(TUẦN 5)

Đề tài: Tìm hiểu các thuật toán Recommendation

Link github: [Recommendation]

❖ Kết quả đạt được

❖ Thuật toán memory-based:

- Hay còn gọi là phương pháp láng giềng (neighborhood-based)
- Xây dựng dự đoán bằng cách sử dụng các bộ các sản phẩm đã được đánh giá trước bởi người dùng.
- Cần phải tính toán một vài độ đo tương tự như là Pearson correlation coefficient, sau đó xác định hàng xóm của người dùng, hệ thống sử dụng thuật toán khác để kết hợp sở thích của các hàng xóm để tính toán đề nghị cho hoạt động người dùng. Sử dụng toàn bộ dữ liệu đánh giá để đưa ra dự đoán.
- Lọc dựa trên user và Lọc dựa trên item là hai cách để tiếp cận lọc cộng tác dựa trên **memory-based**

- Phương pháp giải quyết bài toán:

Để giải quyết bài toán neighborhood-based cần phải qua một vài bước, nhóm em xin phép mượn hình ảnh minh họa từ [blog machinelearningcoban]:

	u_0	u_1	u_2	u_3	u_4	u_5	u_6
i_0	5	5	2	0	1	?	?
i_1	4	?	?	0	?	2	?
i_2	?	4	1	?	?	1	1
i_3	2	2	3	4	4	?	4
i_4	2	0	4	?	?	?	5
	↓	↓	↓	↓	↓	↓	↓
$\bar{u}_j$	3.25	2.75	2.5	1.33	2.5	1.5	3.33

a) Original utility matrix $\mathbf{Y}$ and mean user ratings.

	u_0	u_1	u_2	u_3	u_4	u_5	u_6
i_0	1.75	2.25	-0.5	-1.33	-1.5	0	0
i_1	0.75	0	0	-1.33	0	0.5	0
i_2	0	1.25	-1.5	0	0	-0.5	-2.33
i_3	-1.25	-0.75	0.5	2.67	1.5	0	0.67
i_4	-1.25	-2.75	1.5	0	0	0	1.67

b) Normalized utility matrix $\bar{\mathbf{Y}}$.

	u_0	u_1	u_2	u_3	u_4	u_5	u_6
u_0	1	0.83	-0.58	-0.79	-0.82	0.2	-0.38
u_1	0.83	1	-0.87	-0.40	-0.55	-0.23	-0.71
u_2	-0.58	-0.87	1	0.27	0.32	0.47	0.96
u_3	-0.79	-0.40	0.27	1	0.87	-0.29	0.18
u_4	-0.82	-0.55	0.32	0.87	1	0	0.16
u_5	0.2	-0.23	0.47	-0.29	0	1	0.56
u_6	-0.38	-0.71	0.96	0.18	0.16	0.56	1

c) User similarity matrix $\mathbf{S}$.

	u_0	u_1	u_2	u_3	u_4	u_5	u_6
i_0	1.75	2.25	-0.5	-1.33	-1.5	0.18	-0.63
i_1	0.75	0.48	-0.17	-1.33	-1.33	0.5	0.05
i_2	0.91	1.25	-1.5	-1.84	-1.78	-0.5	-2.33
i_3	-1.25	-0.75	0.5	2.67	1.5	0.59	0.67
i_4	-1.25	-2.75	1.5	1.57	1.56	1.59	1.67

d) $\hat{\mathbf{Y}}$

Predict normalized rating of u_1 on i_1 with $k = 2$

Users who rated i_1 : $\{u_0, u_3, u_5\}$

Corresponding similarities: {0.83, -0.40, -0.23}

$\Rightarrow$ most similar users: $\mathcal{N}(u_1, i_1) = \{u_0, u_5\}$

with **normalized ratings {0.75, 0.5}**

$$\Rightarrow \hat{y}_{i_1,u_1} = \frac{0.83*0.75+(-0.23)*0.5}{0.83+|-0.23|} \approx 0.48$$

e) Example

	u_0	u_1	u_2	u_3	u_4	u_5	u_6
i_0	5	5	2	0	1	1.68	2.70
i_1	4	3.23	2.33	0	1.67	2	3.38
i_2	4.15	4	1	-0.5	0.71	1	1
i_3	2	2	3	4	4	2.10	4
i_4	2	0	4	2.9	4.06	3.10	5

f) Full $\mathbf{Y}$

a) Lấy trung bình các cột,
b) Chuẩn hóa ma trận bằng cách trừ đi trung bình,
c) Tính hệ số tương quan của ma trận chuẩn hóa,

d) Dự đoán trên ma trận chuẩn hóa các vị trí chưa được rate,
e) Diễn giải công thức dự báo rating
f) Chuyển đổi sang giá trị rating thực tế.

Demo Collaborative filtering theo hướng memory-based:

➢ Tìm hiểu về thuật toán SVD:

Một ma trận A_{mxn} có thể phân rã thành:

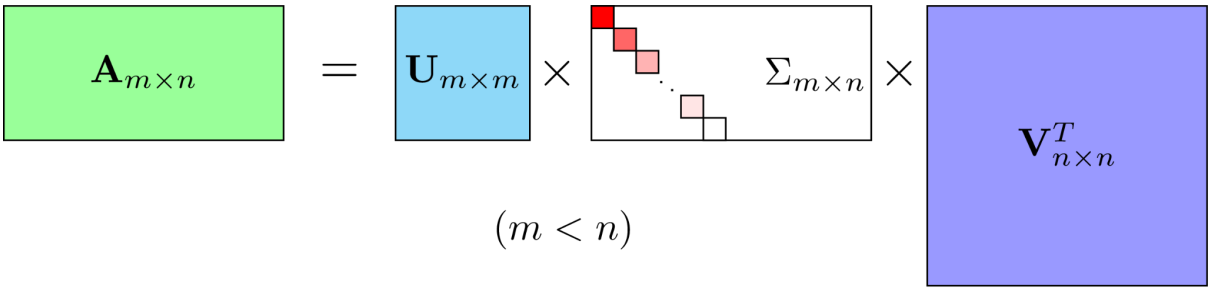
$$M = U \cdot \Sigma \cdot V^T$$

- U: ma trận m x m

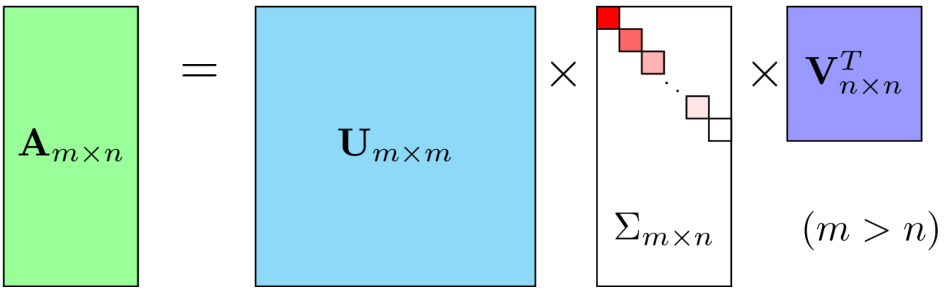
- Σ: ma trận đường chéo không vuông với các phần tử trên nó

- V: ma trận n x n

Mô tả SVD của ma trận A mxn, nhóm xin phép mượn hình ảnh của machinelearningcoban: (mô tả 2 trường hợp là m < n và m > n)



$(m < n)$



$(m > n)$

➢ **Ý nghĩa của SVD trong recommender system:**

Nếu ma trận M là hạng r, có thể chứng minh rằng các ma trận $M \cdot M^T$ và $M^T \cdot M$ đều có hạng r. Trong phân rã giá trị đơn lẻ (SVD giảm), các cột của ma trận U là các hàm riêng của $M \cdot M^T$ và các hàng của ma trận V^T là các yếu tố riêng của $M^T \cdot M$. Điều thú vị là ma trận $M \cdot M^T$ và $M^T \cdot M$ có thể có kích thước khác nhau (vì ma trận M có thể là hình dạng không vuông), nhưng chúng có cùng một tập hợp các giá trị riêng, là bình phương của các giá trị trên đường chéo của Σ.

=> SVD có thể cho thấy nhiều vấn đề từ ma trận M

Ví dụ: Thu thập được một số đánh giá sách là cột và người đọc là hàng, và trọng số là đánh giá của người đọc trên từng cuốn sách. Khi đó, $M \cdot M^T$ sẽ là bảng của person-to-person, nghĩa là tổng số đánh giá mà người này kết hợp với người khác. $M^T \cdot M$ là bảng book-to-book, nghĩa là tổng số xếp hạng nhận được kết hợp với cuốn sách khác nhận được.

➢ **Tập dữ liệu:**

- Link dataset: Ml-latest-small
- 100.000 lượt xếp hạng và 3.600 ứng dụng gắn thẻ được áp dụng cho 9.000 phim của 600 người dùng.

➢ **Demo:**

Link github code: Memory-based-using-SVD

Link video: Video-báo-cáo.