주문수요예측방안-연관분석 (장바구니분석)

예) 기저귀-맥주(미국월마트분석)

1. 고객들은 어떤 상품들을 동시에 구매하는가?
2. 라면을 구매한 고객은 주로 다른 어떤 상품을 구매하는가?

위와 같은 질문에 대한 분석을 토대로 고객들에게 SMS를 보낸다든가, 판촉용 전화를 한다든가 묶음 판매를 기획함.

이와 같은 질문에 대한 답은 연관규칙을 이용하여 구할 수 있습니다. 연관규칙은 상업 데이터베이스에서 가장 흔히 쓰이는 도구로, 어떤 사건이 얼마나 자주 동시에 발생하는가를 표현하는 규칙 또는 조건을 의미.

 # support: 지지도는 품목 A와 B를 동시에 구매할 확률인 P(A∩B)를 나타냅니다

 # confidence: 신뢰도는 품목 A가 구매하고나서, 품목 B 가 구매될 확률

 # lift: 향상도는 A를 구매한 사람이 B를 구매할 확률과 A의 구매와 상관없이 B를 구매할 확률의 비율

 # lift>1 이면 관련도가 높고 lift<1 이면 A구매자가 B를 구매하지 않을 확률이 높음

참고사이트: https://ratsgo.github.io/machine%20learning/2017/04/08/apriori/

연관성 분석

https://m.blog.naver.com/PostView.nhn?blogId=gkenq&logNo=10188110816&proxyReferer=https:%2F%2Fwww.google.com%2F

https://www.youtube.com/watch?v=RR0bgIKD_p8

https://jsideas.net/jdj-association-analysis/

https://www.google.com/search?q=%EC%97%B0%EA%B4%80%EB%B6%84%EC%84%9D&source=lnms&tbm=isch&sa=X&ved=2ahUKEwjNsNjPtIjuAhXTFIgKHTVODGsQ_AUoAXoECAYQAw&biw=1600&bih=806#imgrc=1Vm7kvcakvwKLM

https://hamait.tistory.com/743

*연관분석, 장바구니 분석

*지지도(Support): 전체 집합군에서 [조건] 자료가 포함된 집합수, 비율,
[조건1]자료수 / 전체자료수

*신뢰도(Confidence): [조건1]가 있을때 [조건2]도 같이 있는 확률
[조건1]->[조건2] 라고 하면
[조건1],[조건2] 가 같이나온 자료수/[조건1] 자료수

즉: [조건1],[조건2] 지지도 / [조건1] 지지도

*향상도(Lift:Improvement):
[조건1][조건2]가 같이 나온 자료수/[조건1]자료수/전체자료수

연관분석 - 화장품전문점 패키지 구성방법?

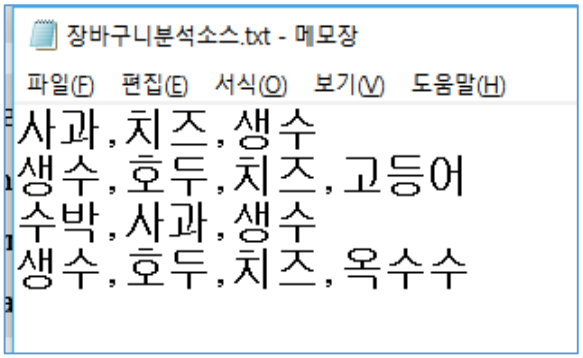
분류	내용
예제 데이터	▪ B화장품전문점에서 판매된 트랜잭션 데이터
변수명	▪ 단일변수 - Nail Polish(매니큐어), Brushes(브러시), - Concealer(컨실러: 피부 결점을 감추어 주는 화장품) - Bronzer(피부를 햇볕에 그을린 것처럼 보이게 하는 화장품) - Lip liner(입술 라이너), Mascara(마스카라: 속눈썹용 화장품) - Eye shadow(아이섀도: 눈꺼풀에 바르는 화장품) - Foundation(파운데이션: 가루분), Lip Gloss(립글로스: 입술 화장품) - Lipstick(립스틱), Eyeliner(아이 라이너: 눈의 윤곽 그림)
분석문제	▪ 전체 트랜잭션 개수와 상품아이템 유형은 몇 개인가? ▪ 가장 발생빈도가 높은 상품아이템은 무엇인가? ▪ 지지도를 10%로 설정했을 때의 생성되는 규칙의 가지수는? ▪ 상품아이템 중에서 가장 발생확률이 높은 아이템과 낮은 아이템은 무엇인가? ▪ 가장 발생가능성이 높은 <2개 상품간>의 연관규칙은 무엇인가? ▪ 가장 발생가능성이 높은 <2개 상품이상에서> <제3의 상품으로>의 연관규칙은?

판매촉진 - 프로모션 효율화 방안

[우체국 쇼핑부문] 쇼핑몰 이용고객을 위한 추천상품 분석

분류	내용
예제 데이터	▪ 우체국 쇼핑에서 판매된 트랜잭션 데이터파일
변수명	▪ 단일변수: 의류(clothes), 냉동식품(frozen), 주류(alcohol,) 야채(veg), 제과(bakery), 육류(meat), 과자(snack), 생활장식(deco)에 대한 거래처리데이터
분석문제	▪ 전체 트랜잭션 개수와 상품아이템 유형은 몇 개인가? ▪ 가장 발생빈도가 높은 상품아이템은 무엇인가? ▪ 지지도를 10%로 설정했을 때의 생성되는 규칙의 가지수는? ▪ 상품아이템 중에서 가장 발생확률이 높은 아이템과 낮은 아이템은 무엇인가? ▪ 가장 발생가능성이 높은 <2개 상품간>의 연관규칙은 무엇인가? ▪ 가장 발생가능성이 높은 <2개 상품이상에서> <제3의 상품으로>의 연관규칙은?

연관성 분석(지지도, 신뢰도, 향상도)




```
dataset=[['사과','치즈','생수'],
         ['생수','호두','치즈','고등어'],
         ['수박','사과','생수'],
         ['생수','호두','치즈','옥수수']]
```

사과를 구매한 고객이 치즈도 함께구매할 연관성에 대해 분석

지지도=P(A∩B) **신뢰도=P(A∩B)/P(A)** **향상도=신뢰도(A,B)/지지도(B)**

▶ 지지도=[사과][치즈]가 같이 나온 자료/전체자료 => 1/4 =>0.25

구매자번호	제품명
1	사과
	치즈
	생수
2	생수
	호두
	치즈
	고등어
3	수박
	사과
	생수
4	생수
	호두
	치즈
	옥수수

▶ 신뢰도=[사과][치즈]가 같이 나온 자료/[사과]자료 => 1/2 =>0.5

구매자번호	제품명
1	**사과**
	치즈
	생수
2	생수
	호두
	치즈
	고등어
3	수박
	사과
	생수
4	생수
	호두
	치즈
	옥수수

▶ 향상도=0.5/0.75=0.6666667

구매자번호	제품명
1	**사과**
	치즈
	생수
2	생수
	호두
	치즈
	고등어
3	수박
	사과
	생수
4	생수
	호두
	치즈
	옥수수

항목별 지지도[Support]

번호	제품명	지지도(자료수/4)	
1	고등어	1	0.25
2	사과	2	0.5
3	생수	4	1
4	수박	1	0.25
5	옥수수	1	0.25
6	치즈	3	0.75
7	호두	2	0.5

연관성 분석(지지도, 신뢰도, 향상도)

미션: 생수를 구매한 사람이 치즈를 구매할 연관성에 대한 분석

지지도=P(A∩B) **신뢰도=P(A∩B)/P(A)** **향상도=신뢰도(A,B)/지지도**

▶ 지지도=

구매자번호	제품명
1	사과
	치즈
	생수
2	생수
	호두
	치즈
	고등어
3	수박
	사과
	생수
4	생수
	호두
	치즈
	옥수수

▶ 신뢰도=

구매자번호	제품명
1	사과
	치즈
	생수
2	생수
	호두
	치즈
	고등어
3	수박
	사과
	생수
4	생수
	호두
	치즈
	옥수수

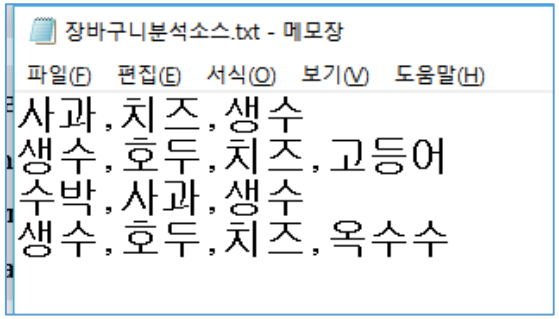
▶ 향상도=

구매자번호	제품명
1	사과
	치즈
	생수
2	생수
	호두
	치즈
	고등어
3	수박
	사과
	생수
4	생수
	호두
	치즈
	옥수수

연관성 분석

```
!pip install  pandas
!pip install  mlxtend
```

```
dataset=[['사과','치즈','생수'],
         ['생수','호두','치즈','고등어'],
         ['수박','사과','생수'],
         ['생수','호두','치즈','옥수수']]
```




```
import pandas as pd
from mlxtend.preprocessing import TransactionEncoder
from mlxtend.frequent_patterns import apriori

te = TransactionEncoder()
te_ary = te.fit(dataset).transform(dataset)
df = pd.DataFrame(te_ary, columns=te.columns_)
frequent_itemsets = apriori(df, min_support=0.3, use_colnames=True)

frequent_itemsets
```

	support	itemsets
0	0.50	(사과)
1	1.00	(생수)
2	0.75	(치즈)
3	0.50	(호두)
4	0.50	(사과, 생수)
5	0.75	(생수, 치즈)
6	0.50	(호두, 생수)
7	0.50	(호두, 치즈)
8	0.50	(호두, 생수, 치즈)

지지도 0.3 이상

참고: min_support=0.0 으로 하면 모든 경우의 수 다 나옴

지지도 0.3 이상 자료

지지도 30% 이상인 자료 8개 있음
고등어(0.25), 수박(0.25), 옥수수(0.25) 나타나지 않음.
사과,치즈도 지지도가 0.25 이므로 나타나지 않음

	support	itemsets
0	0.50	(사과)
1	1.00	(생수)
2	0.75	(치즈)
3	0.50	(호두)
4	0.50	(사과, 생수)
5	0.75	(생수, 치즈)
6	0.50	(호두, 생수)
7	0.50	(호두, 치즈)
8	0.50	(호두, 생수, 치즈)

항목별 지지도[Support]			
번호	제품명	지지도(자료수/4)	
1	고등어	1	0.25
2	사과	2	0.5
3	생수	4	1
4	수박	1	0.25
5	옥수수	1	0.25
6	치즈	3	0.75
7	호두	2	0.5

연관성 분석

```
22 frequent_itemsets = apriori(df, min_support=0.3, use_colnames=True)
23
24 frequent_itemsets
```

[30] :

	support	itemsets
0	0.50	(사과)
1	1.00	(생수)
2	0.75	(치즈)
3	0.50	(호두)
4	0.50	(사과, 생수)
5	0.75	(생수, 치즈)
6	0.50	(호두, 생수)
7	0.50	(호두, 치즈)
8	0.50	(호두, 생수, 치즈)

use_colnames 표시

In [32] :

```
1 frequent_itemsets = apriori(df, min_support=0.3)
2
3 frequent_itemsets
```

Out [32] :

	support	itemsets
0	0.50	(1)
1	1.00	(2)
2	0.75	(5)
3	0.50	(6)
4	0.50	(1, 2)
5	0.75	(2, 5)
6	0.50	(2, 6)
7	0.50	(5, 6)
8	0.50	(2, 5, 6)

신뢰도 0.3 이상 자료 보기

```
from mlxtend.frequent_patterns import association_rules

association_rules(frequent_itemsets, metric="confidence", min_threshold=0.3)
```

	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction
0	(고등어)	(생수)	0.25	1.00	0.25	1.000000	1.000000	0.0000	inf
1	(고등어)	(치즈)	0.25	0.75	0.25	1.000000	1.333333	0.0625	inf
2	(치즈)	(고등어)	0.75	0.25	0.25	0.333333	1.333333	0.0625	1.125000
3	(호두)	(고등어)	0.50	0.25	0.25	0.500000	2.000000	0.1250	1.500000
4	(고등어)	(호두)	0.25	0.50	0.25	1.000000	2.000000	0.1250	inf
5	(사과)	(생수)	0.50	1.00	0.50	1.000000	1.000000	0.0000	inf
6	(생수)	(사과)	1.00	0.50	0.50	0.500000	1.000000	0.0000	1.000000
7	(사과)	(수박)	0.50	0.25	0.25	0.500000	2.000000	0.1250	1.500000
8	(수박)	(사과)	0.25	0.50	0.25	1.000000	2.000000	0.1250	inf
9	(사과)	(치즈)	0.50	0.75	0.25	0.500000	0.666667	-0.1250	0.500000
10	(치즈)	(사과)	0.75	0.50	0.25	0.333333	0.666667	-0.1250	0.750000

참고: min_threshold=0.0 으로 하면 모든 경우의 수 다 나옴

연관성 분석

향상도 0.0이상인 자료보기 (자료모두보기)

```
rules = association_rules(frequent_itemsets, metric="lift", min_threshold=0.0)
rules
```

	antecedents	consequents	antecedent support	consequent support	support	confidenc	lift	everage	conviction
0	(사과)	(고등어)	0.50	0.25	0.00	0.000000	0.000000	-0.1250	0.750000
1	(고등어)	(사과)	0.25	0.50	0.00	0.000000	0.000000	-0.1250	0.500000
2	(생수)	(고등어)	1.00	0.25	0.25	0.250000	1.000000	0.0000	1.000000
3	(고등어)	(생수)	0.25	1.00	0.25	1.000000	1.000000	0.0000	inf
4	(고등어)	(수박)	0.25	0.25	0.00	0.000000	0.000000	-0.0625	0.750000
5	(수박)	(고등어)	0.25	0.25	0.00	0.000000	0.000000	-0.0625	0.750000
6	(고등어)	(옥수수)	0.25	0.25	0.00	0.000000	0.000000	-0.0625	0.750000
7	(옥수수)	(고등어)	0.25	0.25	0.00	0.000000	0.000000	-0.0625	0.750000
8	(고등어)	(치즈)	0.25	0.75	0.25	1.000000	1.333333	0.0625	inf
9	(치즈)	(고등어)	0.75	0.25	0.25	0.333333	1.333333	0.0625	1.125000
10	(호두)	(고등어)	0.50	0.25	0.25	0.500000	2.000000	0.1250	1.500000
11	(고등어)	(호두)	0.25	0.50	0.25	1.000000	2.000000	0.1250	inf
12	(사과)	(생수)	0.50	1.00	0.50	1.000000	1.000000	0.0000	inf
19	(치즈)	(사과)	0.75	0.50	0.25	0.333333	0.666667	-0.1250	0.750000

lift 는 1이 기본임. 1초과인 자료는 묶음시 효과있음(긍정)
1미만인 자료는 묶음시 효과 떨어짐(0.6 일때는 40%의 감소효과, 부정)
(호두),(고등어) 2 => 고등어는 치즈에 많이 종속적임
(치즈),(사과) 0.666667 => 사과는 치즈와 많이 연관안됨
===> 치즈는 3번 팔렸으며 사과에1번종속됨 -> confidenc는 30%