Seoul National University

M1522.001400 Introduction to Data Mining

Spring 2016, Kang

Homework 8: Recommendation System (Chapter 9)

Due: May 30, 09:30 AM

Reminders

- The points of this homework add up to 100.
- Like all homeworks, this has to be done individually.
- Lead T.A.: Minsoo Jung (qtyp456987@gmail.com)
- Please type your answers *in English*. Illegible handwriting may get no points, at the discretion of the graders.
- If you have a question about assignments, please upload your question in eTL.
- If you want to use slipdays or consider late submission with penalties, please note that you are allowed one week to submit your assignment after the due date.

Remember that:

- Whenever you are making an assumption, please state it clearly.

Question 1

Three computers, A, B, and C, have the numerical features listed below:

Feature	A	B	C
Processor Speed	3.06	2.68	2.92
Disk Size	500	320	640
Main-Memory Size	6	4	6

We may imagine these values as defining a vector for each computer; for instance, A's vector is [3.06, 500, 6]. We can compute the cosine similarity between any two of the vectors, but if we do not scale the components, then the disk size will dominate and make differences in the other components essentially invisible. Let us use 1 as the scale factor for processor speed, α for the disk size, and β for the main memory size.

(a) In terms of α and β, compute the cosines of the angles between the vectors for each pair of the three computers.

(b) What are the angles between the vectors if α = β = 1?

(c) What are the angles between the vectors if α = 0.01 and β = 0.5?

(d) One fair way of selecting scale factors is to make each inversely proportional to the average value in its component. What would be the values of α and β, and what would be the angles between the vectors?

Question 2

A certain user has rated the three computers of **Question 1** as follows: A: 4 stars, B: 2 stars, C: 5 stars.

(a) Normalize the ratings by subtracting the average ratings for this user.

(b) Compute a user profile for the user, with components for processor speed, disk size, and main memory size, based on the data of **Question 1**.

Question 3

Figure 1 is a utility matrix, representing the ratings, on a 1–5 star scale, of eight items, a through h, by three users A, B, and C. Compute the following from the data of this matrix.

	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

Figure 1: A utility matrix for **Question 3** and **Question 4**

(a) Treat blank as 0 and ratings of 1 ~ 5 as 1. Compute the Jaccard similarity between each pair of users.

(b) Repeat Part (a), but use the cosine similarity.

(c) Treat ratings of 3, 4, and 5 as 1 and 1, 2, and blank as 0. Compute the Jaccard similarity between each pair of users.

(d) Repeat Part (c), but use the cosine similarity.

(e) Normalize the matrix by subtracting the average value from each nonblank entry for its user.

(f) Using the normalized matrix from Part (e), compute the cosine similarity between each pair of users.

Question 4

We cluster items in the matrix of Figure 1. Do the following steps.

(a) Cluster the eight items hierarchically into four clusters. The following method should be used to cluster. Replace all 3's, 4's, and 5's by 1 and replace 1's, 2's, and blanks by 0. Use the Jaccard distance to measure the distance between the resulting column vectors. For clusters of more than one element, take the distance between clusters to be the minimum distance between pairs of elements, one from each cluster. Note that Jaccard distance = 1 – Jaccard similarity.

(b) Then, construct from the original matrix of Figure 1 a new matrix whose rows correspond to users, as before, and whose columns correspond to clusters. Compute the entry for a user and cluster of items by averaging the nonblank entries for that user and all the items in the cluster.

(c) Compute the cosine similarity between each pair of users, according to your matrix from Part (b).