

04/24/24



Lecture 8: Association Rule Mining

Pilsung Kang

School of Industrial Management Engineering

Korea University



Lecture 8: Association Rule Mining

종속변수 = 우리가 맞추고자 하는 정답이 있는 dataset

Yes = 지도학습

No = 비지도학습

S = 행들이 어떤 속성, 변수들을 가지고 있고 변수들끼리의
최종적으로 종속변수 값이 얼마나 될 것인지를 예측, Y와 X사이
의 관계식을 찾는다.

- 연관 규칙 분석
 - ① 종속변수가 연속형이면 회귀모형
 - ② 종속변수가 범주형이면 블류오형

거의 모든 알고리즘 종속변수가 있는 지도학습 계열, 선형회귀분석,
로지스틱, 의사결정나무, 인공신경망, 양상을 전부 종속변수 존재

Type of Machine Learning/Data Mining

- According to the existence of target (Y) variable
 - ✓ Supervised learning vs. Unsupervised learning

Supervised Learning

A given dataset \mathbf{X} & \mathbf{Y}

	Var. 1	Var. 2	...	Var. d	→	Y
Ins. 1
Ins. 2	$y = f(x)$..
...
Ins. N

Semi-supervised Learning

A given dataset \mathbf{X} & \mathbf{Y}

	Var. 1	Var. 2	...	Var. d	→	Y
Ins. 1
Ins. 2	$y = f(x)$..
...
Ins. N
...
...
...
Ins. M

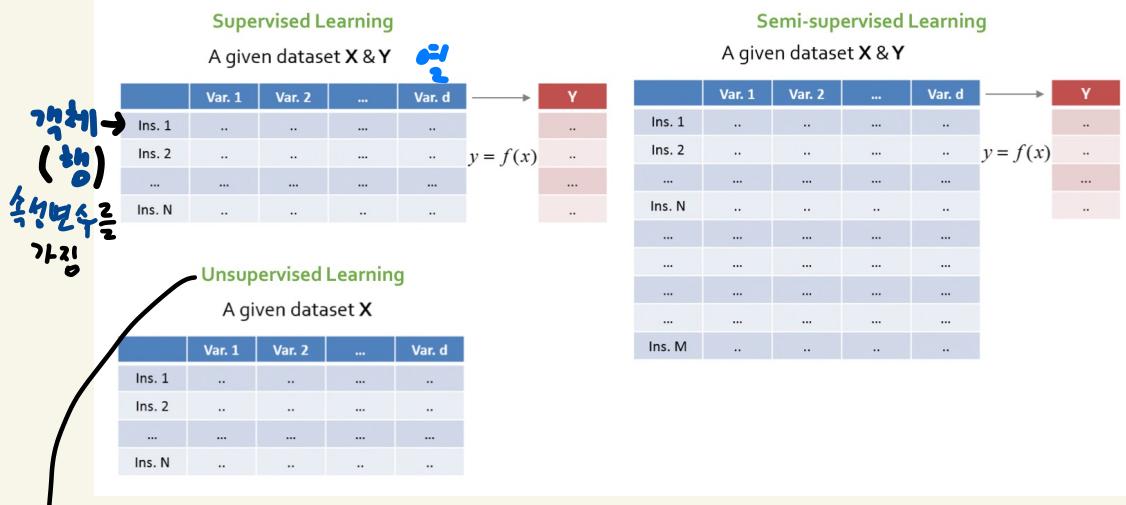
Unsupervised Learning

A given dataset \mathbf{X}

	Var. 1	Var. 2	...	Var. d
Ins. 1
Ins. 2
...
Ins. N

Type of Machine Learning/Data Mining

- According to the existence of target (Y) variable
 - ✓ Supervised learning vs. Unsupervised learning



설명변수만 존재하는 dataset을 이용해서 그 dataset이 가지고 있는
변수간의 관계 또는 각체계간의 관계를 찾아내는 비지도 학습
각체계간에 어떤 관계가 있는지에 대한 설명

Type of Machine Learning/Data Mining

- Unsupervised Learning

$$\mathcal{X} = \{\mathbf{x}_i | i = 1, \dots, n, \mathbf{x}_i \in \mathbb{R}^d\}$$



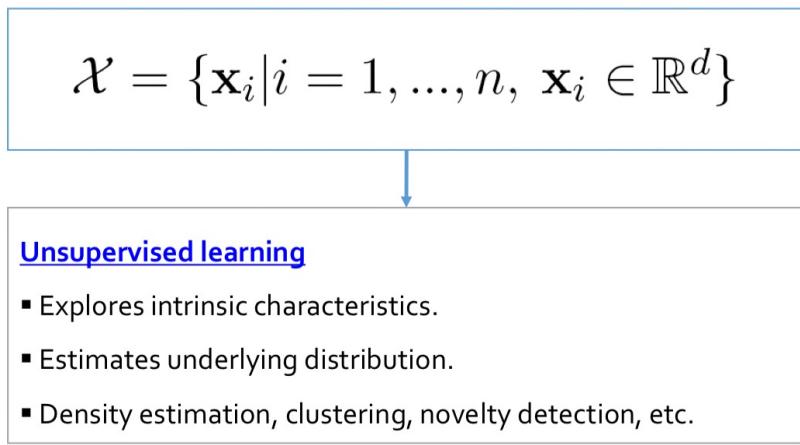
Unsupervised learning

- Explores intrinsic characteristics.
- Estimates underlying distribution.
- Density estimation, clustering, novelty detection, etc.

비지도 학습은 종속변수가 되는 y 가 없지만
 x 만 있다.

Type of Machine Learning/Data Mining

- Unsupervised Learning



비지도 학습은 종속변수가 되는 y 가 없어.
 x 만 있어.

각각의 객체들은 1차원의 수치형 벡터, 벡터를 표현
했고 그것을 n개만큼 가지고 있을 때, 각각의 데이터가 가지고
있는 내재적인 특징을 탐색하는데 단지 또는 레이아웃의 블록을 찾기
한다던지, 일도, 고집화, 이색탐지

↑
비지도 학습의 목적

Type of Machine Learning/Data Mining

비지도학습

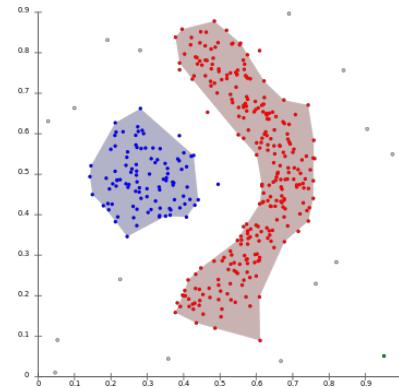
주어진 dataset을 가지고 일도를 추정, 각각의
유사한 객체를 grouping하는 것

- Unsupervised Learning

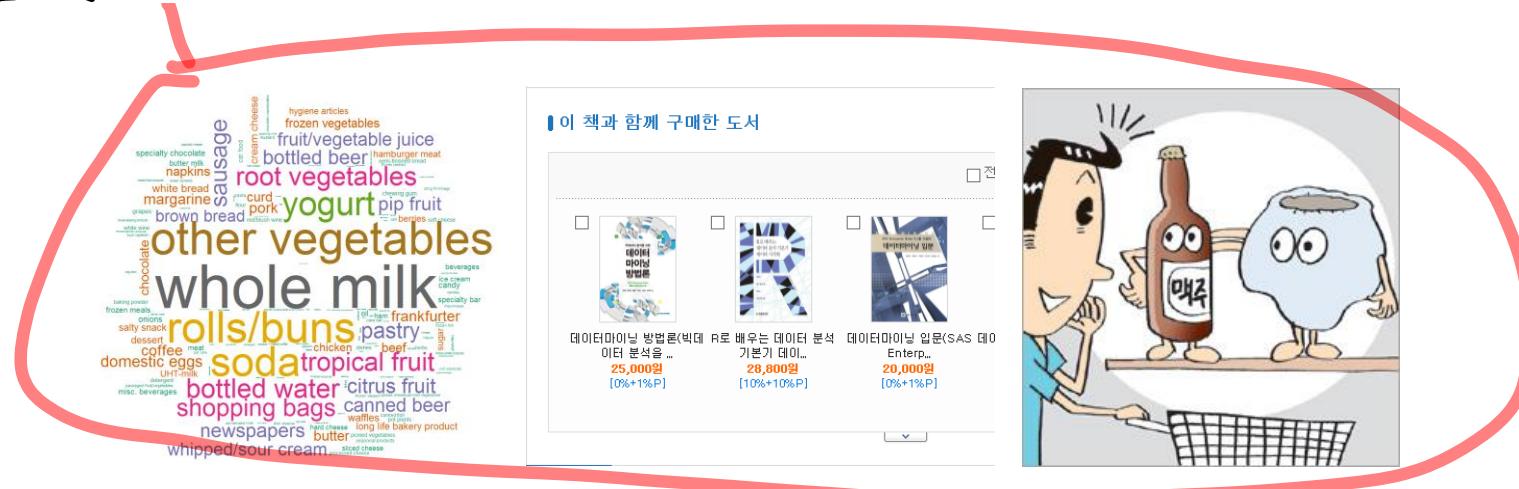
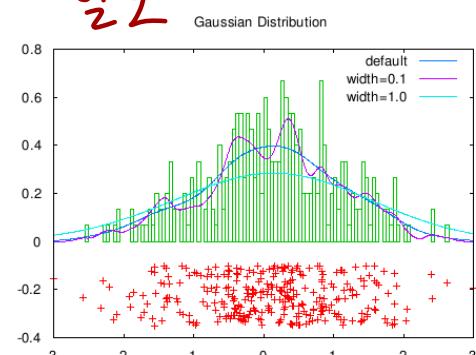
A given dataset X

	Var. 1	Var. 2	...	Var. d
Ins. 1
Ins. 2
...
Ins. N

각 변수들 간에 얼마나 강한 연관성을
가지고 있는가? (연관규칙분석)



일도



Type of Machine Learning/Data Mining

- Supervised Learning *Supervised learning은 X와 Y의 상이 있는 것을 알 수 있다. 훈련 데이터를 잘 추정해주는 관계식 찾기 또는 관계형식 찾기 형 F를 찾는 것*

$$\mathcal{X} = \{\mathbf{x}_i | i = 1, \dots, n, \mathbf{x}_i \in \mathbb{R}^d\}$$

Supervised learning

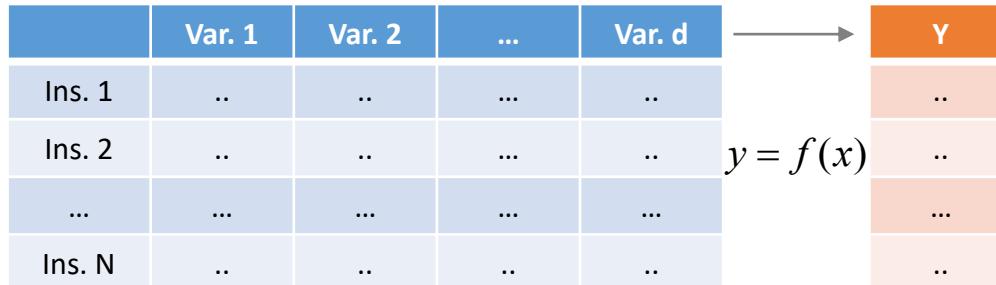
- Finds relations between X and Y.
- Estimate the underlying function $y = f(x)$.
- Classification, regression.

$$\mathcal{Y} = \{y_i | i = 1, \dots, n, y_i = f(\mathbf{x}_i)\}$$

$$y = f(x)$$

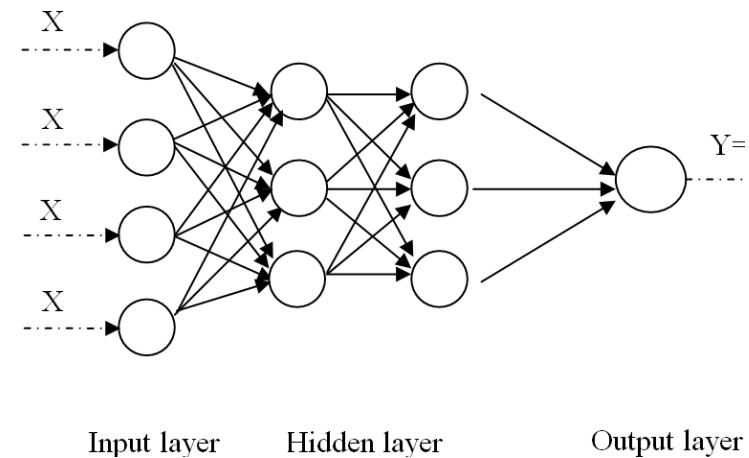
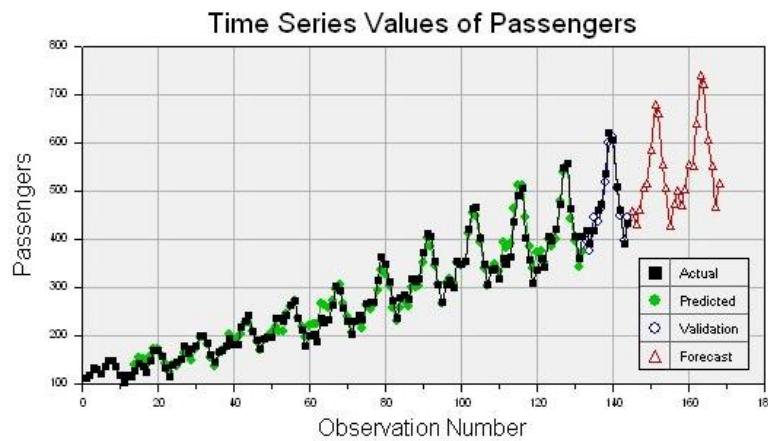
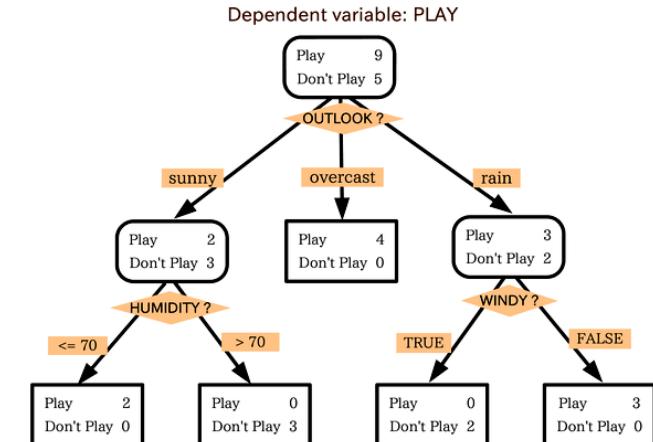
Type of Machine Learning/Data Mining

- Supervised Learning



$$y = f(x)$$

의사 결정 나무

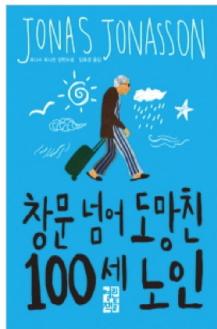


Input layer

Hidden layer

Output layer

Association Rule Mining



크게보기

미리보기

매장 재고 · 위치 >

키워드 Pick

[안내]

양로원

갱단

트렁크

데뷔작

클리우스

핵폭탄

오늘의책 | 무료배송 | 소득층제

창문 넘어 도망친 100세 노인 요나스 요나손 장편소설

요나스 요나손 지음 | 임호경 옮김 | 열린책들 | 2013년 07월 25일 출간

★★★★★ 리뷰 112개 | [리뷰쓰기](#) | 9.0(137)

KBS TV책 -김창완과 책읽기 ▾

정가: 13,800원

판매가: **12,420원** [10%↓ 1,380원 할인]

통합포인트: [기본적립] 690원 적립 [5% 적립] [안내]

[추가적립] 5만원 이상 구매 시 2천원 추가적립

[회원혜택] 우수회원 5만원 이상 구매 시 2~3% 추가적립

추가혜택: [카드/포인트 안내](#) [도서소득공제 안내](#) [추가혜택 더보기](#)

배송비: 무료 [배송비 안내](#)

배송일정: 서울특별시 종로구 세종대로 기준 [지역변경](#)

03월 04일 출고 예정 [배송일정 안내](#)

바로드림: 인터넷으로 주문하고 매장에서 직접 수령 [\[안내\]](#)

이 책의 다른 상품 정보

sam : 한달 3권 9,900원 >

eBook : 9,000원 >

원서/번역서 :
[보유] The Hundred-Year Old
Man Who Climbed Out of the
Window and Disappeared >

주문수량

장바구니 담기

바로구매

바로드림 주문

선물하기

보관함 담기

7

연관 규칙 분석

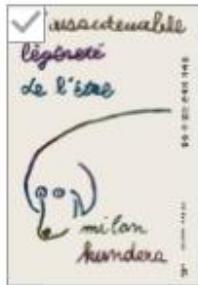
→ 책을 온라인으로 판매하는 사이트로 들어가서 콜릭해서 이 책의 정보를 보는 순간, 우리는 밑에서 이 책을 구매하신 분들이 함께 구매하신 상품입니다 해서 모다른 책을 추천해주는 것을 볼 수 있다.

Association Rule Mining

이 책을 구매하신 분들이 함께 구매하신 상품입니다

전체선택

장바구니 담기



참을 수 없는 존재의 가
벼움(양장본)

13,500원



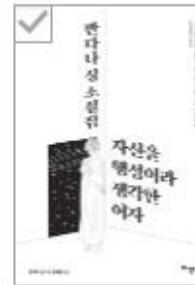
셈을 할 줄 아는
까막눈이 여자(큰글자판)

13,320원



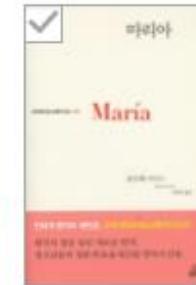
셈을 할 줄 아는
까막눈이 여자

13,320원



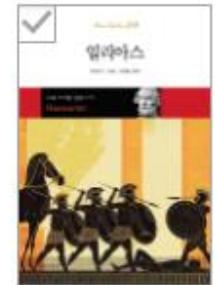
자신을 행성이라 생각한
여자

13,320원



마리아(Maria)(고려대
교 청소년문학 시리즈

11,000원

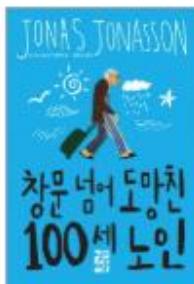


일리아스(클래식 투게더
23)

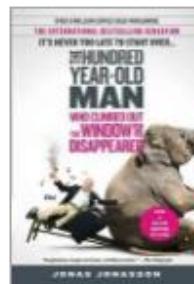
10,620원



이 상품의 꾸러미



창문 넘어 도망친 100세 노인



The 100-Year-Old
Man Who Climbed Out



창문 넘어 도망친 100세 노인 한영판 세트
(도서 2종)

25,640원

18,460원 [28% 할인] | 690원 [4% 적립]

자세히 보기

장바구니 담기

Association Rule Mining

이 책을 구매하신 분들이 함께 구매하신 상품입니다

참을 수 없는 존재의 가벼움(양장본)
13,500원

샘을 할 줄 아는 까막눈이 여자(큰글자판)
13,320원

샘을 할 줄 아는 까막눈이 여자
13,320원

자신을 행성이라 생각한 여자
13,320원

마리아(Maria)(고려대학교 청소년문학 시리즈)
11,000원

일리아스(클래식 투게더 23)
10,620원

전체선택장바구니 담기

이 상품의 꾸러미



창문 넘어 도망친 100세 노인
+
창문 넘어 도망친 100세 노인
The 100-Year-Old Man Who Climbed Out

창문 넘어 도망친 100세 노인 한영판 세트
(도서 2종)

25,640원
18,460원 [28% 할인] | 690원 [4% 적립]

[자세히 보기](#) [장바구니 담기](#)

작가의 책을 먼저보고 작가의 다른 책도 추천리스트에 있는 것
中에서 고등

이 책을 파는 업체에서 입장은 사용자가 어떤 특별한 책을 구경하고 있을 때, 그 책을 구경하는 그 사용자의 취향을 모두 파악해서 비슷한 아이템, 또는 연관성이 높은 책을 추천해 주면 구애로까지 이어진다.

Association Rule Mining

amazon Try Prime

All nespresso

Tax Cen

Departments ▾ Prime ▾ Video ▾ Music ▾

Kitchen & Dining Best Sellers Wedding Registry Small Appliances ▾ Kitchen Tools ▾ Cookware ▾ Bakeware ▾ Cutlery ▾ Dining & Entertaining ▾ Storage & Organization ▾ Event & Party Supplies ▾ Shop by Room

< Back to search results for "nespresso"



Nespresso VertuoLine Coffee and Espresso Maker with Aeroccino Plus Milk Frother, Black

by Nespresso

★★★★★ 684 customer reviews | 163 answered questions

List Price: \$249.00

Price: \$199.05 & FREE Shipping Details

You Save: \$49.95 (20%)

In Stock.

Want it Wednesday, March 1? Order within 6 hrs 36 mins and choose Two-Day Shipping at checkout. Details

Ships from and sold by Amazon.com. Gift-wrap available.

Color: Black



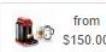
\$199.05



\$161.85



\$229.13



from
\$150.00

- Includes Aeroccino Plus milk frother: rapid one touch preparation of hot or cold milk froth; Items sold separately valued at 398
- New revolutionary Centrifusion technology to gently brew both Coffee and Espresso with one touch of a button
- Capsule recognition and code reading technology for blend-specific parametric brewing: Two capsule sizes, large for Coffee and small for Espresso
- Easy insertion and ejection of capsules; For use with Nespresso VertuoLine capsules only; Not compatible with Nespresso Original Line capsules
- Removable used capsule container holds 13-20 used capsules; Automatic off mode after 9 minutes of inactivity; Fast heat up time 15 seconds

Customers Who Bought This Item Also Bought

Page 1 of 6



DecBros Crystal
Tempered Glass
Nespresso VertuoLine
Storage Drawer Holder...
★★★★★ 737
\$29.99 ✓Prime



Nespresso VertuoLine Best
Seller Assortment, 30
Count
★★★★★ 15
\$42.46 ✓Prime



Nespresso VertuoLine
Coffee Capsules
Assortment - The Best
Sellers: 1 Sleeve of...
★★★★★ 81
\$44.92 ✓Prime



Nespresso VertuoLine
Voltesso Espresso, 10
Count
★★★★★ 17
\$11.00 ✓Prime



Nespresso VertuoLine
Espresso Assortment, 50
Count
★★★★★ 16
\$48.85 ✓Prime



Nespresso VertuoLine
Caramelizzi Coffee, 10
Count
★★★★★ 30
\$11.00 ✓Prime



Nespresso VertuoLine
Odacio Coffee, 10 Count
★★★★★ 21
\$11.00 ✓Prime



Nespresso VertuoLine
Altissio Espresso, 10
Count
★★★★★ 19
\$11.00 ✓Prime



Nespresso VertuoLine
Diavolito Espresso, 10
Count
★★★★★ 16
\$8.60



Nespresso VertuoLine
Intense Assortment, 10
Count (Pack of 4)
★★★★★ 13
\$44.77 ✓Prime

* 커피머신을 구매할때 사람들이 많이 찾던 제품들을 추천해주는 것이다.

Association Rule Mining

Amazon Tax Center

All ▾ nespresso

Departments ▾ Prime ▾ Video ▾ Music ▾

Kitchen & Dining Best Sellers Wedding Registry Small Appliances ▾ Kitchen Tools ▾ Cookware ▾ Bakeware ▾ Cutlery ▾ Dining & Entertaining ▾ Storage & Organization ▾ Event & Party Supplies ▾ Shop by Room

Help Sell Gift Cards & Registry Deals Your Amazon.com Order

[Back to search results for "nespresso"](#)

아마존도 연관상품이 나온다



Nespresso VertuoLine Coffee and Espresso Maker with Aeroccino Plus Milk Frother, Black
by Nespresso ★★★★★ 684 customer reviews | 163 answered questions

List Price: \$249.00
Price: \$199.05 & FREE Shipping. Details
You Save: \$49.95 (20%)

In Stock.
Want it Wednesday, March 12 Order within 6 hrs 36 mins and choose Two-Day Shipping at checkout. Details
Ships from and sold by Amazon.com. Gift-wrap available.

Color: Black

\$199.05 \$161.85 \$229.13 \$159.00

Includes Aeroccino Plus milk frother; rapid one touch preparation of hot or cold milk froth; items sold separately valued at 398
New revolutionary Centrifusion technology to gently brew both Coffee and Espresso with one touch of a button
Capsule recognition and code reading technology for blend-specific parametric brewing: Two capsule sizes, large for Coffee and small for Espresso
Easy insertion and ejection of capsules; For use with Nespresso VertuoLine capsules only. Not compatible with Nespresso Original Line capsules
Removable used capsule container holds 13-20 used capsules; Automatic off mode after 9 minutes of inactivity; Fast heat up time 15 seconds

Customers Who Bought This Item Also Bought

DecoBros Crystal Tempered Glass Nespresso Vertuoline Storage Drawer Holder... ★★★★★ 737 \$25.99 <small>✓ Prime</small>	Nespresso VertuoLine Best Seller Assortment, 30 Count ★★★★★ 15 \$42.46 <small>✓ Prime</small>	Nespresso VertuoLine Coffee Capsules Assortment - The Best Sellers! 1 Sleeve of... ★★★★★ 81 \$44.92 <small>✓ Prime</small>	Nespresso VertuoLine Voltesso Espresso, 10 Count ★★★★★ 17 \$11.00 <small>✓ Prime</small>	Nespresso VertuoLine Espresso Assortment, 50 Count ★★★★★ 16 \$48.85 <small>✓ Prime</small>	Nespresso VertuoLine Caramellozio Coffee, 10 Count ★★★★★ 20 \$11.00 <small>✓ Prime</small>	Nespresso VertuoLine Oacião Coffee, 10 Count ★★★★★ 21 \$11.00 <small>✓ Prime</small>	Nespresso VertuoLine Altissio Espresso, 10 Count ★★★★★ 19 \$11.00 <small>✓ Prime</small>	Nespresso VertuoLine Dlavorito Espresso, 10 Count ★★★★★ 16 \$8.60	Nespresso VertuoLine Intenso Assortment, 10 Count (Pack of 4) ★★★★★ 13 \$44.77 <small>✓ Prime</small>
---	--	---	---	---	---	---	---	--	--

Page 1 of 6

* 커피머신을 구매할때 사람들이 많이 삼킨 제품들을 추천해주는 것이다.

연관규칙분석 추천시스템이 극대화될 수 있는것은 바로 코로나 때울때 offline을 잘안가는데 온라인 선호

온라인 매장에서는 어떤 2개의 item이 잘 팔린다고
불식할 결과 나오았을때 물리적으로 2개의 item을 가까운곳에
배치를 매일 매일 할수가 없으니까, 상품에 대한 진열자체는
굉장히 빠르게 반응해서 불식결과의 빠른반응을 통해서 반영할
수가 있다.

→ 온라인 같은 경우에는 책략/아이템추천은 그 해당하는
아이템을 링크한 바구기쪽면 된다. 매우 유연하게 추천 시스템
결과물을 바로바로 적용할수가 있다.

Association Rule Mining

90년대 주목받은
분야로

- Also known as “Market Basket Analysis”

연관 규칙 분석은 다른 말로
장바구니 분석이라 불린다.



Wall Mart (USA)



E-Mart (Korea)



- ❶ 주목을 받게 된 2 가지 이유
- ❷ 철아트가 아니라 COSCO의 사례이다.
소비자들이 구매하는 영수증을 분석해 봤더니 공장이 재밌는
사실을 발견, 맥주 & 기저귀는 아무런 상관이 없는데
실제로 한꺼번에 많이 구매

Association Rule Mining

90년대 주목받던
방법론

- Also known as "Market Basket Analysis"

연관 규칙을 찾을 때
다른 말로

장바구니 분석이라 불린다.



Wall Mart (USA)



E-Mart (Korea)



사례)

▶ 주목을 받게 된 2 가지 이유

① 월마트가 아니라 OSCO의 사례이다.
소비자들이 구매한 영수증을 분석해 봤더니 굉장히 재밌는
사실을 발견, 맥주 & 기저귀는 아무런 상관이 없는데
실제로 한꺼번에 같이 구매

1) 이것들을 유통상품으로 판매하게 되면 기저귀를 "쏙" 사놓은
공산품 코너에 잘 팔리는 맥주를 가격과 놓고면 기저귀인
사례를 사령이라도 맥주가 땅져서 충동적으로 구매

↳ 1인당 구매 금액을 증가시키는, 아트 입장에서는 수익성
을 높일 수 있는 하나의 장치로 작동.

Cx) 우리나라에는 E-Mart, homeplus가 주거지역에 굉장히 가까운
일정지역에 존재, 미국은 딱 정이가 늘어서 네트워크가 동떨어져
있다.

Association Rule Mining

90년대 주목받은
분야로

- Also known as "Market Basket Analysis"

연관 규칙 분석은 다른 말로

장바구니 분석이라 불린다.



Wall Mart (USA)



E-Mart (Korea)



❶ 주목을 받게 된 2 가지 이유
① 월마트가 아니라 Cosco의 사례이다.
소비자들이 구매한 영수증을 분석해 봤더니 굉장히 재밌는
사실을 발견, 맥주 & 기저귀는 아무런 상관이 없는데
실제로 한꺼번에 같이 구매

가는 날 (시리얼 + 우유)

4부에서 실습 (AB)

시리얼 \Rightarrow 공산품

우유 \Rightarrow 신선식품

A) 원래대로 배치

B) 우유가 있는 냉장고 가는 경계 인기있는 시리얼을 둘에다가
놓기

❷ 3개월의 매출액을 보면 B) 가 구매에 긍정의 증가율이 유의미
하게 증가

❸ 분석을 통하여 똑같은 상품일지라도 어떤 게 배치하는 데 따라 많은
매출을 누릴 수 있다는 것이 연관 규칙 분석의 활용방안

Association Rule Mining

- Goal:

- ✓ Produce rules that define “what goes with what”

연관 규칙 분석의 목적은 일련의 규칙들을 생성하는 것이다.
X라는 item이 구매되면 Y라는 item과 함께 구매되게 되는 것 같은
이미지라는 결과를이다. 즉 ‘생성하는’ 데다

- ✓ “If X was purchased, then Y was also purchased”

각각의 행은 transaction(판매/기록)

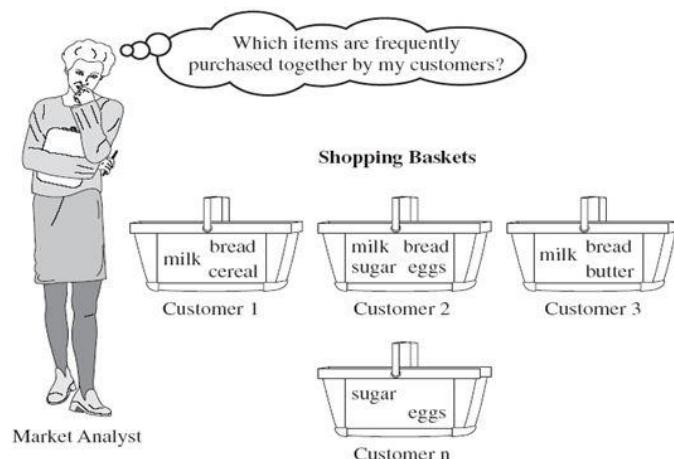
- Features

- ✓ Rows are transactions

각각의 행수증에 있는 item들이
판매되는 빈도를 의미

- ✓ Used in recommendation systems – “Our records show that you bought X, thus you may also like Y”

- ✓ Also called “affinity analysis” or “market basket analysis”



Association Rule Mining

- Goal:

연관 규칙 분석의 목적은 일련의 규칙들을 생성하는 것이다.
X라는 item이 구매되면 Y라는 item도 함께 구매되며 되는 것 같다는
이야기라는 결과를 말이다. 즉 생성되는 규칙이

- ✓ Produce rules that define “what goes with what”

- ✓ “If X was purchased, then Y was also purchased”

- Features

- ✓ Rows are transactions

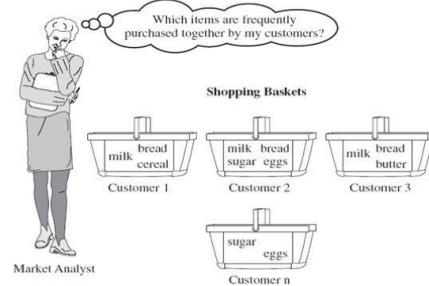
각각의 행은 transaction(판매)이다.

각각의 행은 item들이다.

한데 한데 묶음을 의미

- ✓ Used in recommendation systems – “Our records show that you bought X, thus you may also like Y”

- ✓ Also called “affinity analysis” or “market basket analysis”



11

문서를 해보려니 기울고 있는 데이터를 가지고 판단한다.
X라는 아이템을 단신이 샀을 때, X를 산 경우에 Y를
Y라는 아이템과 함께 사보지 않을래요로 추천

Association Rule Mining

- Dataset for association rule mining

- Each transaction is represented as a record

- Two representations are possible: (1) item list and (2) item matrix

→ 연관 규칙 분석을 위한 dataset
→ 각각의 행들이 transaction
즉, 명수증이다
→ 명수증에 적힌 아이템들의 유무를
놓여놓은 행과 열로 구성된 대상
우리가 놓은 연관 규칙 분석에서는
아이템의 수량을 고려하지 않는다.

①

[Item list]

Transaction ID	Items
1	Bread, Milk
2	Bread, Diaper, Beer, Eggs
3	Milk, Diaper, Beer, Coke
4	Bread, Milk, Diaper, Beer
5	Bread, Milk, Diaper, Coke

②

[Item matrix]

Transaction ID	Bread	Milk	Diaper	Beer	Eggs	Coke
1	1	1	0	0	0	0
2	1	0	1	1	1	0
3	0	0	0	0	0	0
4	0	0	0	0	0	0
5	0	0	0	0	0	1

① 번 transaction에서 빵을 1개, 우유를 1개를 사고, Bread, milk

빵을 10개 구매, Milk를 사도 5개 사도 Bread, milk

↳ 어떤 item을 샀는가가 중요하지, 몇개를 샀는가가 중요한 요소는 아니다.

Association Rule Mining

- Dataset for association rule mining

✓ Each transaction is represented as a record

✓ Two representations are possible: (1) item list and (2) item matrix

→ 연관규칙분석을 위한 dataset
→ 각각의 행들이 transaction
즉, 품목들이 있는
→ 영수증에 있는 아이템들의 구조를
분석해야 하는 대상

우리가 할 연관규칙분석에서는
아이템의 수량을 고려하지 않는다.

①

[Item list]

Transaction ID	Items
1	Bread, Milk
2	Bread, Diaper, Beer, Eggs
3	Milk, Diaper, Beer, Coke
4	Bread, Milk, Diaper, Beer
5	Bread, Milk, Diaper, Coke

②

[Item matrix]

Transaction ID	Bread	Milk	Diaper	Beer	Eggs	Coke
1	1	1	0	0	0	0
2	1	0	1	1	1	0
3	0	0	0	0	0	0
4	0	0	0	0	0	0
5	0	0	0	0	0	1

① 번 transaction에서 빵을 1개, 우유를 1개를 사고, Bread, Milk

빵을 10개 구매, Milk를 사고 5개 사고 Bread, Milk

↳ 어떤 item을 샀는가가 중요하지, 몇개를 샀는가가 중요한 요소는 아니다.

① 아이템 리스트

→ 이 방식은 각각의 transaction ID에다가 등장한 item
을 나열식으로 표현해주는 것

② item matrix

→ transaction이 행인 것은 똑같은데 열의 각각, 即 column
들이 각각의 item으로 치환되어서 해당되는 아이템이
구매되었으면 1 아니면 0. Binary representation

Association Rule Mining

- Dataset for association rule mining

- Each transaction is represented as a record

- Two representations are possible: (1) item list and (2) item matrix

→ 연관 규칙 분석을 위한 dataset
→ 각각의 행들이 **transaction**
즉 명수증이다.
→ 명수증에 적힌 아이템들의 구조를
분석하는 행하는 대상
모든 것을 연관 규칙 분석에 넣는
아이템의 수량을 고려하지 않는다.

①

[Item list]

Transaction ID	Items
1	Bread, Milk
2	Bread, Diaper, Beer, Eggs
3	Milk, Diaper, Beer, Coke
4	Bread, Milk, Diaper, Beer
5	Bread, Milk, Diaper, Coke

②

[Item matrix]

Transaction ID	Bread	Milk	Diaper	Beer	Eggs	Coke
1	1	1	0	0	0	0
2	1	0	1	1	1	0
3	0	0	0	0	0	0
4	0	0	0	0	0	0
5	0	0	0	0	0	1

① 번 transaction에서 빵을 1개, 우유를 1개를 사고, Bread, milk

빵을 10개 구매, Milk를 사고 5개 사고 Bread, milk

↳ 어떤 item을 샀는가가 중요하지, 몇개를 샀는가가 중요한 요소는 아니다.

① 아이템 리스트

→ 이 방식은 각각의 transaction ID에 따라 등장한 item을 나열식으로 표현해주는 것

② item matrix

→ transaction이 행인 것은 똑같은데 열의 각각, 한 column 들이 하나의 item으로 치환되어서 해당되는 아이템이 구애되었으면 1 아니면 0. **Binary representation**

② 실제 계산은 오른쪽으로 하는데, 조금 더 깊게 들어가면 원래 모든 데이터에서는 오른쪽으로 표현하게 되면 데이터는 굉장히 **Sparse** 된다.

Association Rule Mining

• Dataset for association rule mining

✓ Each transaction is represented as a record

✓ Two representations are possible: (1) item list and (2) item matrix

- 연관규칙분석을 위한 dataset
- 각각의 행들이 **transaction** 즉 명수증이다
- 명수증에 적힌 아이템들의 수가로 분석해야 하는 대상
- 우리가 낼 연관규칙분석에서는 아이템의 수량을 고려하지 않는다.

① [Item list]

Transaction ID	Items
1	Bread, Milk
2	Bread, Diaper, Beer, Eggs
3	Milk, Diaper, Beer, Coke
4	Bread, Milk, Diaper, Beer
5	Bread, Milk, Diaper, Coke

② [Item matrix]

Transaction ID	Bread	Milk	Diaper	Beer	Eggs	Coke
1	1	1	0	0	0	0
2	1	0	1	1	1	0
3	0	0	0	0	0	0
4	0	0	0	0	0	0
5	0	0	0	0	0	1

① 번 transaction에서 빵을 1개, 우유를 1개를 사고, Bread, milk

빵을 10개 구매, Milk를 사고 5개 사고 Bread, milk

↳ 어떤 item을 샀는가가 중요하지, 몇개를 샀는가가 중요한 표소는 아니다.

② 실제 계산은 오른쪽으로 하는데, 조금 더 깊게 들어가면 원래 모든 데이터에서는 오른쪽으로 표기하게 되면 데이터는 굉장히 **Sparse** 된다.

Sparcity 가 높아진다. 예를 들면 이 마트에서 다른 모든 제품을 한번에 보면 약 1000개 이상 제품을 판매한다. 그러나 **transaction** 을 포함하는데 있어서 이 허락하는 item의 개수가 10000개가 필요.



Association Rule Mining

- Dataset for association rule mining

- Each transaction is represented as a record

- Two representations are possible: (1) item list and (2) item matrix

- 연관규칙분석을 위한 dataset
- 각각의 행들이 **transaction** 즉 명수들이다
- 명수들에 포함된 아이템들의 유무로 분석해야 하는 대상
- 우리가 낼 연관규칙분석에서는 아이템의 수량을 고려하지 않는다.

① [Item list]

Transaction ID	Items
1	Bread, Milk
2	Bread, Diaper, Beer, Eggs
3	Milk, Diaper, Beer, Coke
4	Bread, Milk, Diaper, Beer
5	Bread, Milk, Diaper, Coke

② [Item matrix]

Transaction ID	Bread	Milk	Diaper	Beer	Eggs	Coke
1	1	1	0	0	0	0
2	1	0	1	1	1	0
3	0	0	0	0	0	0
4	0	0	0	0	0	0
5	0	0	0	0	0	1

① 번 transaction에서 빵을 1개, 우유를 1개를 사고, Bread, milk

빵을 10개 구매, Milk를 사도 5개 사도 Bread, milk

↳ 어떤 item을 샀는가가 중요하지, 몇개를 샀는가가 중요한 요소는 아니다.

→ 아무리 마트에 가서도 많이 사도, 자영업이나 식당을 하지 않는 이상은 기껏해야 10, 20개를 사는데 10, 20개 아이템만 1이고, 나머지는 0이다.

이것을 저장하는데 있어서 비효율적인 matrix

* 데이터 저장과 활용에서는 이러한 **Sparsity** 혹은 **Non-Zero** 저장할 수 있는 베커니즘이 있다.

* 머커니즘

→ 구멍가게

Association Rule Mining

1) 10개 transaction(7개) 놀고

2) 각각의 item들을 이쪽에 구성

- A toy example: a tiny retail market data

① transaction은 누군가 와서 계란, 라면, 총각국에
② transaction은 라면과 햇반

4 어떤 2개의 item이 같이 잘 팔리는가?

Transaction	Item 1	Item 2	Item 3	Item 4
1				
2				
3				
4				
5				
6				
7				
8				
9				
10				

* 애커4증

→ 3영가지

1) 10개 transaction(10개) 냄새

2) 각각의 item들을 이뤄서 구성

• A toy example: a tiny retail market data

◀ 어떤 2개의 item이 같이 잘 팔리는가?

Association Rule Mining

① transaction은 누군가 와서 계란, 라면, 흉내치에
② transaction은 라면과 흉내치

Transaction	Item 1	Item 2	Item 3	Item 4
1				
2				
3				
4				
5				
6				
7				
8				
9				
10				

연관 규칙 분석에서는 이를적으로 어떤 규칙을 찾아낼 수 있는가?
 계란을 사면 라면을 사는 규칙도 가능하는 규칙이지만
 계란을 사면 라면과 콜라를 산다. 또는 라면과
 즉석밥을 사면 콜라와 양파를 산다는 것으로 가능하는
 규칙

▶ 단 하나의 제약 조건은 조건절과 결과절이 disjoint되는 것.

(공통된 아이템이 없어)

item set이라는 개념은 (라면, 콜라)를 사면 (밥, 흉내치)를 사는 것을 의미한다.

조건절과 결과절을 disjoint하는 것은 (라면, 콜라)를 사면 (라면, 밥)을 사면, 이 경우에는 라면이라는 item이 조건절과 결과절에 둘다 들어가 있기 때문에 이런 안된다.

(라면, 콜라) → (밥, 흉내치) ok

(라면, 콜라) → (라면, 밥) X (no)

◀ item 집합을 활용하기 때문에 어마어마한 개수의 조합이 나온다.

예시의 총 4개의 item안을 가지고 이를 적으로

Association Rule Mining

$X \rightarrow Y$
인과관계 X

• Terminology

- ✓ Antecedent – “IF” part
- ✓ Consequent – “THEN” part
- ✓ Item set – the items comprising the antecedent or consequent
- ✓ Antecedent and consequent are **disjoint** (have no items in common)

① IPAD \Rightarrow Apple pencil

연관규칙분석에서는 Apple pencil을 42 IPAD로 살자.

② Apple pencil \Rightarrow IPAD

조건집 = X라는 item을 구매하면
결과집 = Y를 구매

• Generating rules

- ✓ Many rules are possible (e.g., for transaction 1)

- If egg is bought, then noodle is also bought
- If egg and noodle are bought, then tuna is also bought
- If tuna is bought, then egg is also bought, etc.

연관규칙분석으로 찾은 규칙, X를 구매하면
Y를 구매가 인과관계를 찾아주는것을 나타나.

데이터상으로 찾을때 X라는게 기준으로 잡으면 Y로
함께 빈번하게 구매되는 것과 같다.

(Generalization (인과관계)를 파악보면 ① 전자가 맞지안 연관규칙분석에서는 ①, ② 전자를 후자는 아우친다)
업계 품종에 평가표가 높기만하면 기계적으로 찾아낸다.

Association Rule Mining

$X \rightarrow Y$
인과관계 X

Terminology

✓ Antecedent – “IF” part

✓ Consequent – “THEN” part

✓ Item set – the items comprising the antecedent or consequent

✓ Antecedent and consequent are disjoint (have no items in common)

① IPAD \Rightarrow APPLE PENCIL

연관규칙분석에서는 APPLE PENCIL을 4로 IPAD로 한다.

② APPLE PENCIL \Rightarrow IPAD

조건집 = X라는 item을 구매하면
결과집 = Y를 구매

Generating rules

✓ Many rules are possible (e.g., for transaction 1)

▪ If egg is bought, then noodle is also bought

▪ If egg and noodle are bought, then tuna is also bought

▪ If tuna is bought, then egg is also bought, etc.

연관규칙분석으로 찾은 규칙, X를 구매하면
Y를 구매가 인과관계를 찾아주는것은 아니다.

데이터상으로 찾을때 X라는게 기준으로 강의면 Y로
함께 빈번하게 구매하는 것과 같다.

(crosstabulation (인과관계)를 따져보면 ① 전자가 맞지만 연관규칙분석에서는 ①, ② 전자를 두자들 아무거나
없이 풀수에 평가표가 놓기만하면 기계적으로 찾수된다.

연관규칙분석에서는 조건집 결과집간에 공통히 계산방법 복잡하지는
않아서 나온다면 item set에서 나온다.

item set은 조건집, 결과집을 구성하는 item들의 집합

X가 될수있는 경우도 있고 Y가 될수있는 경우도 있는데 X는 최소 1개에서부터 최대
5개까지 될수있다. 그래서 조건집에서 6개를 다 써버리면 결과집이 될수있다.

X가 1일때 결과집은 최소 1개의 item에서부터 최대 5개, X가 2개일때
최대 4개, X=3은 최대 3개.

X Y N=r4/c

1 < 5

$$6C_1 \times 5C_1 = 30$$

2 < 4

$$6C_1 \times 5C_2 = 60$$

:

5 - 1

조건집에서 item이 1개면 $6C_1 \times 5C_1 = 30$ 개
가능한 연관규칙이 수백개나되니 모든 연관조합을
다 따지는것은 불가능

9라는 item을 사면, 6라는 item을 사더라라는 이규칙
에 대해서 어떻게 이 규칙이 얼마나큼 유용한지
평가하기위한 지표를 3가지 사용

Association Rule Mining

Performance Measures for the rule $A \rightarrow B$

- Support

$$\text{support}(A \rightarrow B) = P(A) \text{ or } P(A, B)$$

교과서적

맞은 SW에서 많이 사용

✓ Used to find the frequent item sets

✓ The higher the support, the higher the chance of applying the rule

라면 \rightarrow 냉면
Support = $P(\text{라면}) = \frac{8}{10} = 0.8$

Transaction	Item 1	Item 2	Item 3	Item 4
1	라면		라면чи킨	
2	라면	라면치킨		라면치킨
3	라면	라면치킨	라면치킨	
4	라면	라면치킨	라면치킨	라면치킨
5	라면	라면치킨	라면치킨	라면치킨
6	라면	라면치킨	라면치킨	라면치킨
7	라면	라면치킨	라면치킨	라면치킨
8	라면	라면치킨	라면치킨	라면치킨
9	라면	라면치킨	라면치킨	라면치킨
10	라면	라면치킨	라면치킨	라면치킨

① 지지도 = Support

정말로 일관된 조건절이 발생시킬 확률, 맞은 SW에서 2개만

함께 발생시킬 확률을 계산한다.

어떤 축천을 속행하려면 조건절에 해당하는 4라는 item을 누군가 있고
있어야 컵을 커피머신을 보고 있어야지 컵을 축천 하나를 기호가 생기는지,
그러니 지지도 Support가 높다는 얘기는 해당하는 규칙을 여러번 활용할 수 있는
가능성이 있는 것이다.

Performance Measures for the rule $A \rightarrow B$

• Support

$$\text{support}(A \rightarrow B) = P(A) \text{ or } P(A, B)$$

교과서적

실제 SW에서의 적용 예

✓ Used to find the frequent item sets

✓ The higher the support, the higher the chance of applying the rule

라면 \rightarrow 냄비
 $\text{Support} = P(\text{라면}) = \frac{8}{10} = 0.8$

Transaction	Item 1	Item 2	Item 3	Item 4
1	라면		라면	
2		라면		
3		라면		
4	라면		라면	
5	라면		라면	
6		라면		
7		라면		
8	라면		라면	라면
9	라면		라면	
10	라면			

① 지지도 = Support

정말로 라면과 함께 냄비를 구매한 거래는 2건이었고, 전체 거래 수를 확률을 계산한다.

어떤 조건을 충족하여 조건집에 해당하는 9라는 item을 누군가 보고 있거나 캡슐 커피어선을 보고 있거나 캡슐을 충전해줄 기회가 생기는데, 그래서 지지도 Support 가 높다는 얘기는 해당하는 규칙을 여러번 활용할 수 있는 가능성이 있는 것이다.

규칙이 냉동식 차원에서 나오지 다른 카테고리 혹은 다른 차원에서 높으면 높을수록 좋다는 얘기

ex) 만약 라면을 사면 즉석밥도 산다. 여기서 $P(\text{즉석밥} | \text{라면})$ 은 Support는 1/2로 $P(\text{라면})$ 이 된다. 라면은 총 10개 transaction에서 8번 등장했기에 0.8

Association Rule Mining

Performance Measures for the rule $A \rightarrow B$

- Confidence

$$\text{confidence}(A \rightarrow B) = \frac{P(A, B)}{P(A)}$$

→ 2개의 item을 일렬적으로
B라는 item이 A와 얼마나 함께
같이 가게 되는가를 계산을
해낸다.

- The conditional probability of B given A
- Used to generate meaningful rules

Transaction	Item 1	Item 2	Item 3	Item 4
1				
2				
3				
4				
5				
6				
7				
8				
9				
10				

각연 → 밥

$$P(A) = \frac{8}{10} \quad P(A, B) = \frac{3}{10}$$

Confidence

$$\frac{\frac{3}{10}}{\frac{8}{10}} = \frac{3}{8} = 37.5\%$$

② 신뢰도 = Confidence

→ A라는 조건의 item이
구매가 되었을 때, A와 B가 동시에 구매될 확률

Association Rule Mining

Performance Measures for the rule $A \rightarrow B$

② 신뢰도 = Confidence

→ A라는 조건의 item이

구매가 되었을 때, A와 B가 동시에 구매될 확률

- Confidence

$$\text{confidence}(A \rightarrow B) = \frac{P(A, B)}{P(A)}$$

→ 2개의 item을 살았을 때
B라는 item이 A와 동시에 구매되는 확률
같이 구매가 되는가를 계산을
해준다.

✓ The conditional probability of B given A

✓ Used to generate meaningful rules

Transaction	Item 1	Item 2	Item 3	Item 4
1	薯条	可乐	汉堡	
2	薯条	可乐		
3	薯条			
4	薯条	可乐		
5	薯条	可乐		
6	薯条			
7	薯条	可乐		
8	薯条	可乐	汉堡	
9	薯条	可乐	汉堡	
10	水果			

라면 → 냄비

$$P(A) = \frac{8}{10} \quad P(A, B) = \frac{3}{10}$$

$$\text{Confidence } \frac{\frac{3}{10}}{\frac{8}{10}} = \frac{3}{8} = 37.5\%$$

라면을 사면 즉석밥을 산다는 규칙에 대해서는 $P(A) = \frac{8}{10} = 0.8$
 $P(B)$ 를 계산하면 라면과 냄비가 같은 번호에 구매가 되니까
2, 4, 7 transaction이 있다.

$P(A, B)$ 는 $\frac{3}{10}$ 이 된다. Confidence 는 $(\frac{3}{10})(\frac{8}{10}) = 3/8 = 37.5\%$ 된다.

→ 1끼이 우승할 이유면 라면을 사고객들 중에서 37.5%가 즉석밥을 함께 구매를 했겠지.

Association Rule Mining

Performance Measures for the rule $A \rightarrow B$

- ~~Confidence~~ lift

$$\text{lift}(A \rightarrow B) = \frac{P(A, B)}{P(A) \cdot P(B)}$$

✓ Used to determine the usefulness of generated rules

- Lift = 1: A and B are statistically independent
- Lift > 1: Positive relationship between A and B
- Lift < 1: Negative relationship between A and B

Transaction	Item 1	Item 2	Item 3	Item 4
1				
2				
3				
4				
5				
6				
7				
8				
9				
10				

③ lift(리프팅)

- 높은 도는 계산식이 따로 있다. 그 중 높은 도는 계산식을 통해 도를 계산한다.
- 같은 item
- 어떤 사건 A와 B가 통계적으로 독립이면 2 사건이 함께 발생할 확률은 각각의 발생 확률을 따로 계산을 해서 더한다.

Performance Measures for the rule A → B

• ~~Confidence~~ lift

$$\text{lift}(A \rightarrow B) = \frac{P(A, B)}{P(A) \cdot P(B)} \frac{\text{분자}}{\text{분모}}$$

✓ Used to determine the usefulness of generated rules

- Lift = 1: A and B are statistically independent
- Lift > 1: Positive relationship between A and B
- Lift < 1: Negative relationship between A and B

Transaction	Item 1	Item 2	Item 3	Item 4
1				
2				
3				
4				
5				
6				
7				
8				
9				
10				

③ lift(상관도)

- 두 아이템은 계산식이 따로있나. 그 중 러닝에 설명을 하기 위해 바꿨다.
- 2개의 item
- 어떤 사건 A와 B가 통계적으로 독립이면 2 사건이 함께 발생할 확률은 각각이 발생할 확률을 따로 계산을 해서 곱하는다.

lift가 1이 되면 A라는 item과 B라는 item이 통계적으로 독립. (아무런 연관성이 없다)

반면에 lift가 1보다 크면 A라는 item이 B라는 item과 실질적인 데이터상에서 함께 발생할 확률을 함께 발생하는 정도가 독립으로 가정했을 때의 대상 정도보다 더 높기 때문에 두 사이에 긍정적인 연관성이 있다.

Performance Measures for the rule A → B

• ~~Confidence~~ lift

$$\text{lift}(A \rightarrow B) = \frac{P(A, B)}{P(A) \cdot P(B)}$$

✓ Used to determine the usefulness of generated rules

- Lift = 1: A and B are statistically independent
- Lift > 1: Positive relationship between A and B
- Lift < 1: Negative relationship between A and B

Transaction	Item 1	Item 2	Item 3	Item 4
1	Food		Canned Food	
2		Food		
3		Food		
4	Food	Food		Food
5	Food	Food		
6		Food		
7		Food		
8	Food	Food	Food	
9	Food			Canned Food
10	Food	Food	Food	

③ lift(리프팅)

- 품목은 계산식이 따로있나. 그걸 러닝에 설명을 넣기 위해 바꿨다.
- 2개의 item
- 어떤 사건 A와 B가 통계적으로 독립이면 2 사건이 품목에 발생할 확률은 각각이 발생할 확률을 따로 계산을 해서 곱하는다.

17

라면을 사면 즉석밥을 산다는 규칙에 의해서 $P(A)$ 는 $\frac{8}{10}$ 이고 $P(B)$ 를 계산하면 B 같은 경우인 즉석밥만 사는 경우니까 2번, 4번, 7번 $\frac{3}{10}$ 이 된다.

$$P(A, B) = \frac{3}{10} \quad \text{lift} = \frac{\frac{3}{10}}{\frac{8}{10} \times \frac{3}{10}} = 1.25$$

1.25라는 이야기는 2개 item이 완벽하게 독립이라고 가정했을 때 라면과 즉석밥은 1.25 배 정도 실질적으로 끌어 3배가 됐다.

* lift가 크면 끌어온 규칙은 효과적이다.

Confidence가 1인 경우면 사실은 A라는 item을 구매하면 B는 무조건 구매를 해니까 좋은 규칙이 아니라는 질문을 할 수 있다.

Performance Measures for the rule A → B

• ~~Confidence~~ lift

$$\text{lift}(A \rightarrow B) = \frac{P(A, B)}{P(A) \cdot P(B)} \frac{\text{분자}}{\text{분모}}$$

✓ Used to determine the usefulness of generated rules

- Lift = 1: A and B are statistically independent
- Lift > 1: Positive relationship between A and B
- Lift < 1: Negative relationship between A and B

Transaction	Item 1	Item 2	Item 3	Item 4
1	bread		juice	
2				
3		juice		
4	bread			juice
5	bread	juice		
6				
7				
8	bread			
9		juice		
10	bread		juice	

③ lift(분자)

- 분자는 계산식이 따로있나. 그중 러닝에 설명을 하기 위해 바꿨다.
- 2개의 item
- 어떤 사건 A와 B가 통계적으로 독립이면 2사건이 함께 발생할 확률은 각각이 발생할 확률을 따로 계산을 해서 곱하는다.

17

en lift가 필요한가? lift가 필요한 이유는 confidence이며 A를 구매하면 B를 구매하기, 이 규칙에 대해서 만약에 B가 구매아이템일 때는 confidence가 1이 나온다.

$$\text{Conf} \quad \frac{P(A, B)}{P(A)} \quad A \rightarrow B$$

만약 학교 앞 슈퍼집에 가서 액젓 1개, 노가리 1개를 주문했다. 대량의 경우 슈퍼집에 가서 노가리만 시키고, 액젓을 시키지 않는 경우는 없기 때문에 결국 $P(A, B)$ 와 똑같은 확률값을 가지게 된다. 다시 말해 액젓라는 item이 그 애장에서 대부분하는 기호 item이기 때문이다. confidence가 1이면 두상 조건절이 발생하면 결론절이 발생하는것은 맞는데 그것만 가지고 좋은 규칙이라고 생각된다.

↑ 결국 lift까지 같이 봐야 한다.

• ~~Confidence~~ lift

$$\text{lift}(A \rightarrow B) = \frac{P(A, B)}{P(A) \cdot P(B)}$$

✓ Used to determine the usefulness of generated rules

- Lift = 1: A and B are statistically independent
- Lift > 1: Positive relationship between A and B
- Lift < 1: Negative relationship between A and B

Transaction	Item 1	Item 2	Item 3	Item 4
1			■	
2				
3		■		
4	■			
5			■	
6				■
7				
8	■			
9				
10		■		

③ lift (상관도)

- 향상되는 계산식이 따로 있다. 그 중 더 쉽게 설명을 하기 위해 나누었다.
- 2개의 item
- 어떤 사건 A와 B가 동계적으로 독립이면 2 사건이 함께 발생하는 확률은 각각의 발생하는 확률을 따로 계산을 해서 곱한다.

lift 가 1 이 되면 A라는 item과 B라는 item이 독립적
으로 독립. (아무런 연관성이 없다.)

반면에 lift 가 1 보다 크면 A라는 item이 B라는 item
과 실질적인 데이터상에서 함께 발생하는 확률을 함께
발생하는 비도가 독립으로 가정했을 때의 예상 비도보다
더 높기 때문에 둘 사이에 긍정적인 연관성·1 있다.

리스트가 1보다 작으면 반대로 기대했던 향수보다도 일정정
하는 것이기 때문에 부정적인 연관성이 된다.

관련 → 넓

$$P(A) = \frac{8}{10} \quad P(B) = \frac{3}{10}$$

Association Rule Mining

Q) 연관 규칙을 어떻게 하면 효과적으로 만들 수 있는가?

- How to generate an effective association rules?

- ✓ Ideally, create all possible combinations of items and see what rules are effective and what rules are not.
- ✓ Computation time grows exponentially as the number of items increases.

우리가 엄청난 컴퓨팅파워를 가지고 있으면 그냥 item set 기준으로 아까 보여줬던 모든 조합에 대해서
支撑을 계산하고 support, confidence, lift를 계산을 하면 된다.

- Brute-force approach (모든 경우의 수에 대해서 탐색한다)

- ✓ List all possible association rules
- ✓ Compute the support and confidence for each rule
- ✓ Prune rules that fail the **minsup** and **minconf** threshold.
- ✓ Computationally prohibitive!

item 개수가 늘어나면 가능한 조합의 규칙의 수가 기하급수적으로 늘어나기 때문에 계산 곤란증에서 불가능하다.

Association Rule Mining

→ minimum support는 유저의 hyperparameter이다.

- A priori algorithm
→ Apriori; 알고리즘은 빈발되는지 빨圣地는 아이템 집합을 고려해

✓ Consider only “frequent item sets”

✓ “support” ← 사용자간 지정하는 hyperparameter

- Criterion for item set frequency $P(A)$
- # (%) of transactions that include both the antecedent and the consequent
- Support for the item set {egg, noodle} is 4 out of transactions, or 40%

* ✓ Support of an itemset never exceeds the support of its subsets, which is known as anti-monotone property of support.

* 연관규칙을 추천하는데 있어서 서포트가 minimum support 이상을 만족하고 있지 않거나 그 규칙은 처음부터 고려해온에서 제외되었거나 하는 것을 의미

→ Minimum Support는 유저의 hyperparameter이다.

- A priori algorithm → Apriori 알고리즘은 빈발한게 빈약한는 아이템집합을 고려해
 - ✓ Consider only "frequent item sets"
 - ✓ "support" ← 사용자가 지정하는 hyperparameter
 - Criterion for item set frequency P(A)
 - # (%) of transactions that include both the antecedent and the consequent
 - Support for the item set {egg, noodle} is 4 out of transactions, or 40%

* ✓ Support of an itemset never exceeds the support of its subsets, which is known as anti-monotone property of support.

* 연관규칙을 추출하는데 있어서 서포트가 minimum support 이상을 만족하고 있지 않거나 그 규칙은 처음부터 고려 대상에서 제외되었던 것을 의미

* 이제 어떤 효과를 가지고 있는지 어떤 item이 고려될 때는 minimum support 값은 넘지 못하면 그 item은 부른집합으로 갖는 모든 item도 고려기-정하는 minimum support를 넘지 못한다.

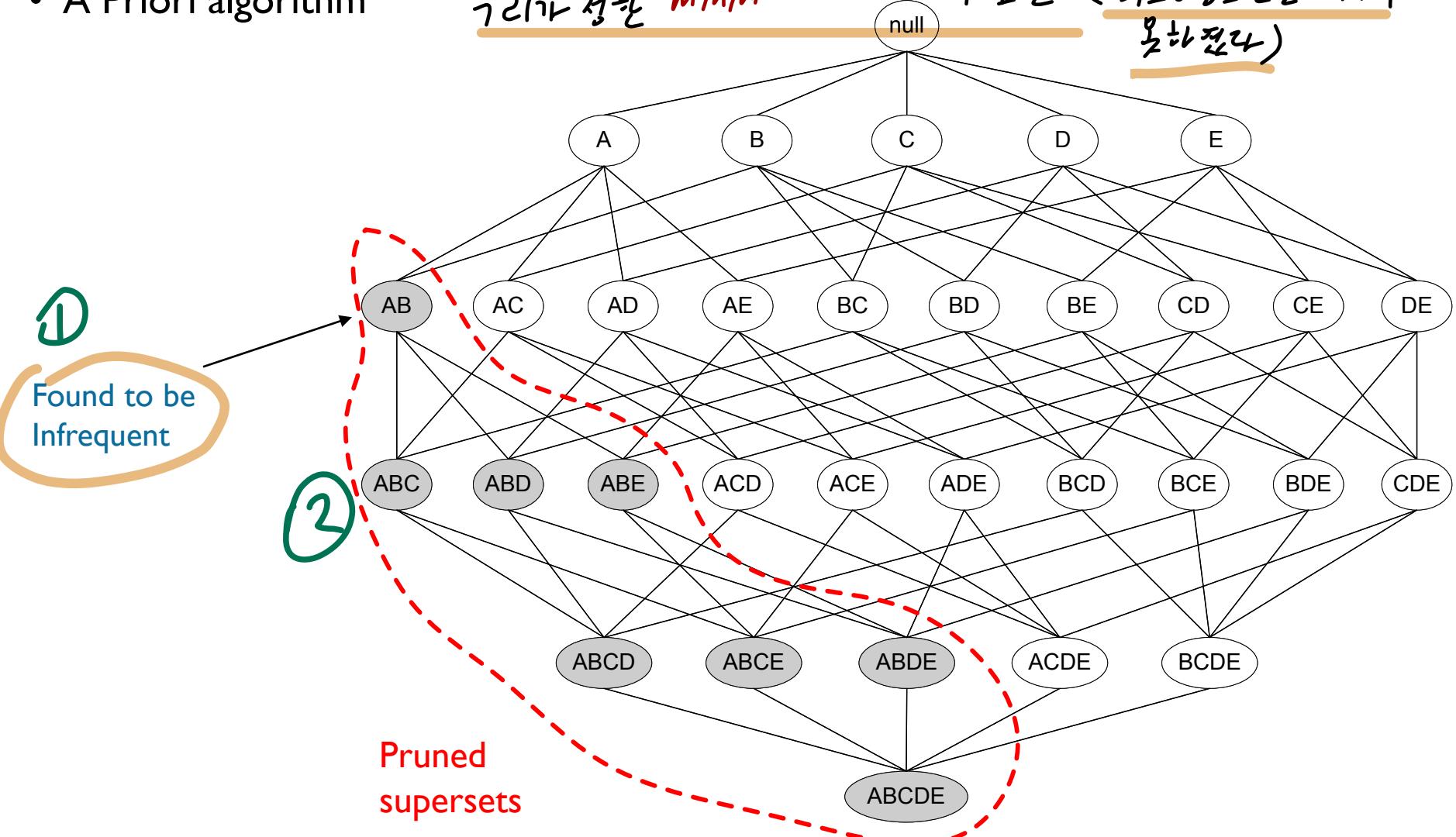
minimum support를 넘지 못하는 조합의 Superset은 전부 minimum support를 넘지 못한다.

Association Rule Mining

① 그 량이 AB 라는 item이 빈집합하지 않는다.

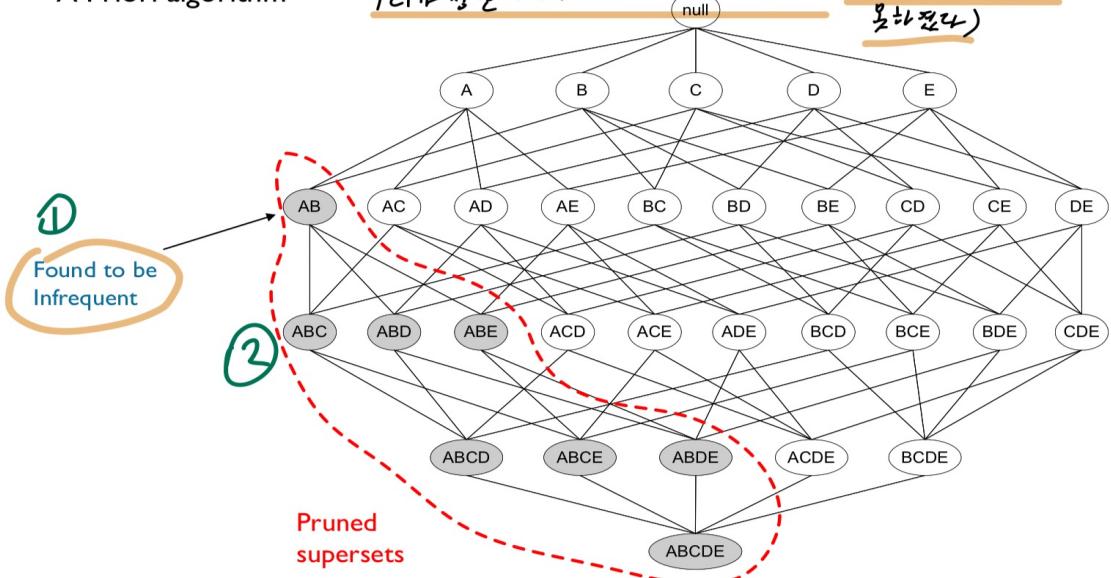
우리가 정한 $minimum\ support$ 의 조건 (최소 등장 요건을 차지하는
무수집단)

- A Priori algorithm



- A Priori algorithm

① 그림에서 AB라는 item이 빈출하지 않다.
우리가 정하는 minimum support의 조건 (최소 빈출율을 채우지 못한다)



② 그러면 우리는 a, b라는 조합을 부른집합으로 갖는 모든집합
들에 대해서 앞으로 계산을 할 필요가 없어진다는 의미

↑
실제로는 하나를 찾게되면 그것을 부른집합으로
가지는 상기 모든 집합이 다 놀라가 버리기 때문에
매우 빠르고 효율적으로 계산을 할 수 있다.

Association Rule Mining

Minimum Support

$$\frac{2}{10} = 20\% \text{ 지정}$$

- Generating frequent item sets

✓ Users set a minimum support criterion: e.g. 2 transactions or 20%

Transaction	Item 1	Item 2	Item 3	Item 4
1				
2				
3				
4				
5				
6				
7				
8				
9				
10				

Association Rule Mining

10개의 transaction이 있다.

- Generating frequent item sets

✓ Generate the list of one-item sets that meets the support criterion



8 (80%)

5 (50%)

5 (50%)

3 (30%)

2 (20%)

1 (10%)

라면 80%.

→ 창치 20%.

양파는 못넣는다.

✓ Onion is removed because it does not meet the minimum support criterion

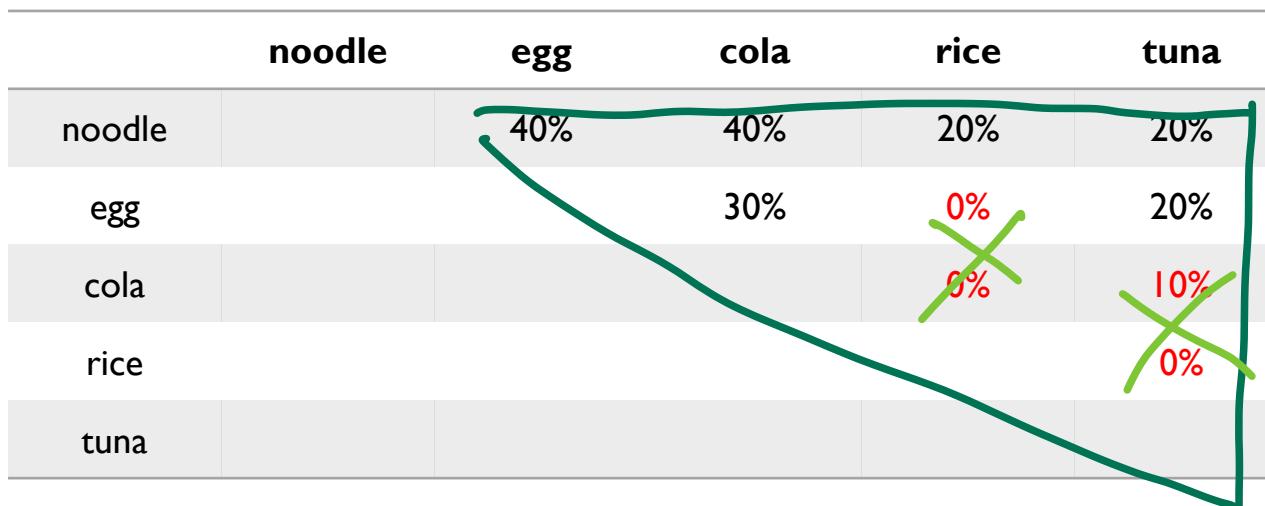
우선 각 개별적인 item들이 대해서 support를 계산합니다.

그러면 높으로는 2개짜리 item 투보리를 선정할 때 양파가 들어간 조합은 아예 차등 부터 고려자체를 하지 않았다는 뜻입니다.

Association Rule Mining

- Generating frequent item sets

✓ Use the life of one-item sets to generate list of two-item sets that meet the support criterion



✓ {noodle, egg}, {noodle, cola}, {noodle, rice}, {noodle, tuna}, {egg, cola}, {egg, tuna} are frequent two-item sets

☞ 그 21연 item 3개를 뒤집어 보면 그 21연 item 3개를 뒤집을 때는 이 3개를 아래 칸에 넣어야 하지 않겠라는 이야기입니다.

Association Rule Mining

- Generating frequent item sets

✓ Use the life of one-item sets to generate list of two-item sets that meet the support criterion

	noodle	egg	cola	rice	tuna
noodle	40%	40%	20%	20%	
egg		30%	0%	20%	
cola			0%		10%
rice				0%	
tuna					

✓ {noodle, egg}, {noodle, cola}, {noodle, rice}, {noodle, tuna}, {egg, cola}, {egg, tuna} are frequent two-item sets

☞ 그려면 item 3개짜리 조합을 찾을 때는 이 조합을 아래처럼 헤아려야 한다.
하지 않겠라는 이야기입니다.

→ 아이템이 처음 6개였는데 5개로 줄었다.

→ 5개를 봐서 대체로 조합은 예상각형 $\frac{5}{2}$ 에 해당하는 10개를
설 갯수는 총 조합 10개 된다.

근데 비중에서 즉석밥과 콜라와 찹시, 즉석밥과 찹시,
다른 조합들을 우리가 가정하는 20% 기준치를 넘지 못할 것이다.

Association Rule Mining

- Generating frequent item sets

✓ Use the list of two-item sets to generate the three-item sets.

✓ Continue up through k-item sets.

이것을 계속하면 item의 사양과 함께, 만족하는 것을 만족하는 것, 2개일 때 만족하는 것, 3개일 때 만족하는 것
이 정수로 깃털이 있다면 20/19 만족하는 것들이 없어진다.

Set-size	Item 1	Item 2	Item 3	...	Item 6
1	noodle				
1	egg				
1	cola				
1	rice				
1	tuna				$X \rightarrow Y$
2	noodle	egg			$20 \rightarrow 20C, 19C,$
2	noodle	cola			$\Rightarrow 20X19$
2	noodle	rice			
...			

$$20 \rightarrow 20C, 19C, \\ \Rightarrow 20X19$$

→ 그러면 앞서서 2개가 이를 적으로 가능했던 모든 item의 조합의 수보다는 훨씬 적은 조합의 수가 나타난다.
 → 이렇게 된 조합들이 대체로 그려면 이 조합들의 지금 만약 20개가 4였다고 가정하면 X가 등장하면 Y가 등장하는 이 규칙은 각각 20개에서 하나가 만들 수 있고, 19개에서 하나가 만들어질 수 있으므로 20×19 만큼의 조합률 지지도와 확률과 비율로 계산.

Association Rule Mining

Pseudo Code 2721

- A Priori algorithm
 - ✓ Let $k=1$
 - ✓ Generate frequent itemsets of length k
 - ✓ Repeat until no new frequent itemsets are identified
 - Generate length $(k+1)$ candidate itemsets from length k frequent itemsets
 - Prune candidate itemsets containing subsets of length k that are infrequent
 - Count the support of each candidate by scanning the DB
 - Eliminate candidates that are infrequent, leaving only those that are frequent

Association Rule Mining

- Confidence

- ✓ The % of antecedent transactions that also have the consequent item set
- ✓ E.g. "if noodle is purchased, then egg is also purchased"

$$\text{support}(\text{noodle}) = P(\text{noodle}) = \frac{8}{10}, \quad \text{support}(\text{egg}) = P(\text{egg}) = \frac{5}{10}$$

$$\text{confidence}(\text{noodle} \rightarrow \text{egg}) = \frac{P(\text{noodle}, \text{egg})}{P(\text{noodle})} = \frac{4/10}{8/10} = 0.5(50\%)$$

$\text{lift}(\text{noodle} \rightarrow \text{egg})$

$$= \frac{\text{confidence}(\text{noodle} \rightarrow \text{egg})}{\text{support}(\text{egg})} = \frac{\frac{P(\text{noodle}, \text{egg})}{P(\text{noodle})}}{\frac{P(\text{noodle}, \text{egg})}{P(\text{noodle}) \times P(\text{egg})}} = \frac{P(\text{noodle}, \text{egg})}{P(\text{noodle}) \times P(\text{egg})}$$

$$= \frac{\frac{4}{10}}{\frac{8}{10} \times \frac{5}{10}} = 1$$



이 일은 라면과 달걀을 통계적으로 상관이 있는 것을 의미.
그냥 우연히도 놓게 되어 있을 확률과 실제로 놓게 되어 될 확률이 같다.

Association Rule Mining

- Confidence

- ✓ The % of antecedent transactions that also have the consequent item set
- ✓ E.g. "if noodle is purchased, then egg is also purchased"

$$\text{support}(\text{noodle}) = P(\text{noodle}) = \frac{8}{10}, \quad \text{support}(\text{egg}) = P(\text{egg}) = \frac{5}{10}$$

$$\text{confidence}(\text{noodle} \rightarrow \text{egg}) = \frac{P(\text{noodle}, \text{egg})}{P(\text{noodle})} = \frac{4/10}{8/10} = 0.5(50\%)$$

$\text{lift}(\text{noodle} \rightarrow \text{egg})$

$$= \frac{\text{confidence}(\text{noodle} \rightarrow \text{egg})}{\text{support}(\text{egg})} = \frac{\frac{P(\text{noodle}, \text{egg})}{P(\text{noodle})}}{\frac{P(\text{egg})}{P(\text{noodle}) \times P(\text{egg})}} = \frac{P(\text{noodle}, \text{egg})}{P(\text{noodle}) \times P(\text{egg})}$$

$$= \frac{\frac{4}{10}}{\frac{8}{10} \times \frac{5}{10}} = 1$$

* 이 일은 라면과 달걀을 통제적으로 상관이 있는 것을 의미.
그냥 우연히도 함께 구매될 확률과 실제로 함께 구매될 확률이 같았다.

26

Confidence를 계산할 때 noodle을 사면 달걀을 살 때는 것을 보면, Support가 달걀 고정으로 높았을 때는 80%가 나오고 Confidence로 계산하면 조건부 확률이 50%가 나온다. lift는 1이 나온다.

Association Rule Mining

- Generated rules

✓ Set the support to 20%. (최소 지지도 20%, 필수 조건)

✓ Set the confidence to 70%. (최소 신뢰도 or Confidence는 70%를 option으로 놓으면 8가지 규칙이 만들어진다)

Rule #	Antecedent (a)	Consequent	Support	Confidence	Lift
1	tuna=>	egg, noodle	2	100	2.5
2	tuna=>	egg	2	100	2
3	noodle, tuna=>	egg	2	100	2
4	rice=>	noodle	3	100	1.25
5	egg, tuna=>	noodle	2	100	1.25
6	tuna=>	noodle	2	100	1.25
7	cola=>	noodle	5	80	1
8	egg=>	noodle	5	80	1

- Generated rules

✓ Set the support to 20%. (최소 지지도 20%, 필수옵션)

✓ Set the confidence to 70%. (최소 신뢰도 or Confidence는 70%를 option으로 놓으면 8 가지 규칙이 만들어진다)

Rule #	Antecedent (a)	Consequent	Support	Confidence	Lift
1	tuna=>	egg, noodle	2	100	2.5
2	tuna=>	egg	2	100	2
3	noodle, tuna=>	egg	2	100	2
4	rice=>	noodle	3	100	1.25
5	egg, tuna=>	noodle	2	100	1.25
6	tuna=>	noodle	2	100	1.25
7	cola=>	noodle	5	80	1
8	egg=>	noodle	5	80	1

① 쟁반을 사면 달걀과 라면을 산다. 이것이 2 건 있었고,
 100%. Confidence라는 것은 쟁반을 사는 순간 지금까지 과거
 거래내역을 보면 달걀과 라면은 항상 함께 같아졌다는
 얘기. lift는 2.5라는 것은 실제로 쟁반이라는 item과 달걀,
 라면 아이템 집합이 통계적으로 독립인 경우에 비하여
 기대독립일 경우를 가정하고 산출했을 때 구매한 수보다
 실질적으로 늘어난 수가 2.5배 더 높았다는 이야기다.

② 연관규칙에서 각각의 규칙들에 대해서 효율성을 상대적 우월성을 어떻게
 평가하는가? ①번 규칙과 ②번 규칙은 누가 더 우수한가?

① > ② 반론의 여지가 없이 ①번 규칙이 더 좋다. 왜냐하면 support, confidence
 lift가 더 높기 때문이다.

- Generated rules

- ✓ Set the support to 20%. (최소 지지도 20%, 필수 조건)
- ✓ Set the confidence to 70%. (최소 신뢰도 or Confidence는 70%를 option으로 놓으면 8 가지 규칙이 만들어진다)

Rule #	Antecedent (a)	Consequent	Support	Confidence	Lift
1	tuna=>	egg, noodle	2	100	2.5
2	tuna=>	egg	2	100	2
3	noodle, tuna=>	egg	2	100	2
4	rice=>	noodle	3	100	1.25
5	egg, tuna=>	noodle	2	100	1.25
6	tuna=>	noodle	2	100	1.25
7	cola=>	noodle	5	80	1
8	egg=>	noodle	5	80	1

(3) (3) 번 규칙과 (4) 번 규칙을 비교하면 사실 계산, 판단이 어렵다. 상황에 따라서 판단해야 한다. 왜냐하면 (3) 번 규칙은 (4) 번 규칙보다 lift가 높다. 하지만 (4) 번 규칙은 (3) 번 규칙보다 support가 크다.

3 4
lift support

이것을 활용관점에서 보면 (4) 번 규칙을 (3) 번 규칙보다 사용해볼 가능성을 놓지만 사용했을 때 효과를 볼 가능성이 높다.

- Generated rules

✓ Set the support to 20%. (최소 지지도 20%, 필수옵션)

✓ Set the confidence to 70%. (최소 신뢰도 or Confidence는 70%를 option으로 놓으면 8 가지 규칙이 만들어진다)

Rule #	Antecedent (a)	Consequent	Support	Confidence	Lift
1	tuna=>	egg, noodle	2	100	2.5
2	tuna=>	egg	2	100	2
3	noodle, tuna=>	egg	2	100	2
4	rice=>	noodle	3	100	1.25
5	egg, tuna=>	noodle	2	100	1.25
6	tuna=>	noodle	2	100	1.25
7	cola=>	noodle	5	80	1
8	egg=>	noodle	5	80	1

④ 광고관점에서 보면 광고 노출에 반도는 낮지만, 광고를 노출시켰을 때 대출까지 이어지는 가능성은 ④ 번 규칙이 훨씬 높다. 다만 ③ 번 규칙은 광고 자체를 노출시킬 가능성이 높은 규칙이라는 것이다. 무조건 매출이 중요한 것이 아니라 어떤 상황에서는 단순히 홍보관점에서는 꼭 구매를 하지 않더라도 사용자들에게 보여주는 것만으로 충분한 목적달성을 할 경우에는 ③ 번 규칙이 ④ 번 규칙보다 더 좋을 수 있다.

* 대전제는 어떤 하나의 규칙보다 유월하려면 Support, Confidence, Lift 이 3개의 값이 모두 전부 다 커야 한다.

Association Rule Mining

- Summary
 - ✓ Produce rules on associations between items from a database of transactions
 - ✓ Widely used in recommender systems
 - ✓ Most popular method is A-priori algorithm
 - ✓ To reduce computation, consider only “frequent” item sets (=support)
 - ✓ Performance is measured by confidence and lift



ANY
questions?