 2y_3))^2 + (3 - (0.5 + y_3))^2 = (3.2 - 2y_3)^2 + (2.5 - y_3)^2$$

$$(1 - 2y_4)^2 + (3 - y_4)^2$$

We want the value of y that minimizes the sum, so we take the derivative and set that equel to 0, as:

$-2 \times \{(3.2 - 4y_1) + (0.4 - y_1)\} = 0$ -> $y_1 = 0.72$

$-2 \times \{(0.5 - y_2) + (0.8 - y_2)\} = 0$ -> $y_2 = 0.65$

$-2 \times \{(6.4 - 4y_3) + (2.5 - y_3)\} = 0$ -> $y_3 = 1.78$

$-2 \times \{(2 - 4y_4) + (3 - y_4)\} = 0$ -> $y_4 = 1$

$\therefore y_1 = 0.72, \; y_2 = 0.65, \; y_3 = 1.78, \; y_4 = 1$