ASSIGNMENT 3 PRESENTATION

JINGXUAN FENG S3843790

SUGGESTED APPROACH TO COLLABORATIVE FILTERING (CF)

- Combining user-based and item-based approaches
- Why combining the two approaches?
- Use correlation inside a user-item matrix
- Possible issue of either similar users or items or both at the same time

LOCATING SIMILAR NEIGHBOURS

- First step is to locate similar neighbours for both users and items using the following:

$$Sim(a,u) = \frac{\sum_{i \in I(a) \cap I(u)} (r_{a,i} - \overline{r}_a) \cdot (r_{u,i} - \overline{r}_u)}{\sqrt{\sum_{i \in I(a) \cap I(u)} (r_{a,i} - \overline{r}_a)^2} \cdot \sqrt{\sum_{i \in I(a) \cap I(u)} (r_{u,i} - \overline{r}_u)^2}}, \quad (1)$$

$$Sim(i,j) = \frac{\sum_{u \in U(i) \cap U(j)} (r_{u,i} - \overline{r}_i) \cdot (r_{u,j} - \overline{r}_j)}{\sqrt{\sum_{u \in U(i) \cap U(j)} (r_{u,i} - \overline{r}_i)^2} \cdot \sqrt{\sum_{u \in U(i) \cap U(j)} (r_{u,j} - \overline{r}_j)^2}}, \quad (2)$$

- Significance rating between 0 – 1 & corresponding equations:

$$Sim'(a,u) = \frac{Min(|I_a \cap I_u|, \gamma)}{\gamma} \cdot Sim(a,u), \quad (4)$$

$$Sim'(i,j) = \frac{Min(|U_i \cap U_j|, \delta)}{\delta} \cdot Sim(i,j), \quad (5)$$

OVERCOMING FLAWS OF TOP-N NEIGHBOUR SELECTION

- If selected neighbours are dissimilar with the current user, prediction of missing data will be inaccurate.
- Threshold eta is used to overcome the flaws of Top-N neighbour selection

$$S(u) = \{u_a | Sim'(u_a, u) > \eta, u_a \neq u\}, \tag{6}$$

$$S(i) = \{i_k | Sim'(i_k, i) > \theta, i_k \neq i\}, \tag{7}$$

FURTHER FORMULAS USED

- Where S(u) = 0, the missing value is calculated by:

$$P(r_{u,i}) = \bar{i} + \frac{\sum_{i_k \in S(i)} Sim'(i_k, i) \cdot (r_{u,i_k} - \bar{i}_k)}{\sum_{i_k \in S(i)} Sim'(i_k, i)}. \qquad (10)$$

- If the length of $S(i)$ is equal to zero, the missing value can be calculated by:

$$P(r_{u,i}) = \bar{u} + \frac{\sum_{u_a \in S(u)} Sim'(u_a, u) \cdot (r_{u_a,i} - \bar{u}_a)}{\sum_{u_a \in S(u)} Sim'(u_a, u)}. \qquad (9)$$

- If both $S(i)$ and $S(u)$ are not equal to zero, the missing value can be by:

$$P(r_{u,i}) = \lambda \times (\bar{u} + \frac{\sum_{u_a \in S(u)} Sim'(u_a, u) \cdot (r_{u_a,i} - \bar{u}_a)}{\sum_{u_a \in S(u)} Sim'(u_a, u)}) + (1 - \lambda) \times (\bar{i} + \frac{\sum_{i_k \in S(i)} Sim'(i_k, i) \cdot (r_{u,i_k} - \bar{i}_k)}{\sum_{i_k \in S(i)} Sim'(i_k, i)}), \qquad (8)$$

- If both $S(i)$ and $S(u)$ are equal to zero the prediction of missing data $P(r_{u,i})$ is defined as follows:

$$P(r_{u,i}) = 0. \qquad (11)$$

WHY IS THE SOLUTION PROVIDED VIABLE?

- The method detailed in the report can tackle the missing value problem in CF
- With only 5 parameters, it is possible to tune these parameters
- Alleviates the potential negative influences from bad predictions

MY CODE IMPLEMENTATION

- Implemented the equations using some methods from Numpy
- 2 methods to calculate the similarity score between two users and two items
  - Sim_U (a , u , GAMMA)
  - Sim_I (I , j , DELTA)
- 2 methods to find similar users and items
  - Sim_users (u_indx , GAMMA , ITA)
  - Sim_items (i_indx , DELTA , THETA)
- If / else statements used to implement equations 8-11
  - P_u (u_indx , i_indx , GAMMA , ITA)
  - P_i (u_indx , i_indx , DELTA , THETA)
  - P_ui (u_indx , i_indx , LAMBDA , DELTA , THETA , GAMMA , ITA)
- Searching the 2D array for missing value using 2 for loops.

REFERENCES

- [1] H. Ma, I. King and M. R. Lyu, "Effective Missing Data Prediction for Collaborative Filtering", p. 8, 2007. [Accessed 11 June 2021].
- [2] H. Singh, "Methods in Python - A Key Concept of Object Oriented Programming", Analytics Vidhya, 2020. [Online]. Available: https://www.analyticsvidhya.com/blog/2020/11/basic-concepts-object-oriented-programming-types-methods-python/. [Accessed: 12- Jun- 2021].