전자상거래 데이터 분석

ANOVA 분석과 추천시스템

빅데이터처리및응용 1조

CONTENTS

조원 및 역할

- 주로 맡은 업무들은 다음과 같습니다.

지 승 찬

전처리, ANOVA,
발표 자료 준비

김 현 규

ANOVA 구현 및 해석

박 정 재

추천 시스템 구축

정 호 빈

고객 분석, 발표 담당

데이터 소개

- 실제 전자상거래에서의 사용자 행동 데이터
- 데이터 크기 : 13.67GB (변수9개, 관측치 1억개 이상)
- 특징 : 1) 전자상거래에서 쌓이는 데이터가 변수
  2) numeric, categorical 변수 존재
  3) 전처리가 되어있는 데이터가 아닌 날 것의 데이터
  4) 대용량 데이터
- 목표 : 1) ANOVA 분석으로 사용자의 event 수에 따라 평균구매가격의 차이가 존재하는지 파악
  2) 추천시스템 모델로 사용자에게 구매하지 않은 물건 중 좋아할 만한 상품들을 추천

공통적으로 사용된 데이터 전처리 요약

- 결측치를 행 기준으로 삭제
- Category_code값을 '.' 기준으로 split, 각 특성으로 새로운 범주열 생성
- NaN을 N으로 변환해 프로그램이 인식하게 함

공통적으로 사용된 데이터 전처리 요약

- 유효한 열을 선정하여 category to numeric (label encoding) 진행
- 두 개의 전처리 한 csv 파일을 합치고 생성하여 데이터 크기를 14GB에서 GB로 줄여 40% 이상 효율적으로 만들어 뒤에서 진행하는 ANOVA 분석과 ALS 추천 모델을 더 빠르게 진행가능하게 했음

데이터 전처리 (생성된 데이터)

- 전처리하여 얻은 데이터 ('.' 기준 split)

category_code_type1	category_code_type2	category_code_type3	category_code_type4

데이터 전처리 (생성된 데이터)

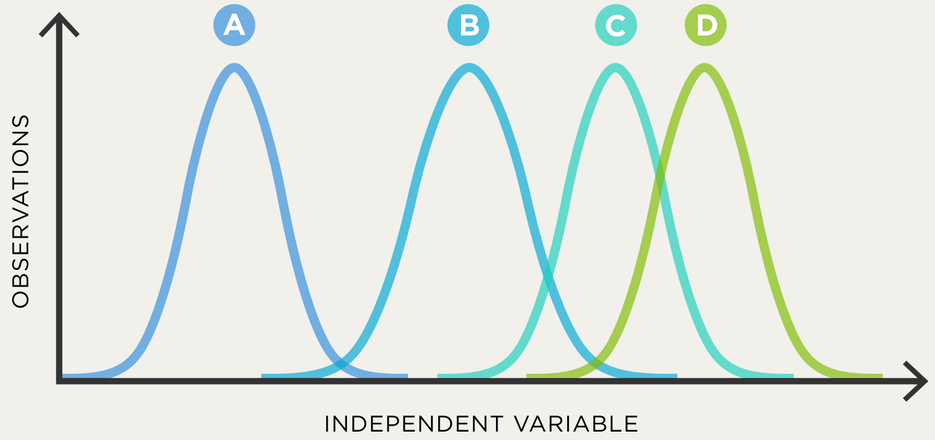
- 전처리하여 얻은 데이터 (label encoding)

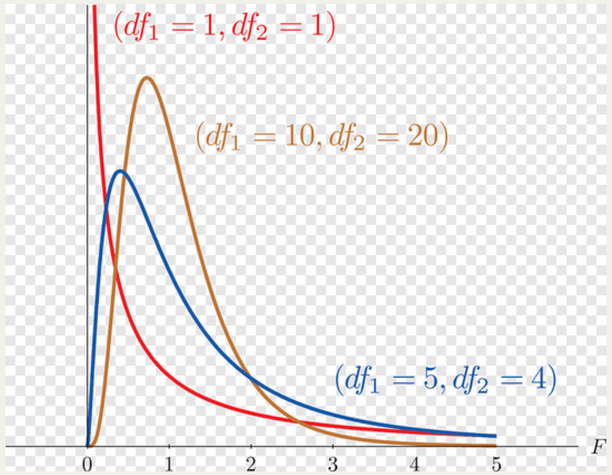
event_time	event_type	product_id	category_id	brand	price	user_id	user_session	category_code_type1	category_code_type2	category_code_type3	category_code_type4
2019-10-01 00:00:00	1	3900821	2053013552326770905	1	33.2	554748717	9333dfbd-b87a-470...	1	1	0	1
2019-10-01 00:00:01	1	1307067	2053013558920217191	2	251.74	550050854	7c90fc70-0e80-459...	2	0	0	2
2019-10-01 00:00:04	1	1004237	2053013555631882655	3	1081.98	535871217	c6bd7419-2748-4c5...	3	0	0	3
2019-10-01 00:00:05	1	1480613	2053013561092866779	4	908.62	512742880	0d0d91c2-c9c2-4e8...	2	0	0	4
2019-10-01 00:00:10	1	28719074	2053013565480109009	5	102.71	520571932	ac1cd4e5-a3ce-422...	4	3	0	5
2019-10-01 00:00:11	1	1005011	2053013555631882655	8	900.64	530282093	50a293fb-5940-41b...	3	0	0	3
2019-10-01 00:00:11	1	2900536	2053013554776244595	7	51.46	555158050	b5bdd0b3-4ca2-4c5...	1	4	0	6
2019-10-01 00:00:11	1	1004545	2053013555631882655	6	566.01	537918940	406c46ed-90a4-478...	3	0	0	3
2019-10-01 00:00:13	1	3900746	2053013552326770905	9	102.38	555444559	98b88fa0-d8fa-4b9...	1	1	0	1
2019-10-01 00:00:16	1	13500240	2053013557099889147	10	93.18	555446365	7f0062d8-ead0-4e0...	5	5	0	7
2019-10-01 00:00:18	1	1801995	2053013554415534427	9	193.03	537192226	e3151795-c355-4ef...	3	6	0	8
2019-10-01 00:00:18	1	10900029	2053013555069845885	11	58.95	519528062	901b9e3c-3f8f-414...	1	4	0	9
2019-10-01 00:00:19	1	1005135	2053013555631882655	3	1747.79	535871217	c6bd7419-2748-4c5...	3	0	0	3
2019-10-01 00:00:19	1	1306631	2053013558920217191	12	580.89	550050854	7c90fc70-0e80-459...	2	0	0	2
2019-10-01 00:00:20	1	4803399	2053013554658804075	13	33.21	555428858	8a6afed4-77f8-40c...	3	7	0	10
2019-10-01 00:00:20	1	1003306	2053013555631882655	3	588.77	555446831	6ec635da-ea15-4a5...	3	0	0	3
2019-10-01 00:00:22	1	1480714	2053013561092866779	4	921.49	512742880	0d0d91c2-c9c2-4e8...	2	0	0	4
2019-10-01 00:00:23	1	6200260	2053013552293216471	15	47.62	538645907	7d9a8784-7b6c-426...	1	1	0	11
2019-10-01 00:00:23	1	1004739	2053013555631882655	14	197.55	519530528	9882d21f-2c5f-496...	3	0	0	3
2019-10-01 00:00:24	1	1003306	2053013555631882655	3	588.77	555446831	6ec635da-ea15-4a5...	3	0	0	3

only showing top 20 rows

ANOVA (분산분석)

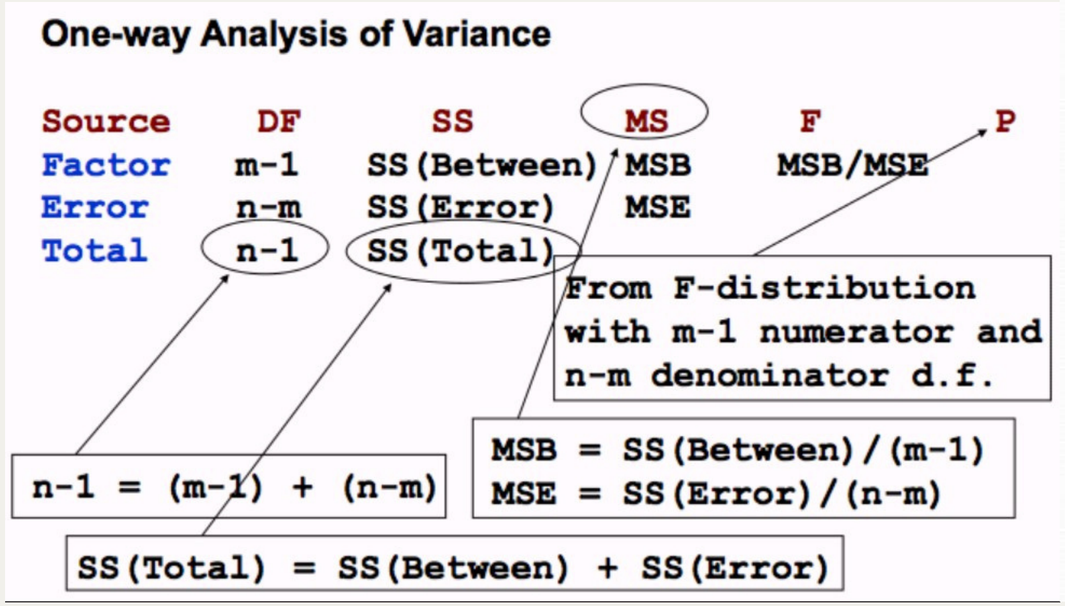
- 세 개 이상 다수의 집단의 모평균을 서로 비교하고자 할 때 집단 내 분산과 집단 간 분산의 비가 따르는 F분포를 이용하여 가설검정하는 방법
- 예) 한, 중, 일 국가 간의 학습 기술에 따른 성적 비교를 할 때 ANOVA를 사용 가능





ANOVA (분산분석)

- 자주 쓰이는 ANOVA의 형태 중 하나인 One-Way ANOVA를 pyspark로 구현



Source of Variance	Degree of Freedom (df)	Sum Square (SS)	Mean Square (MS)	F-ratio
Between Groups (Treatment)	k-1	$SSB = \sum_{j=1}^{k}\left(\frac{T_j^2}{n_j}\right) - \frac{T^2}{n}$ $SSB = \sum_{j=1}^{k} n_j(\overline{X}_j - \overline{X}_t)^2$	$MSB = \frac{SSB}{k-1}$	$F = \frac{MSB}{MSW}$
Within Groups (Error)	n-k	$SSW = \sum_{j=1}^{k}\sum_{i=1}^{n} X_{ij}^2 - \sum_{j=1}^{k}\left(\frac{T_j^2}{n_j}\right)$ $SSW = \sum_{j=1}^{k}\sum_{i=1}^{n}(X_{ij} - \overline{X}_j)^2$	$MSW = \frac{SSW}{n-k}$	
Total	n-1	$SST = \sum_{j=1}^{k}\sum_{i=1}^{n} X_{ij}^2 - \frac{T^2}{n}$ $SST = \sum_{j=1}^{k}\sum_{i=1}^{n}(X_{ij} - \overline{X}_t)^2$		

- SST = SSB + SSW

k: number of groups n: number of samples
df: degree of freedom

ANOVA (분산분석)

- 이번 프로젝트에서는 각 사용자에게서 발생한 event 수의 수준을 사분위수 기준으로 나누어 얻은 4개의 사용자 그룹 간 구매 가격의 모평균에 유의한 차이가 있는지 확인하였음
- Levene의 등분산 검정, 사후분석(post-hoc analysis)은 구현을 생략하였음(구체적인 동작방식이 학부 통계 수준을 뛰어넘음)

ANOVA (분산분석)

- 귀무가설 H0: event 수에 따라 4개 그룹 간 구매가격에 차이가 없다.
- 대립가설 H1: 적어도 한 쌍의 그룹에서 구매가격에 차이가 있다.
- F-검정통계량 = 집단 내 분산과 집단 간 분산의 비율
- F값은 두 개의 자유도를 가짐
- 이를 기반으로 p-value를 계산하여 유의수준과 비교

$H_0 = \mu_1 = \mu_2 = \ldots = \mu_k$

H_1 = At least two means differ

Statistical test is F test

$$F = \frac{MSB}{MSW}$$

Where: MSB is the *Mean Square Between*

MSW is the *Mean Square Within*

ANOVA (분산분석)

- 전처리된 데이터로부터 pyspark.sql.functions에 있는 메서드과 sql 쿼리를(groupby, left outer join 등) 적절히 여러 개 이용하여 사용자에 따라 event 수와 평균구매가격들을 담은 DataFrame을 추출
- 이후 중복되는 열은 삭제하였음

ANOVA (분산분석)

- event 수에 대해 udf를 새로 정의하여 사분위수와의 관계에 따라 high, intermediate, low, lowest로 categorize하였음
- 이후 직접 구현한 one_way_anova 함수에 이 DataFrame을 넣음

ANOVA (분산분석)

event_time	event_type	product_id	category_id	category_code	brand	price	user_id	user_session
2019-10-01 00:00:00	view	44600062	2103807459595387724	null	shiseido	35.79	541312140	72d76fde-8bb3-4e0...
2019-10-01 00:00:00	view	3900821	2053013552326770905	appliances.enviro...	aqua	33.2	554748717	9333dfbd-b87a-470...
2019-10-01 00:00:01	view	17200506	2053013559792632471	furniture.living_...	null	543.1	519107250	566511c2-e2e3-422...
2019-10-01 00:00:01	view	1307067	2053013558920217191	computers.notebook	lenovo	251.74	550050854	7c90fc70-0e80-459...
2019-10-01 00:00:04	view	1004237	2053013555631882655	electronics.smart...	apple	1081.98	535871217	c6bd7419-2748-4c5...
2019-10-01 00:00:05	view	1480613	2053013561092866779	computers.desktop	pulser	908.62	512742880	0d0d91c2-c9c2-4e8...
2019-10-01 00:00:08	view	17300353	2053013553853497655	null	creed	380.96	555447699	4fe811e9-91de-46d...
2019-10-01 00:00:08	view	31500053	2053013558031024687	null	luminarc	41.16	550978835	6280d577-25c8-414...
2019-10-01 00:00:10	view	28719074	2053013565480109009	apparel.shoes.keds	baden	102.71	520571932	ac1cd4e5-a3ce-422...
2019-10-01 00:00:11	view	1004545	2053013555631882655	electronics.smart...	huawei	566.01	537918940	406c46ed-90a4-478...
2019-10-01 00:00:11	view	2900536	2053013554776244595	appliances.kitche...	elenberg	51.46	555158050	b5bdd0b3-4ca2-4c5...
2019-10-01 00:00:11	view	1005011	2053013555631882655	electronics.smart...	samsung	900.64	530282093	50a293fb-5940-41b...
2019-10-01 00:00:13	view	3900746	2053013552326770905	appliances.enviro...	haier	102.38	555444559	98b88fa0-d8fa-4b9...
2019-10-01 00:00:15	view	44600062	2103807459595387724	null	shiseido	35.79	541312140	72d76fde-8bb3-4e0...
2019-10-01 00:00:16	view	13500240	2053013557099889147	furniture.bedroom...	brw	93.18	555446365	7f0062d8-ead0-4e0...
2019-10-01 00:00:17	view	23100006	2053013561638126333	null	null	357.79	513642368	17566c27-0a8f-450...
2019-10-01 00:00:18	view	1801995	2053013554415534427	electronics.video.tv	haier	193.03	537192226	e3151795-c355-4ef...
2019-10-01 00:00:18	view	10900029	2053013555069845885	appliances.kitche...	bosch	58.95	519528062	901b9e3c-3f8f-414...
2019-10-01 00:00:19	view	1306631	2053013558920217191	computers.notebook	hp	580.89	550050854	7c90fc70-0e80-459...
2019-10-01 00:00:19	view	1005135	2053013555631882655	electronics.smart...	apple	1747.79	535871217	c6bd7419-2748-4c5...

only showing top 20 rows

결과

event 수	평균 price
high	91.28000000000002
high	0.0
high	0.0
intermediate	738.615
low	0.0
high	0.0
low	0.0
low	0.0
high	38.31
high	129.1866666666667
high	0.0
high	0.0
low	0.0
high	249.34
high	0.0
intermediate	0.0
low	0.0
high	0.0
low	0.0
lowest	0.0

only showing top 20 rows

SQL 쿼리와 udf를 이용해
ANOVA 분석에 사용할
DataFrame 추출

ANOVA (분산분석)

- one_way_anova 함수는 DataFrame과 그룹을 나타내는 열 이름, 구매 가격을 나타내는 열 이름을 입력받아 집단 내 제곱합, 집단 간 제곱합, F 검정통계량, 두 개의 자유도를 반환하는 함수로 구현함
- ANOVA를 시행한 결과, p=0.000***으로 유의수준 0.001에서 그룹 간 평균의 차이가 유의하였음

```
sswg = 1042506262O6.43079, ssbg=6704092852.707795, F_statistic=89337.40087733661, df_1=3, df_2=4167669
p-value = 1.1102230246251565e-16
```

ANOVA (분산분석)

- 직접 구현하지는 않았으나, post-hoc analysis(사후분석)을 수행하여 구체적으로 어느 그룹 쌍에서 유의한 평균 차이가 났는지 확인할 수 있음
- 이 분석 결과는 전자상거래 사이트를 운영하는 회사의 입장에서 기존 사용자와 신규 사용자 중 더 평균 구매 가격이 낮은 쪽을 위한 프로모션 및 홍보 진행을 위한 근거로 유용하게 사용할 수 있으리라 기대됨

추천 시스템

- ALS 추천 모델을 적용하여 특정 사용자에게 선호할만한 제품을 추천하는 함수를 새로 구현함.
- ALS 추천 시스템을 이용하기 위해서 이벤트 타입 별로 임의의 가중치 점수를 부여함. (event type: view=1, cart=0.2, purchase=0.3)

추천 시스템

- 10월, 11월 데이터를 합친 후 결측치 행 제거, 불필요한 열을 제거 하여 DataFrame 생성. (user_session, event_time열 제거)

```
root
 |-- event_type: string (nullable = true)
 |-- product_id: integer (nullable = true)
 |-- category_id: long (nullable = true)
 |-- category_code: string (nullable = true)
 |-- brand: string (nullable = true)
 |-- price: double (nullable = true)
 |-- user_id: integer (nullable = true)
```

추천 시스템

- 유저별 각 제품에 대한 가중치 총점을 구하여 새로운 DataFrame 생성함.
- 기존의 DataFrame에서 user_id, product_id가 중복된 요소들 제거함.
- 이후 기존 DataFrame에 user_id, product_id값이 같은것 끼리 join함

(다음 슬라이드에 결과 사진)

추천 시스템

user_id	product_id	category_code	brand	price	total_score
31198833	1005065	electronics.smart...	xiaomi	218.54	0.1
88309646	1003906	electronics.smart...	huawei	123.07	0.1
96041329	5100850	electronics.clocks	huawei	154.16	0.1
116566414	2701022	appliances.kitche...	leadbros	198.2	0.1
116566414	2701889	appliances.kitche...	beko	360.08	0.1
116566414	2702522	appliances.kitche...	dauscher	185.31	0.2
126473256	17800342	computers.desktop	zeta	66.9	0.1
128968633	1802004	electronics.video.tv	yasin	253.49	0.1
149382035	3900652	appliances.enviro...	ariston	118.38	0.1
199915639	4700330	auto.accessories....	ibox	131.28	0.1
213763705	2500577	appliances.kitche...	electrolux	398.72	0.1
213763705	2700588	appliances.kitche...	indesit	293.03	0.1
214859623	1004767	electronics.smart...	samsung	246.96	0.2
219406386	13200475	furniture.bedroom...	permanence	864.86	0.1
224421853	3900805	appliances.enviro...	shivaki	76.94	0.1
226242984	1005135	electronics.smart...	apple	1728.73	0.1
229226898	1005121	electronics.smart...	apple	941.77	0.1
229356564	3900493	appliances.enviro...	ariston	285.33	0.7
237973968	4803977	electronics.audio...	samsung	107.16	0.1
240522111	4804055	electronics.audio...	apple	190.45	0.5

only showing top 20 rows

추천 시스템

- 모델의 학습은 user_id, product_id, total_score값만 포함된 Dataframe 생성하여 진행함.
- ALS 추천 모델을 만들고 RMSE 평가 지표를 사용하여 모델을 평가함.
- RMSE는 0.445 정도로 준수한 수준을 보임.

```
RMSE: 0.44508719566614777
```

추천 시스템

- 제품 id DataFrame과 특정 유저가 본 제품 관련 DataFrame 생성함.
- 이 둘을 product_id열을 통해 join하여 이 사용자가 보지 못한 제품들을 추출함.
- 이후 만들어진 ALS모델을 적용하여 총점을 예측하고 내림차순으로 제품정보와 같이 보여줌.

추천 시스템

- ex) user_id가 219406386인 유저가 볼만한 제품 상위 5개 추천

product_id	user_id	prediction	category_code	brand	price
5100789	219406386	0.68104655	electronics.clocks	garmin	236.5
1701356	219406386	0.5485222	computers.peripherals.monitor	asus	549.8
44800012	219406386	0.4855878	sport.tennis	babolat	164.48
100009895	219406386	0.47270906	computers.components.power_supply	corsair	137.47
3901170	219406386	0.46167257	appliances.environment.water_heater	ariston	179.63

추천 시스템

- 제품과 사용자 간 상호작용 정보가 존재했다면 더 좋은 결과를 보였을 것으로 예상함.
- 이 추천 시스템을 통해 각 유저들에 대한 선호도를 파악하여 전자상거래 사이트 운영자들의 이익 창출과 전자상거래 이용자들에게도 도움이 될 거라 기대됨.

결론 및 기대 효과

- ANOVA test 결과, event 수(활동 수에 비례)에 따라 평균구매가격의 모평균 차가 유의하였다($p<0.001$).
- 이 분석 결과는 전자상거래 사이트를 운영하는 회사의 입장에서 기존 사용자와 신규 사용자 중 더 평균 구매 가격이 낮은 쪽을 위한 프로모션 및 홍보 진행을 위한 근거로 유용하게 사용할 수 있으리라 기대된다.

결론 및 기대 효과

- ALS 추천 모델을 적용하여 특정 사용자에게 선호할만한 제품을 추천하는 함수를 직접 구현하였다.
- 이 추천 시스템을 통해 각 유저들에 대한 선호도를 파악하여 전자상거래 사이트 운영자들의 이익 창출과 전자상거래 이용자들에게도 도움이 될 거라 기대된다.

THANK YOU

감사합니다