R 교육 세미나

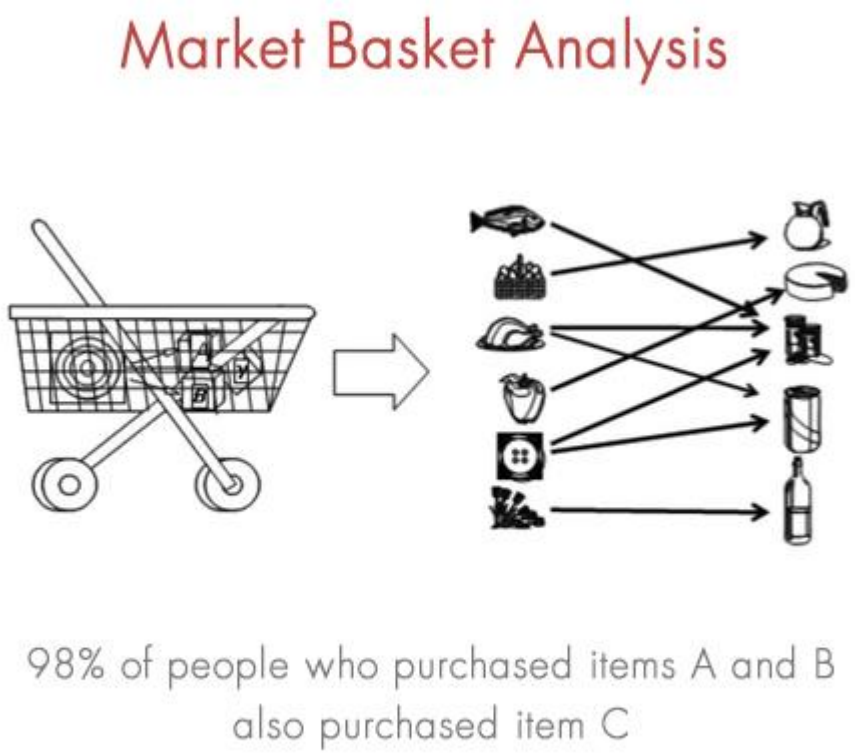
ToBig's 8기 류호성

Association rules analysis

연관성 분석

contents

Unit 01 | 연관성 분석이란?

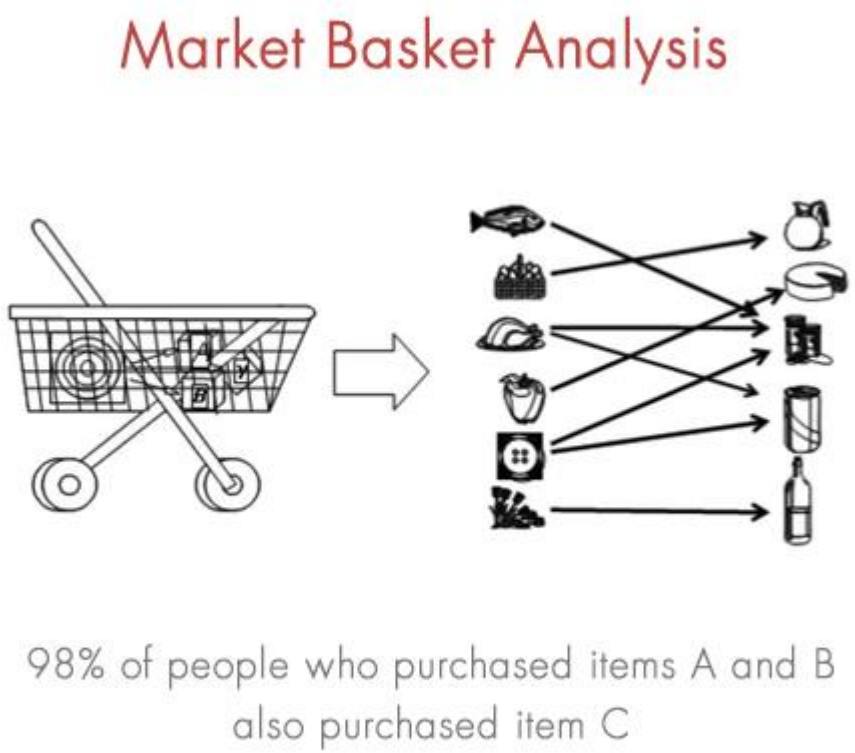


연관성 분석..?
장바구니 분석..?

Q. 계란을 구입하는 사람은 콜라도 구입할까?

→ Yes? Or No?

Unit 01 | 연관성 분석이란?



연관성 분석..?
장바구니 분석..?

Q. 계란을 구입하는 사람은 콜라도 구입할까?

→ Yes? Or No?

위의 답을 찾기 위해서는 마트에 온 다른 사람들의 구입목록을 확인해서 연관 규칙을 찾으면 알 수 있을 것이다.

연관성 분석

Y(종속변수,타겟변수)가 없는 상태에서 데이터 속에 숨겨져 있는 패턴, 규칙을 찾아내는 비지도학습(Unsupervised Learning)의 하나로 데이터 간의 **연관성 및 상관관계**를 표현하는 규칙을 찾아내는 것

Unit 01 | 연관성 분석이란?

" If A then B "

연관 규칙이란 특정 사건(A)이 발생했을 때
함께 (빈번하게) 발생하는 또 다른 사건의 규칙(B)을 의미한다.

Ex) (맥주) → (기저귀) : 맥주를 사는 고객은 기저귀도 같이 산다.
(당근,양파) → (계란) : 당근과 양파를 사는 고객은 계란도 같이 산다.

일반적으로 A를 LHS(Left-hand-side) , B를 RHS(Right-hand-side) 라 지칭한다.

Unit 02 | 연관성 분석 척도

지지도(Support)

두 품목 A와 B의 지지도는 전체 거래항목 중 A와 B가 동시에 포함되는 거래의 비율

지지도는 0~1 값을 가지며, 1에 가까울수록 관련이 높다.

$$P(A \cap B) = \frac{A, B\text{가 동시에 거래된 횟수}}{\text{전체 거래 횟수}} \text{, } Support(A, B)$$

지지도 높다 : 계란과 콜라를 함께 사는 경우가 흔함

Unit 02 | 연관성 분석 척도

신뢰도(Confidence)

품목 A를 구매했을 때, 품목 B가 함께 구매될 확률

신뢰도는 0~1 값을 가지며, 1에 가까울수록 관련이 높다.

A를 구매한 고객 중 B도 구매한 고객의 비율을 알 수 있다!

$$P(B \mid A) = \frac{P(A \cap B)}{P(A)} = \frac{A, B\text{가 동시에 거래된 횟수}}{A\text{가 거래된 횟수}} = \frac{Support(A, B)}{Support(A)}, Confidence(A \Rightarrow B)$$

신뢰도 높다 : 계란을 사는 경우 콜라도 사는 비율이 높다

Unit 02 | 연관성 분석 척도

향상도(Lift)

품목 A가 주어지지 않았을 때의 품목 B의 확률 대비
품목 A가 주어졌을 때의 품목 B의 확률의 증가 비율 이다.

$$\frac{P(B \mid A)}{P(B)} = \frac{P(A \cap B)}{P(A)P(B)} = \frac{A, B\text{가 동시에 거래된 횟수}}{A\text{가 거래된 횟수}} = \frac{Confidence(A \Rightarrow B)}{Support(B)}, Lift(A \Rightarrow B)$$

향상도 높다 : 계란을 사는 경우 그렇지 않은 경우에 비해 콜라도 함께 사는 경우가 많다

Unit 02 | 연관성 분석 척도

향상도(Lift)

$$\frac{P(B \mid A)}{P(B)} = \frac{P(A \cap B)}{P(A)P(B)} = \frac{A,B\text{가 동시에 거래된 횟수}}{A\text{가 거래된 횟수}} = \frac{Confidence(A \Rightarrow B)}{Support(B)}, Lift(A \Rightarrow B)$$

향상도 < 1 : 두 품목 간의 음의 연관 — 음의 상관 : A를 사면, 보통 B를 사지 않는다.
향상도 = 1 : 두 품목 서로 독립 — 독립 : A를 사는 것과 상관없이 B를 산다.
향상도 > 1 : 두 품목 간의 양의 상관 — 양의 상관 : A를 사면, 보통 B를 산다.

양의 상관이 있어야 하므로, Lift 값이 1보다 커야 된다.

Unit 02 | 연관성 분석 척도

고객	구매품목
1	계란, 고구마
2	우유, 계란, 기저귀
3	계란, 콜라
4	계란, 고구마, 콜라
5	기저귀, 고구마

EX) 계란(A) → 고구마(B)

지지도(Support)

$$P(A \cap B) = \frac{2}{5}$$

신뢰도(Confidence)

$$P(B \mid A) = \frac{2/5}{4/5} = \frac{2}{4}$$

향상도(Lift)

$$\frac{P(B \mid A)}{P(B)} = \frac{2/4}{3/5} = \frac{5}{6}$$

Unit 02 | 연관성 분석 척도

고객	구매품목
1	계란, 고구마
2	우유, 계란, 기저귀
3	계란, 콜라
4	계란, 고구마, 콜라
5	기저귀, 고구마

EX) 계란(A) → 콜라(B)

지지도(Support)

$$P(A \cap B) = \frac{2}{5}$$

신뢰도(Confidence)

$$P(B \mid A) = \frac{2/5}{4/5} = \frac{2}{4}$$

향상도(Lift)

$$\frac{P(B \mid A)}{P(B)} = \frac{2/4}{2/5} = \frac{5}{4}$$

Unit 02 | 연관성 분석 척도

연관성 분석 규칙 선택 과정

1 Minimum Support / Minimum Confidence
특정 지지도(Support) 와 신뢰도(Confidence) 이하의 Rule은 Screening out 시키게끔 한다

2 향상도(Lift) sorting (큰 수부터)
향상도(Lift)가 클수록 양의 연관관계가 있으니깐, 향상도를 내림차순으로 Sorting 해서 Rule을 평가한다

- 관심이 있는 품목이 있다면, lhs나 rhs에 있는 rule만을 subset으로 선별해서 보기도 한다

연관성 분석 규칙 선택 과정

앞서 3가지 선택 기준으로 rule을 선별했을 때, 설명 가능하고, 활용 가능한 rule 만이 유의미한 규칙이다.

	설명가능	활용가능	사례
유용한 rule	O	O	기저귀,남성 → 맥주
사소한 rule	O	X	상식적인 규칙: 컴퓨터→ 프린터
설명할 수 없는 rule	X	X	해석 불가능 : 기초화장품 → 자동차 와이퍼

Unit 02 | 연관성 분석 척도

알쓸신척

IS(Interect-Support) 측도 : 향상도(Lift)와 지지도(Support)의 곱에 제곱근을 취한 값

$$IS(A \Rightarrow B) = \sqrt{Lift(A \Rightarrow B) \times Support(A,B)} = \frac{Support(A,B)}{\sqrt{Support(A) \times Support(B)}}$$

➡ 향상도와 지지도가 모두 높을수록 IS 값이 커짐. 둘 중 하나만 낮거나 높은 경우를 screening out 시키고, 둘 다 높은 rule만 선별할 수 있다

교차 지지도(Cross Support) : 최대지지도에 대한 최소지지도의 비율

$$r(X) = \frac{\min\{s(i1), s(i2), \ldots, s(im)\}}{\max\{s(i1), s(i2), \ldots, s(im)\}}$$

➡ 항목집합 $X = \{i1, i2, \ldots im\}$에 대해서 의미 없는 연관규칙의 생성을 방지하기 위해 이용한다. $r(X)$의 값이 매우 작으면 항목집합 X에서 생성되는 연관규칙이 의미가 없을 가능성이 크다.

Unit 03 | 연관성 분석 알고리즘

앞에서 배운 규칙대로 연관성 분석을 하면 되나…?
아니다!! ㅜㅜ 왜냐하면 따져봐야 하는 경우의 수가 너무 많다!

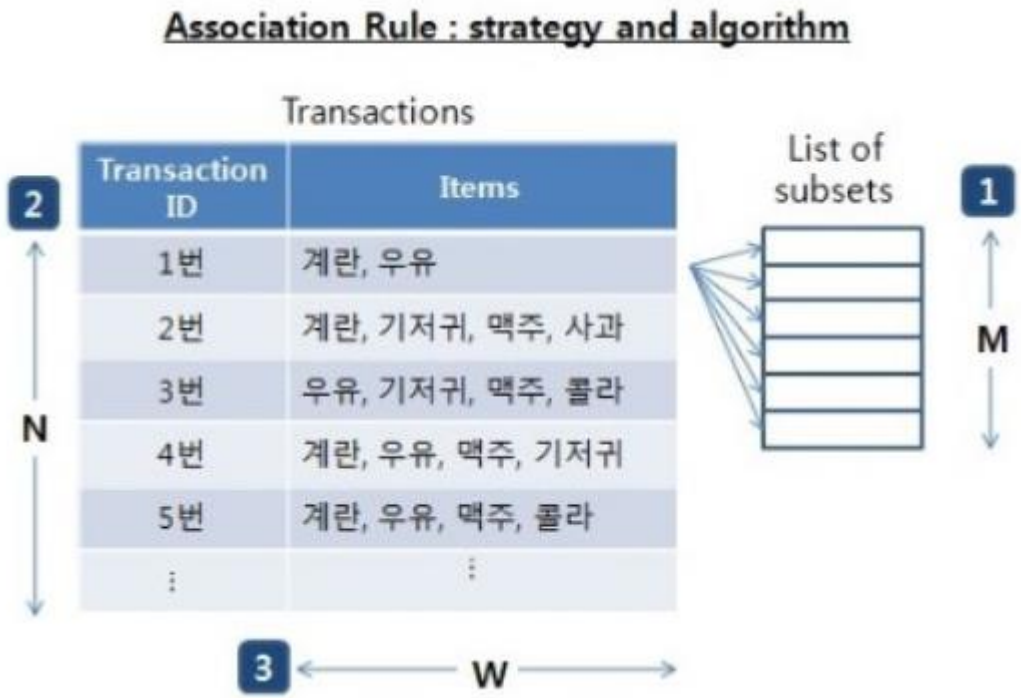
품목수가 k개라면, 모든 가능한 부분 집합의 개수는 공집합을 제외하고 $2^k - 1$

품목수가 k개라면, 모든 가능한 연관 규칙의 개수는 $3^k - 2^k + 1$

품목수가 증가할 때마다, 연산량은 **지수적으로 증가**

→ **빠르고 효율적인 연관 규칙 계산 알고리즘 필요**

Unit 03 | 연관성 분석 알고리즘



1 모든 가능한 항목집합 개수(M)를 줄이는 방식 ➡ Apriori algorithm

2 Transaction 개수(N)를 줄이는 방식 ➡ DHP Algorithm

3 비교하는 수(W)를 줄이는 방식 ➡ FP-growth Algorithm

Apriori Algorithm

: 모든 가능한 항목 집합 개수(M)을 줄이는 방식

최소지지도(Minimum Support) 이상을 갖는 항목집합 = 빈발항목집합
모든 항목 집합 대신 빈발항목집합만을 찾아내러 연관 규칙을 계산하는 방법이다.

Apriori Algorithm 원리

1 한 항목집합이 빈발(frequent)하다면 이 항목집합의 부분집합은 역시 빈발 항목집합이다.
Ex) (계란,콜라) 의 지지도가 0.3보다 크면, 부분집합인 (계란), (콜라)의 지지도도 0.3보다 크다

2 한 항목집합이 비빈발(Infrequent)하다면 이 항목집합을 포함하는 모든 부분집합은 비빈발 항목집합이다.
Ex) (계란,콜라) 의 지지도가 0.2보다 작으면, (계란+콜라+α)의 지지도도 0.2보다 작다

Unit 03 | 연관성 분석 알고리즘

Ex) 최소 지지도 : 0.5

최소 지지도
= 0.5 * (고객수)
= 0.5 * 4
= 2

고객	품목
1	A,C,D
2	B,C,E
3	A,B,C,E
4	B,E

items set	Support
{A}	2
{B}	3
{C}	3
{D}	1
{E}	3

items set	Support
{A,B}	1
{A,C}	2
{A,E}	1
{B,C}	2
{B,E}	3
{C,E}	2

Unit 03 | 연관성 분석 알고리즘

items set	Support
{A,B}	1
{A,C}	2
{A,E}	1
{B,C}	2
{B,E}	3
{C,E}	2

items set	Support
{B,C,E}	2

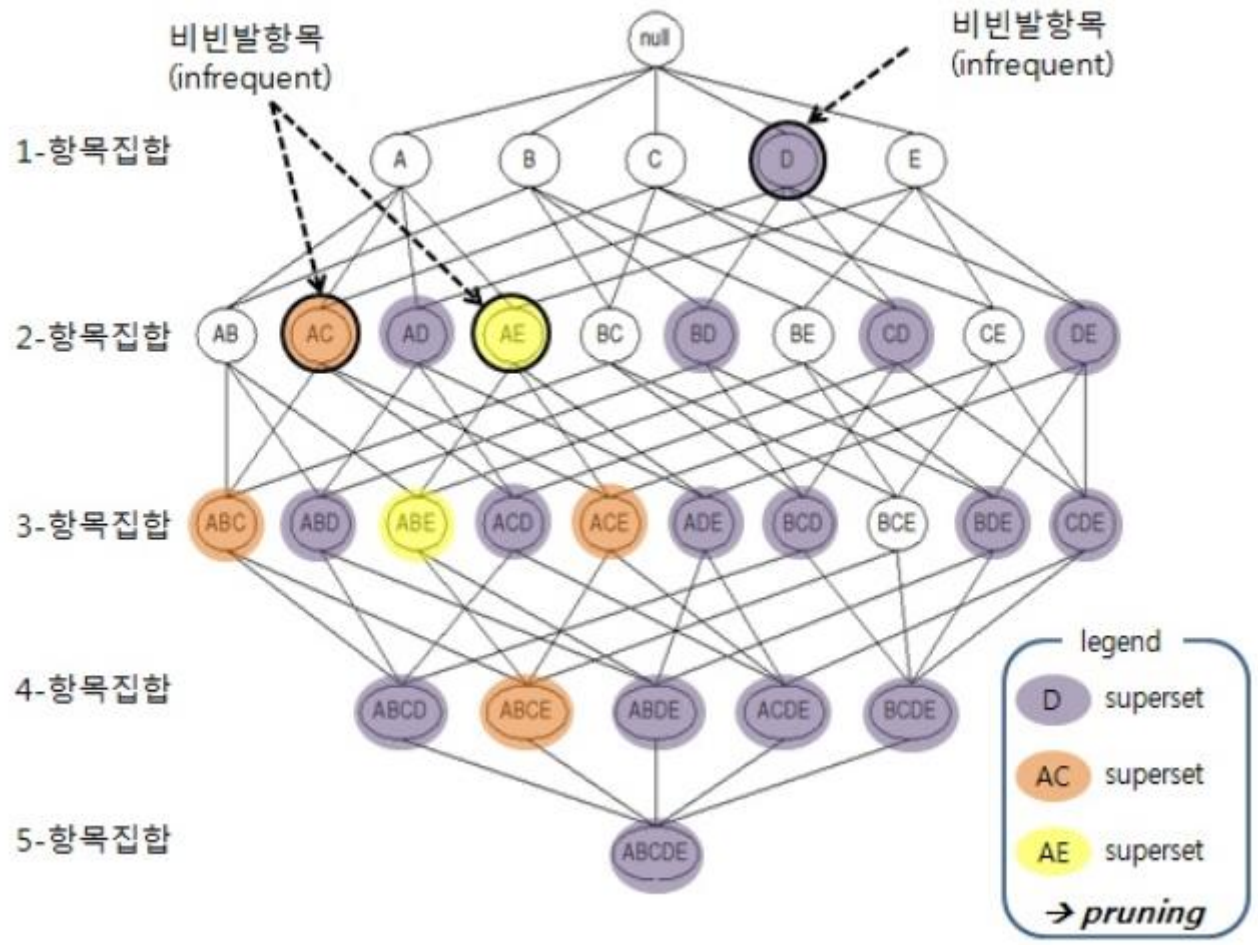
품목 수 5개일 때,
모든 가능한 부분집합의 수:
$2^5 - 1 = 31$

Apriori 규칙 시행 후:
4+4+1 = 9

→ 계산량 감소 효과

Unit 03 | 연관성 분석 알고리즘

{D}, {A, C}, {A, E}가 infrequent item set 일 경우 superset pruning 예시



Unit 04 | 마무리

연관성 분석의 장점

- 결과가 분명하다(If-then 규칙)
- 변수가 많은 경우 쉽게 사용할 수 있다.
- 계산이 용이하다.
- 강력한 비목적성 분석

연관성 분석의 단점

- 품목 수 증가에 따라 계산량이 폭등한다.
- 자료의 속성에 제한이 있다.(연속형 변수 제한)
- 적절한 품목을 결정하기 어렵다.
- 거래가 드문 품목에 대한 정보 부족

들어주셔서 감사합니다.