

# DS 12주차 수업 팀과제

연관성 분석

팀 AOA

경영학과 2014109050 박한솔  
경영학과 2015126062 신지수

# 01

## 연관 규칙의 여러가지 평가 척도

### 1. 지지도(support)

거래 집합에서 해당 항목집합의 상대적 발생 빈도를 나타낸다. 항목  $x$ 의 지지도는  $x$ 의 거래 수 나누기 전체 거래수로 계산 가능하다.

### 2. 신뢰도(confidence)

선행조건을 포함하는 모든 거래들 중 규칙의 결과가 발생할 가능성을 나타낸다.  $X \rightarrow Y$  신뢰도는  $X$ 와  $Y$ 가 함께 발생할 때의 지지도 나누기 항목  $X$ 의 지지도로 계산된다.

### 3. 향상도(lift)

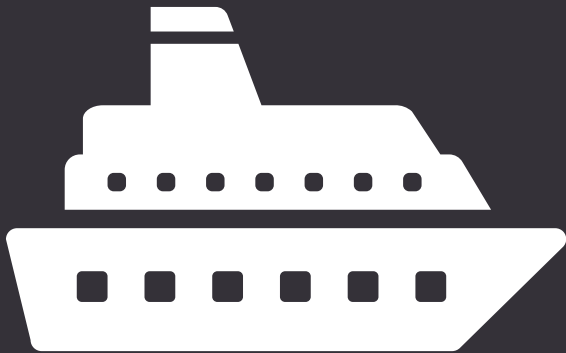
연관규칙( $X \rightarrow Y$ )로 인해 항목  $Y$ 를 포함하고 있는 거래의 비율이 증가한 정도를 나타낸다.  $X$ 와  $Y$ 가 함께 발생할 때의 지지도를 항목  $X$ 의 지지도와  $Y$ 의 지지도의 곱으로 나눠 계산한다.

### 4. IS(Interest-Support)척도

향상도와 지지도의 곱에 제곱근을 취한 값으로, 향상도와 지지도가 모두 높은 경우에만 높은 값을 가지므로, 향상도 도는 지지도 한쪽만 높은 rule은 제외할 수 있음

## 02

# Titanic Dataset



타이타닉 데이터셋을 활용하여 승객의 나이, 성별, 요금 등 속성간의 연관성을 분석해보고자 한다.

속성을 입력할 때는 모두 1또는 0 이진형으로 변환해주어야 한다.

## 03

## 연관성 분석 모델 생성

Row No.	Survived	Age	Passenger Class	Sex
1	Yes	29	First	Female
2	No	2	First	Female
3	No	30	First	Male
4	No	25	First	Female
5	Yes	48	First	Male
6	Yes	63	First	Female
7	No	39	First	Male
8	Yes	18	First	Female
9	Yes	26	First	Female
10	Yes	80	First	Male

〈기존 데이터셋〉

Male	Female	Age(0-19)	Age(20-39)	Age(40-59)	Age(60-)
0	1	0	1	0	0
0	1	1	0	0	0
1	0	0	1	0	0
0	1	0	1	0	0
1	0	0	0	1	0
0	1	0	0	0	1
1	0	0	1	0	0
0	1	1	0	0	0
0	1	0	1	0	0

〈속성을 분리한 이후의 데이터셋〉

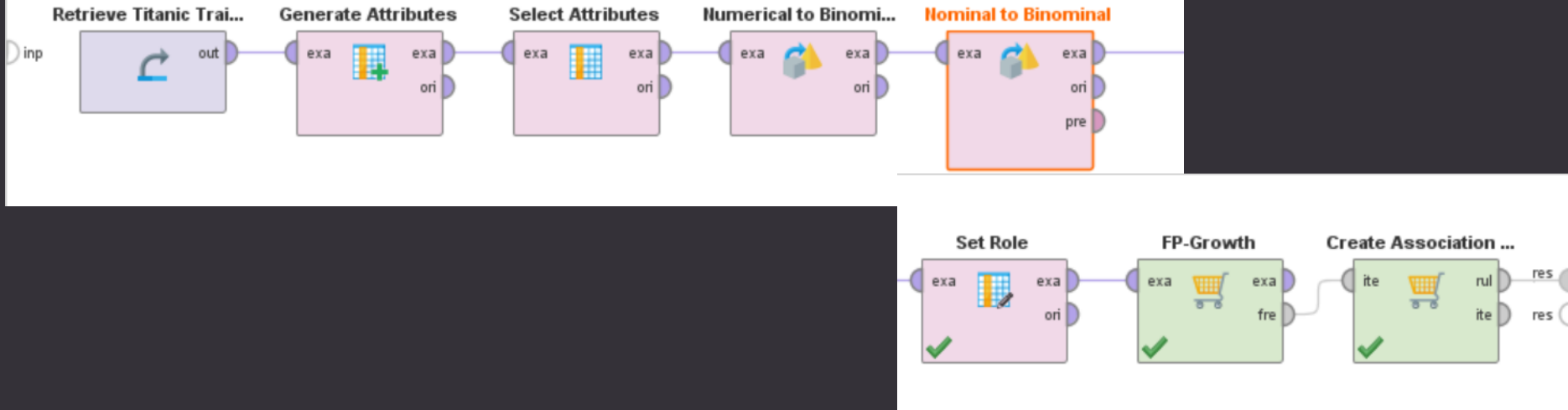
먼저 원 데이터셋의 속성을 분리해주는 작업을 해야 한다.

Generate attribute 오퍼레이터를 통해 범위 또는 클래스 일치 여부에 따라 속성을 분리하였고 분리한 속성에 적용될 때 1 값을 부여하였다.

# 04

## 연관성 분석 모델 생성

Process




〈연관성 분석 모델 생성 process〉

FP-Growth에서 아이템 셋을 생성하고, support(지지도) 수준을 결정하여 일정 수준 이상의 지지도를 갖는 아이템 셋만 채택한다.

Create Association 오퍼레이터에서는 confidence, lift 등의 기준으로 규칙들을 추출할 수 있다.

## 05

## 연관성 분석 모델 생성

 FP-Growth

input format items in... ⓘ

positive value ⓘ

min require... support ⓘ

min support 0.95 ⓘ

criterion lift ⓘ

min criterion v... 1.2 ⓘ

gain theta 2.0 ⓘ

laplace k 1.0 ⓘ

No.	Premises	Conclusion	Support	Confidence	Lift
7	PClass-3rd	Male, Age(20-39), Survive...	0.231	0.432	1.232
8	PClass-3rd	Fare(-10), Male, Age(20-3...	0.231	0.432	1.232
9	Fare(-10), PClass-3rd	Male, Age(20-39), Survive...	0.231	0.432	1.232
10	Survived-No	Age(20-39), PClass-3rd	0.276	0.446	1.202
11	Survived-No	Fare(-10), Age(20-39), PC...	0.276	0.446	1.202
12	Fare(-10), Survived-No	Age(20-39), PClass-3rd	0.276	0.446	1.202

〈FP-Growth와 Create Association Rules 파라미터〉


〈프로세스 결과 일부〉

FP-Growth 파라미터에서 개별 아이템 셋 선정 시 support를 0.95 이상인 것들로 채택하고,  
Create Association Rules 파라미터에서는 lift를 기준으로 하여 1.2가 넘는 Rule들을 추출하였다.

프로세스를 돌리면 오른쪽과 같이 77개의 연관 규칙이 도출되었다.

## 06

## 연관성 분석 모델 생성

 FP-Growth

input format  ⓘ

positive value  ⓘ

min require...  ⓘ

min support  ⓘ

criterion  ⓘ

min confidence  ⓘ

gain theta  ⓘ

laplace k  ⓘ

No.	Premises	Conclusion	Support	Confidence	Lift
7	Male, Age(20-39)	Survived-No	0.350	0.829	1.340
8	Male, Age(20-39)	Fare(-10), Survived-No	0.350	0.829	1.340
9	Fare(-10), Male, Age(20-39)	Survived-No	0.350	0.829	1.340
10	Age(20-39), Survived-No, PClas...	Male	0.231	0.838	1.292
11	Age(20-39), Survived-No, PClas...	Fare(-10), Male	0.231	0.838	1.292
12	Fare(-10), Age(20-39), Survived...	Male	0.231	0.838	1.292

〈FP-Growth와 Create Association Rules 파라미터〉

〈프로세스 결과 일부〉

다음으로, FP-Growth 파라미터는 그대로 유지하고(support = 0.95),  
Create Association Rules 파라미터에서는 confidence를 기준으로 하여 0.8이 넘는 Rule들을 추출하였다.  
프로세스를 돌리면 오른쪽과 같이 40개의 연관 규칙이 도출되었다.

No.	Premises	Conclusion	Support	Confidence	Lift ↑
69	Fare(-10), Male, Age(20-39)	Survived-No	0.350	0.829	1.340
66	Fare(-10), Male	Survived-No	0.528	0.815	1.316
76	Male, PClass-3rd	Survived-No	0.316	0.853	1.377
57	Fare(-10), Female	Survived-Yes	0.261	0.742	1.948
73	Male, Age(20-39), PClass-3rd	Survived-No	0.231	0.851	1.375

위와 같이 5개의 규칙을 선정해보았다. support 수치, lift 수치, confidence 수치가 높은 것을 위주로 선정하였고, 우리가 관심 있는 정보는 어떤 경우에 survive가 yes이고 no인지(사람의 생사에 연관이 있는 속성이 무엇인지)이므로 Conclusion가 survived-Yes 또는 No 단일로 되어있는 것으로 선정하였다.

특히 흥미로웠던 점은, 남성인 경우 10,000불 미만의 요금을 지불, 나이가 20-30대, 3등급 좌석을 이용하였을 때 생존율이 낮다는 것과 10,000불 미만의 탑승료를 지불한 여성의 경우 생존율이 높다는 것이었다.