

데이터사이언스 기말고사 예시 문제

1. 주어진 Users-to-Ratings matrix를 활용하여 문제를 푸시오.

	I1	I2	I3	I4	I5
U1	5	4	3	3.25	1
U2	2	5	.	3	4
U3	3	5	4	.	.
U4	.	.	5	5	5
U5	1	2	3	4	5

$$\text{sim}(u_1, u_2) = \frac{1.75 \times -1.5 + 0.95 \times 1.5}{\sqrt{(3.0625 + 0.5625) \times (2.25 + 0.5)}} = -0.4086$$

$$\text{sim}(u_1, u_4) = 0$$

$$\text{sim}(u_1, u_5) = -0.9354$$

$$u' = (1.75, 0.95, -0.25, -2.25)$$

$$u'' = (-1.5, 1.5, -0.5, 0.5)$$

$$v' = (-1, 1, 0)$$

$$u''' = (0, 0, 0)$$

$$u'''' = (-2, -1, 0, 1.2)$$

- A. PCC (Pearson correlation coefficient) 유사도 수식을 쓰고, U1과 U3사이의 유사도를 계산 하시오.
- B. 사용자 기반의 "CF algorithm with mean" 알고리즘의 수식을 쓰고, U1의 I4에 대한 예측 선호도 값(Predicted Rating)을 측정 하시오. I4를 rating한 사용자 중 유사한 2명의 사용자 사용함

$$\text{PCC}(u, v) = \frac{\sum_{i: r_{u,i} \neq \text{null} \wedge r_{v,i} \neq \text{null}} (r_{u,i} - \bar{r}_u) \cdot (r_{v,i} - \bar{r}_v)}{\sqrt{\sum_{i: r_{u,i} \neq \text{null} \wedge r_{v,i} \neq \text{null}} (r_{u,i} - \bar{r}_u)^2} \sqrt{\sum_{i: r_{v,i} \neq \text{null} \wedge r_{u,i} \neq \text{null}} (r_{v,i} - \bar{r}_v)^2}}$$

$$= \frac{-1.75 + 0.95}{\sqrt{3.0625 + 0.5625} \sqrt{2.25 + 0.5}} = -0.2959$$

$$\text{sim}(U1, U2) = -0.4086, \text{sim}(U1, U4) = -1, \text{sim}(U1, U5) = -0.9860$$

$$3.25 + \frac{-0.4086 \times -0.5 - 0.9860}{(-0.4086 - 0.9860)} = 3.25 + 0.5605$$

$$\hat{r}_{u,i} = \mu_u + \frac{\sum_{v \in N_i^+(u)} \text{sim}(u, v) (r_{v,i} - \mu_v)}{\sum_{v \in N_i^+(u)} \text{sim}(u, v)} = 3.8105$$

$$\text{sim} : \text{PCC!}$$

$$\text{sim}(u_1, u_2) = -0.4086$$

$$\text{sim}(u_1, u_3) = -1$$

$$\text{sim}(u_1, u_4) = 0$$

$$\text{sim}(u_1, u_5) = -0.9354$$

2. PMF (Probabilistic Matrix Factorization) 수식을 쓰고, 기존 SVD (Singular Value Decomposition)의 문제를 PMF 수식 상에서 어떻게 해결하는지 서술 하시오. SVD는 2배 더 많은 문제.

$$\text{PMF} = \min \sum_{r_{u,i} \in R_{train}} y_{u,i} (r_{u,i} - \hat{r}_{u,i})^2 + \lambda (\|p_u\|^2 + \|q_i\|^2)$$

user based > collaborative filtering

3. Memory-based CF와 Model-based CF와 관련된 문제 입니다.

- A. Memory-based CF의 특징을 간단히 서술하고, Memory-based CF의 2가지 방식에 대한 차이점을 서술하시오.
- B. Model-based CF의 특징을 간단히 서술하고, Memory-based CF와 비교했을 때 Model-based CF의 장점을 2가지 이상 서술하시오.

User
: 다른 유저의 행렬을 이용해 점수 예측

Item
: U는 Item 정보 이용해 점수 예측!

Memory bCF: ① 가장 유사한 사용자/상품 기반의 추천. 추천된 라트코. → 전파에 대한 설명 가능.

* Memory 단점

① 데이터 sparse → 성능 저하 심함.

데이터 늘면
sim 측정 다함.
+ 많아질.
O(N^2) + d

② 데이터가 커질수록 유연하게 적용이 힘들.

② 사용자 기반 기법.

③ 새로운 데이터에 대한 이용도 낮음.

④ 데이터 종류에 관계없이 추천 가능.

sim 순위 높은 상위 유저/item 행렬 바탕으로 추천 제공

sum 측정비용 (시간) 절약.
* Model 비해 정확성 많이 떨어짐.

= 5개인

* Model 광범 (Memory 대비)

→ 예측 정확도 향상

→ 차원 축소로 연산량 감소

→ implicit 레이어 모델에 추가 됨.

4. 추천 시스템의 성능 평가와 관련된 문제입니다.

	I1	I2	I3	I4	I5	I6	I7	I8	I9	I10
U1	5	4	3	.	1	3	4	2	1	5

- U1의 평점 행렬 -

실제 만족 4개

	I1	I2	I3	I4	I5	I6	I7	I8	I9	I10
U1	4.5	5.3	3.2	1.4	2.3	2.2	4.6	2.1	1.9	3.8

- Algorithm1을 통해 생성된 예측 평점 -

	I1	I2	I3	I4	I5	I6	I7	I8	I9	I10
U1	2.3	2.5	3.5	0.5	1.5	4.8	4.3	4.1	2.3	4.8

- Algorithm2를 통해 생성된 예측 평점 -

A. Precision과 Recall의 차이를 서술하시오.

threshold 사용. (1~5 정도이면
4~5)

여기서 thres.

Precision = 추천된 상품 (n개) 중 실제 사용자와 만족하는 비율.

Recall = 실제 사용자와 좋아하는 상품 중

B. K-랭크의 Precision을 구하는 것을 P@K (Precision at K)라고 한다. Algorithm 1과 2의

추천결과에 존재하는

P@5를 구하시오. (Threshold = 4, 즉 4 이상의 item이 추천 되었을 때 제대로 추천 된 것으로 본다.)

상품.

= 1. relevant.

AI 1: 5

AI 2: 0

C. Algorithm1과 2의 5등까지의 NDCG를 구하시오 (NDCG@5).

AI 1: 2, 7, 1, 10, 3, 5, 6, 8, 9, 4 → $r_1=4$ $r_2=4$ $r_3=5$ $r_4=5$ $r_5=3$

AI 2: 6, 10, 7, 8, 3, 9, 2, 1, 5, 4 → $r_1=3$ $r_2=5$ $r_3=4$ $r_4=2$ $r_5=3$

D. Algorithm1과 Algorithm2 중 어떤 알고리즘이 더 좋은 알고리즘이라고 할 수 있는가?

AI 1

$$AI 1 = 4 + \frac{4}{\log_2 2} + \frac{5}{\log_2 3} + \frac{5}{\log_2 4} + \frac{3}{\log_2 5} = 12.6694$$

$$AI 2 = 3 + \frac{5}{\log_2 2} + \frac{4}{\log_2 3} + \frac{2}{\log_2 4} + \frac{3}{\log_2 5} = 9.9691$$

$$IDCG = 5 + \frac{5}{\log_2 2} + \frac{4}{\log_2 3} + \frac{4}{\log_2 4} + \frac{3}{\log_2 5} = 12.9691$$

NDCG@5 < AI 1: 0.9769

AI 2: 0.7687

DCG / IDCG

recall

AI 1: 3/5

AI 2: 4/5

precision

AI 1:

AI 2:

5. Group Recommendation 관련 문제입니다.

	I1	I2	I3	I4	I5
U1	5	4	3	.	1
U2	2	5	.	3	4
U3	3	5	4	.	.
U4	.	.	5	5	5
U5	1	2	3	4	5

add mean
 11 16 15 12 15
 2.75 4 3.75 4 3.75

A. Additive Utilitarian을 통해 그룹 선호도를 측정 하시오.

I1 I2 I3 I4 I5
 score 11 16 15 12 15
 ranking 5 1 2 4 2

B. Average를 통해 그룹 선호도를 측정 하시오. Average가 똑 같은 상품이 있다면 그 2개의 상품 중 어떤 상품을 해당 그룹에 더 높은 순위로 측정해야 하는지 설명하시오.

I1 I2 I3 I4 I5
 score 2.75 4 3.75 4 3.75
 ranking 5 1 3 1 3

평균 랭크 랭킹이 같은 상품은 Average without misery 적용 가능.

I2, I4에서 threshold 2.3 값이면, I2 랭크가 4.67로 1등.

I2 추천.

