CS-E4650 - Assignment 4

Arttu Häkkinen - 596077

November 27, 2021

Task 3

The original ratings table:

	$m1$	$m2$	$m3$	$m4$	$m5$	$m6$
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

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

a)

	$u1$	$u2$	$u3$	$u4$	$u5$	$u6$
mean rating	2.0	3.0	3.0	4.0	3.6	3.3

Table 2: Mean movie ratings for each user u1-u6.

The mean ratings for each user not including the missing values in the calculations are in the table 2 above. Script to calculate these mean values can be found from the zip file.

b)

I calculated pairwise similarities between users using modified Pearson correlation (eq. 18.12 in Aggarwal book). In the calculations for each user pair, I used only the co-rated movie ratings. The source code for the calculations can be

	$u1$	$u2$	$u3$	$u4$	$u5$	$u6$
$u1$	1 (5)	-	-	-	-	-
$u2$	0.845 (5)	1 (6)	-	-	-	-
$u3$	0.715 (5)	0 (6)	1 (6)	-	-	-
$u4$	1 (3)	1 (4)	0.426 (4)	1 (4)	-	-
$u5$	-0.816 (4)	-0.559 (5)	-0.589 (5)	-0.866 (3)	1 (5)	-
$u6$	-0.721 (3)	-0.721 (3)	-0.577 (3)	1 (2)	1 (2)	1 (3)

Table 3: Pairwise similarities in co-rated movie ratings between users using modified Pearson correlation. The number of co-rating in the parenthesis.

found from the zip file (programming language: Python, libraries used: pandas and numpy).

The resulting pairwise similarity matrix is reported in table 3 in a triangle similarity matrix form.

c)

There are in total 7 missing ratings in the original table:

- u1 rating of m5
- u4 rating of m1
- u4 rating of m6
- u5 rating of m4
- u6 rating of m3
- u6 rating of m5
- u6 rating of m6

We will predict these ratings using $K = 2$ nearest neighbors with the similarity requirement of $r \geq 0.5$ with the following equation

$$\tilde{x_j} = \frac{\sum_{\boldsymbol{y} \in NN_{\boldsymbol{x}}} r(\boldsymbol{x}, \boldsymbol{y}) \cdot (y_j - \mu_{\boldsymbol{y}})}{\sum_{\boldsymbol{y} \in NN_{\boldsymbol{x}}} r(\boldsymbol{x}, \boldsymbol{y})} + \mu_{\boldsymbol{x}}$$

Predict u1 rating for m5:

The best two similarities between u1 and other users are $r(u1, u4) = 1$ and $r(u1, u2) = 0.845$. So, the predicted u1 rating for m5 is

$$\frac{1 \cdot (5 - 4.0) + 0.845 \cdot (4 - 3.0)}{1 + 0.845} + 2.0 = 3.0$$

Predict u4 rating for m1:

The best three similarities between u4 and other users are $r(u4, u1) = r(u4, u2) = r(u4, u6) = 1$. Now we have to choose two of these. Choosing u1 and u2. So, the predicted u4 rating for m5 is

$$\frac{1 \cdot (3 - 2.0) + 1 \cdot (4 - 3.0)}{1 + 1} + 4.0 = 5.0$$

Let's see if this is different when choosing e.g. u1 an u6:

$$\frac{1 \cdot (3 - 2.0) + 1 \cdot (1 - 3.3)}{1 + 1} + 4.0 = 3.35$$

So, it clearly matters which two of the three nearest neighbors we choose.

Predict u4 rating for m6:

The best three similarities between u4 and other users are $r(u4, u1) = r(u4, u2) = r(u4, u6) = 1$. Now we have to choose two of these. Choosing u1 and u2 because u6 is missing the rating for m6. So, the predicted u4 rating for m6 is

$$\frac{1 \cdot (2 - 2.0) + 1 \cdot (2 - 3.0)}{1 + 1} + 4.0 = 3.5$$

Predict u5 rating for m4:

The only pairwise similarity between u5 and other user such that $r \geq 0.5$ is $r(u5, u6) = 1$. So, the prediction cannot be made using $K = 2$ nearest neighbors. (With $K = 1$ nearest neighbors the prediction would be $\frac{1 \cdot (5 - 3.3)}{1} + 3.6 = 5.3$.)

Predict u6 rating for m3:

The best two similarities between u1 and other users are $r(u6, u4) = 1$ and $r(u6, u5) = 1$. So, the predicted u6 rating for m3 is

$$\frac{1 \cdot (4 - 4.0) + 1 \cdot (5 - 3.6)}{1 + 1} + 3.3 = 4.0$$

Predict u6 rating for m5:

The best two similarities between u1 and other users are $r(u6, u4) = 1$ and $r(u6, u5) = 1$. So, the predicted u6 rating for m5 is

$$\frac{1 \cdot (5 - 4.0) + 1 \cdot (5 - 3.6)}{1 + 1} + 3.3 = 3.5$$

Predict u6 rating for m6:

The best two similarities between u1 and other users are $r(u6, u4) = 1$ and $r(u6, u5) = 1$. The rating from u4 for m6 is missing, so we cannot predict u6 rating for m6 with $K = 2$ nearest neighbors. (With $K = 1$ nearest neighbors the prediction would be $\frac{1 \cdot (3 - 3.6)}{1} + 3.3 = 2.7$.)