
추천시스템

CF와 MF

김지선

Contents

- 추천이란?
- 유사도의 종류
- 협업 필터링(CF/MF)
- 무비렌즈(MovieLense) 데이터로 CF 실습

추천의 활용 사례

- 쿠팡의 물건 추천
- 넷플릭스의 콘텐츠 추천
- 유튜브의 추천 알고리즘
- 결측치 처리

유사도의 종류

- 자카드 유사도
- 코사인 유사도
- 유클리디안 유사도

자카드 유사도

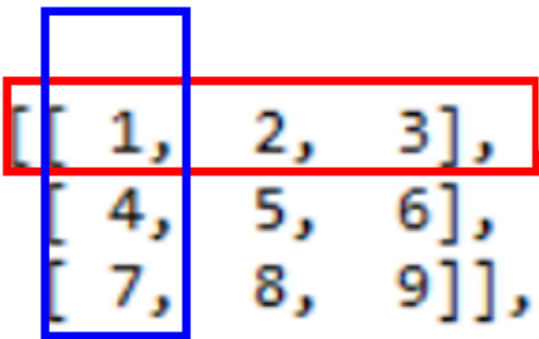
- 가장 간단한 유사도 계산 방식으로
합집합 중 교집합의 비율로 유사도를 측정

$$J(A, B) = \frac{|A \cap B|}{|A \cup B|} = \frac{|A \cap B|}{|A| + |B| - |A \cap B|}.$$

What is Vector?

- 컴퓨터에서 벡터란 1차원 배열의 데이터
- 보통 한 행 또는 한 열을 뜻함

```
array([[ 1,  2,  3],  
       [ 4,  5,  6],  
       [ 7,  8,  9]])
```



The diagram illustrates a 3x3 array. A red rectangular box highlights the first row, containing the values [1, 2, 3]. A blue rectangular box highlights the first column, containing the values [1, 4, 7].

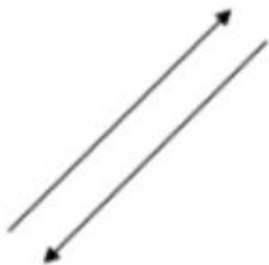
Row Vector

Column Vector

코사인 유사도

- 두 개의 벡터 값에서 코사인 각도로 유사도를 측정하는 방법
- 범위는 -1에서 1사이의 값

$$\text{similarity} = \cos(\theta) = \frac{A \cdot B}{\|A\| \|B\|}$$



코사인 유사도 : -1



코사인 유사도 : 0

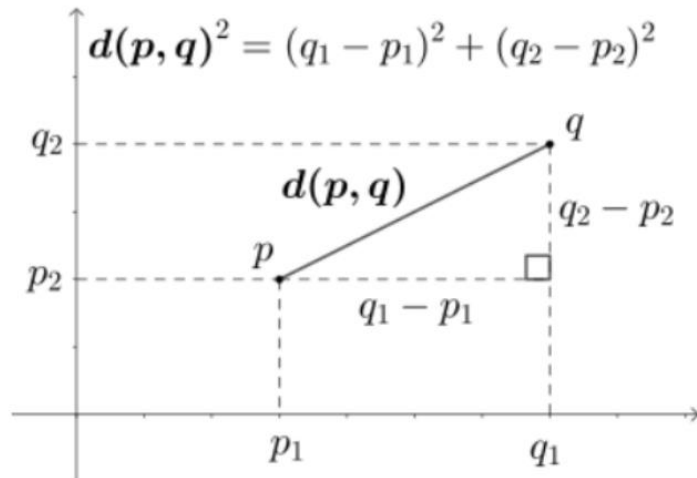


코사인 유사도 : 1

유클리디안 유사도

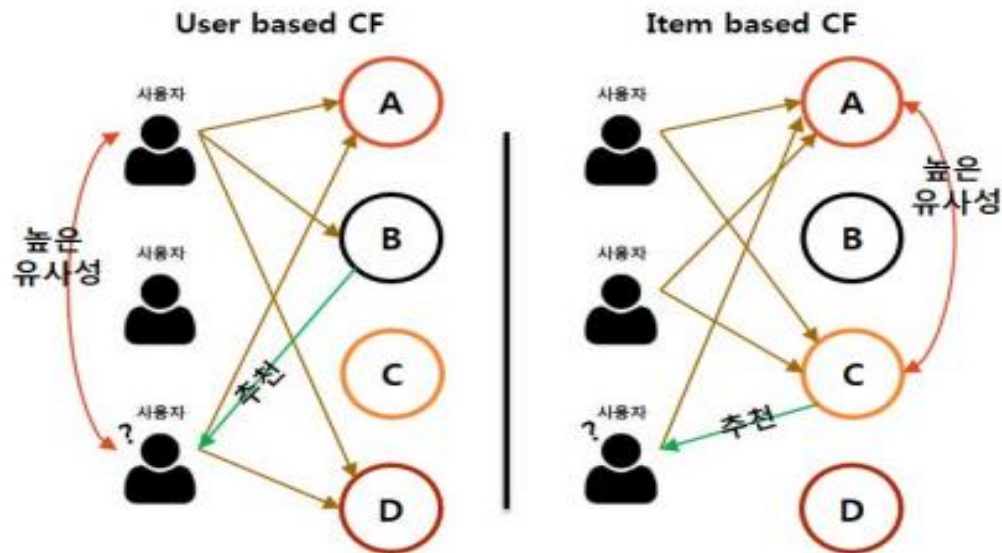
- 유클리디안 거리라고도 불림
- 두 점 사이의 거리를 구하기 위해 피타고라스 정리 이용

$$L_2 = \sqrt{(p_1 - q_1)^2 + (p_2 - q_2)^2 + \dots + (p_n - q_n)^2} = \sqrt{\sum_{i=1}^n (p_i - q_i)^2}$$



협업 필터링이란

- 협업필터링(CF, Collaborative Filtering)은
- 나와 비슷한 취향의 사람들이 좋아하는 것은 나도 좋아할 가능성이 높다는 개념을 이용



협업 필터링 종류

- Memory Based Approach
 - Item-based CF
 - User-based CF
- Model Based Approach
 - Matrix-Factorization (행렬 분해)
 - Neural Network (신경망)

Memory Based Approach

What is User and Item?

- 유사도를 기반으로 추천을 할 때
 - User가 기준이면 User-based CF
 - Item이 기준이면 Item-based CF
 - Item은 보통 칼럼이 됨

Model Based Approach

MF(Matrix Factorization)

- MF란 행렬 분해를 이용해 결측치를 예측하는 추천방식이다.
- Netflix에서 높은 성능으로 유명해짐

Diagram illustrating Matrix Factorization (MF) equation:

$$R \approx \hat{R} = U \cdot \Sigma \cdot V^T$$

Dimensions and components:

- R : $m \times n$
- \hat{R} : $m \times n$
- U (User Latent Matrix): $m \times k$
- Σ : $k \times k$
- V^T (Item Latent Matrix): $k \times n$

Legend:

- m : User 수
- n : Item 수
- k : 잠재 벡터 크기

MF의 동작방식

- 아래 목적함수를 최소화 시키는 방향으로 SGD사용하여 학습
- SGD(Stochastic Gradient Descent)

$$\min_{P, Q} \sum_{\text{observed } r_{u,i}} (r_{u,i} - \hat{r}_{u,i})^2 + \lambda(\|p_u\|^2 + \|q_i\|^2)$$

Matrix Factorization

- 원본 Rating Matrix를 두 행렬 P , Q 로 분해
- 두 행렬은 초기엔 Null값/Random 값으로 채움
- 두 행렬을 내적인 값이 원본에 가까워지도록 함
 - Gradient Dscent에서 Cost를 측정
 - SGD(Sta~ Gradient Dscent)를 이용하여 P , Q 를 구함
 - K 는 하이퍼 파라미터.
- 결국 구한 P , Q 행렬을 곱하면 빈 값의 예측 가능

무비렌즈 데이터로 영화 추천 실습

<https://grouplens.org/datasets/movielens/latest/>

무비렌즈 데이터로 영화 추천 실습

- 추천시스템 웹페이지 HTML/CSS