

연관분석 (장바구니 분석)

TECHNOLOGY & PROGRAMMER

목차

- 1 주문수요예측방안-연관분석 (장바구니분석) 정의
 - 1. 연관분석(Association Analysis)개념
 - 2. 연관분석 예시
 - 3. 장바구니 분석 데이터셋의 구성
- 2 연관분석 원리
 - 1. 동시출현관계
 - 2. X와 Y의 방향성
 - 3. 척도를 통한 예측
- 3 연관분석 장단점
- 4 실전 파이썬 연관규칙(Apriori algorithm) 알고리즘

Part1. 주문수요예측방안-연관분석 (장바구니분석) 정의

1. 연관분석(Association Analysis)개념

- 고객들은 어떤 상품들을 동시에 구매하는가?
- 라면을 구매한 고객은 주로 다른 어떤 상품을 구매하는가?

위와 같은 질문에 대한 분석을 토대로 고객들에게 **SMS**를 보낸다든가, 판촉용 전화를 한다든가 묶음 판매를 기획할수 있음.

이와 같은 질문에 대한 답은 연관규칙을 이용하여 구할 수 있으며

연관규칙은 상업 데이터베이스에서 가장 흔히 쓰이는 도구로, 어떤 사건이 얼마나 자주 동시에 발생하는가를 표현하는 규칙 또는 조건을 의미(데이터 내부의 연관성, 즉 상품과 상품간의 상호관계 또는 종속관계를 찾아내 분석함)

월마트 사례

- 90년대 중반 월마트에서는 매주 수요일 저녁마다 기저귀와 맥주의 매출이 함께 높아졌다고 함
- 마트측에서 맥주와 기저귀 진열대를 일부러 가까운 곳에 붙여 놓았더니 놀랍게도 두 제품의 매출이 전날보다 5배나 상승

Part1. 주문수요예측방안-연관분석 (장바구니분석) 정의

2. 연관분석 예시

연관분석 - 화장품전문점 패키지 구성방법?

구매고객을 위한 추천상품

분류	분석 내용
예제 데이터	<ul style="list-style-type: none"> ■ B화장품전문점에서 판매된 트랜잭션 데이터
변수명	<ul style="list-style-type: none"> ■ 단일변수 <ul style="list-style-type: none"> - 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	<ul style="list-style-type: none"> ■ 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] 쇼핑물 이용고객을 위한 추천상품 분석

분류	내용
예제 데이터	<ul style="list-style-type: none"> ■ 우체국 쇼핑에서 판매된 트랜잭션 데이터파일
변수명	<ul style="list-style-type: none"> ■ 단일변수: <ul style="list-style-type: none"> 의류(clothes), 냉동식품(frozen), 주류(alcohol), 야채(veg), 제과(bakery), 육류(meat), 과자(snack), 생활장식(deco)에 대한 거래처리데이터
분석문제	<ul style="list-style-type: none"> ■ 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?

Part1. 주문수요예측방안-연관분석 (장바구니분석) 정의

2. 연관분석 예시

유통업	고객들이 함께 구매할 상품 제안
호텔/숙박업	고객들이 특정 서비스를 받은 후 어떤 서비스를 다음으로 원하는지 선제안 가능
금융사	기존 금융 서비스 내역을 통해 대출 같은 특정 서비스를 받을 가능성이 높은 고객 발굴
보험사	정상적인 청구 패턴과 다른 패턴을 보이는 고객을 찾아 추가 조사 실시가능

Part1. 주문수요예측방안-연관분석 (장바구니분석) 정의

3. 장바구니 분석 데이터셋의 구성

	사과	치즈	생수	호두	고등어	옥수수	수박
고객1	1	1	1				
고객2		1	2	3	1		
고객3	1		1				1
고객4		2	3	1		1	

예) csv또는 txt 데이터셋

사과,치즈,생수
치즈,생수,호두,고등어
사과,생수,수박
치즈,생수,호두,옥수수

예) 데이터프레임 □ list 로 변경해야 작업가능

고객	상품	고객	상품
1	사과	3	사과
1	치즈	3	생수
1	생수	3	수박
2	치즈	4	치즈
2	생수	4	생수
2	호두	4	호두
2	고등어	4	옥수수

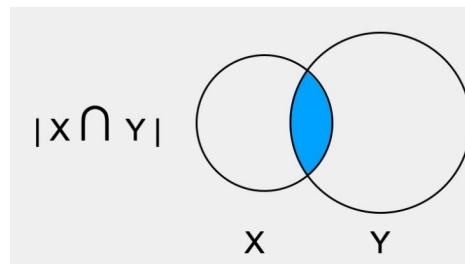
[[사과,치즈,생수],
[치즈,생수,호두,고등어],
[사과,생수,수박],
[치즈,생수,호두,옥수수]]

Part2. 연관분석 원리

1. 동시출현관계

동시출현 빈도는 연관성 분석에 있어 가장 중요한 기본 지표로서 구매고객 데이터셋의 사과를 X, 생수를 Y라고 표현한다면 이 두개를 함께 구매한 고객 1과 고객 3은 총 2명이기 때문에 동시 발생빈도가 2가 됨

	사과	치즈	생수	호두	고등어	옥수수	수박
고객1	1	1	1				
고객2		1	2	3	1		
고객3	1		1				1
고객4		2	3	1		1	



참고

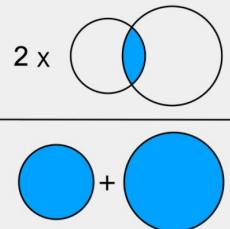
- 각 집단의 개별적인 이해, 합집합일때의 이해, 그리고 각 집단의 수치들을 하나의 지표로 표현하기 위한 대표값을 확인하는 것은 연관성을 이해하는데 참고할 수 있는 정보임.
- 이러한 정보를 얻기 위한 다이스계수, 자카드계수, 코사인계수가 있음

Part2. 연관분석 원리

1. 동시출현관계

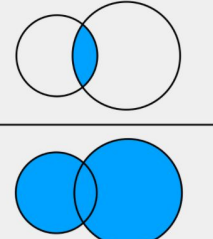
다이스(Dice) 계수

X를 구매한 사람과 Y를 구매한
사람의 합계

$$\frac{2x|X \cap Y|}{|X| + |Y|}$$


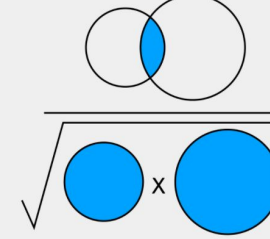
자카드(Jaccard) 계수

X 또는 Y를 하나라도
구매한 사람의 수

$$\frac{|X \cap Y|}{|X \cup Y|}$$


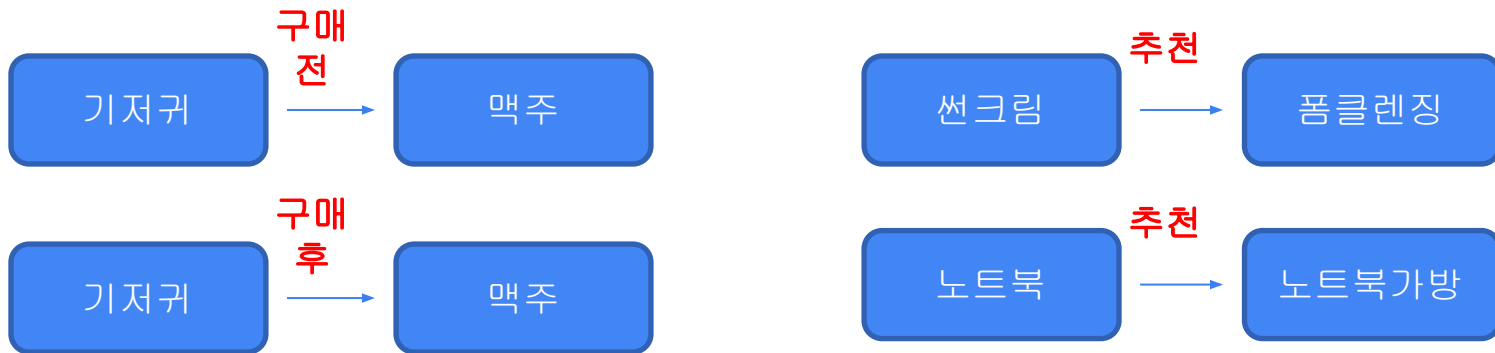
코사인(Cosine) 계수

X를 구매한 사람과 Y를 구매한
사람의 기하평균

$$\frac{|X \cap Y|}{\sqrt{|X| |Y|}}$$


Part2. 연관분석 원리

2. X와 Y의 방향성



[고려대상, 연관규칙에서 향상도]

- A→B의 연관 규칙에서 임의로(Random) B가 구매되는 경우에 비해 A와의 관계가 고려되어 구매되는 경우의 비율
- '항목 A'와 '항목 B'의 향상도 = 3.xxx 이라면 B만 구매할때보다 A를 구매했을때 B를 구매할 확률이 3.xxx배 높음으로 해석
- 즉, 연관 규칙이 오른쪽 항목을 예측하기 위한 얼마나 향상되었는가를 표현하는 값임.

Part2. 연관분석 원리

3. 척도를 통한 예측

- 지지도

(Support)

전체 데이터에서 [조건] 자료가 포함된 집합수, 비율, 조건 1]자료수 / 전체자료수

- 신뢰도

(Confidence)

[조건 1]과 있을 때 [조건 2]도 같이 있는 확률 □ [조건 1],[조건 2] 지지도 / [조건 1] 지지도
X를 구매한 사람이 Y도 구매할 확률

- 향상도(Lift:Improvement) □ 1은 연관성 없음.

1이상이어야 연관성이 있음으로

[조건 1][조건 2]가 같이 나온 자료수/[조건 1]자료수/전체자료수

장바구니분석소스.txt - 메모장

파일(F) 편집(E) 서식(O) 보기(V) 도움말(H)

사과, 치즈, 생수
생수, 호두, 치즈, 고등어
수박, 사과, 생수
생수, 호두, 치즈, 옥수수

	사과	치즈	생수	호두	고등어	옥수수	수박
고객1	0	0	0				
고객2		0	0	0	0		
고객3	0		0				0
고객4		0	0	0		0	

항목	고객몇명이 구매했는지	지지도 (빈도수/고객수4)
고등어	1	0.25
사과	2	0.5
생수	4	1
수박	1	0.25
옥수수	1	0.25
치즈	3	0.75
호두	2	0.5
사과,치즈	1	0.25
생수,치즈	3	0.75
치즈,생수,호두	2	0.5

Part2. 연관분석 원리

3. 척도를 통한 예측

사과를 구매한 고객이 치즈도 함께구매할 연관성에 대해 분석

지지도= $P(A \cap B)$

신뢰도= $P(A \cap B)/P(A)$

향상도= $\text{신뢰도}(A,B)/\text{지지도}(B)$

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

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

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

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

▶ 향상도= $0.5/0.25=0.6666667$

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

Part2. 연관분석 원리

3. 척도를 통한 예측

- 신뢰도는 데이터 추출의 기준점을 제공해줄수 있음

	사과	치즈	생수	호두	고등어	옥수수	수박
고객1	○	○	○				
고객2		○	○	○	○		
고객3	○		○				○
고객4		○	○	○		○	

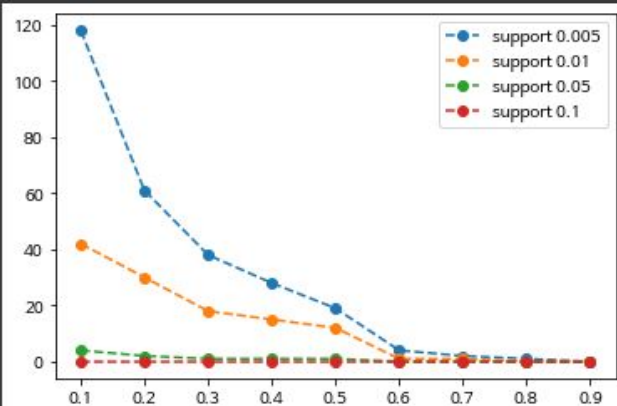
A	B	$P(A \cap B) / P(A)$	%
생수	고등어	1/4	25%
생수	사과	2/4	50%
생수	수박	1/4	25%
생수	옥수수	1/4	25%
생수	치즈	3/4	75%
생수	호두	2/4	50%

	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction
0	(생수)	(고등어)	1.0	0.25	0.25	0.25	1.0	0.0	1.0
7	(생수)	(사과)	1.0	0.50	0.50	0.50	1.0	0.0	1.0
13	(생수)	(수박)	1.0	0.25	0.25	0.25	1.0	0.0	1.0
14	(생수)	(옥수수)	1.0	0.25	0.25	0.25	1.0	0.0	1.0
17	(생수)	(치즈)	1.0	0.75	0.75	0.75	1.0	0.0	1.0
19	(생수)	(호두)	1.0	0.50	0.50	0.50	1.0	0.0	1.0

Part2. 연관분석 원리

3. 척도를 통한 예측

```
1 #####
2 ## 지지도 10%, 5%의 경우 생성되는 규칙이 매우 적음
3 ## 지지도 0.5%의 경우 생성되는 규칙이 너무 많음
4 ## 적정선인 지지도 1%를 선택하고자함.
5 ## 최소 50%의 신뢰도에서 15개 정도의 규칙이 생성되므로
6 ## 그 이상으로 신뢰도를 선택하고자함.
7 #####
8
9 con_list=[0.1,0.2,0.3,0.4,0.5,0.6,0.7,0.8,0.9]
10 support_list=[0.005, 0.01, 0.05, 0.1]
11
12 for y,name in zip(total,support_list):
13
14     plt.plot(con_list,y,'o',linestyle='dashed', label='support ' + str(name*100) + '%')
15     plt.legend()]
16
17 #####
18 ### 최소 지지도 1%(0.01), 최소 신뢰도 50%(0.05) 이상인 연관 규칙들을 생성
19 #####
20
21 frequent_itemsets = apriori(df, min_support=0.01, use_colnames=True)
22 rules= association_rules(frequent_itemsets, metric="confidence", min_threshold=0.05)
23 rules.head()
```



Part2. 연관분석 원리

3. 척도를 통한 예측

	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction
0	(생수)	(고등어)	1.0	0.25	0.25	0.25	1.0	0.0	1.0
7	(생수)	(사과)	1.0	0.50	0.50	0.50	1.0	0.0	1.0
13	(생수)	(수박)	1.0	0.25	0.25	0.25	1.0	0.0	1.0
14	(생수)	(옥수수)	1.0	0.25	0.25	0.25	1.0	0.0	1.0
17	(생수)	(치즈)	1.0	0.75	0.75	0.75	1.0	0.0	1.0
19	(생수)	(호두)	1.0	0.50	0.50	0.50	1.0	0.0	1.0
30	(생수)	(치즈, 고등어)	1.0	0.25	0.25	0.25	1.0	0.0	1.0
36	(생수)	(호두, 고등어)	1.0	0.25	0.25	0.25	1.0	0.0	1.0
49	(생수)	(사과, 수박)	1.0	0.25	0.25	0.25	1.0	0.0	1.0
55	(생수)	(치즈, 사과)	1.0	0.25	0.25	0.25	1.0	0.0	1.0
60	(생수)	(치즈, 옥수수)	1.0	0.25	0.25	0.25	1.0	0.0	1.0
66	(생수)	(호두, 옥수수)	1.0	0.25	0.25	0.25	1.0	0.0	1.0
73	(생수)	(치즈, 호두)	1.0	0.50	0.50	0.50	1.0	0.0	1.0
92	(생수)	(치즈, 호두, 고등어)	1.0	0.25	0.25	0.25	1.0	0.0	1.0
106	(생수)	(치즈, 호두, 옥수수)	1.0	0.25	0.25	0.25	1.0	0.0	1.0

Part3. 연관분석 장단점

장점

- 탐색적인 기법으로 조건 반응으로 표현되는 연관성분석의 결과를 쉽게 이해할 수 있다
- 비목적성 분석기법으로 분석 방향이나 목적이 특별히 없는 경우 목적변수가 없으므로 유용하게 활용됨
- 사용이 편리한 분석 데이터의 형태로 거래 내용에 대한 데이터를 변환 없이 그 자체로 이용할 수 있는 간단한 자료 구조를 갖는다
- 분석을 위한 계산이 간단(품목이 많지 않다는 전제 하에)

단점

- **품목수가 증가하면 분석에 필요한 계산은 기하급수적으로 늘어난다**
 - 유사한 품목을 한 범주로 일반화
 - 연관 규칙의 신뢰도 하한을 새롭게 정의해 빈도수가 작은 연관규칙은 제외
- **너무 세분화한 품목을 갖고 연관성 규칙을 찾으면 의미없는 분석이 될 수도 있다**
 - 적절히 구분되는 큰 범주로 구분해 전체 분석에 포함시킨 후 그 결과 중에서 세부적으로 연관규칙을 찾는 작업을 수행
- **거래량이 적은 품목은 당연히 포함된 거래수가 적을 것이고, 규칙 발견 시 제외하기 쉬움**
 - 그 품목이 관련성을 살펴보고자 하는 중요한 품목이라면 유사한 품목들과 함께 범주로 구성하는 방법 등을 통해 연관성 규칙의 과정에 포함시킬 수 있음

Part3. 연관분석 장단점

	특징	장점	단점
1세대 Apriori 알고리즘	<ul style="list-style-type: none">▪ 부분집합의 개수를 줄이는 방식▪ 최소지지도보다 큰 지지도값을 갖는 빈발항목집합(frequent item set) 대해서만 연관규칙을 계산	<ul style="list-style-type: none">• 알고리즘 구현과 이해가 쉬움	아이템 개수가 많아지면 계산 복잡도 증가
2세대 FP-Growth 알고리즘	<ul style="list-style-type: none">▪ Apriori 알고리즘의 약점을 보완▪ 후보 빈발항목집합을 생성하지 않고▪ FP-Tree(Frequent Pattern Tree)를 만든 후 분할 정복 방식을 통해 빈발항목집합을 추출	<ul style="list-style-type: none">• Apriori 알고리즘 보다 빠르게 빈발항목집합을 추출• 데이터베이스를 스캔하는 횟수가 작고 빠른 속도로 분석 가능	
3세대 FPV	<ul style="list-style-type: none">▪ 메모리를 효율적으로 사용해 SKU 레벨의 연관성분석 가능		

Part4. 파이썬 연관규칙(Apriori algorithm) 알고리즘

1. 정의

1993년 애가왈(Agarwal)과 스리칸트(Srikant)의 해 제안됨.

연관규칙 알고리즘은 지지도가 일정 '이하'인 아이템을 포함하는 조합은 처음부터 신뢰도를 계산하지 않음

```
from mlxtend.preprocessing import TransactionEncoder
from mlxtend.frequent_patterns import apriori

from apyori import apriori
```

주의

- 리스트 구조만 연관규칙 모듈 사용가능

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

2. 시각화(networkx 그래프)

A network graph illustrating relationships between various food items. The central node is 'Coffee' (yellow). Other nodes include 'Soup', 'Scene', 'Toast', 'Cake', 'Cookies', 'Hot chocolate', 'Medicine', 'Alfajores', 'Pastry', 'Bread', 'NONE', 'Sandwich', 'Muffin', 'Juice', and 'Spanish Brunch'. Edges connect 'Coffee' to most other nodes, and there are additional connections between 'Test', 'Bread', and 'NONE'.