# Association Rules
# Market Basket Analysis

Gerome Yoo

2020

# INDEX

# Machine Learning

## Supervised Learning : 지도 학습

- Target Value to estimate
- Classification
- Regression

## Unsupervised Learning : 비지도학습

- No Target Value to estimate
- Clustering
- Association Rules

# RECOMMENDATION : 상품추천

추천은 그 자체로 환전성이 있기 때문에 가치가 있다.
넷플릭스의 경우 대여되는 영화의 2/3가 추천을 통해 발생했으며,
구글 뉴스(Google News)의 경우 38% 이상이 추천을 통해서
조회가 발생하는 것으로 알려져 있다.
또한 아마존의 경우에도 추천을 통해 판매가 전체 매출액의 35%를
넘는다.

- 넷플릭스의 빅데이터, 인문학적 상상력과의 접점, 조영신, KISDI 동향 Focus -

# RECOMMENDATION : 상품추천

**Recommendation**

- Associtaion Rules
- Sequence Analysis
- Collaborative Filtering
- Content-based Recommendation
- Who-which Model

# Association Rule Analysis : 연관분석

- Retail, Wholesale & Distribution Industry:
  Market basket Analysis : 장바구니 분석
- Unsupervised Learning
- Collaborative Filtering
- Content-based Recommendation
- Who-which Model

# Association Rule

**Purpose : 목적**

- 데이터 간의 연관법칙을 찾는 방법
- 특정 사건이 발생하였을 때
  (빈번하게) 발생하는 또 다른 사건의 규칙(Rule)

# DEFINITION : 정의

**Rule**

- 규칙
- 인과관계로서 표현되나 인과관계가 아님.
- Mathematically
  Rule : if Condition then Result $\Longleftrightarrow$ if $A$ then $B$

# DEFINITION : 정의

**Item**

- 상품 ex) 야채, 과자
- Denote Item space as $I$
- Denote a Item as $i_k$
- Mathematically
  $I = \{i_1, i_2, \cdots, i_n\}$

# DEFINITION : 정의

**Transaction**

- 구매내역, 구매로그
- Denote Transaction space as $T$
- Denote a Transaction as $t_j$
- Mathematically
  $T = \{t_1, t_2, \cdots, t_m\}$
  where $t_j = \{i_2, i_3, \cdots, i_k\}$

# 3 IMPORTANT MEASURES FOR ANALYSIS : 측도

## Support : 지지도

- $P(A \cap B)$
- 항목 A와 B가 동시에 포함된 사례 수 ÷ 전체 사례 수
- 전체 거래에서 A와 B 종목을 모두 매수한 비중
- 얼마나 의미 있는 규칙인지 판단하는 지표

# 3 IMPORTANT MEASURES FOR ANALYSIS : 측도

**Confidence : 신뢰도**

- $P(B|A) = P(A \cap B)/P(A)$
- 항목 A를 포함하는 거래 중 항목 B가 동시에 포함된 사례 수
- A 종목 매수 시 B 종목을 매수한 비중
- 채워넣어야함 ㅇ ㅇ

# 3 IMPORTANT MEASURES FOR ANALYSIS : 측도

**Lift : 향상도**

- $\frac{P(B|A)}{P(B)} = \frac{P(A \cap B)}{P(A)P(B)} = \frac{P(A|B)}{P(A)}$
- 항목 A를 포함하는 거래 중 항목 B가 동시에 포함된 사례 수
- A 종목 매수 시 B 종목을 매수한 비중
- 채워넣어야함 ㅇ ㅇ

# Additional Measures for Analysis : 측도

**IS Measure**

- 항목 A와 B가 동시에 포함된 사례 수 ÷ 전체 사례 수
- Unsupervised Learning
- Collaborative Filtering
- Content-based Recommendation
- Who-which Model

# Why not Brute-Force?

## Complexity for Brute-Force

- 항목 A와 B가 동시에 포함된 사례 수 ÷ 전체 사례 수
- Unsupervised Learning
- Collaborative Filtering
- Content-based Recommendation
-

$$R = \sum_{k=1}^{d-1} \left[ \binom{d}{k} \times \sum_{j=1}^{d-k} \binom{d-k}{j} \right]$$
$$= 3^d - 2^{d+1} + 1$$

# Why not Brute-Force?

**Proof**

$$\text{Recall } (a+b)^n = \sum_{x=0}^{n} \binom{n}{x} a^x b^{n-x} : \text{Binomial Theorem}$$

$$R = \sum_{k=1}^{d-1} \left[ \binom{d}{k} \times \underline{\sum_{j=1}^{d-k} \binom{d-k}{j}} \right]$$

# Why not Brute-Force?

**Proof**

$$
\begin{aligned}
\text{Then } & \sum_{j=1}^{d-k} \binom{d-k}{j} \\
&= \sum_{j=1}^{d-k} \binom{d-k}{j} 1^j 1^{d-k-j} + \binom{d-k}{0} 1^0 1^{d-k} - \binom{d-k}{0} 1^0 1^{d-k} \\
&= \sum_{j=0}^{d-k} \binom{d-k}{j} 1^j 1^{d-k-j} - \binom{d-k}{0} 1^0 1^{d-k} \\
&= (1+1)^{d-k} - 1 \\
&= 2^{d-k} - 1
\end{aligned}
$$

# Why not Brute-Force?

## Proof

$$
\begin{aligned}
\text{Then } R &= \sum_{k=1}^{d-1} \binom{d}{k} (2^{d-k} - 1) \\
&= \sum_{k=1}^{d-1} \binom{d}{k} 2^{d-k} - \sum_{k=1}^{d-1} \binom{d}{k} \\
&= \sum_{k=0}^{d} \binom{d}{k} 2^{d-k} - \binom{d}{d} 2^0 1^d - \binom{d}{0} 2^d 1^0 \\
&- \sum_{k=0}^{d} \binom{d}{k} - \binom{d}{d} 1^0 1^d - \binom{d}{0} 1^d 1^0 \\
&= \{(2+1)^d - 1 - 2^d\} - \{2^d - 2\} \\
&= 3^d - 2^{d+1} + 1
\end{aligned}
$$

# Analysis with R

## Related Packages

- arules
- arulesSequences
- arulesViz
- arulesCBA
- arulesNBMiner
- RSarules

# Analysis with R

## Arules Package

- Arules package provide pre-built function with Apriori Algorithm
- The package provide measure of Support, Confidence and Lift
- Does not provide IS measure and Cross Support.

# Analysis with R

## Arules Viz

- Arules package provide pre-built function with Apriori Algorithm
- The package provide measure of Support, Confidence and Lift
- Does not provide IS measure and Cross Support.