데이터사이언스 기말고사 예시 문제

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

	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

A. PCC (Pearson correlation coefficient) 유사도 수식을 쓰고, U1과 U3사이의 유사도를 계산하시오.

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

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 사용자를 사용함

sim(U1, U2) = -0.4086, sim(U1, U4) = -1, sim(U1, U5) = -0.9860

$$P_{u,i} = \bar{u} + \frac{\sum_{v \in N_i^k(u)} sim(u,v)(r_{v,i} - \bar{v})}{\sum_{v \in N_i^k(u)} sim(u,v)}$$

2. PMF (Probabilistic Matrix Factorization) 수식을 쓰고, 기존 SVD (Singular Value Decomposition)의 문제를 PMF 수식 상에서 어떻게 해결하는지 서술하시오.

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

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.

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 -

	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.

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

Recall = 실제 사용자가 좋아하는 상품 중 추천리스트에 포함되는 상품.

B. K-랭크의 Precision을 구하는 것을 P@K (Precision at K)라고 한다. Algorithm 1과 2의 P@5를 구하시오. (Threshold = 4, 즉 4 이상의 item이 추천 되었을 때 제대로 추천 된 것으로 본다.)

Al 1 : $\frac{2}{5}$

Al 2 : 0

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

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

Al 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?

DCG

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

Al 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 < Al 1 : 0.9769

Al 2 : 0.7687

$\frac{DCG}{IDCG}$

Al 1

recall

Al 1 : $\frac{3}{10}$

Al 2 : $\frac{4}{10}$

precision

Al 1 :

Al 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	11	16	15	12	15
mean	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 점수 랭킹이 같은 상품은 Average without misery 적용 가능.

I2, I4에서 threshold 2.5 잡으면, I_2 점수가 4.67로 높음.

I_2 추천.