14강 연관성분석 I

통계·데이터과학과 장영재 교수

한국방송통신대학교

목차

1. 연관성분석이란

01 연관성분석이란

1 연관성분석의 정의

- 연관성분석(Association Analysis)은 연관성 규칙을 통해 하나의 거래나 사건에 포함되어 있는 둘 이상의 품목 간 상호 연관성을 발견해 내는 것
  - 연관규칙이란 어떤 사건이 얼마나 자주 동시에 발생하는가를 표현하는 규칙 또는 조건
    - 장바구니분석(market basket analysis)은 장바구니에 들어 있는 품목 간의 관계를 알아보는 분석

01 연관성분석이란

2 연관규칙의 특징

연관규칙은 사건과 사건 간의 관계에 관심

- 목표변수 없이 특성의 조합으로 규칙을 표현한다. 따라서 연관성분석은 자율학습에 속함
  - 연관규칙이 서로 영향을 주지 않으므로 하나의 고객이 여러 개의 규칙에 해당될 수 있음
  - 마케팅에 이용할 수 있음

01 연관성분석이란

2 연관규칙의 특징

- 연관규칙을 통하여 일반적으로 "A가 발생하면 B가 발생한다."는 규칙을 도출하게 되는데 이를 '$A \rightarrow B$'로 표현

- 많은 품목들의 관계 속에서 의미 있는 결과를 찾기 위해서는 결과해석에 앞서 연관성의 내용이 일반화할 수 있는 내용인지 판단할 수 있도록 각 연관규칙을 비교평가할 수 있는 기준이 필요
  - 객관적이고 일반성 있는 연관관계를 규명하기 위한 정량화된 평가 기준이 필요
  - 연관규칙을 평가하기 위한 첫 단계는 동시구매표 작성

01 연관성분석이란

2 연관규칙의 특징

<예제 8-1>

어느 식료품점에서 다섯 개 품목의 거래 내역이 아래 표와 같다고 할 때, 다섯 개 품목 이름을 행과 열 이름으로 하고 교차점에 동시 구매 횟수가 나타나는 동시구매표를 작성

거래	품목
1	고기류, 쌀, 상추
2	고기류, 상추, 스낵과자
3	쌀, 상추, 스낵과자
4	고기류, 스낵과자, 탄산음료
5	쌀, 상추, 탄산음료

01 연관성분석이란

2 연관규칙의 특징

풀이

단계 1. 행과 열 이름을 다섯 개 품목의 이름으로 하는 행렬구조를 생성

단계 2. 각 교차점에 공통적으로 구매한 거래 수를 기입. 자신과 자신의 교차점에는 해당 품목의 총 구매 수를 기입

단계 3. 동시구매표로부터 간단한 규칙을 파악

구매품목	고기류	쌀	상추	스낵과자	탄산음료
고기류	3	1	2	2	1
쌀	1	3	3	1	1
상추	2	3	4	2	1
스낵과자	2	1	2	3	1
탄산음료	1	1	1	1	2

2. 연관성분석의 종류

02 연관성분석의 종류

1 동시구매 품목의 연관성분석

(1) 지지율(support)

- 지지율(support)은 연관규칙의 유용성을 평가하는 측도
- 연관규칙 '$A \rightarrow B$'의 지지율은 전체 거래 중 A와 B가 동시에 포함된 거래의 비율

ex) 대형할인점의 1백만 건의 거래 중에서 1만 건의 거래가 A와 B를 모두 포함한 경우 연관규칙 '$A \rightarrow B$'의 지지율은 1%

02 연관성분석의 종류

1 동시구매 품목의 연관성분석

(1) 지지율(support)

- 확률을 개념을 적용하면 아래 식과 같이 표현

$$\text{지지율} = P(S_A \cap S_B) = \frac{A\text{와 } B\text{를 모두 구매한 거래 수}}{\text{전체 거래 수}}$$

ex) 지지율의 계산

거래	품목
1	고기류, **쌀**, **상추**
2	고기류, 상추, 스낵과자
3	**쌀**, **상추**, 스낵과자
4	고기류, 스낵과자, 탄산음료
5	**쌀**, **상추**, 탄산음료

02 연관성분석의 종류

1 동시구매 품목의 연관성분석

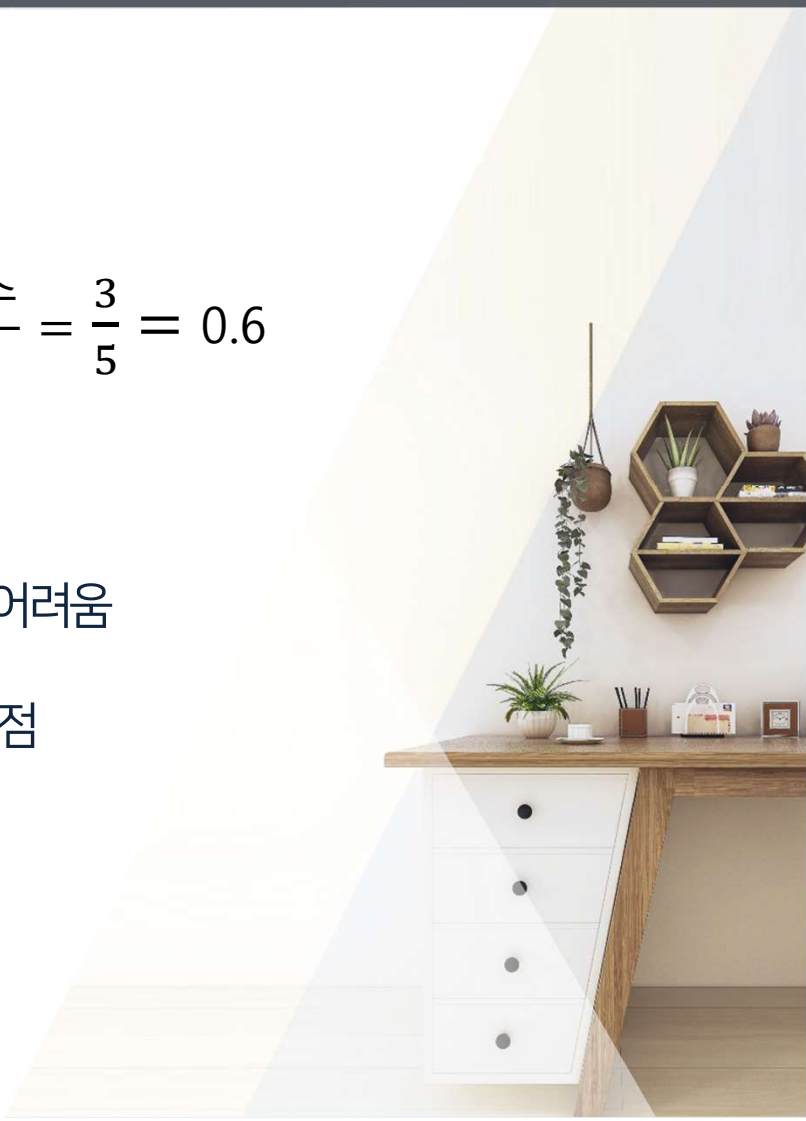
(1) 지지율(support)

$$'쌀 \rightarrow 상추'의\ 지지율 = \frac{쌀과\ 상추를\ 모두\ 구매한\ 거래\ 수}{전체\ 거래\ 수} = \frac{3}{5} = 0.6$$

지지율의 단점

1) 표본수가 적은 경우 연관관계에 대한 통계적 유의성을 증명하기 어려움

2) 투자한 시간, 비용에 비해 판매량을 증가시키는 기여도가 작다는 점



02 연관성분석의 종류

1 동시구매 품목의 연관성분석

(2) 신뢰도(Confidence)

- 연관성분석에 있어서 신뢰도는 원인이 발생할 때 결과가 발생할 가능성을 나타냄
- 지지율의 경우 기준이 되는 사건(구매)이 전체집합인데 비하여 신뢰도는 기준이 되는 사건을 특정 품목을 구매한 것에 한정
- '$A \rightarrow B$'의 신뢰도는 품목 A가 구매되었을 때, 품목 B가 추가로 구매될 확률을 의미

$$'A \rightarrow B'\text{의 신뢰도} = \frac{P(S_A \cap S_B)}{P(S_A)} = \frac{A\text{와 } B\text{를 모두 구매한 거래 수}}{A\text{를 구매한 거래 수}}$$

02 연관성분석의 종류

1 동시구매 품목의 연관성분석

(2) 신뢰도(Confidence)

ex) 어느 식료품점에서 구매자들의 전체 장바구니 수가 150개였고 쌀과 고기류가 함께 들어가 있는 장바구니가 30개였다면, 이는 '쌀 → 고기류'의 지지율이 20%임을 의미. 그런데, 쌀이 들어가 있는 장바구니만을 추렸더니 100개였고 그 100개의 장바구니 중 고기류가 들어가 있는 장바구니는 30개였다면, '쌀 → 고기류' 의 신뢰도는 이 비율

$$\frac{30}{100} = 0.3(30\%)$$

- 연관규칙 '$A \rightarrow B$'의 신뢰도는 A가 발생했을 때 B가 발생할 조건부 확률 $P(S_B \mid S_A)$와 같음. 따라서 '$A \rightarrow B$'의 신뢰도와 '$B \rightarrow A$'의 신뢰도는 일반적으로 같지 않음

02 연관성분석의 종류

1 동시구매 품목의 연관성분석

(3) 향상도(lift)

- 향상도는 규칙을 모를 때에 비하여 규칙을 알 때에 판매가 얼마나 향상되는가를 나타냄

  - 즉, 향상도는 품목 B를 연관규칙과 관계없이 판매하는 것에 비하여 연관규칙을 알고 A를 구매한 고객에 대하여 B를 판매하는 경우 판매가 얼마나 증가하는가를 나타냄

- 향상도의 개념이 필요한 이유는 신뢰도의 제약 때문

  - 신뢰도는 연관규칙이 실제로 유용한지 아니면 임의로 나타난 결과인지 알 수 없다는 단점

02 연관성분석의 종류

1 동시구매 품목의 연관성분석

(3) 향상도(lift)

- 향상도는 '$A \rightarrow B$'의 신뢰도를, 독립 가정 하에서의 신뢰도(A가 B에 영향을 미치지 않는 경우)인 전체에서 B가 포함된 거래의 비중으로 나눈 것을 의미

$$'A \rightarrow B'\text{의 향상도} = \frac{P(S_B \mid S_A)}{P(S_B)} = \frac{P(S_A \cap S_B)}{P(S_A)P(S_B)}$$

$$\frac{P(S_A \cap S_B)}{P(S_A)P(S_B)} = \frac{\text{품목 } A,\ B\text{를 동시에 포함한 거래 수} \times \text{전체 거래 수}}{\text{품목 } A\text{ 포함 거래 수} \times \text{품목 } B\text{ 포함 거래 수}}$$

- 향상도가 1에 가까우면 A와 B가 확률적으로 독립에 가까움을 의미하고 향상도가 1보다 크면 A를 구매하는 경우 B를 구매할 가능성이 높다는 것을 의미(양의 상관관계)

02 연관성분석의 종류

2 시차 연관성분석

- 시차연관성분석은 일반적인 연관성분석과 유사하나 연관규칙을 고려할 때 순서를 고려한다는 점에서 차이가 있음
  - 일반적으로 비대칭

$$'A \rightarrow B'\text{의 지지율} = \frac{\text{품목 } A\text{를 구매한 후 } B\text{를 구매한 거래 수}}{\text{전체 거래 수}}$$

$$'A \rightarrow B'\text{의 신뢰도} = \frac{\text{품목 } A\text{를 구매한 후 } B\text{를 구매한 거래 수}}{A\text{를 포함한 전체 거래 수}}$$

$$'A \rightarrow B'\text{의 향상도} = \frac{'A \rightarrow B'\text{의 신뢰도}}{B\text{를 포함한 거래의 비중}}$$

3. 연관성분석의 절차

03 연관성분석의 절차

1 품목과 수준의 선택

- 장바구니 내의 구매 품목은 매우 다양하며, 분류 수준도 차이가 나는 여러 품목이 섞여 있는 경우가 빈번히 발생
  - 복잡한 품목에 대해서는 품목분류표가 있으면 유용
  - 분석의 목적에 부합되는 수준과 품목에 대한 후보군이 정해졌다면 이들을 변화시켜가면서 여러 번의 연관규칙을 검토하고 유용한 연관규칙을 찾게 됨
- 연관규칙에서는 품목의 개수가 많아질수록 고려해야 할 규칙수가 크게 증가하므로 품목수가 너무 많지 않도록 고려
- 품목 외에도 고객의 인구통계적 자료를 활용할 수도 있음(고객 정보)

03 연관성분석의 절차

2 연관규칙의 생성

- 품목수의 증가에 따라 연관규칙의 수도 지수적으로 증가

  - 지지율이 낮더라도 신뢰도가 높은 경우 유용한 연관규칙을 찾을 가능성
    - 여러 조합의 지지율과 신뢰도의 하한을 정하여 연관규칙을 찾도록
  - 품목의 구매순서 정보 등이 있다면, 시차연관규칙을 도출

03 연관성분석의 절차

3 연관규칙의 분석

- 지지율, 신뢰도, 향상도에 근거한 유용한 연관규칙 후보군을 선정

- 연관규칙 후보들에 대한 이유를 점검
  - 일부 규칙은 실제로 유의하지 않으면서도 우연히 유의하게 나타날 수도 있음

4. 연관성분석의 장단점

04 연관성분석의 장단점

1 연관성분석의 장점

- 연관성분석은 여타 데이터마이닝 기법에 비해 이해와 적용이 용이

  - 연관규칙이 '조건 → 반응' 형태로 표현되므로 쉽게 이해할 수 있고, 바로 실제에 적용하기가 용이
  - 연관성분석 알고리즘은 비교적 단순한 자료 형태와 계산이 간단하다는 장점
  - 목표변수 없는 직관적인 분석
  - 본격적인 데이터마이닝에 들어가기에 앞서 거대 자료의 탐색 도구로서 매우 유용하게 사용될 수도 있음

04 연관성분석의 장단점

2 연관성분석의 단점

- 연관성분석은 품목수의 증가에 따라 계산량이 증가한다는 단점

  - 유사한 품목을 한 범주로 일반화해야 하는데, 적절한 품목 결정이 쉽지 않음
  - 수 천 가지의 품목을 모두 분석에 그대로 사용할 경우 수많은 연관 규칙이 발견되므로 유용한 규칙을 한정하여 찾기 어렵다는 제약
  - 연속형 변수를 사용하여 연관규칙을 구하기 힘들며 거래가 드문 품목에 대한 정보를 찾기도 어렵다는 단점도 존재

다음시간 안내

15강.
연관성분석 II