Item Based KNN Collaborative Filtering

J. Milmore

University of Massachusetts Amherst
Department of Mathematics and Statistics

April 6, 2020

Item Based CF

Memory Based Collaborative Filtering: Predict a rating for user u and movie m by using previous ratings

- Item Based: $sim(m, m')$

Problem Formulation

Rating Matrix

- Rows represent ratings given to each movie
- Columns represent ratings given by each user
- $R_{m,u}$ represents rating given to movie m by user u

$$R = \begin{array}{c c} & \begin{array}{ccc} u_1 & u_2 & u_* \end{array} \\ \begin{array}{c} m_* \\ m_2 \\ m_3 \\ m_4 \\ m_5 \end{array} & \begin{pmatrix} 2 & 5 & \color{red}{?} \\ 3 & ? & 3 \\ ? & 4 & 2 \\ 3 & ? & 4 \\ 2 & 5 & ? \end{pmatrix} \end{array}$$

Goal

- Predict R_{m_*,u_*}, the missing rating for user u_* and movie m_* using the existing ratings

Method

- Filter rating matrix to just contain rows of movies rated by u_*, along with the movie of interest, m_*

$$
\begin{array}{c} \\ m_* \\ m_2 \\ m_3 \\ m_4 \\ m_5 \end{array}
\begin{array}{c} \begin{array}{ccc} u_1 & u_2 & u_* \end{array} \\ \begin{pmatrix} 2 & 5 & \color{red}{?} \\ 3 & ? & 3 \\ ? & 4 & 2 \\ 3 & ? & 4 \\ 2 & 5 & ? \end{pmatrix} \end{array}
\rightarrow
\begin{array}{c} \\ m_* \\ m_2 \\ m_3 \\ m_4 \end{array}
\begin{array}{c} \begin{array}{ccc} u_1 & u_2 & u_* \end{array} \\ \begin{pmatrix} 2 & 5 & \color{red}{?} \\ 3 & ? & 3 \\ ? & 4 & 2 \\ 3 & ? & 4 \end{pmatrix} \end{array}
$$

Method (Cont.)

- Replace missing ratings by the average rating of the user

$$\begin{array}{c c} & \begin{array}{ccc} u_1 & u_2 & u_* \end{array} \\ \begin{array}{c} m_* \\ m_2 \\ m_3 \\ m_4 \end{array} & \begin{pmatrix} 2 & 5 & \color{red}{?} \\ 3 & ? & 3 \\ ? & 4 & 2 \\ 3 & ? & 4 \end{pmatrix} \end{array} \rightarrow \begin{array}{c c} & \begin{array}{ccc} u_1 & u_2 & u_* \end{array} \\ \begin{array}{c} m_* \\ m_2 \\ m_3 \\ m_4 \end{array} & \begin{pmatrix} 2 & 5 & \bar{u}_* \\ 3 & \bar{u}_2 & 3 \\ \bar{u}_1 & 4 & 2 \\ 3 & \bar{u}_2 & 4 \end{pmatrix} \end{array}$$

Method (Cont.)

- Adjust each rating by the average rating of the user

$$\begin{array}{c} \\ m_* \\ m_2 \\ m_3 \\ m_4 \end{array} \begin{array}{c} \begin{array}{ccc} u_1 & u_2 & u_* \end{array} \\ \begin{pmatrix} 2 & 5 & \bar{u_*} \\ 3 & \bar{u_2} & 3 \\ \bar{u_1} & 4 & 2 \\ 3 & \bar{u_2} & 4 \end{pmatrix} \end{array} \rightarrow \begin{array}{c} \\ m_* \\ m_2 \\ m_3 \\ m_4 \end{array} \begin{array}{c} \begin{array}{ccc} u_1 & u_2 & u_* \end{array} \\ \begin{pmatrix} 2-\bar{u_1} & 5-\bar{u_2} & 0 \\ 3-\bar{u_1} & 0 & 3-\bar{u_*} \\ 0 & 4-\bar{u_2} & 2-\bar{u_*} \\ 3-\bar{u_1} & 0 & 4-\bar{u_*} \end{pmatrix} \end{array}$$

We will further refer to this *adjusted* rating matrix simply as R

- Find the k movies most **similar** to m_* and **predict** (m_*, u_*) to be the weighted average of the ratings given by u_* to these k movies

Scoring Similar Movies

Adjusted Cosine Similarity Metric

$$sim(m_1, m_2) = \frac{\sum_{u \in A}(R_{m_1,u} - \bar{R}_u)(R_{m_2,u} - \bar{R}_u)}{\sum_{u \in A}(R_{m_1,u} - \bar{R}_u)^2 \sum_{u \in A}(R_{m_2,u} - \bar{R}_u)^2}$$

Where $A = U(m_1) \cap U(m_2) = \{u\colon\ u \text{ has rated both } m_1 \text{ and } m_2\ \}$

Calculating Similarity

- Want to calculate the similarity between m_* and all

  $m_i \in M(u_*)$

- Break into subproblems. For each $m_i \in M(u_*)$

  1. Calculate $\sum_{u \in A}(R_{m_*,u} - \bar{R}_u)(R_{m_i,u} - \bar{R}_u)$
  2. Calculate $\sum_{u \in A}(R_{m_*,u} - \bar{R}_u)^2$
  3. Calculate $\sum_{u \in A}(R_{m_i,u} - \bar{R}_u)^2$

Step 1

- Calculate $\sum_{u \in A} (R_{m_*,u} - \bar{R_u})(R_{m_i,u} - \bar{R_u})$
  - Multiply adjusted rating matrix by row vector of ratings for m_*

$$\begin{array}{cc} & \begin{array}{ccc} u_1 & u_2 & u_* \end{array} \\ \begin{array}{c} m_* \\ m_1 \\ m_2 \\ m_3 \end{array} & \begin{pmatrix} 2 - \bar{u_1} & 5 - \bar{u_2} & 0 \\ 3 - \bar{u_1} & 0 & 3 - \bar{u_*} \\ 0 & 4 - \bar{u_2} & 2 - \bar{u_*} \\ 3 - \bar{u_1} & 0 & 4 - \bar{u_*} \end{pmatrix} \end{array} \times \begin{pmatrix} 2 - \bar{u_1} & 5 - \bar{u_2} & 0 \end{pmatrix}^T$$

Step 1 (Cont.)

Resulting Vector

$$
\begin{array}{c}
m_*, m_* \\
m_1, m_* \\
m_2, m_* \\
m_3, m_*
\end{array}
\begin{pmatrix}
(2-\bar{u_1})^2 + (5-\bar{u_2})^2 \\
(3-\bar{u_1})(2-\bar{u_1}) \\
(4-\bar{u_2})(5-\bar{u_2}) \\
(3-\bar{u_1})(2-\bar{u_1})
\end{pmatrix}
$$

- Each entry is the sum of products of adjusted ratings between m_* and $m_i \in M(u_*)$
- The only movie ratings considered are from users that have rated both m_* and m_i
- Each entry equal to $\sum_{u \in A}(R_{m_*,u} - \bar{R}_u)(R_{m_i,u} - \bar{R}_u)$

Step 2

Calculate $\sum_{u \in A}(R_{m_*,u} - \bar{R_u})^2$

- Transpose the adjusted rating matrix: R^T
- Transform the vector of ratings m_* into a matrix of same shape as R^T such that each column represents m_*: M_*
- Multiply M_* by R^T elementwise, then divide by R^T elementwise

Step 2 (Cont.)

$$
\begin{array}{cc}
 & \begin{array}{cccc} m_* & m_* & m_* & m_* \end{array} \\
M_* \quad \begin{array}{c} u_1 \\ u_2 \\ u_* \end{array} & \begin{pmatrix} 2-\bar{u}_1 & 2-\bar{u}_1 & 2-\bar{u}_1 & 2-\bar{u}_1 \\ 5-\bar{u}_2 & 5-\bar{u}_2 & 5-\bar{u}_2 & 5-\bar{u}_2 \\ 0 & 0 & 0 & 0 \end{pmatrix} *
\end{array}
$$

$$
\begin{array}{cc}
 & \begin{array}{cccc} m_* & m_1 & m_2 & m_3 \end{array} \\
R^T \quad \begin{array}{c} u_1 \\ u_2 \\ u_* \end{array} & \begin{pmatrix} 2-\bar{u}_1 & 3-\bar{u}_1 & 0 & 3-\bar{u}_1 \\ 5-\bar{u}_2 & 0 & 4-\bar{u}_2 & 0 \\ 0 & 3-\bar{u}_* & 2-\bar{u}_* & 4-\bar{u}_* \end{pmatrix} \div
\end{array}
$$

$$
\begin{array}{cc}
 & \begin{array}{cccc} m_* & m_1 & m_2 & m_3 \end{array} \\
R^T \quad \begin{array}{c} u_1 \\ u_2 \\ u_3 \end{array} & \begin{pmatrix} 2-\bar{u}_1 & 3-\bar{u}_1 & 0 & 3-\bar{u}_1 \\ 5-\bar{u}_2 & 0 & 4-\bar{u}_2 & 0 \\ 0 & 3-\bar{u}_* & 2-\bar{u}_* & 4-\bar{u}_* \end{pmatrix}
\end{array}
$$

Step 2 (Cont.)

$$
= \begin{array}{c} \\ u_1 \\ u_2 \\ u_* \end{array}
\begin{array}{c}
\begin{array}{cccc} m_*{\sim}m_* & m_*{\sim}m_1 & m_*{\sim}m_2 & m_*{\sim}m_3 \end{array} \\
\begin{pmatrix} 2-\bar{u_1} & 2-\bar{u_1} & 0 & 2-\bar{u_1} \\ 5-\bar{u_2} & 0 & 5-\bar{u_2} & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}
\end{array}
$$

In the resulting matrix, entry (u_i, m_i) is the adjusted rating of user u_i on movie m_* such that u_i has rated both m_i and m_*, 0 otherwise.

- Now multiply the row vector of ratings for movie m_* by the resulting matrix

Step 2 (Cont.)

$$
\begin{pmatrix} 2-\bar{u_1} & 5-\bar{u_2} & 0 \end{pmatrix} *
\begin{array}{c c}
 & \begin{array}{cccc} m_*\sim m_* & m_*\sim m_1 & m_*\sim m_2 & m_*\sim m_3 \end{array} \\
\begin{array}{c} u_1 \\ u_2 \\ u_* \end{array} & \begin{pmatrix} 2-\bar{u_1} & 2-\bar{u_1} & 0 & 2-\bar{u_1} \\ 5-\bar{u_2} & 0 & 5-\bar{u_2} & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}
\end{array}
$$

$$
= \begin{pmatrix} (2-\bar{u_1})^2 + (5-\bar{u_2})^2 & (2-\bar{u_1})^2 & (5-\bar{u_2})^2 & (2-\bar{u_1})^2 \end{pmatrix}
$$

Resulting Vector

- Each entry represents the sum of squares of the adjusted ratings of m_* given by users that have rated both m_* and m_i
- Each entry equal to $\sum_{u \in A} (R_{m_*,u} - \bar{R}_u)^2$

Step 3

Calculate $\sum_{u \in A}(R_{m_i,u} - \bar{R_u})^2$

- Transform the vector of ratings m_* into a matrix of same shape as R^T such that each column represents m_*: M_*
- Transpose the adjusted rating matrix: R^T
- Multiply R^T by M_* elementwise, then divide by M_* elementwise

Step 3 (Cont.)

$$
R_T \quad \begin{array}{c} \\ u_1 \\ u_2 \\ u_* \end{array} \begin{array}{c} \begin{array}{cccc} m_* & m_1 & m_2 & m_3 \end{array} \\ \begin{pmatrix} 2-\bar{u_1} & 3-\bar{u_1} & 0 & 3-\bar{u_1} \\ 5-\bar{u_2} & 0 & 4-\bar{u_2} & 0 \\ 0 & 3-\bar{u_*} & 2-\bar{u_*} & 4-\bar{u_*} \end{pmatrix} \end{array} *
$$

$$
M_* \quad \begin{array}{c} \\ u_1 \\ u_2 \\ u_* \end{array} \begin{array}{c} \begin{array}{cccc} m_* & m_* & m_* & m_* \end{array} \\ \begin{pmatrix} 2-\bar{u_1} & 2-\bar{u_1} & 2-\bar{u_1} & 2-\bar{u_1} \\ 5-\bar{u_2} & 5-\bar{u_2} & 5-\bar{u_2} & 5-\bar{u_2} \\ 0 & 0 & 0 & 0 \end{pmatrix} \end{array} \div
$$

$$
M_* \quad \begin{array}{c} \\ u_1 \\ u_2 \\ u_* \end{array} \begin{array}{c} \begin{array}{cccc} m_* & m_* & m_* & m_* \end{array} \\ \begin{pmatrix} 2-\bar{u_1} & 2-\bar{u_1} & 2-\bar{u_1} & 2-\bar{u_1} \\ 5-\bar{u_2} & 5-\bar{u_2} & 5-\bar{u_2} & 5-\bar{u_2} \\ 0 & 0 & 0 & 0 \end{pmatrix} \end{array} *
$$

Step 3 (Cont.)

$$
= \begin{array}{c} \\ u_1 \\ u_2 \\ u_* \end{array}
\begin{array}{c}
\begin{array}{cccc} m_* \sim m_* & m_1 \sim m_* & m_2 \sim m_* & m_3 \sim m_* \end{array} \\
\begin{pmatrix} 2 - \bar{u}_1 & 3 - \bar{u}_1 & 0 & 3 - \bar{u}_1 \\ 5 - \bar{u}_2 & 0 & 4 - \bar{u}_2 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}
\end{array}
$$

In the resulting matrix, entry (u_i, m_i) is the adjusted rating of user u_i on movie m_i such that u_i has rated both m_i and m_*, 0 otherwise.

- Now multiply the rating matrix by the above matrix

Step 3 (Cont.)

$$
\begin{array}{c|c}
 & \begin{array}{ccc} u_1 & u_2 & u_* \end{array} \\
\begin{array}{c} m_* \\ m_1 \\ m_2 \\ m_3 \end{array} & \begin{pmatrix} 2-\bar{u}_1 & 5-\bar{u}_2 & 0 \\ 3-\bar{u}_1 & 0 & 3-\bar{u}_* \\ 0 & 4-\bar{u}_2 & 2-\bar{u}_* \\ 3-\bar{u}_1 & 0 & 4-\bar{u}_* \end{pmatrix} \times
\end{array}
$$

$$
\begin{array}{c|c}
 & \begin{array}{cccc} m_*{\sim}m_* & m_1{\sim}m_* & m_2{\sim}m_* & m_3{\sim}m_* \end{array} \\
\begin{array}{c} u_1 \\ u_2 \\ u_* \end{array} & \begin{pmatrix} 2-\bar{u}_1 & 3-\bar{u}_1 & 0 & 3-\bar{u}_1 \\ 5-\bar{u}_2 & 0 & 4-\bar{u}_2 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}
\end{array}
$$

Step 3 (Cont.)

$$
\begin{array}{c|cccc}
 & m_* & m_1 & m_2 & m_3 \\
\hline
m_* & (2-\bar{u_1})^2+(5-\bar{u_2})^2 & - & - & - \\
m_1 & - & (3-\bar{u_1})^2 & - & - \\
m_2 & - & - & (4-\bar{u_2})^2 & - \\
m_3 & - & - & - & (3-\bar{u_1})^2
\end{array}
$$

Resulting Matrix

- Entries along the diagnal represent the sum of squares of the adjusted ratings of m_i given by users that have rated both m_* and m_i
- Entries along the diagnal represent $\sum_{u \in A}(R_{m_i,u} - \bar{R}_u)^2$

Prediction

- Able to calculate $sim(m_*, m_i)$ for all $m_i \in M(u_*)$ in a few simple matrix operations
- For all $m_i \in M(u_*)$
  - Step 1 provides us $\sum_{u \in A}(R_{m_*,u} - \bar{R}_u)(R_{m_i,u} - \bar{R}_u)$
  - Step 2 proviedes us $\sum_{u \in A}(R_{m_*,u} - \bar{R}_u)^2$
  - Step 3 provideees us $\sum_{u \in A}(R_{m_i,u} - \bar{R}_u)^2$ along the diagnals of a matrix
  - Combine to get $sim(m_*, m_i) = \frac{\sum_{u \in A}(R_{m_*,u} - \bar{R}_u)(R_{m_i,u} - \bar{R}_u)}{\sum_{u \in A}(R_{m_*,u} - \bar{R}_u)^2 \sum_{u \in A}(R_{m_i,u} - \bar{R}_u)^2}$
- Now can choose the k movies most similar to m_* (that are rated by u_*) to predict R_{m_*,u_*}

Prediction

Predict rating of user u for movie m by taking weighted average of the ratings given by the k movies most similar to m that have rating from u.

$$P_{m,u} = \frac{\sum_{m' \in N_u^K(m)} R_{m',u} sim(m, m')}{\sum_{m' \in N_u^K(m)} |sim(m, m')|}$$

$$N_u^K(m) = \{m' : m' \text{ belongs to the } k \text{ most similar movies of } m \text{ and } u \text{ has rated } m'\}$$