추천 엔진 평가 방법

추천 엔진의 성능을 평가하는 다양한 방법을 알아보자

머신 러닝 모델 평가 방법 (1)

- 먼저 모델 평가에 사용할 지표 결정
  - Confusion matrix, AUC-ROC, F1 점수
  - RMSE, MAE, 로그 손실 (Log Loss), ...
- 다음으로 평가 방법 결정
  - 홀드 아웃 테스트 (Train & Test 혹은 Train & Validation)
    - 보통 70:30 혹은 75:25 혹은 80:20을 사용
  - 교차 검증 (Cross-Validation 혹은 K-Fold 테스트)
    - 일반적으로 홀드 아웃 방식보다 오버피팅 이슈가 없음

머신 러닝 모델 평가 방법 (2)

- 특수한 교차 검증 방식: LOOCV (Leave One Out Cross Validation)
  - 교차 검증에서 폴드 수가 트레이닝 데이터 레코드 수와 동일한 경우
  - 즉 테스트를 한 예제를 대상으로 하는 것! 가장 좋은 방법이지만 시간이 오래 걸림
  - 훈련 데이터에 총 N개의 예제가 있다면
    - N-1개로 학습후 나머지 1개로 테스트하는 것을 N번 반복

추천 엔진 평가

	scifi1	scifi2	scifi3	comedy1	comedy2	comedy3
user1	4.0	5.0	3.0	NaN	2.0	1.0
user2	5.0	3.0	3.0	2.0	2.0	NaN
user3	1.0	NaN	NaN	4.0	5.0	4.0
user4	NaN	2.0	1.0	4.0	NaN	3.0
user5	1.0	NaN	2.0	3.0	3.0	4.0

테스트셋

추천 엔진 평가 방법

- 평점 기반
  - 평점을 예측하고 실제 평점과 비교
  - 추천 엔진에서는 보통 RMSE 혹은 MAE를 사용
- Top-N 추천 정확도 기반
  - 사용자별로 일부 높은 평점 레코드(사용자, 아이템, 평점)를 따로 빼놓고 나중에 추천되는 아이템들과의 일치율을 계산
  - LOOCV (Leave One Out) 테스트 방법과 병행하여 사용
    - scikit-learn과 surprise 라이브러리에서 지원
      - from surprise.model_selection import LeaveOneOut
    - LOOCV는 교차검증 방식으로 테스트셋에 사용자별로 오직 하나의 평점만을 남김

Top N 추천 정확도 (1)

- 모델 방식의 추천이고 N이 10이라고 가정하고 사용자별로 아래를 반복
  - 평점 데이터에서 이 사용자의 모든 데이터를 찾는다
  - 여기서 한 평점 레코드를 빼서 (LeaveOneOut) 테스트 셋에 추가
  - 나머지 레코드들을 훈련 셋에 추가
- 만들어진 훈련 셋으로 모델을 학습
- 훈련에 사용되지 않은 모든 레코드들 (build_anti_testset)을 가지고 평점 예측
  - 기본적으로 평점 정보가 없는 모든 사용자ID와 아이템ID 레코드들
    - 여기에는 테스트셋의 레코드들도 들어감
- Top N 추천 정확도
  - 사용자별로 테스트셋의 아이템 중 평점이 높은 것들 중에 추천된 Top 10개에 포함된 것의 비율 계산 후 평균 계산

Top N 추천 정확도 (2)

- surprise의 LeaveOneOut 모듈을 사용하여 사용자별로 평점 정보를 하나씩 테스트셋으로 저장하고 이를 나중에 Top N 추천 정확도 계산에 사용

사용자 ID	(영화ID, 평점)
1	(1, 3.5), (2, 4.5), (3, 5.0)
2	(2, 3.5), (3, 4.5), (4, 3.0), (6, 4.5)
3	(1, 4.0), (2, 3.5), (5, 4.5), (7, 3.5), (9, 4.5)

사용자 ID	훈련용 데이터	테스트용 데이터
1	(1, 3.5), (2, 4.5)	(3, 5.0)
2	(2, 3.5), (3, 4.5), (4, 3.0)	(6, 4.5)
3	(1, 4.0), (2, 3.5), (5, 4.5), (7, 3.5)	(9, 4.5)

Top N 추천 정확도 (3)

- Top N 추천 정확도: 모든 사용자들의 추천 정확도 평균
  - 추천 위치에 따라 가중치를 줬다면 이를 Top N 추천 NDCG(Normalized Discounted Cumulative Gain) 정확도라 부름
- 실습:
  - “SVD 추천 엔진"에서 개발한 SVD 모델을 대상으로 Top-N 정확도 계산
  - surprise에서 제공해주는 LeaveOneOut 모듈을 사용
    - Top-N 정확도 계산을 쉽게 해줌!
  - 실습 링크

추천 엔진 개발 교훈

추천 엔진 개발시 주의점

- Cold Start: 특히 CF 기반 추천의 경우
  - 사용자 데이터: 사용자가 서비스를 처음 사용하기 시작하는 경우
  - 아이템 데이터: 아이템이 처음 서비스에 노출되기 시작하는 경우
- 확장성 (Scalability)
  - 몇 천만의 사용자를 처리할 수 있나?
    - 서비스가 성장하면 사용자 수는 대폭 커질 수 있음
  - 몇 천만의 아이템을 처리할 수 있나?
    - 사용자의 수에 비해 아이템의 수는 상대적으로 성장 폭이 작음
  - 모델링에 걸리는 시간 뿐만 아니라 서빙시 시간도 중요
    - 넷플릭스 프라이즈 1등 모델 예

추천 엔진 개발시 고려할 점 (1)

- 데이터의 부족
  - 명시적 혹은 암묵적인 레이블 데이터의 부족
    - 리뷰/평점 데이터와 클릭 데이터의 크기는 상대적으로 작음
    - 평점 데이터는 조작 가능 (악의적인 해킹)
- 다양한 아이템의 추천
  - 관점 혹은 선호도 편향화 심화
    - 예) 유투브 추천 -> 정치적인 편향의 심화
    - 신선한 아이템 추천이 필요
  - 새로운 아이템을 어떻게 노출시킬지 고민이 필요
    - The rich gets richer 현상을 어떻게 타파할 것인가?

추천 엔진 개발시 고려할 점 (2)

- 인프라 필요
  - 필요한 데이터가 수집 저장되어야 하고 이를 처리할 수 있는 인프라가 필요
    - 예) 사용자가 무슨 코스를 보았고 무엇을 클릭했나
  - 인프라 없이 계속적인 개발과 개선은 불가능
  - 기술적인 관점에서는 Spark이 많이 사용됨
- 개인 정보 보호
  - 많은 추천이 개인 정보들을 취합하여 만들어짐
    - GDPR(유럽연합)과 CCPA (미국 캘리포니아)등의 법률 존재
      - 내 개인정보 삭제 권리 - “the right to be forgotten”
  - 이 과정에서 의도치 않은 개인정보 노출 가능
    - 검색어 자동완성 예: 개인의 이름 입력에 “이혼(Divorce)”라는 추천

- 유데미에서 추천엔진 개발기 공유
- 추천 엔진 (혹은 머신러닝) 개발시 알아야하는 개인정보 보호 상식 공유
- 추천 엔진 개발시 교훈 공유

UpZen

keeyong@gmail.com

A/B 테스트란?

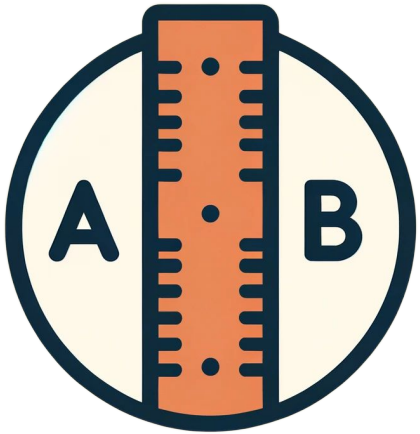
추천 엔진의 성능을 실제 사용자에 대상으로 검증하는 A/B 테스트에 대해서 알아보자

A/B 테스트: 온라인 실험이라고 부르기도 함

- A/B 테스트란 다수의 그룹으로 구성
  - 하나의 컨트롤 그룹과 하나 이상의 테스트 그룹
- 가설이 꼭 필요! 인프라도 필요!

A/B 테스트란? (1)

- 실제 사용자를 대상으로 새로운 기능이나 변경을 객관적으로 검증하는 방법
- 테스트 시작 전에 미리 어떤 지표를 가지고 테스트의 성패 여부를 정할지 결정함
  - 가설의 중요성!



A/B 테스트란? (2)

- 한번에 하나의 새로운 기능이나 변화를 테스트해야함
  - 동시에 2가지 이상을 테스트할 경우 결과를 해석할 수 없음
- 작은 수의 사용자들에게 먼저 노출시킴으로써 위험부담을 줄임
  - 작게 시작하고 지표를 모니터하면서 점차적으로 노출 비율을 높임
- A/B 테스트 인프라 없이는 테스트를 할 수도 없고 분석도 불가능!
  - 모든 엔지니어링 팀의 도움이 필요 (프런트엔드, 백엔드, 데이터)

A/B 테스트 방식 설명 (1)

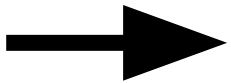
- A/B 테스트 가설 세우기: 앞서 설명
- 사용자를 같은 크기와 같은 속성의 두 그룹으로 나누기 (치우침이 없어야 함)
  - 기존 기능에 노출될 사용자 vs. 새로운 기능에 노출될 사용자

A	B
User1	User3
User4	User6
User7	User8
User10	User13
...	...

A/B 테스트 방식 설명 (2)

- 이 사용자들의 다양한 행동을 기록
  - 사용자별로 어떤 아이템을 보았고 클릭했고 구매했고 소비했고 리뷰를 했는지 기록
- 두 개의 그룹별로 다양한 지표 계산 후 비교
  - 두 그룹 간의 지표 차이가 통계적으로 유의미한지 꼭 확인
    - 통계적 지식이 필요
  - 시간이 지나면서 어떤 흐름이 있는지 확인

A	B
User1	User3
User4	User6
User7	User8
User10	User13
...	...



	A	B
아이템 클릭률	10%	12%
아이템 평균 판매액	$21	$23
아이템 평균 평점	3.7	**4.1**
...		
...		

태블로 기반 A/B 테스트 대시보드 예제

분석 기간 선택

다양한 필터