**연관분석(Association Analysis)**

**연관분석(Association Analysis = Market Basket Analysis)**

**Theory**

연관분석(Assocation analysis)은 기업의 데이터베이스에서 상품의 구매, 서비스 등 일련의 거래 또는 사건들 간의 규칙을 발견하여 IF-THEN의 구조로 분석 결과의 연관성을 파악하는 데이터마이닝 방법론이다. 연관규칙분석은 거래(Transaction)와 항목(Item)으로 구성되어 있는 경우 분석이 가능하다.

▷ Run-test : 표본집단이 무작위로 선별되었는지를 검사하는 Test, 귀무가설 = 무작위로 선별했다.

**연관규칙의 측도**

- **지지도(support)** : 한 거래 항목 안에 A와 B를 동시에 포함하는 거래의 비율. 지지도는 A와 B가 함께 등장할 확률이다. 전체 거래의 수를 A와 B가 동시에 포함된 거래수를 나눠주면 구할 수 있다.

- **신뢰도(confidence)** : 항목 A가 포함하는 거래에 A와 B가 같이포함될 확률. 신뢰도는 조건부 확률과 유사하다. A가 일어났을 때 B의 확률이다. A의 확률을 A와 B가 동시에 포함될 확률을 나눠주면 구할 수 있다. -> 해당 구매 폼목들 간의 연관성

- **향상도(lift)** : A가 주어지지 않을 때의 품목 B의 확률에 비해 A가 주어졌을 때 품목 B의 증가 비율. B의 확률이 A가 일어났을 때 B의 확률을 나눴을 때 구할 수 있다. lift 값은 1이면 서로 독립적인 관계이며1보다 크면 두 품목이 서로 양의 상관관계, 1보다 작으면 두 품목이 서로 음의 상관관계이다. A와 B가 독립이면 분모, 분자가 같기 때문에 1이 나온다. -> A를 샀을 때 B를 구매할 확률 증가

**Apriori Algorithm**

최소 지지도보다 큰 지지도 값을 갖는 품목의 집합을 **빈발항목집합(frequent item set)**이라고 한다. Apriori 알고리즘은 모든 품목집합에 대한 지지도를 전부 계산하는 것이 아니라, 최소 지지도 이상의 빈발항목집합을 찬은 후 그것들에 대해서만 연관규칙을 계산하는 것이다.

Apriori 알고리즘은 구현과 이해가 쉽다는 장점이 있으나, 지지도가 낮은 후보 집합 생성 시 아이템의 개수가 많아지면 계산 복잡도가 증가한다는 문제점을 가지고 있다.

**▷ One-Hot Encoding -> Run-test(샘플의 무작위성 검증) -> 빈발항목집합(frequent item set) 생성 ->**

**연관규칙 계산(Apriori) -> 중복가지치기(Pruned) -> Confidence or Lift에 따라 상위 n개 규칙 추출.**

**텍스트이(가) 표시된 사진

자동 생성된 설명** **테이블이(가) 표시된 사진

자동 생성된 설명**