연관 규칙 분석(Association Rule Analysis)
a.k.a
장바구니 분석(Market Basket Analysis)

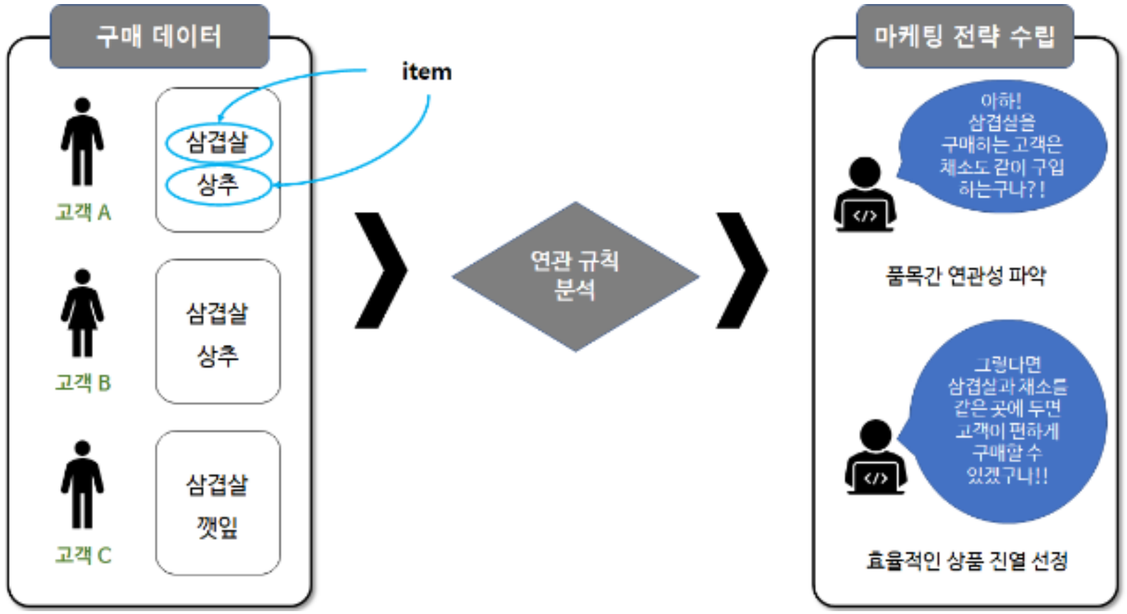
염수지

연관 규칙 분석이란 무엇인가?

01 항목(item)들 관계를 If-Then 형식으로 찾아나가는 분석기법, 일종의 규칙 기반 학습 방법이다.

02 대용량의 DB에서 기존에 발견할 수 없었던 항목간의 관계를 탐색할 수 있는 장점이 있다.

03 어떤 서비스를 원하는지 미리 파악하거나 특정 상품을 추천해주고 싶은 경우에도 사용한다.



01 연관 규칙 분석 방법

연관 규칙 분석 측도

대형 마트의 거래 내역이 아래와 같다.
구매 행렬로 바꾸면 우측과 같다.

고객 번호	품목
1	삼겹살, 상추
2	삼겹살, 상추, 사이다
3	삼겹살, 깻잎
4	닭고기, 샤워 타올
5	닭고기, 콜라, 사이다

	삼겹살	상추	사이다	깻잎	닭고기	샤워 타올	콜라
삼겹살	3	2	1	1	0	0	0
상추	2	2	1	0	0	0	0
사이다	1	1	2	0	1	0	1
깻잎	1	0	0	1	0	0	0
닭고기	1	0	1	0	2	1	1
샤워 타올	0	0	0	0	1	1	0
콜라	0	0	1	0	1	0	1

01 연관 규칙 분석 방법

연관 규칙 분석 측도

1. 신뢰도(Confidence): X를 포함하는 거래 내역 중,
Y가 포함된 비율이 높아야 한다.
비율은 확률을 의미한다.

$$P(Y|X) = P(X \cap Y)/P(X)$$

"삼겹살을 사는 사람은 상추도 구입한다"

$$P(\text{상추}|\text{삼겹살}) = \frac{2}{5}/\frac{3}{5} = \frac{2}{3}$$

"삼겹살을 사는 사람은 사이다도 구입한다"

$$P(\text{사이다}|\text{삼겹살}) = \frac{1}{5}/\frac{3}{5} = \frac{1}{3}$$

	삼겹살	상추	사이다	깻잎	닭고기	샤워 타올	콜라
삼겹살	3	2	1	1	0	0	0
상추	2	2	1	0	0	0	0
사이다	1	1	2	0	1	0	1
깻잎	1	0	0	1	0	0	0
닭고기	1	0	1	0	2	1	1
샤워 타올	0	0	0	0	1	1	0
콜라	0	0	1	0	1	0	1

01 연관 규칙 분석 방법

연관 규칙 분석 측도

2. 지지도(Support): X와 Y를 동시에 포함하는 비율이 높아야 한다.

$$P(X \cap Y)$$

"상추를 구입하는 사람은 사이다도 구입한다"의 신뢰도

$$P(\text{사이다}|\text{상추}) = \frac{1}{5} / \frac{2}{5} = \frac{1}{2}$$

$$P(\text{상추}|\text{삼겹살}) = \frac{2}{5} / \frac{3}{5} = \frac{2}{3}$$

"상추를 구입하는 사람은 사이다도 구입한다"의 지지도

$$P(\text{삼겹살}, \text{상추}) = \frac{2}{5}$$

	삼겹살	상추	사이다	깻잎	닭고기	샤워 타올	콜라
삼겹살	3	2	1	1	0	0	0
상추	2	2	1	0	0	0	0
사이다	1	1	2	0	1	0	1
깻잎	1	0	0	1	0	0	0
닭고기	1	0	1	0	2	1	1
샤워 타올	0	0	0	0	1	1	0
콜라	0	0	1	0	1	0	1

01 연관 규칙 분석 방법

연관 규칙 분석 측도

3. 향상도(Lift): 지지도와 신뢰도 만으로 충분한가?
주어진 규칙이 의미가 있는지 확인한다.

$$P(Y|X)/P(Y)$$

향상도 값이 1이면	X와 Y는 아무런 관계가 없음.
1보다 크면	X가 Y의 발생할 확률을 X를 고려하지 않았을 경우보다 증가 -> X가 Y 발생 예측에 도움.
1보다 작으면	위의 경우와 반대 -> 감소 예측에 도움

"삼겹살을 사는 사람은 상추도 구입한다"의 향상도

$$P(\text{상추}|\text{삼겹살})/P(\text{상추}) = \frac{2}{3}/\frac{2}{5} = \frac{5}{3}$$

	삼겹살	상추	사이다	깻잎	닭고기	샤워 타올	콜라
삼겹살	3	2	1	1	0	0	0
상추	2	2	1	0	0	0	0
사이다	1	1	2	0	1	0	1
깻잎	1	0	0	1	0	0	0
닭고기	1	0	1	0	2	1	1
샤워 타올	0	0	0	0	1	1	0
콜라	0	0	1	0	1	0	1

01 연관 규칙 분석 방법

연관 규칙 분석 측도

4. 레버리지(Leverage): 향상도가 비율을 이용한다면
레버리지는 차이를 이용한다.

$$P(X \cap Y) - P(X)P(Y)$$

레버리지 값이 0에 가깝다면 X와 Y는 독립, 관련 없음
레버리지가 양수면 -> X가 Y 발생 예측에 도움
레버리지가 음수면 -> 감소 예측에 도움

"삼겹살을 사는 사람은 상추도 구입한다"의 레버리지

$$P(\text{삼겹살}, \text{상추}) - P(\text{삼겹살})P(\text{상추}) = \frac{2}{5} - \frac{3}{5}\frac{2}{5} = \frac{4}{25}$$

	삼겹살	상추	사이다	깻잎	닭고기	샤워 타올	콜라
삼겹살	3	2	1	1	0	0	0
상추	2	2	1	0	0	0	0
사이다	1	1	2	0	1	0	1
깻잎	1	0	0	1	0	0	0
닭고기	1	0	1	0	2	1	1
샤워 타올	0	0	0	0	1	1	0
콜라	0	0	1	0	1	0	1

01 연관 규칙 분석 방법

연관 규칙 분석 측도

5. Conviction: 어떤일이 생기지 않을 확률

$$P(Y^c)/P(Y^c|X)$$

6.All-Confidence: 높은 지지도를 갖는 규칙 고려

$$\text{All-Confidence}(X,Y) = P(X \cap Y)/\max[P(X), P(Y)]$$

7. Collective Strength: 지지도, 신뢰도에서 벗어난 새로운 측도

$$CS(X,Y) = \frac{1 - V(X,Y)}{1 - E(V(X,Y))} \frac{E(V(X,Y))}{1 - V(X,Y)}$$

8. Cosine Similarity

$$\cos(X,Y) = \frac{P(X,Y)}{\sqrt{P(X)\sqrt{P(Y)}}}$$

	삼겹살	상추	사이다	깻잎	닭고기	샤워 타올	콜라
삼겹살	3	2	1	1	0	0	0
상추	2	2	1	0	0	0	0
사이다	1	1	2	0	1	0	1
깻잎	1	0	0	1	0	0	0
닭고기	1	0	1	0	2	1	1
샤워 타올	0	0	0	0	1	1	0
콜라	0	0	1	0	1	0	1

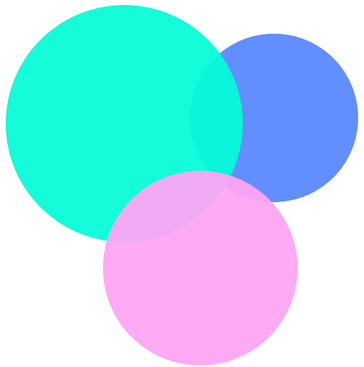
02 연관 규칙 분석 절차(Apriori 알고리즘)

1. 빈발 품목 집합 생성

각 품목별로 발생 횟수(비율)가 특정 값 이상인 품목을 모아놓은 집합

- 최소 지지도를 설정한다.
- p = 1,2,3 ... 에 대하여 p개의 품목을 갖는 품목 집합 중에서 최소 지지도를 넘는 품목을 찾는다.
- 위에서 찾은 품목 중에서 최소 지지도를 넘는 p + 1개 품목 집합을 찾는다.
- 과정을 반복하면 최소 지지도가 넘는 빈발 품목 집합을 찾을 수 있다.

(최소지지도를 설정하는 절대적 방법은 없으며 순전히 분석자의 몫)



2. 연관 규칙 생성

빈발 품목 집합의 공집합을 제외한 모든 부분집합을 고려하고 이 중에서 최소 신뢰도를 넘는 연관 규칙을 찾는다.

03 고려 사항

1. 유용한 연관 규칙 선별

너무 뻔하거나 연관성을 찾기 힘든 규칙들은 유용하지 않다.

2. 적절한 품목 선택

어떤 품목을 선택하는가는 분석의 목적에 따라 달라진다.

3. 연관 규칙 발굴

연관규칙을 선택할 때 어떻게 표현 되는지도 중요하다.
조건은 구체적이고 결과는 더 단순하게 함으로써 해석이 쉽고 의미 있는 규칙을 발굴한다.

4. 계산 문제

품목 수가 증가하면 계산량은 엄청나게 증가하게 되어 많은 시간이 소요된다.
이러한 문제점 때문에 최소 지지도보다 작은 규칙들은 선택하지 않는 가지치기를 적용한다.

04 예제 with Python