연관 규칙 마이닝

학번 : 2017316003

이름 : 고지명

Titanic 데이터에는 성별과 생사 여부에 대한 데이터가 포함되어 있습니다. apriori() 함수를 이용하여 클래스별 각 성별의 생존율을 분석하세요.

주의!

분석 코드와 실행 결과는 캡쳐하여 과제 파일에 포함하여 제출하세요.

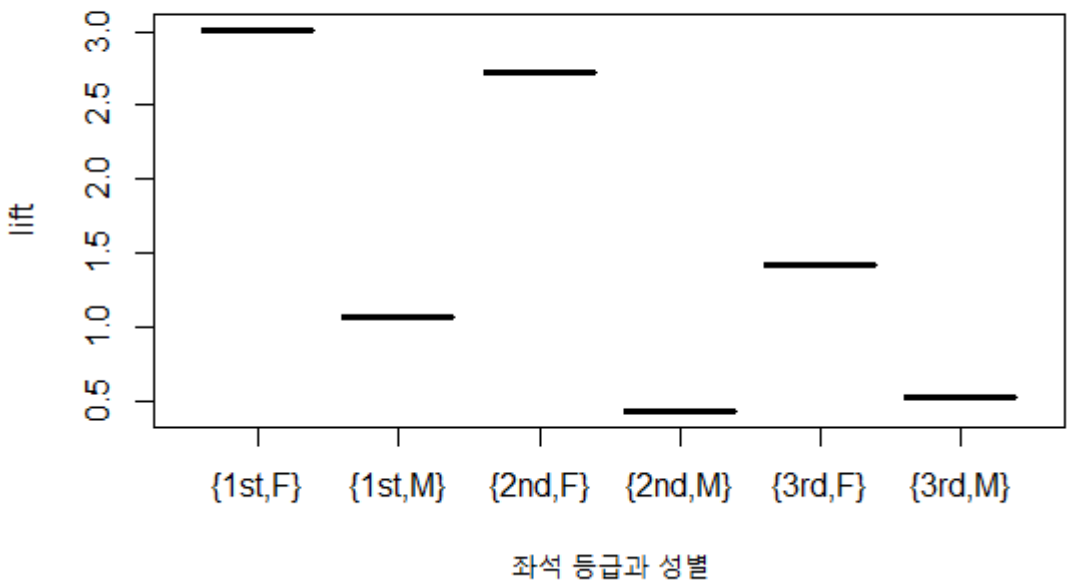
분석 결과를 해석하는 간단한 코멘트를 추가하세요.

```
> #   Titanic dataset으로 연관규칙 생성
> ##################################################
>
> #패키지 설치: arules(연관규칙분석 패키지)
> library(arules)
필요한 패키지를 로딩중입니다: Matrix

다음의 패키지를 부착합니다: 'arules'

The following objects are masked from 'package:base':

    abbreviate, write

>
> #데이터셋을 확인합니다.
> str(Titanic)
 'table' num [1:4, 1:2, 1:2, 1:2] 0 0 35 0 0 0 17 0 118 154 ...
 - attr(*, "dimnames")=List of 4
  ..$ Class   : chr [1:4] "1st" "2nd" "3rd" "Crew"
  ..$ Sex     : chr [1:2] "Male" "Female"
  ..$ Age     : chr [1:2] "Child" "Adult"
  ..$ Survived: chr [1:2] "No" "Yes"
```

```
> #raw data를 데이터 프레임으로 전환합니다.
> df <- as.data.frame(Titanic)
>
> # 알고리즘에 적용할 수 있도록 데이터를 변경합니다.
> titanic.raw <- NULL
> for(i in 1:4){
+     titanic.raw <- cbind(titanic.raw, rep(as.character(df[, i]), df$Freq))
+ }
>
> #문자열 형인 titanic.raw를 데이터 프레임으로 변경합니다.
> titanic.raw <- as.data.frame(titanic.raw,
+                              stringsAsFactors = TRUE)
>
> #titanic.raw 데이터에 변수(Column)명을 추가합니다.
> #           ("Class", "Sex", "Age", "Survived")
> names(titanic.raw) <- names(df)[1:4]
>
> #연관 규칙 생성
> ## apriori 인자 설명
> ## data 인자: titinic.raw
> ## parameter 인자: 규칙에 포함되는 최소 길이, 최소 지지도, 최소 신뢰도
> ## appearance 인자: 원하는 아이템[list(rhs=c("Survived=Yes")만 노출)]
> ## lhs(Left Hand Side): 왼쪽에 규칙을 구성
> rules <- apriori(titanic.raw, control=list(verbose=F),
+             parameter = list(minlen=3, supp=0.0015, conf=0.13),
+             appearance = list(rhs=c("Survived=Yes"),
+             lhs=c("Class=1st", "Class=2nd", "Class=3rd",
+             "Sex=Male", "Sex=Female"),
+             default="none"))
>
> #quality(): SOM(Self-Organizing Map) 알고리즘을 활용해 결괏값으로부터
> #           몇몇 Quality 지표를 생성합니다.
> #두 자릿수로 제한된 퀄리티 지표를 산출합니다.
```

```
> quality(rules) <- round(quality(rules), digits = 2)
>
> #inspect(): Transaction Object의 연관 규칙 내용 확인합니다.
> #           (lift 내림차순으로 정렬된)
> result <- inspect(sort(rules, by="lift"))
    lhs                         rhs           support confidence coverage lift
[1] {Class=1st,Sex=Female} => {Survived=Yes} 0.06    0.97       0.07     3.01
[2] {Class=2nd,Sex=Female} => {Survived=Yes} 0.04    0.88       0.05     2.72
[3] {Class=3rd,Sex=Female} => {Survived=Yes} 0.04    0.46       0.09     1.42
[4] {Class=1st,Sex=Male}   => {Survived=Yes} 0.03    0.34       0.08     1.07
[5] {Class=3rd,Sex=Male}   => {Survived=Yes} 0.04    0.17       0.23     0.53
[6] {Class=2nd,Sex=Male}   => {Survived=Yes} 0.01    0.14       0.08     0.43
    count
[1] 141
[2]  93
[3]  90
[4]  62
[5]  88
[6]  25
```

```
> #One More Thing!
> ##sex에 성별과 등급 정보를 담는데
> #   , 필요한 정보만 남기기 위해 gsub을 통해 값을 정제합니다.
> sex = result[, 1]
> sex = gsub("Class=","",sex)
> sex = gsub("Sex=","",sex)
> sex = gsub("Male","M",sex)
> sex = gsub("Female","F",sex)
> sex = as.factor(sex)
>
> ##lift: {등급, 성별}의 신뢰도/{살아남은 경우}
> #       즉, 단순히 살아남은 경우보다 {등급, 성별}이 주어졌을 때
> #       살아남을 확률이 얼마나 증가했는가를 나타냅니다.
> lift = as.numeric(result[, 7])
>
> #plot 함수를 통해 데이터를 시각화합니다!
> plot(sex, lift, xlab="좌석 등급과 성별", ylab="lift")
>
```



분석 결과 코멘트

- **여성은 모든 클래스에서 남성보다 생존율이 높았습니다.**
- **Class 가 높은 여성일수록, 생존 확률이 증가했습니다.**
- **반면, 남성의 경우 Class 가 1 인 남성은 생존율과 독립적인 관계를 이루었으며, Class 가 2, 3 인 남성의 경우 생존율이 음의 상관관계로 Class 2, 3 에 탑승한 남성일수록 생존율이 떨어진다고 해석할 수 있습니다.**