

3-4 연관 분석

연관분석

연관 규칙 분석

기업의 데이터베이스에서 상품의 구매, 서비스 등 일련의 거래 또는 사건들 간의 규칙을 발견하기 위해 적용

- 장바구니 분석: 장바구니에 무엇이 같이 들어 있는지에 대한 분석
- 선결 분석: A를 산 다음 B를 산다.

연관 규칙의 형태

조건과 반응의 형태로 구성

연관 규칙의 척도

산업의 특성에 따라 지지도, 신뢰도, 향상도 값을 잘 보고 규칙을 선택해야 함

주요 지표

지표명	수식 또는 정의	의미 및 해석	해석 기준
지지도 (Support)	$P(A \cap B)$	A와 B가 함께 발생한 비율	높을수록 일반적 규칙
신뢰도 (Confidence)	$P(A \rightarrow B) = P(A \cap B) / P(A)$	항목 A를 포함하는 거래 중 항목 A와 항목 B가 같이 포함될 확률	A가 발생했을 때 B도 발생할 확률 (정확도)
향상도 (Lift)	$P(A \cap B) / (P(A) \times P(B))$	A, B가 우연일 확률 대비 몇 배 자주 발생?	1보다 크면 유의미

주요지표와 정리

연관 분석(Association Rule Mining)은 데이터에서 **"A를 산 사람은 B도 살 가능성이 높다"** 같은 규칙을 찾는 기법


- 이 분석의 핵심 지표는 ****지지도(Support), 신뢰도(Confidence), 향상도(Lift)****이며, 각각 규칙의 빈도, 정확도, 유용성을 의미
- 특히 Lift는 "우연이 아니라 진짜 관계냐?"를 판단하는 핵심 기준
- 이를 바탕으로 마케팅, 추천 시스템, 진단 지원 등에 폭넓게 활용

연관 분석 주요 지표 정리

1. 지지도 (Support)

$$\text{Support}(A \rightarrow B) = P(A \cap B)$$

- **정의:** A와 B가 동시에 발생한 비율
- **해석:** 전체 거래 중 A와 B가 **같이 등장한 비율**
- **목적:** 규칙이 **얼마나 자주 등장하는가**
- **예시:** 전체 1,000건 중 A와 B가 같이 있는 게 50건이면 $\rightarrow \text{Support} = 0.05$

|  지지도는 "얼마나 흔한 규칙인가?"를 보는 기준이다.

2. 신뢰도 (Confidence)

$$\text{Confidence}(A \rightarrow B) = \frac{P(A \cap B)}{P(A)} = P(B|A)$$

- **정의:** A가 발생했을 때 B도 함께 발생할 확률
- **해석:** A를 구매한 사람이 **얼마나 자주 B도 구매했는가**
- **목적:** 규칙의 **정확도 또는 신뢰 수준**
- **예시:** A가 100번 나왔고 그중 A와 B가 함께 나온 게 60번이면 $\rightarrow \text{Confidence} = 0.6$

|  신뢰도는 "A이면 B일 가능성이 얼마나 높나?"를 나타낸다.

3. 향상도 (Lift)

$$\text{Lift}(A \rightarrow B) = \frac{\text{Confidence}(A \rightarrow B)}{P(B)} = \frac{P(A \cap B)}{P(A) \cdot P(B)}$$

- **정의:** A와 B가 **우연히 발생한 수준**과 비교해 얼마나 더 자주 같이 발생하는지 측정
- **해석:** $Lift > 1 \rightarrow$ 양의 상관 관계
- **목적:** 규칙의 **유의미성(의미 있는 상관 관계인가?)**
- **예시:** Confidence가 0.6이고 $P(B)=0.3$ 이면 $\rightarrow Lift = 2.0 \rightarrow$ **B는 A 없이 나올 확률의 2배**

💬 Lift는 이 규칙이 그냥 우연인지, 진짜 연관성이 있는지를 말해주는 결정적인 지표야.

해석 정리

지표	의미	좋아야 할 방향	활용 포인트
Support	얼마나 자주 발생?	너무 낮으면 폐기	자주 등장하는 규칙 찾기
Confidence	얼마나 정확한가?	높을수록 좋음	추천 정확도 기준
Lift	얼마나 우연을 넘는가?	1보다 클수록 좋음	규칙의 실질 가치 판단

실전 예시

규칙: {우유} \rightarrow {빵}

- Support = 0.1 \rightarrow 전체 중 10%가 우유+빵 동시 구매
- Confidence = 0.4 \rightarrow 우유 구매자 중 40%가 빵도 샀음
- Lift = 1.8 \rightarrow 빵은 우유 없이도 많이 팔리지만, 우유 있으면 **1.8배 더 팔림** \rightarrow 추천 가치 있음!

연관 규칙의 장점

- 탐색적인 기법으로 조건값으로 표현되는 연관성 분석 결과를 쉽게 이해
- 강력한 비목적성 분석기법으로 분석 방향이나 목적이 특별히 없는 경우 목적 변수 없이도 유용하게 활용
- 데이터의 형태도 거래 내역에 대한 데이터로 복잡하지 않아 자료를 이용할 수 있는 간단한 자료구조를 갖는다. 분석을 위한 계산이 간단

연관 규칙의 단점

- 품목의 수가 증가하면 분석에 필요한 계산은 기하급수적으로 늘어남
 - 이를 개선하기 위해 유사한 품목을 한 범주로 묶음
 - 연관규칙의 신뢰도만을 해석할 경우에 실제 존재하는 관계는 의외로 약한 연관 규칙은 제외됨
 - 너무 세분화된 품목을 조건 및 결과 규칙으로 찾으면 의미 없는 분석이 될 수 있음
 - 거래가 자주 발생하지 않는 희귀한 품목들 간의 연관 규칙은, 규칙 발견 시 제외됨
 - 이를 해결하기 위해, 품목의 연관성을 살펴보고 중요한 품목이면 유사한 품목들과 함께 범주로 구성하는 방법 등을 통해 연관성 규칙의 과정에 포함시킬 수 있음
-

순차 패턴

동시에 구매할 가능성이 큰 상품군을 찾아내는 연관성 분석에 시간이라는 개념을 포함시켜 순차적으로 구매 가능성이 큰 상품군을 찾아내는 방법

연관성 분석에 시간의 순서를 추가해서 각기 고객으로부터 발생한 구매 시점에 대한 정보가 포함

최근 연관성 분석 동향

Apriori 알고리즘(1세대)

- 모든 가능한 품목 부분집합의 개수를 줄이는 방식으로 작동하는 것이 Apriori 알고리즘
 - 최소 지지도보다 큰 지지도 값을 갖는 일반 항목 집합이라고 한다.
 - Apriori 알고리즘은 품목 집합에 대한 지지도는 전부 계산하는 것이 아니라, 최근 지지도 이상의 빈발 항목 집합만 찾은 후 그것에 대해서만 연관 규칙을 계산하는 것이다.
 - 구하려는 대상이 많다는 장점이 있으나, 지지도가 낮은 항목 집합 생성 시 아이템셋 수가 많아지고 계산 복잡도가 증가한다는 문제를 가짐
-

FP-Growth 알고리즘(Frequent Pattern, 2세대)

- 거래내역에 포함되는 항목 전체 집합을 줄이는 방식으로 작동
- Apriori 알고리즘은 candidate 집합 생성을 통해 연산을 수행한 반면 FP-Growth 알고리즘은 압축된 FP-Tree를 이용한 분석으로 보다 빠르게 빈발 항목 집합을 추출하는 방법
- Apriori 알고리즘보다 효율성이 뛰어난 것으로 알려져 있음

- 데이터베이스 스캔 횟수가 작아 빠른 속도로 분석 가능

실습 데이터셋

데이터 adult

✓ 기본 정보

- 데이터셋 이름: adult (또는 census income)
- 목적: 개인의 특성(인구통계학적 정보 등)을 이용하여 연소득이 ≤50K 또는 >50K 인지 예측
- 관측치 수: 약 32,000개
- 변수 수: 15개 (14개 입력 변수 + 1개 출력 변수)

변수명	설명	타입
age	나이 (숫자)	numeric
workclass	고용 형태 (예: Private, Self-emp, Gov 등)	categorical
fnlwgt	인구 가중치	numeric
education	교육 수준 (예: Bachelors, HS-grad 등)	categorical
education.num	교육 수준을 수치로 표현	numeric
marital.status	결혼 상태	categorical
occupation	직업군 (예: Tech-support, Sales 등)	categorical
relationship	가족 관계	categorical
race	인종 (예: White, Black 등)	categorical
sex	성별	categorical
capital.gain	자본 이득	numeric
capital.loss	자본 손실	numeric
hours.per.week	주당 근무 시간	numeric
native.country	출신 국가 (예: United-States, Mexico 등)	categorical
income	수입 수준 (예: ≤50K , >50K)	categorical (목표 변수)

종료

