천재교육

머신러닝 (Machine Learning) 이해와 실습

K - 디지털 아카데미

머신러닝 교육과정 강의 일정

1일차 > 머신러닝

- > 머신러닝의 개요
- > [실습] 데이터 분석의 시작 Numpy, Pandas 활용하기

2일차 > Classification (분류)

- > 분류를 평가하는 지표 알아보기
- > 분류 알고리즘 (결정트리, 앙상블, 랜덤포레스트 등) 익히기
- > [실습] 분류를 통한 밀크T 만료및탈퇴회원 예측(이탈 회원 예측)

3일차 > Regression (회귀)

- > 회귀와 경사 하강법
- > 로지스틱 회귀와 소프트맥스 회귀
- > [실습] 로지스틱 회귀를 통한 문항별 정오답 예측

4일차 > 차원 축소와 Clustering(군집화)

- > PCA, LDA
- > K-means, DBSCAN 등 다양한 클러스터링 기법 알아보기
- > [실습] 밀크T중학 회원수준 군집화(GMM)

5일차 > 추천시스템과 최종 프로젝트

- > 추천 시스템
- > 최종 프로젝트

추천시스템

01 추천시스템

02 협업필터링

03 [실습] 교육데이터 소개 및 적용

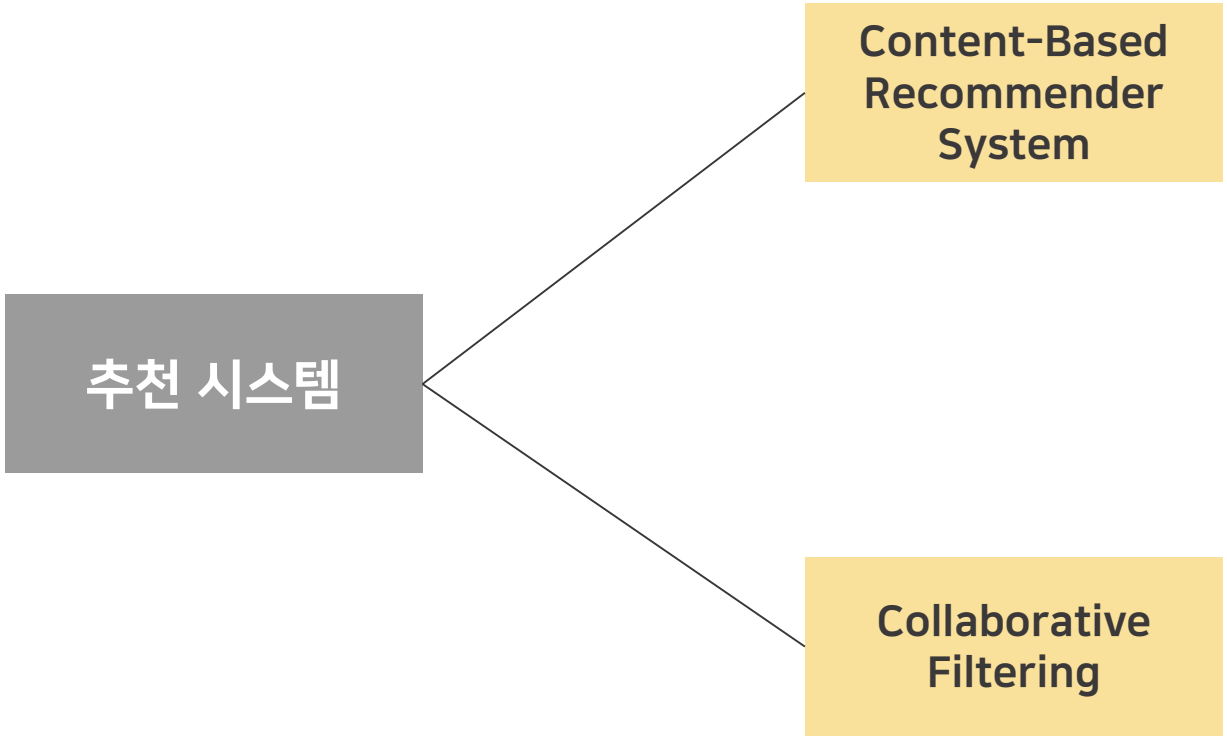
머신
러닝.

추천시스템?

추천시스템 (Recommender System) 은
특정 사용자가 관심을 가질만한 정보(영화, 음악, 책, 뉴스, 이미지, 웹페이지 등) 를 추천하는 것
추천 시스템에는 협업 필터링 기법을 주로 사용한다.

넷플릭스에서 시청한 영상과 유사한 영상을 추천하고, 쿠팡 같은 온라인마켓에서 사용자가 과거에 구매하거나 관심을 가진 물건과 유사한 물건을 보여주는 활용

추천시스템의 종류


추천 시스템
Content-Based Recommender System
Collaborative Filtering

콘텐츠 기반 추천 시스템

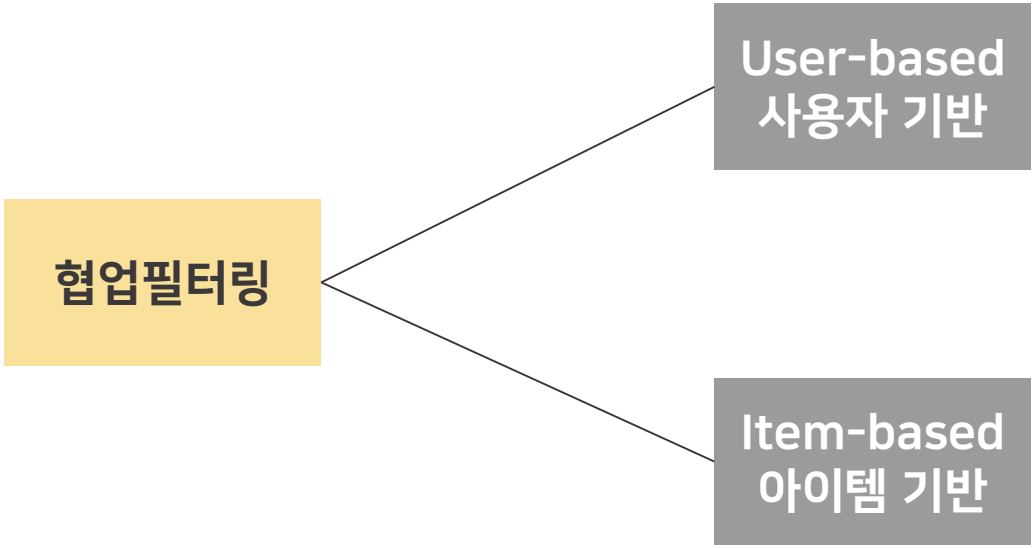
콘텐츠 기반 추천 시스템(Content-Based Recommender System)은?

아이템에 대한 세부 정보를 토대로 사용자가 과거에 소비했던 콘텐츠와 유사한 콘텐츠를 추천해주는 방식

	액션	어드벤쳐	코미디	판타지	스릴러
영화 1	0	1	1	1	0
영화 2	0	0	1	0	1
영화 3	1	1	0	0	0
영화 4	0	1	1	0	0

협업 필터링

협업필터링이란,
추천 시스템에서 많이 쓰이는 방법 중 하나로 사용자의 행동방식에 의존하여 추천하는 시스템



코사인 유사도 (Cosine Similarity)

$$similarity = cos(\Theta) = \frac{A \cdot B}{||A||\,||B||} = \frac{\sum_{i=1}^{n} A_i \times B_i}{\sqrt{\sum_{i=1}^{n}(A_i)^2} \times \sqrt{\sum_{i=1}^{n}(B_i)^2}}$$

0	0	1	1	0

1	0	1	1	1

$$= \frac{0+0+1+1+0}{\sqrt{2} * \sqrt{4}} = \frac{1}{\sqrt{2}} = 0.7$$

코사인 유사도 (Cosine Similarity)

$$similarity = cos(\Theta) = \frac{A \cdot B}{||A||\,||B||} = \frac{\sum_{i=1}^{n} A_i \times B_i}{\sqrt{\sum_{i=1}^{n}(A_i)^2} \times \sqrt{\sum_{i=1}^{n}(B_i)^2}}$$

1	1	1	1	0

0	1	1	1	1

$$= \frac{0+1+1+1+0}{\sqrt{4} * \sqrt{4}} = \frac{3}{4} = 0.75$$

협업 필터링 : User-based

	Item A	Item B	Item C	Item D	Item E
User 1	3	4	4		1
User 2	4	4	4	2	
User 3	1	2	4	5	
User 4	1		3	5	

User가 각 Item에 매긴 평점을 보고 매긴 평점이 서로 유사한 사용자를 찾음.
→ 나와 성향이 비슷한 사용자 찾기

협업 필터링 : Item-based

	User 1	User 2	User 3	User 4	User 5
Item A	3	2			1
Item B	4	4	1	2	3
Item C	4	4		2	
Item D		1		1	

좋아하는 사용자가 서로 유사한 아이템을 찾음.
→ B를 좋아하면 C도 좋아하던데? → 아이템 추천.

모든 유저에 대한 아이템 평점을 표현하기에 버겁기 때문에 실제로 User-based보다 Item-based가 많이 사용됨.

교육에서의 추천?

좋아하는 아이템?

➔ 좋아하는 문항?

➔ 맞힌 문항?
➔ 틀린 문항?

취향이 비슷한 사용자 찾기
➔ 틀린 문제가 유사한 학생 찾기

A를 좋아하면 B도 좋아하던데? → A를 구매한 고객에게 B아이템 추천.
➔ A문항을 틀린 사용자에게 B를 추천

교육 데이터 소개

교육 AI를 적용하기 위해 사용할 수 있는 데이터

ex. 학생별 문항 정오답 데이터 / 학생별 인강 이용기록 데이터

교육 데이터 : ASSISTment 2009-2010

미국 수학 중등과정의 학습이력 데이터셋

	order_id	assignment_id	user_id	assistment_id	problem_id	original	correct
0	33022537.0	277618	64525	33139	51424	1.0	1.0
1	33022709.0	277618	64525	33150	51435	1.0	1.0
2	35450204.0	220674	70363	33159	51444	1.0	0.0

데이터 건수	525,535 건
사용자 수	4,218 명
기간	2009 ~ 2010년 (날짜기록 없음)
개념(Skill) 수	110 개

https://sites.google.com/site/assistmentsdata/home/2009-2010-assistment-data

교육 데이터 : ASSISTment 2009-2010

Features	설명
order_id	이력 Log를 기반으로 시간순서대로 매긴 ID
user_id	학생 ID
problem_id	문항 ID
correct	정오답 (맞힘 : 1 / 틀림 : 0)
attempt_count	이 문제에 대한 학생의 시도 횟수
skill_id	Skill(개념) ID
skill_name	Skill(개념) 이름
answer_id	학생이 선택한 보기 ID
answer_text	학생이 입력한 답(주관식)

교육 데이터 : EdNet

산타토익 AI서비스를 만든 뤼이드의 데이터셋
: 사용자의 컨텐츠 수강 이력 및 풀이이력

KT1	문제풀이 로그 (최종제출한 답만 포함)
KT2	문제풀이 로그 (제출 전 선택한 답까지 포함)
KT3	문제풀이 로그 + 문항해설 + 강의시청
KT4	문제풀이 로그 + 문항해설 + 강의시청 + 결제 + 쿠폰사용 등

https://github.com/riiid/ednet

교육 데이터 : EdNet

51. The bank is _____ tomorrow.

(A) open
(B) short
(C) true
(D) poor

	질문이 주어진 순간 타임스탬프	문제 묶음 ID	문항 ID	학생의 답변	풀이 시간
	timestamp	solving_id	question_id	user_answer	elapsed_time
0	1565313803824	1	q5467	a	17000
1	1565313821053	2	q4470	a	13000
2	1565313840802	3	q3710	a	17000
3	1565313869883	4	q6173	d	26000
4	1565313906637	5	q4080	a	34000

교육 데이터 : EdNet

51. The bank is _____ tomorrow.

(A) open
(B) short
(C) true
(D) poor

timestamp	action_type (행동타입)	item_id (차시 ID)	cursor_time (행동 순간)	source (선택한소스 (문제풀이/강의위치))	user_answer (학생의 답변)	platform
1573801116701	play_audio	b911	0.0	sprint	NaN	mobile
1573801132281	pause_audio	b911	14104.0	sprint	NaN	mobile
1573801132483	respond	q911	NaN	sprint	a	mobile
1573801135798	submit	b911	NaN	sprint	NaN	mobile
1573801135880	enter	e911	NaN	sprint	NaN	mobile
1573801139924	play_audio	b911	0.0	sprint	NaN	mobile
1573801145936	pause_audio	b911	5845.0	sprint	NaN	mobile
1573801145936	play_audio	b911	845.0	sprint	NaN	mobile
1573801147152	pause_audio	b911	1866.0	sprint	NaN	mobile
1573801147152	play_audio	b911	311.0	sprint	NaN	mobile
1573801149244	pause_audio	b911	2210.0	sprint	NaN	mobile
1573801149244	play_audio	b911	0.0	sprint	NaN	mobile
1573801151074	pause_audio	b911	1602.0	sprint	NaN	mobile

교육 데이터 : AI허브 수학분야 학습자 역량 측정 데이터

A090001576														
566	566	556	566	566	566	570	570	569	569	569	570	570	571	
0	1	1	0	1	0	1	1	1	1	1	1	1	1	
A090001667														
1493	1493	1494	1493	1494	1493	1494	1493	1493	1500	1498	1498	1500	1499	
0	1	1	0	1	1	1	0	1	0	0	0	1	0	
A090001382														
1093	1093	1093	1093	1091	1078	1093	1078	1088	1091	1092	1088	1088	1087	
1	1	1	1	1	1	1	1	1	1	1	1	1	0	

평가 ID


```
{"learnerID": "A090000914", "learnerProfile": "M;S01;9", "testID": "A090000001",
"assessmentItemID": "A090001001", "answerCode": "0", "Timestamp": "2021-05-30 11:56:11"}
```


문항 ID

https://aihub.or.kr/aihubdata/data/view.do?currMenu=115&topMenu=100&dataSetSn=133

교육에서의 추천시스템 (사내 PoC)

- ✓ 유사한 문제를 틀린 학생 찾기
- ✓ 틀린 사용자가 유사한 문제 찾기 : 유사한 메타를 가지는 문항인지 검증
- ✓ 틀린문제가 유사한 학생 Top10을 찾아, 그 학생들이 틀린 문제를 추천하는 방식

[실습]

협업 필터링을 적용해 유사한 문항 찾기
협업 필터링을 적용해 유사한 학생 찾기

최종 과제

01 데이터셋 1

02 데이터셋 2

03 데이터셋 3

머신
러닝.

데이터셋 1

중학 실력Test 데이터셋

총 250,190건

	idx	guid	st_year	Test_Name	Subject	Subject2	Grade	Grade2	Term	Test_Level	take_date	test_time	Test_Jumsu	QuizNum	QuizCode	QuizYN	QuizAnswer
0	521308	9728f3ec-0b86-4d5e-99c2-4412fc09d3a7	2022	수수행대비	MM	MM	8	8	01;02	04;03	2022-08-19 20:01:02.977	1305	85	20	40094723	N	④
1	521308	9728f3ec-0b86-4d5e-99c2-4412fc09d3a7	2022	수수행대비	MM	MM	8	8	01;02	04;03	2022-08-19 20:01:02.977	1305	85	19	40095385	Y	⑤
2	521308	9728f3ec-0b86-4d5e-99c2-4412fc09d3a7	2022	수수행대비	MM	MM	8	8	01;02	04;03	2022-08-19 20:01:02.977	1305	85	18	40094651	Y	6
3	521308	9728f3ec-0b86-4d5e-99c2-4412fc09d3a7	2022	수수행대비	MM	MM	8	8	01;02	04;03	2022-08-19 20:01:02.977	1305	85	17	40095295	Y	②
4	521308	9728f3ec-0b86-4d5e-99c2-4412fc09d3a7	2022	수수행대비	MM	MM	8	8	01;02	04;03	2022-08-19 20:01:02.977	1305	85	16	40095274	Y	①

데이터셋 2

월간 학습자 통계 데이터

총 500,000건

	userid	gender	membertype_codename	grade_codename	memberstatus	memberstatus_codename	memberstatus_change	status_null_count	statusgroup_10_count
0	0001809c-1725-4ccd-86b0-d02ed0937a83	M	초등	초2	11.0	학습생(정)	-,11,-,11,-,11,-,11,-,11,-,11,-,11,-,11	13	0
1	00028ac1-a0ab-486f-bfdd-de2b0bf70980	F	초등	초4	11.0	학습생(정)	-,11,-,11,-,11,-,11,-,11,-,11,-,11,-,11,-,11	15	0
2	00181cb5-7afd-4cb9-ac7c-37aa66796167	F	초등	초2	11.0	학습생(정)	-,11,-,11,-,11,-,11,-,11	10	0
3	001ad7ff-3db5-4705-a036-2d9b6260957d	M	초등	초3	11.0	학습생(정)	-,11,-,11	2	0
4	002a7014-ee46-4a0e-85e6-389214ca3421	M	초등	초3	11.0	학습생(정)	11,-,11,-,11,-,11,-,11	8	0

데이터셋 3

학생별 AI진단평가 과목별 점수

총 2,816건

	응시기간	시험명	학년	응시일자	전체	국어	영어	수학	과학	사회	역사
0	2021. 12 ~2022. 12	AI진단평가	중2	2021-12-23 00:00:00	338	68	93.0	71.0	NaN	53.0	53.0
1	2021. 12 ~2022. 12	AI진단평가	중2	2021-12-23 00:00:00	50	18	0.0	9.0	NaN	23.0	NaN
2	2021. 12 ~2022. 12	AI진단평가	중2	2021-12-23 00:00:00	65	29	36.0	NaN	NaN	NaN	NaN
3	2021. 12 ~2022. 12	AI진단평가	중2	2021-12-24 00:00:00	276	75	72.0	76.0	NaN	38.0	15.0
4	2021. 12 ~2022. 12	AI진단평가	중2	2021-12-26 00:00:00	29	29	0.0	0.0	NaN	0.0	0.0

감사합니다