04 - 연관분석

1. 개요 및 측도

(1) 개요

- 장바구니 분석 ~~ 패턴
  ⇒ 의미 있는 규칙

  조건 따른 결과 형태

- 탐색적 기법의 일종

  조건 반응에 의해 표현

  비지도학습 유형

- But,

  품목 수 ↑ → 분석계산 ↑ (기하급수적)

  너무 세분화된 품목으로 연관규칙 찾기
  ⇒ 의미 X 될수도

~~~~

(2) 측도

① 지지도 (support)

- 전체 거래 中 A, B 두 품목이 동시에 포함된 거래 비율

  지지도 높다 ⇒ 그 두 item이 같이 잘 팔림
~~~~

$$P(A \cap B) = \frac{A, B \text{ 동시 거래 수}}{\text{전체 거래 수}}$$

② 신뢰도

· 어떤 하나의 품목 구매됐을 때

다른 품목 하나가 구매될 확률

⟹ 조건부확률

※ $A \to B$ 와 $B \to A$ 는

서로 다르다.

$$\text{신뢰도}(A \to B) = P(B|A) = \frac{P(A \cap B)}{P(A)} = \frac{A\text{와 } B\text{가 동시 포함 거래 수}}{A\text{가 포함된 거래 수}}$$

$$\text{신뢰도}(B \to A) = P(A|B) = \frac{P(B \cap A)}{P(B)} = \frac{B\text{와 } A\text{가 동시 포함 거래 수}}{B\text{가 포함된 거래 수}}$$

③ 향상도

· 품목 A가 주어지지 X을 때

품목 B가 구매될 확률 대비

품목 A가 구매될 때

품목 B가 구매될 확률

※ 향상도$(A \to B)$, 향상도$(B \to A)$

서로 같다.

향상도 (A → B)

$$= \frac{\text{신뢰도}(A \to B)}{P(B)} = \frac{P(A \cap B)}{P(A)P(B)}$$

향상도 (B → A)

$$= \frac{\text{신뢰도}(B \to A)}{P(A)} = \frac{P(B \cap A)}{P(A)P(B)}$$

ex)

		구매 확률 (리리도)
치킨	100	$\Rightarrow \frac{100+300+100+100}{1000} = 0.6$
콜라	200	$\Rightarrow \frac{200+300+100+100}{1000} = 0.7$
사이다	100	$\Rightarrow \frac{100+100+100+100}{1000} = 0.4$
치킨, 콜라	300	$\Rightarrow \frac{300+100}{1000} = 0.4$
콜라, 사이다	100	$\Rightarrow \frac{100+100}{1000} = 0.2$
사이다, 치킨	100	$\Rightarrow \frac{100+100}{1000} = 0.2$
치킨, 콜라, 사이다	100	$\Rightarrow \frac{100}{1000} = 0.1$
전체	1000	

---

2. 연관분석의 알고리즘과 특징

(1) 알고리즘

① apriori

- 지지도 사용 → 빈발 item 집합 판별

⇒ 계산 복잡도 감소 (낮은 지지도 품목 ⇒ 의미 X)

② apriori 절차

i) 최소 지지도 설정

ii) ↳ 보다 큰 지지도 갖는 단일품목 선별

iii) ↳ 으로 그 지지 품목으로 생성되는 연관규칙 (A → B) 中
최소 지지도 이상의 연관규칙 탐색

iv) iter
3가지 이상 품목에 대한 연관규칙 생성

③ FP-Growth

- why? data set 큰 경우,
모든 item set 다 생성하는 것
⇒ 비효율적

- 지지도 낮은 것 ——→ 놓은 것

⇒ 빈도 수 높은 item set 생성
(상향식)

(2) 연관분석 특징

- 장점
  - 결과가 단순, 분명
  - 분석 위한 계산 간단
  - 목적변수 X
    → data srch 로도 가능

· 단점

- 품목 세분화에 어려움

- 품목 수 ↑ ⟶ 계산량 ↑↑

- 거래 미발생 품목은 분석 불가