Shiny로 나만의 추천 시스템 만들기

Movielens 데이터를 이용한 영화 추천

이영록

동기

- 각각의 영화 플랫폼이 나의 선호를 온전히 반영하기에 한계가 있음
- 영화 플랫폼에서 제공하는 추천 시스템은 나의 오늘의 선호를 충분히 반영하지 못함

목표

- 나의 일반적인 성향을 나타내는 utility 함수 학습
- 내가 오늘 선호하는 영화를 utility에 반영
- 필터링을 통해 원하는 영화를 간추림

데이터

MovieLens genome score data

movieId	title	action	based_on_a_book	bleak	comedy	criterion	drama
25	Leaving Las Vegas (1995)	0.06800	0.37750	0.73600	0.03100	0.32650	0.81750
50	Usual Suspects, The (1995)	0.74500	0.23000	0.23550	0.48175	0.56200	0.76575
58	Postman, The (Postino, Il) (1994)	0.06875	0.34425	0.09200	0.05275	0.45100	0.33725
356	Forrest Gump (1994)	0.65375	0.61325	0.12625	0.65675	0.26500	0.89425
364	Lion King, The (1994)	0.31600	0.20500	0.11225	0.31250	0.25000	0.44675
539	Sleepless in Seattle (1993)	0.17650	0.15600	0.04300	0.48100	0.21225	0.72850
1172	Cinema Paradiso (Nuovo cinema Paradiso) (1989)	0.11075	0.14600	0.24825	0.07750	0.73250	0.80750

- 1만 3천편 이상의 영화, 1천개 이상의 feature (본 데모에서는 20개의 feature만 사용)
- 데이터 출처: MovieLens 25M Dataset (https://grouplens.org/datasets/movielens/)
- Feature 생성에 대한 설명: “The Tag Genome: Encoding Community Knowledge to Support Novel Interaction” (http://files.grouplens.org/papers/tag_genome.pdf)

이산선택모형

Discrete Choice Model

$$u_{ij} = \beta_{action} x_{j,action} + \beta_{comedy} x_{j,comedy} + \cdots + \epsilon_{ij}$$

Utility Coefficient Feature Unobservable Utility

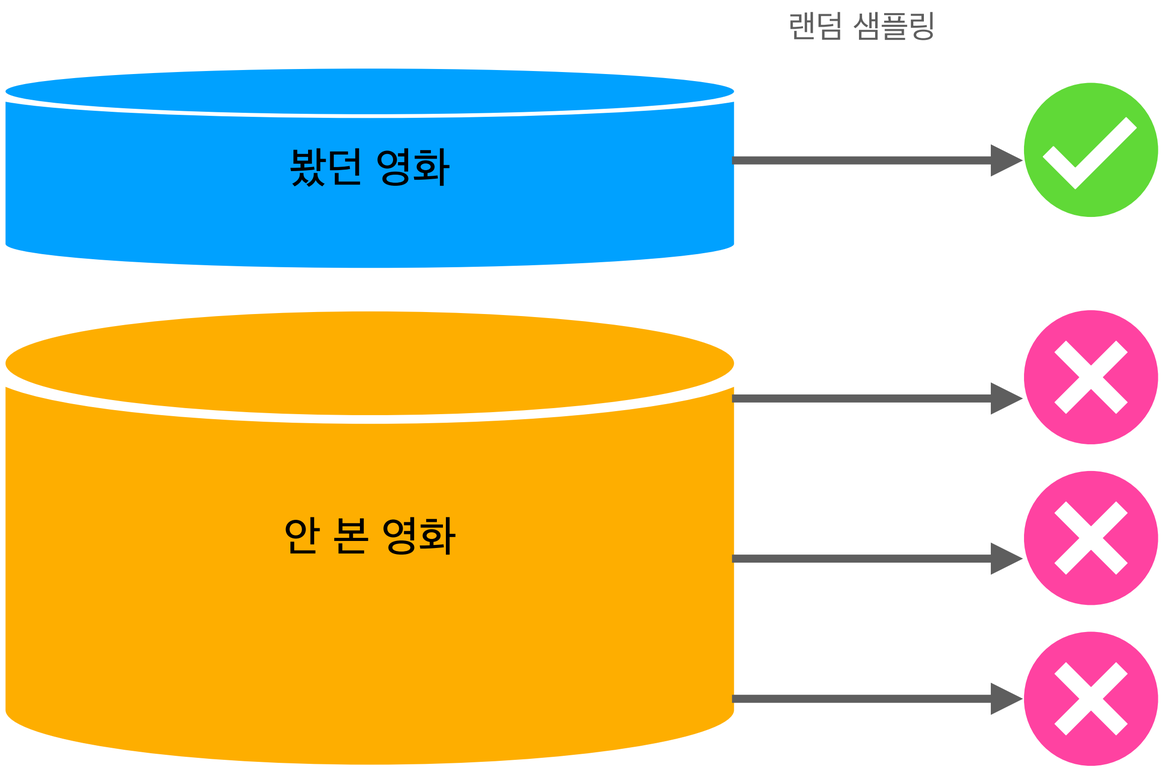
- i: 선택 기회
- j: 영화
- 최종 선택: $\arg\max_j u_{ij}$ (각 i번째 선택 기회에서 제시된 영화 중 utility가 가장 높은 영화를 사용자가 선택한다고 가정)

앱 실행 전 학습데이터

Training Data

- 기존에 각 플랫폼에서 제공한 선택 기회에 대한 데이터를 내가 갖고 있지 않음
- 전체 영화 리스트를 알고 있고, 그 중 내가 본 영화 리스트를 알고 있음
- 이를 이용하여 가상의 학습 데이터를 생성하여 utility 함수를 학습

앱 실행 전 학습

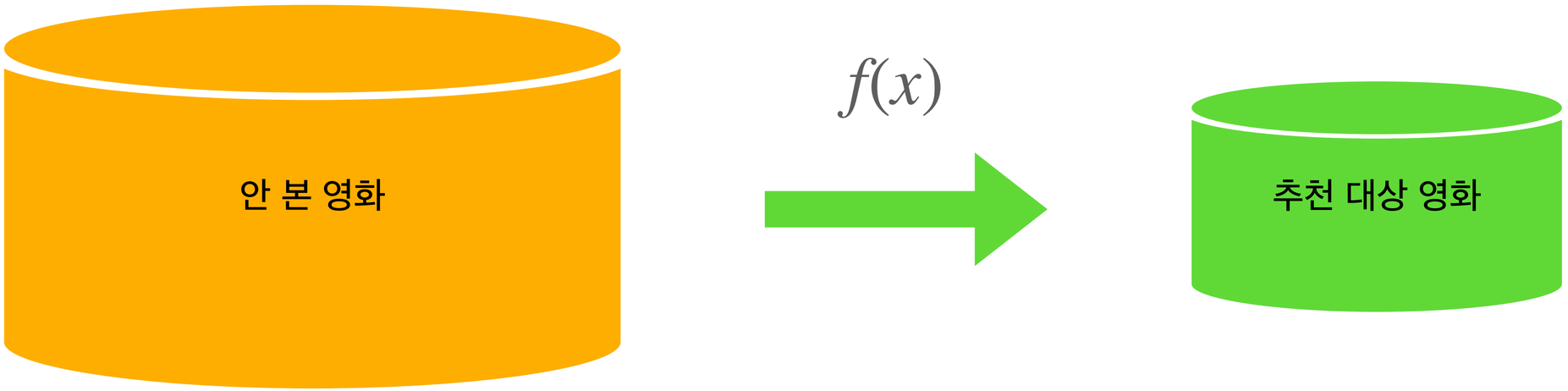


- 한 편의 이미 본 영화와 J개의 안 본 영화를 랜덤 샘플링
- 본 영화가 선택되었다고 레이블링
- 이 과정을 N회 반복하여 N개의 선택 기회를 생성
- 이산선택모형을 생성된 데이터에 대해 학습

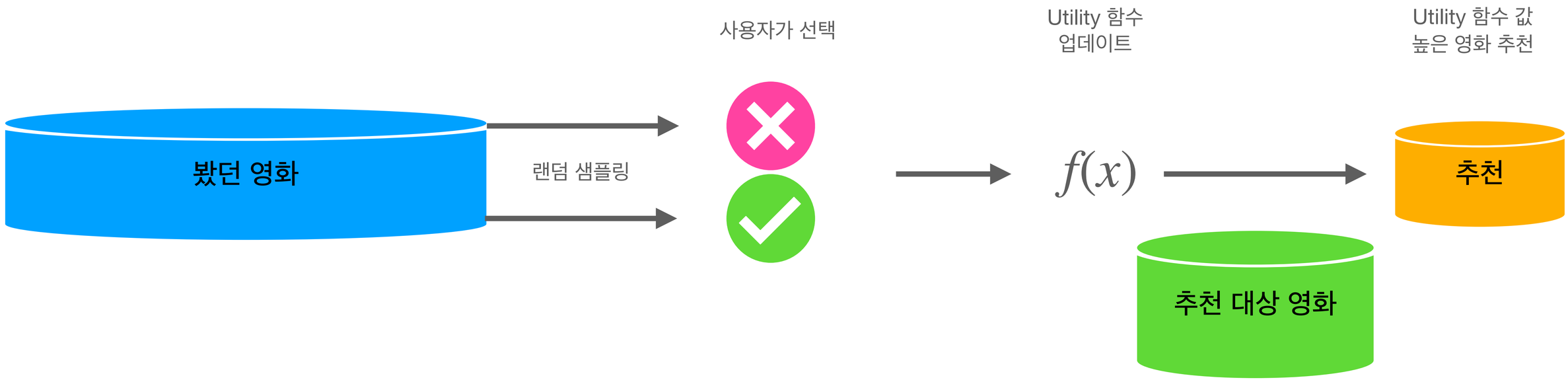
학습된 Utility 함수 $f(x)$ 예

term	estimate
<chr>	<dbl>
loneliness	14.7
suspense	13.4
funny	7.67
enigmatic	7.63
relationships	2.26
family	1.61
action	1.47
dramatic	1.19
based_on_a_book	0.767
drama	-0.0522
criterion	-1.48
violence	-1.70
comedy	-3.03
bleak	-4.02
obsession	-6.66
intimate	-7.34
tense	-7.66
murder	-11.0
independent_film	-11.3
horror	-32.3

추천 대상 영화 필터링



앱 사용 시 Utility 함수 재학습



Demo

Utility 함수 학습에 영향을 미치는 요인

- 히스토리 데이터의 양
- 어떤 feature를 사용하고, 어떻게 transformation할 것인지
- 각 선택 데이터에 대한 가중치
- 랜덤 샘플링 vs Adaptive sampling

많은 사용자를 서포트하기 위해 고려할 사항

- 각 사용자의 utility 함수를 batch process를 통해 미리 학습
- 앱 사용 시 보다 utility 함수 업데이트를 보다 가볍게 할 필요
- Utility 함수 학습 서버를 Shiny 서버와 분리
- WebAssembly

Q&A