**연관성 분석의 원리&이론**

**1.기본개념**

연관성규칙발견 : 하나의 거래나 사건에 포함되어 있는 둘 이상의 품목들의 상호 관련성을 발견하는 것

Ex) ‘신발을 구매하는 고객의 10%는 양말을 동시에 구입한다.’ ‘최근에 구좌정리와 이율상담을 요구한 고객은 이 후 한 달 이내에 거래를 중단할 가능성이 일반 고객의 2배이다.’ -> 조건과 반응(if-then)의 형식

* 다수의 거래내역 각각에 포함된 품목의 관찰을 통해 발견된 규칙은 쿠폰의 배포, 상품의 판매, 상품의 배열 등과 같은 기업업무에서 마케팅 의사결정에 도움을 준다.
* 장바구니 분석 : 고객이 동시에 구매한 장바구니를 살펴봄으로써 거래되는 상품들 의 관련성을 발견, 분석
* 일반적으로 ‘A’->’B’ 즉, ‘A를 사면 B를 살 것이다’라는 직관적인 형태로 결과 도출 A는 lhs(left-hands side), B를 rhs(right-hands side)라 부름
* 원인과 결과의 직접적인 인과관계로 생각해서는 안되며, 두 개 또는 그 이상의 품목들 사이의 상호의 관련성으로 해석해야 한다.

**2.연관분석 지표의 개념**

1. 지지도support:

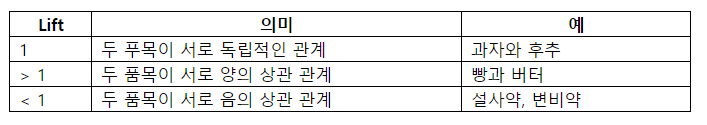
* 거래 내역 중 품목 A와 B를 동시에 구매할 확률을 나타냄
* ->클수록 A와 B의 관련도가 높음 -> 지지도가 높을수록 A와 B를 동시에 구매할 확률 높음

1. 신뢰도confidence:

* 품목 A가 구매되었을 때, 품목 B가 구매될 확률을 나타냄
* ->클수록 A와 B의 관련도가 높음-> 70%~80%이상이면 A 구매시 B가 동시에 구매될 확률 높음

1. 향상도Lift

* 지지도와 신뢰도 간의 규칙이 의미가 있는지
* A를 구매한 사람이 B를 구매할 확률과 A의 구매와 상관없이 B를 구매할 확률의 비
* A->B의 연관규칙에서 ‘임의로 B가 구매되는 경우’에 비해 ‘A와 의 관계가 고려되어 구매되는 경우’의 비율
* 1이 넘으면 A를 구매한 사람이 B를 구매할 확률 높음/1이면 서로 독립적임(관계없음)/ 1보다 적으면 A를 구매한 사람이 B를 구매할 확률이 낮음



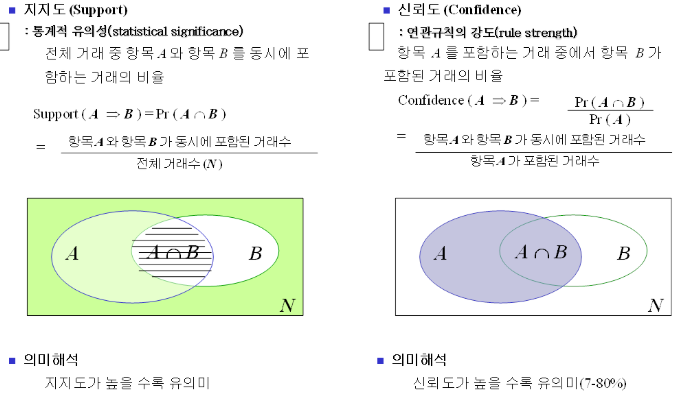
- support(지지도): support(A) = P(A) 빈발 아이템 집합을 판별하는 데 사용

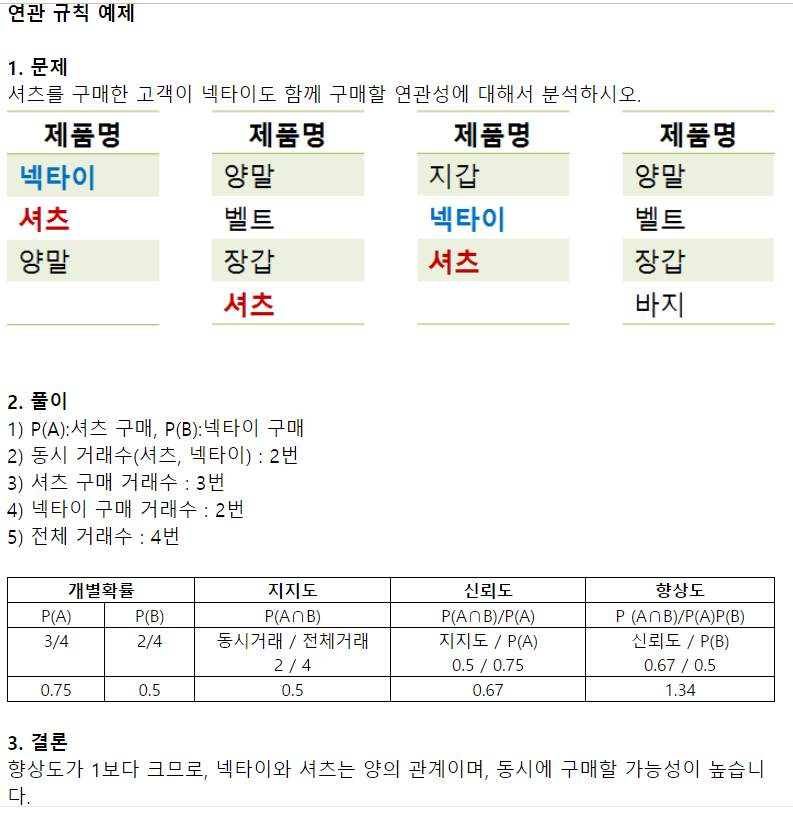
- confidence(신뢰도): confidence(A→B) = P(A, B)/P(A) 아이템 집합 간의 연관성 강도를 측정하는 데 사용

- lift(향상도): lift(A→B) = P(A,B)/P(A)\*P(B) 생성된 규칙이 실제 효용가치가 있는지를 판별하는 데 사용

**[출처]** [연관규칙 분석 해석(지지도, 신뢰도, 향상도)](https://blog.naver.com/name1929/221569252721)|**작성자** [감귤](https://blog.naver.com/name1929)

**3.연관분석 지표 수식**





**응용**

* 둘 또는 그 이상의 항목들 사이의 수많은 연관성 규칙 중에서 지지도, 신뢰도, 향상도에 근거해 일반화 할 수 있는 의미 있는 규칙을 탐색하는 방법
* 교차판매, 매장진열, 카탈로그 디자인, 소프트웨어 번들링, 첨부 우편물, 보험의 부정행위 적발 등에 유용하게 사용
* Ex)
* 어떤 제품과 관련된 선택사양들을 고객의 성향에 따라 정리한다면 비슷한 성향을 갖는 신규고객에게 동시에 그 둘을 제공해 구매의사를 높임
* 예금구좌, CD, 투자 서비스, 카드 대출 등의 은행 상품의 관련성을 파악한다면 고객들이 원하는 또 다른 상품의 형태를 파악해 고객의 구매를 유도
* 보험금 청구 내용이 이전의 내용과 비교해 관련성이 적은 특이한 상황이라면 보험사기를 의심하고 좀 더 면밀한 조사를 요구

분석에 oriented 된 통계전문가, 기계학습 전문가의 경우 신뢰도(confidence)와 향상도(lift)가 높은 rule을 눈에 불을 켜고 찾고, 선호하는 경향이 있습니다. 그런데 매출과 이익을 책임져야 하는 사업부 현업의 경우는 연관규칙을 보는 view가 조금 다를 수 있습니다. "그 rule을 적용하면 **기대할 수 있는 매출 증가분**이 얼마나 되는데요?"라는 질문이 사업부 현업이 던지는 질문인데요, 이 질문에 만족할 만한 답을 주려면 **'지지도(Support)'가 높아서 전체 거래 건수 중에서 해당 rule이 포함된 거래건수가 많아야지만이 해당 rule을 가지고 마케팅전략을 수립해서 실전에 적용했을 때 높은 매출 증가를 기대**할 수 있게 됩니다. 즉, 아무리 신뢰도(confidence)와 향상도(lift)가 높아도 지지도(support)가 빈약해서 전체 거래 중에 가뭄에 콩나듯이 나오는 거래유형의 rule이라면 사업부 현업은 아마 무시할 겁니다. 현업을 빼고 분석가만 참여한 연관규칙 분석이 위험하거나 아니면 실효성이 떨어질 수 있는 이유입니다. 그리고 지지도(support)가 매우 낮으면 몇 개 소수이 관측치의 치우침만으로도 신뢰도나 향상도가 크게 영향을 받게 되어 '우연'에 의한 규칙이 잘못 선별될 위험도 있습니다.

**연관성규칙 발견의 장단점**

**연관성규칙발견의 장점**

* 탐색적인 기법 : ‘조건 => 반응’ 의 규칙의 형태를 가지고 있어 이해가 쉽고 적용이 용이
* 강력한 비목적성 분석기법 : 대부분의 데이터마이닝 기법과 달리 뚜렷한 목적변수 없이도 적용이 쉽다.
* 사용 편리한 분석데이터의 형태 : 특별한 변환 없이 간단히 사용이 가능한 데이터구조를 가지고 있다.
* 계산의 용이성 : 대용량의 데이터인 경우 계산의 수가 크게 증가하기는 하지만 분석을 위한 계산은 아주 간단하다.

**연관성규칙발견의 단점**

* 상당한 수의 계산과정 : 연관성을 관찰하고자 하는 항목이 증가하면 계산의 수가 크게 증가한다.
* 적절한 항목의 결정 : 실제 불필요한 항목들이 많이 존재한다.
* 항목의 비율차이 : 거래량이 적은 품목의 경우 거래수가 적기 때문에 연관성규칙발견 과정 중 제외될 가능성이 있다.