[Week15] 파머완 9장

CHAP 09 추천시스템

01 추천 시스템의 개요와 배경

추천시스템 유형

- 콘텐츠 기반 필터링 방식
- 협업 필터링 방식
    - 최근접 이웃 협업 필터링
    - 잠재 요인 협업 필터링

02 콘텐츠 기반 필터링 추천 시스템

- 사용자가 특정한 아이템을 매우 선호하는 경우, 그 아이템과 비슷한 콘텐츠를 가진 다른 아이템을 추천하는 방식

03 최근접 이웃 협업 필터링

- 협업 필터링 : 사용자가 아이템에 매긴 평점 정보나 상품 구매 이력과 같은 사용자 행동 양식만을 기반으로 추천을 수행하는 것
- 사용자-아이템 평점 매트릭스와 같은 축적된 사용자 행동 데이터를 기반으로 사용자가 아직 평가하지 않은 아이템을 예측 평가하는 것
- 평가한 다른 아이템을 기반으로 평가하지 않은 아이템의 예측 평가를 도출 ⇒ 사용자-아이템 평점 행렬 데이터에 의지
- 사용자-아이템 평점 행렬
  : 행(개별 사용자), 열(개별 아이템), 값(평점) → 판다스 pivot_table()
  : 다차원 행렬 & 희소 행렬
- 최근접 이웃 협업 필터링 = 메모리 협업 필터링
  - 사용자 기반 : 특정 사용자와 유사한 다른 사용자를 TOP-N으로 선정해 이 TOP-N 사용자가 좋아하는 아이템 추천 → 특정 사용자와 타 사용자 간 유사도 기반
  - 아이템 기반 : 아이템 속성과는 상관없이 사용자들이 아이템을 좋아하는지/싫어하는지

의 평가 척도가 유사한 아이템을 추천 → 행과 열 서로 반대
⇒ 아이템 기반 정확도가 더 높음

04 잠재 요인 협업 필터링

잠재 요인 협업 필터링의 이해

- 사용자-아이템 평점 매트릭스 속에 숨어 있는 잠재요인을 추출해 추천 예측을 할 수 있게 하는 기법
- 행렬 분해 : 대규모 다차원 행렬을 SVD와 같은 차원 감소 기법으로 분해하는 과정에서 잠재요인을 추출하는 방법
- 잠재요인을 기반으로 저차원 밀집 행렬의 사용자-잠재요인 행렬과 잠재요인-아이템 행렬로 분해 → 내적 곱으로 결합을 통해 예측 평점 생성

행렬 분해의 이해

- 다차원의 매트릭스를 저차원으로 분해하는 기법 → SVD, NMF 등
- 평점 행렬 R = 사용자 행 M X 아이템 열 N→ 사용자-잠재요인행렬 P = M X K & 잠재요인-아이템행렬 Q.T = K X N ** K = 잠재요인의 차원 수
- R = P * Q.T

확률적 경사 하강법을 이용한 행렬 분해

- P와 Q 행렬로 계산된 예측 R 행렬값이 실제 R 행렬값과 가장 최소의 오류를 가질 수 있도록 반복적인 비용 함수 최적화를 통해 P와 Q를 유추해내는 것
- $\acute{p_u} = p_u + \eta(e_{(u,i)} * q_i - \lambda * p_u)$
- $\acute{q_i} = q_i + \eta(e_{(u,i)} * p_u - \lambda * q_i)$
  - p_u : P 행렬의 사용자 u행 벡터
  - q_i^t : Q 행렬의 아이템 i행의 전치 벡터
  - $r_{(u,i)}$: 실제 R 행렬의 u행, i열에 위치한 값
  - $\hat{r}_{(u,i)}$: 예측 $\hat{R}$ 행렬의 u행, i열에 위치한 값 = $p_u * q_i^t$
  - $e_{(u,i)}$: u행, i열에 위치한 실제 행렬 값과 예측 행렬 값의 차이 오류 = $r_{(u,i)} - \hat{r}_{(u,i)}$

    - η : SGD 학습률
    - λ : L2 규제 계수
- L2 규제를 반영해 실제 R 행렬 값과 예측 R 행렬 값의 차이를 최소화하는 방향성을 가지고 P행렬과 Q행렬에 업데이트 값을 반복적으로 수행하면서 최적화된 예측 R 행렬을 구하는 방식

05 콘텐츠 기반 필터링 실습 - TMDB 5000 영화 데이터 세트

장르 콘텐츠 유사도 측정

- 문자열로 변환된 칼럼을 count 기반 피처 벡터화
- 피처 벡터화 행렬로 변환한 데이터 세트를 코사인 유사도를 통해 비교
- 유사도가 높은 영화 중 평점이 높은 순으로 영화 추천

06 아이템 기반 최근접 이웃 협업 필터링 실습

- 아이템 기반 협업 필터링에서 개인화된 예측 평점을 구하는 식

$$\hat{R}_{u,i} = \sum^{N}(S_{i,N} * R_{u,N}) / \sum^{N}(|S_{i,N}|)$$

- $\hat{R}_{u,i}$: 사용자 u, 아이템 i의 개인화된 예측 평점 값
- $S_{i,N}$: 아이템 i와 가장 유사도 높은 Top-N개 아이템의 유사도 벡터
- $R_{u,N}$: 사용자 u의 아이템 i와 가장 유사도가 높은 Top-N개 아이템에 대한 실제 평점 벡터
- N : 아이템의 최근접 이웃 범위 계수 → 특정 아이템과 유사도가 가장 높은 다른 아이템 추출에 사용

07 행렬 분해를 이용한 잠재 요인 협업 필터링 실습

08 파이썬 추천 시스템 패키지 - Surprise

Surprise 패키지 소개

- `pip install scikit-surprise`

Surprise 주요 모듈 소개

Dataset

API명	내용
Dataset.load_builtin(name='ml-100k')	무비렌즈 아카이브 FTP 서버에서 무비렌즈 데이터 내려받음
Dataset.load_from_file(file_path, reader)	OS 파일에서 데이터 로딩할 때 사용
Dataset.load_from_df(df, reader)	판다스의 DataFrame에서 데이터 로딩

Reader 클래스의 주요 생성 파라미터

- line_format(string) : 칼럼을 순서대로 나열, 입력된 문자열을 공백으로 분리해 칼럼으로 인식
- sep(char) : 칼럼을 분리하는 분리자, 디폴트 '\t'
- rating_scale(tuple, optional) : 평점 값의 최소 ~ 최대 평점 설정

Surprise 추천 알고리즘 클래스

클래스명	설명
SVD	행렬 분해를 통한 잠재 요인 협업 필터링을 위한 SVD 알고리즘
KNNBasic	최근접 이웃 협업 필터링을 위한 KNN 알고리즘
BaselineOnly	사용자 Bias와 아이템 Bias를 감안한 SGD 베이스라인 알고리즘

SVD 클래스 입력 파라미터

- n_factors = 100 : 잠재 요인 K의 개수, 커질수록 정확도가 높아질 수 있으나 과적합 문제 발생
- n_epochs = 20 : SGD 수행 시 반복 횟수
- biased(bool) = True : 베이스라인 사용자 편향 적용 여부

베이스라인 평점

- 개인의 성향을 반영해 아이템 평가에 편향성 요소를 반영하여 평점을 부과하는 것
- 전체 평균 평점 + 사용자 편향 점수 + 아이템 편항 점수 공식
- 전체 평균 평점 = 모든 사용자의 아이템에 대한 평점을 평균한 값
- 사용자 편향 점수 = 사용자별 아이템 평점 평균 값 - 전체 평균 평점
- 아이템 편향 점수 = 아이템별 평점 평균 값 - 전체 평균 평점