연관성 분석(Association Analysis) / 장바구니 분석(Market Basket Analysis) :

전설처럼 회자 되는 **맥주와 기저귀 ~**

연관분석(Association Analysis)

- 상품과 상품사이에 어떤 연관이 있는지 찾는 룰기반의 알고리즘
- 거래형식의 데이터에서 아이템 클러스터 파악

여기서 연관이란~

- 함께 구매 되었는지의 정도(얼마나 자주)
- A상품 구매후 B 아이템을 구매하는지

위 규칙을 찾아내는 것으로 어떤 상품들이 한 바구니 안에 담기는지 살펴보는 모습과 비슷하다고 해서 흔히 장바구니 분석(**Market Basket Analysis**)이라고 많이들 불리고 있음

- 전설처럼 회자되는 일화로 월마트에서 맥주를 구매할때 기저귀를 같이 구매하는 경향이 크다는 것을 확인
- 이후 이 둘을 함께 진열하는 전략을 세움

연관분석(Association Analysis)

- 연관규칙의 활용
  - ✓ 교차판매(Cross Selling)
  - ✓ 묶음판매(Bunding)
  - ✓ 상품진열(Inventory Display)
  - ✓ 쿠폰 제공
  - ✓ 온라인 상품 추천

- 측정도구
  - ✓ 지지도(support)
  - ✓ 신뢰도(confidence)
  - ✓ 향상도(lift)

연관분석(Association Analysis)

Global KU Frontier Spirit

- 연관규칙 생성 과정(비지도 학습방법)

  ✓ **지지도(support)**
  - 빈발 아이템 세트(frequent item set,많이 팔리는 물건들)를 근거로 후보규칙들의 집합 결정
  - 사전에 정의한 지지도(support) 값에 의해 후보군 결정

  ✓ **신뢰도(confidence)**
  - 규칙의 불확실성을 평가하기 위해 신뢰도 사용
  - 최소기준 신뢰도 값을 설정
  - 조건부 확률

  ✓ **향상도(lift)**
  - 임의로 조합된 규칙과의 비교를 통해 해당 규칙이 얼마나 실제적인 연관성을 가지는지 파악

  ✓ 대표 알고리즘 : Apriori Algorithm / FP-Growth Algorithm

● 후보 규칙생성

- ✓ 아이템 사이의 규칙들을 if-then 형식으로 표시
- ✓ 아이템 세트(item sets) : 아이템 들의 집합
  - If 조건부 then 결론부
  - {Item set A} => {Item set B}
- ✓ If 계란 then 맥주 : {계란} => {맥주}
- ✓ If 계란 and 맥주 then 기저귀 : {계란, 맥주} => {기저귀}
- ✓ 빈발 아이템 집합들(frequent item sets)
  - Apriori 알고리즘을 이용하여 빈도수가 높은 조합 생성
  - 지지도(support)를 이용

연관분석(Association Analysis)

연관규칙(빵) -> (버터)
X Y

User \ item	라면	버터	빵	우유	콜라
user_1	-	O	O	O	-
user_2	-	O	-	O	O
user_3	-	O	O	-	O
user_4	O	-	-	O	O
user_5	O	O	O	-	-
	2/5=40%	4/5=80%	3/5=60%	3/5=60%	3/5=60%

✓ 지지도(support) s(X→Y) = X와 Y를 모두 포함하는 거래 수 / 전체 거래 수 = n(X∪Y) / N

✓ 신뢰도(Confidence) c(X→Y) = X와 Y를 모두 포함하는 거래 수 / X가 포함된 거래 수 = n(X∪Y) / n(X)

✓ 향상도(Lift) = $\frac{P(Y|X)}{P(Y)} = \frac{P(X,Y)}{P(X)P(Y)} = \frac{X,Y\text{동시 포함 거래수} \times \text{전체거래수}}{X\text{포함거래수} \times Y\text{ 포함 거래수}}$

연관분석(Association Analysis)

연관규칙(빵) -> (버터)
X Y

User \ item	라면	버터	빵	우유	콜라
user_1	-	O	O	O	-
user_2	-	O	-	O	O
user_3	-	O	O	-	O
user_4	O	-	-	O	O
user_5	O	O	O	-	-
	2/5=40%	4/5=80%	3/5=60%	3/5=60%	3/5=60%

✓ 지지도(support) s(X→Y) = X와 Y를 모두 포함하는 거래 수 / 전체 거래 수 = n(X∪Y) / N

n(1번,3번,5번) / N => 3 / 5 = 0.6

연관분석(Association Analysis)

연관규칙(빵) -> (버터)
X Y

User \ item	라면	버터	빵	우유	콜라
user_1	-	O	O	O	-
user_2	-	O	-	O	O
user_3	-	O	O	-	O
user_4	O	-	-	O	O
user_5	O	O	O	-	-
	2/5=40%	4/5=80%	3/5=60%	3/5=60%	3/5=60%

✓ 신뢰도(Confidence) c(X→Y) = X와 Y를 모두 포함하는 거래 수 / X가 포함된 거래 수 = n(X∪Y) / n(X)

n(1번,3번,5번) / n(1번,3번,5번) => 3 / 3 = 1

= 구매비율(빵, 버터) / 구매비율(빵)

60% / 60% = 1

연관분석(Association Analysis)

연관규칙(빵) -> (버터)
X Y

User \ item	라면	버터	빵	우유	콜라
user_1	-	O	O	O	-
user_2	-	O	-	O	O
user_3	-	O	O	-	O
user_4	O	-	-	O	O
user_5	O	O	O	-	-
	2/5=40%	4/5=80%	3/5=60%	3/5=60%	3/5=60%

✓ 향상도(Lift) = $\frac{P(Y|X)}{P(Y)} = \frac{P(X,Y)}{P(X)P(Y)} = \frac{X,Y\text{동시 포함 거래수} \times \text{전체거래수}}{X\text{포함거래수} \times Y\text{ 포함 거래수}}$ -> 15 / 12 = 1.25

✓향상도lift(빵 -> 버터) = $\frac{\text{구매비율(빵,버터)}}{\text{구매비율(빵)}\times\text{구매비율(버터)}} = \frac{60}{60 \times 80} \times 100 = 1.25$

Global KU Frontier Spirit

실 습

```
from mlxtend.frequent_patterns import association_rules
association_rules(itemset, metric="confidence", min_threshold=0.1)
```

	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction
0	(버터)	(빵)	0.8	0.6	0.6	0.75	1.25	0.12	1.6
1	(빵)	(버터)	0.6	0.8	0.6	1.00	1.25	0.12	inf

User \ item	기저귀	맥주	바나나	우유
user_1	O	O	O	O
user_2	O	-	O	-
user_3	-	-	O	O
user_4	O	O	O	O
user_5	O	O	-	-
user_6	-	-	O	O
user_7	O	O	O	-
user_8	-	-	O	-

정말 기저귀를 구매한 사람이 맥주도 함께 구매한게 맞아?

User \ item	기저귀	맥주	바나나	우유
user_1	O	O	O	O
user_2	O	-	O	-
user_3	-	-	O	O
user_4	O	O	O	O
user_5	O	O	-	-
user_6	-	-	O	O
user_7	O	O	O	-
user_8	-	-	O	-
	5/8=62%	4/8=50%	7/8=87%	4/8=50%

User \ item	기저귀	맥주	바나나	우유
user_1	O	O	O	O
user_2	O	-	O	-
user_3	-	-	O	O
user_4	O	O	O	O
user_5	O	O	-	-
user_6	-	-	O	O
user_7	O	O	O	-
user_8	-	-	O	-
	5/8=62%	4/8=50%	7/8=87%	4/8=50%
기저기와 함께 팔린 상품		4건	4건	2건

연관분석(Association Analysis)

User \ item	기저귀	맥주	바나나	우유
user_1	O	O	O	O
user_2	O	-	O	-
user_3	-	-	O	O
user_4	O	O	O	O
user_5	O	O	-	-
user_6	-	-	O	O
user_7	O	O	O	-
user_8	-	-	O	-
	5/8=62%	4/8=50%	7/8=87%	4/8=50%
기저기와 함께 팔린 상품		4건	4건	2건

$$\text{신뢰도(기저귀} \rightarrow \text{맥주)} = \frac{\text{구매비율(기저귀, 맥주)}}{\text{구매비율(기저귀)}} = \frac{50\%}{62\%} = 0.8$$

$$\text{신뢰도(기저귀} \rightarrow \text{바나나)} = \frac{\text{구매비율(기저귀, 바나나)}}{\text{구매비율(기저귀)}} = \frac{50\%}{62\%} = 0.8$$

$$\text{신뢰도(기저귀} \rightarrow \text{우유)} = \frac{\text{구매비율(기저귀,우유)}}{\text{구매비율(기저귀)}} = \frac{25\%}{62\%} = \mathbf{0.4}$$

연관분석(Association Analysis)

User \ item	기저귀	맥주	바나나	우유
user_1	O	O	O	O
user_2	O	-	O	-
user_3	-	-	O	O
user_4	O	O	O	O
user_5	O	O	-	-
user_6	-	-	O	O
user_7	O	O	O	-
user_8	-	-	O	-
	5/8=62%	4/8=50%	7/8=87%	4/8=50%
기저기와 함께 팔린 상품		4건	4건	2건

$$\text{신뢰도(기저귀} \rightarrow \text{맥주)} = \frac{\text{구매비율(기저귀, 맥주)}}{\text{구매비율(기저귀)}} = \frac{50\%}{62\%} = 0.8$$

$$\text{신뢰도(기저귀} \rightarrow \text{바나나)} = \frac{\text{구매비율(기저귀, 바나나)}}{\text{구매비율(기저귀)}} = \frac{50\%}{62\%} = 0.8$$

바나나는 항상 잘
나가는 상품 이지는 않을까?

$$\text{신뢰도(기저귀} \rightarrow \text{우유)} = \frac{\text{구매비율(기저귀,우유)}}{\text{구매비율(기저귀)}} = \frac{25\%}{62\%} = \mathbf{0.4}$$

상품의 **판매 빈도** 까지 고려하는 지표는 없을까?

$$\text{향상도}Lift(\text{기저귀} \rightarrow \text{맥주}) = \frac{\text{구매비율(기저귀, 맥주)}}{\text{구매비율(기저귀)} \times \text{구매비율(맥주)}} = \frac{50\%}{62\% \times 50\%} = 1.6$$

$$\text{향상도}Lift(\text{기저귀} \rightarrow \text{바나나}) = \frac{\text{구매비율(기저귀, 바나나)}}{\text{구매비율(기저귀)} \times \text{구매비율(바나나)}} = \frac{50\%}{62\% \times 87\%} = 0.91$$

$$\text{향상도}Lift(\text{기저귀} \rightarrow \text{우유}) = \frac{\text{구매비율(기저귀, 우유)}}{\text{구매비율(기저귀)} \times \text{구매비율(우유)}} = \frac{50\%}{62\% \times 50\%} = 0.8$$

위 방법이 사실상
추천 알고리즘의 시작 ~

잘 팔리는데는 이유가 있다.