연관분석 /군집 분석

연관분석

연관규칙분석이란 어떤 두 아이템 집합이 번번히 발생하는가를 알려주는 일련의 규칙들을 생성하는 알고리즘입니다.
경영학에서 장바구니 분석**(Market Basket Analysis)**으로 널리 알려져 있는 방법론인데요,

소비자들의 구매이력 데이터를 토대로 **“X** 아이템을 구매하는 고객들은 **Y** 아이템 역시 구매할 가능성이 높다”는 식의 결론을 내는 알고리즘입니다.

인터넷 쇼핑을 할 때 어떤 상품을 고르면 그 상품을 구매한 사람들이 선택한 다른 상품을 제안해준다던지 하는 컨텐츠 기반 추천**(contents-based recommendation)**의 기본이 되는 방법론입니다.

연관 분석 알고리즘

총 6개 물품을 산다고 가정

(빵,밥,계란,김치,참치,김)

ID	ITEMS
1	빵,밥,김
2	계란,김치,밥
3	참치,김
4	참치,계란,김치
5	빵,김치
6	김,계란,
7	참치,밥
8	계란,빵
9	김치,계란,김
10	밥,계란,김치

연관 분석 알고리즘

각 고객마다 상품을 **구입 했으면 1**의 값을 **아니면 0**의 값을 주는 **행렬**을 만들겠습니다.

고객	빵	계란	김치	참치	김	밥
1	1	0	0	0	1	1
2	0	1	1	0	0	1
3	0	0	0	1	1	0
4	0	1	1	1	0	0
5	1	0	1	0	0	0
6	0	1	0	0	1	0
7	0	0	0	1	0	1
8	1	1	0	0	0	0
9	0	1	1	0	1	0
10	0	1	1	0	0	1

연관 분석 알고리즘

앞에와 같은 행렬을 희소행렬이라고 합니다. 이제 이러한 고객에 대한 데이터를 가지고 규칙을 만드는것이 연관분석 알고리즘 입니다.

이제 데이터 간 규칙을 만들겠습니다. 예를 들면 만약 고객이 밥을 산 사람은 김치를 산다. 만약 계란을 산 사람은 빵을 안산다와 같은 규칙들을 만드는데

여기서 조건절(**Antecedent**)은 위 예시의 규칙에서 '만일 ~라면'에 해당하는 부분입니다. 결과절(**Consequent**)은 그 뒷부분에 해당하는 내용입니다. 아이템 집합(**Item set**)이란 조건절 또는 결과절을 구성하는 아이템들의 집합입니다.

'만약 계란을 산 사람은 빵을 산다'라는 규칙이 있으면 '계란을 산다가' 조건절 '빵을 산다가' 결과절이 됩니다.

여기서 한가지 조건이 있는데 조건절에 있는 물품은 당연히 결과절에 들어가면 안되겠습니다.

연관 분석 알고리즘

이제 어떤 규칙이 데이터들의 특성을 잘 나타내는지에 대한 기준을 알아보겠습니다.

첫번째로 **‘지지도’**라는 개념이 있습니다.

지지도**(support)**는 데이터 전체에서 해당 물건을 고객이 구입한 확률입니다.

표에 데이터를 기준으로 하면 **support(빵)**은 **= 3/10**이 됩니다.

두번째로 **‘신뢰도’(confidence)**란 개념이 있습니다.

신뢰도는 어떤 데이터를 구매했을때 **해당 물건도 동시에 구매할 확률**입니다.

confidence(빵 →밥)은 **1/3**이 됩니다.

연관 분석 알고리즘

마지막 개념으로 향상도(lift)라는 개념이 있습니다.

향상도는 두 물건의 구입여부가 독립인지 판단하는 개념입니다.

lift(A→B) =P(A,B)/P(A)·P(B)

lift(빵 -> 밥)은

(1/10)/(3/10)*(3/10) = 10/9이 되겠습니다.

만약 향상도의 값이 **1**이라면 두 개의 사건은 완전히 독립이다. 즉 상관이 없는 것을 의미합니다. 향상도가 **2**라면 두 사건은 독립인것보다 **2**배의 연관성을 가진다고 볼 수 있겠습니다.

연관 분석 알고리즘

규칙의 효용성은 지지도, 신뢰도, 향상도 세 가지를 모두 반영해 평가하게 됩니다.
임의의 규칙1이 규칙2보다 효과적인 규칙이라는 이야기를 하려면 세 지표 모두 클 경우에만 그렇다고 결론을 내릴 수 있습니다..

연관 규칙 생성

규칙을 생성하는데 가장 좋은 방법은 모든 경우의 수를 모두 계산하는 것이지만 아이템의 숫자가 많아지면 많아질수록 계산의 양이 많아지기 때문에 (아이템 갯수) * (아이템 갯수-1)

많이 나온 아이템 집합을 이용하여 계산하는 **Apriori algorithm**를 사용합니다.

주요한 개념은 다음과 같습니다

표에서 볼수 있듯이 빵에 대한 지지도는 **0.3** 입니다.

그럼 (빵,밥)에 대한 지지도는 당연히 **0.3**을 넘을수 없습니다. 실제로는 어떤 기준의 지지도를 넘지 못하는 집합이 많을 껍니다.

그럼 분석하는 사람은 기준점의 지지도를 정해 놓고 만약 빵의 지지도가 기준점의 지지도를 넘지 못하면 빵이 포함된 모든 지지도를 제외하는 방식이 **Apriori algorithm**의 주요 알고리즘입니다.

연관 분석 알고리즘

각 고객마다 상품을 구입 했으면 1의 값을 아니면 0의 값을 주는 행렬을 만들겠습니다.

고객	빵	계란	김치	참치	김	밥
1	1	0	0	0	1	1
2	0	1	1	0	0	1
3	0	0	0	1	1	0
4	0	1	1	1	0	0
5	1	0	1	0	0	0
6	0	1	0	0	1	0
7	0	0	0	1	0	1
8	1	1	0	0	0	0
9	0	1	1	0	1	0
10	0	1	1	0	0	1

연관 분석 예

최소 지지도 설정=0.35

지지도(빵) = 0.3,지지도(계란)=0.6,지지도(김치)=0.5,지지도(참치)=0.3,지지도(김)=0.4,지지도(밥)=0.4

지지도가 0.35이하인 빵과 참치,밥이 들어가는 모든 집합은 제외 하겠습니다.

이후 아이템 2개의 집합을 계산합니다.

구분	계란	김치	김	~~밥~~
계란		0.4	0.2	~~0.2~~
김치			0.1	~~0.2~~
김				~~0.1~~
~~밥~~				

연관 분석 예

이제 기준 지지도를 넘는게 계란과 김치 밖에 없기 때문에 가장 큰 연관 규칙은

‘계란을 산 사람은 김치를 산다가 되겠습니다’

군집 분석

군집은 각 개체들의 유사성들을 분석해서 높은 대상 끼리 분류하고 군집에 속한 개체성의 유사성과 서로 다른 그룹간의 상이성을 분류하는 통계분석 방법입니다.

군집분석의 기본 가정은 군집내의 속한 객체들의 특성은 동질적이고 서로 다른 군집에 속한 그룹은 서로 이질적이어야 한다는 것입니다.

개별 군집의 특성은 각 특성에 대한 평균값으로 나타낼수 있으며 이것을 집단의 프로필이라 합니다.

군집 분석의 어려움

군집 분석의 문제점 중 하나는 군집의 형태가 다양하다는 것입니다.

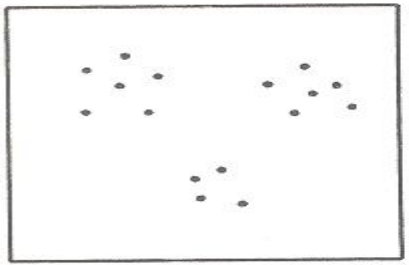
a의 경우는 쉽게 판별이 가능하지만

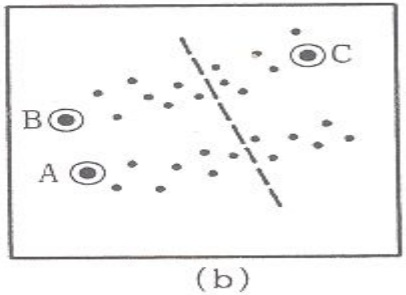
(b)의 경우에는 **b**와**c**가 같은 군집에 속하지만

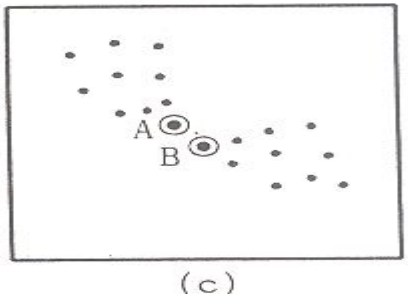
b와 **a**의 거리가 더 가까운것을 볼수있습니다.

(c)와 같은 경우에는 **a**와**b**중에 어느 것을 선택

해야 하는지 애매함이 있을 수 있습니다.







군집분석의 유사성 계산

관찰치들의 유사성 측정은 방법에 따라 거리의 유사성으로 나타낼 수 있습니다.

거리의 값이 작을수록 서로의 유사성한것을 의미하고 유사성은 값이 클수록 두 관찰치가 유사함을 의미합니다.

거리의 척도 중요한 몇가지를 말하자면

(1)유클리드 거리:실제로 우리가 많이 봐왔던 피타고라스의 정리를 이용한 거리입니다.

$$\|\mathbf{p}-\mathbf{q}\| = \sqrt{(\mathbf{p}-\mathbf{q})\cdot(\mathbf{p}-\mathbf{q})} = \sqrt{\|\mathbf{p}\|^2 + \|\mathbf{q}\|^2 - 2\mathbf{p}\cdot\mathbf{q}}.$$

(2)마할라노비스거리:확률분포상의 거리를 의미하며

$$mahalanobis(p,q) = (p-q)\Sigma^{-1}(p-q)^T$$ 중간의 항은 공분산 행렬의 역행렬이고 **T**는 변환 행렬이다.

군집분석시 유사성의 척도

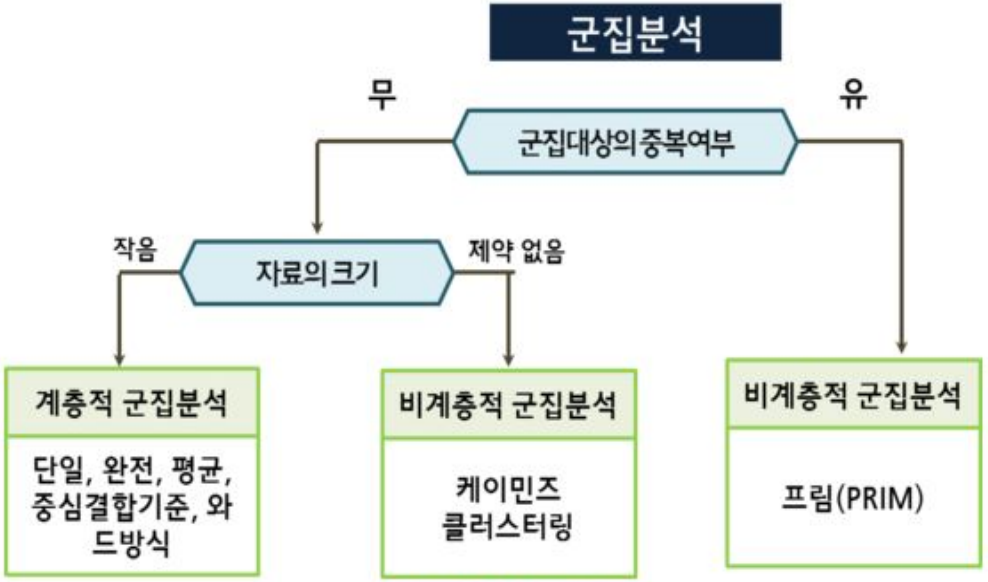
두 개체의 유사성 계산

S(ij)는 일반적으로 두 개체에 대한 변수들 사이의 상관 계수를 주로 사용하며 식은 다음과 같습니다.

$$s_{ij} = \frac{\sum_{k=1}^{p}(X_{ik} - \overline{X}_{i.})(X_{jk} - \overline{X}_{j.})}{\sqrt{\sum_{k=1}^{p}(X_{ik} - \overline{X}_{i.})^2}\sqrt{\sum_{k=1}^{p}(X_{jk} - \overline{X}_{j.})^2}} \quad \text{여기서,} \quad \overline{X}_{i.} = \frac{1}{p}\sum_{k=1}^{p} X_{ik}$$

군집 분석의 종류

군집 분석의 종류는 다음과 같습니다



[군집 분석의 종류]

군집 분석의 종류

계층적 군집 분석:개별 대상간의 거리에 의하여 가장 가까이에 있는 대상들로 부터 시작하여 나무 모양의 계층을 만들어 가는 방법 자료의 크기가 크면 분석하기 어렵다.

최단연결법: $d(U, V) = \min[d(x, y) | x \in U, y \in V]$ 다음과 같이 정의 되며 각 군집간의 최단거리로 정의하여 군집간의 유사성이 큰 군집을 묶어 나가는 방법 계산이 편리하여 속도가 빠르지만 군집이 몇개의 개체로 구성되어있을 경우 결과가 부적합 할수 있다.

최장연결법: $d(U, V) = \max[d(x, y) | x \in U, y \in V]$ 다음과 같이 정의 되며 각 군집간의 최대거리로 군집간의 유사성이 큰 군집을 묶어 나가는 방법입니다. 최단연결법과는 상반된 방법이며 최장연결법은 군집의 응집성에 중점을 둔다고 할수있다.

최단 분석/최장 분석 방법

군집분석의 종류

비 계층적 군집분석

구하고자 하는 군집의 갯수를 정하고 나서 설정된 군집에 가장 가까운 개체를 하나씩 포함해 가는 방식으로 쉽게 많은 자료를 분류 할수 있으나 군집의 수를 미리 정해주어야 하는 어려움이 있다.

가장 대표적인 방법으로는 **k-means**방법이 있습니다.

k-means 방법은 순차적으로 군집화 방법을 사용하기 때문에 순차적 군집 분석이라고 합니다.

순차적 군집 분석

군집의 중심이 정해지고 사전에 지정된 값의 거리안에 있는 값들은 모두 같은 군집으로 분류 됩니다. 한군집이 형성되고 난 다음에 새로운 군집의 중심이 결정되면 새로운 중심 일정 거리안에 있는 데이터들은 모두 같은 군집으로 결정됩니다.

이러한 과정이 모든 데이터가 최종적으로 군집화 될때까지 진행됩니다.

k-means 군집

가정:각 군집은 하나의 중심을 가집니다.