Recommender system task solutions

Table 1 presents movie ratings by 6 users on 6 movies. The latex source of the table is available on the course page (mratingstable.tex). The ratings are between 1 (didn't like at all) to 5 (fantastic movie) and 0 means a missing rating (the user hasn't watched the movie). The users are notated $u1, \ldots, u6$ and movies $m1, \ldots, m6$. The task is to apply recommender systems for rating prediction using neighbourhood-based collaborative filtering (Aggarwal 18.5.2 and an example in the lecture).

a) **Calculate mean ratings per user. Use all non-missing ratings in the calculation.**
The row means are
$\mu(u1) = 2.000$
$\mu(u2) = 3.000$
$\mu(u3) = 3.000$
$\mu(u4) = 4.000$
$\mu(u5) = 3.600$
$\mu(u6) = 3.333$

b) **Calculate required pairwise similarities between users[1] using a modified Pearson correlation r ("Pearson" in Aggarwal Equation 18.2). Use the mean values calculated in part a. Remember that the correlation is calculated only over co-rated movies.**

User-user similarities (number of common ratings in parenthesis):

	u1	u2	u3	u4	u5	u6
u1	1.000 (5)	0.816 (5)	0.707 (5)	1.000 (3)	-0.811 (4)	-0.721 (3)
u2	0.816 (5)	1.000 (6)	0.000 (6)	1.000 (4)	-0.559 (5)	-0.721 (3)
u3	0.707 (5)	0.000 (6)	1.000 (6)	0.316 (4)	-0.589 (5)	-0.557 (3)
u4	1.000 (3)	1.000 (4)	0.316 (4)	1.000 (4)	-0.684 (3)	-0.371 (2)
u5	-0.811 (4)	-0.559 (5)	-0.589 (5)	-0.684 (3)	1.000 (5)	0.905 (2)
u6	-0.721 (3)	-0.721 (3)	-0.557 (3)	-0.371 (2)	0.905 (2)	1.000 (3)

c) **Predict missing ratings using two nearest neighbours ($K = 2$) and an extra requirement that the similarity is $r \geq 0.5$. Tell if**

[1] Note: similarity between $u2$ and $u3$ is not needed, so 14 similarities.

the movie is recommended to the user (if the user would like it more than average).
Report if some prediction cannot be made (not enough sufficiently similar neighbours with required ratings).

$u1$: nearest neighbours are $u4$ and $u2$ (and both $r \geq 0.5$). Predicted rating to $m5$ is $3.000 > \mu(u1)$, so recommend.

$u4$: nearest neighbours $u1$ and $u2$ (and $r \geq 0.5$). For $m1$ the prediction is $5.000 > \mu(u4)$, so recommend.
For $m6$ the prediction is $3.500 < \mu(u4)$, don't recommend.

$u5$: Only $u6$ sufficiently close neighbour, predictions cannot be made.
$u6$: Only $u5$ sufficiently close neighbour, predictions cannot be made.

(Extra note: $u5$ and $u6$ have only 2 common ratings, so the r value is not very reliable.)

d) **Consider the item-based way of predicting the missing ratings of movies $m3$ and $m4$ with adjusted cosine similarity, as suggested in Aggarwal 18.5.2.2. Why it is not a good solution here? Suggest an alternative item-based solution that could be used instead (no need to calculate the actual predictions).**

Aggarwal suggest to choose the most similar items with adjusted cosine similarity. However, it doesn't work here at all. Ratings for $m3$ and $m4$ are indentical (if neither is missing), i.e., all users have liked them equally much. This means that they should have maximal similarity. However, adjusted cos-sim evaluates similarity as 0 (very dissimilar). The reason is that users have given average ratings to movies $m3$ and $m4$ and subtracting the user means (mean-centering) produces zero vectors, whose dot product is zero.

One solution is to use Pearson correlation coefficient for similarity between items. It gets value 1.0, i.e., perfect similarity. (Extra note: here the mean values of two items' ratings are the same, so using a modified Pearson doesn't cause any difference. If this was not the case, the similarity could be smaller.)

Table 1: Movie ratings (scale 1–5) by 6 users ($u1$–$u6$) on 6 movies ($m1$–$m6$). Special value 0 means a missing rating.

	$m1$	$m2$	$m3$	$m4$	$m5$	$m6$
u1	3	1	2	2	0	2
u2	4	2	3	3	4	2
u3	4	1	3	3	2	5
u4	0	3	4	4	5	0
u5	2	5	5	0	3	3
u6	1	4	0	5	0	0