

데이터과학을 위한 **R**프로그래밍

13주차. 연관규칙과 로지스틱모형



이혜선 교수

포항공과대학교 산업경영공학과



목차

13주차. 연관규칙과 로지스틱모형

1차시 연관규칙 I

2차시 연관규칙 II

3차시 로지스틱 회귀모형



13주차

2차시

연관규칙 II

● 연관규칙 - 데이터 설명(Groceries)

☑ Groceries data("arules" 패키지에 탑재되어 있는 데이터)

- data("Groceries")으로 불러옴
- 실제 식료품점에서 1개월(30일)치의 transaction 데이터
- 9835트랜잭션 / 169항목
- 밀도가 0.026라고 되어 있는데, 9835×169 cell 중에서 2.6%의 cell에 거래가 발생해 숫자가 차 있다는 뜻임
- Element(itemset/transaction) length distribution
: 하나의 거래 장바구니(row 1개 당)에 item의 개수 별로 몇번의 거래가 있었는지 나타냄

● 연관규칙 - 데이터 설명(Groceries)

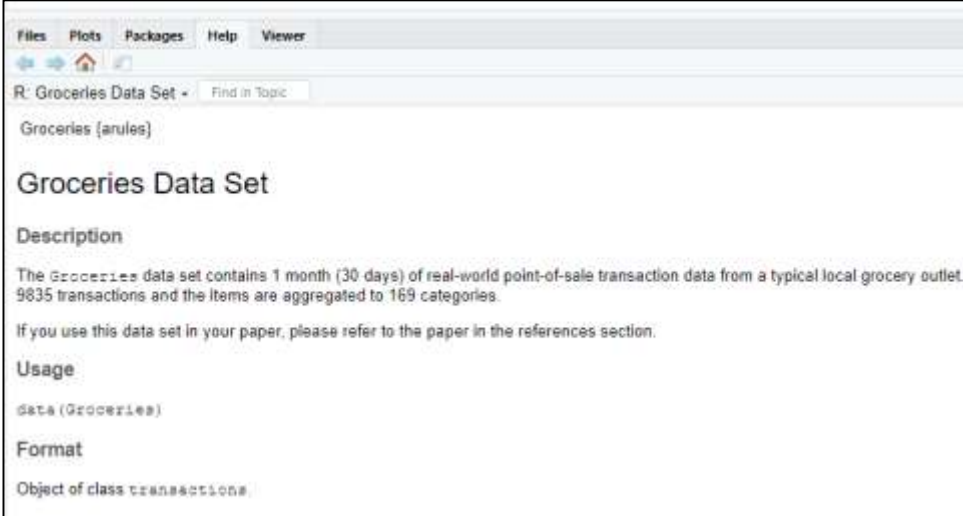
☑ Groceries data – transaction data

```
# association rule analysis package
# install.packages("arules")
library(arules)

#association rule analysis
data("Groceries")

help("Groceries")

summary(Groceries)
```



The screenshot shows the R help page for the 'Groceries' data set. The window title is 'R: Groceries Data Set'. The menu bar includes 'Files', 'Plots', 'Packages', 'Help', and 'Viewer'. Below the menu bar, there is a search bar with the text 'Find in Topic'. The main content area is titled 'Groceries (arules)' and 'Groceries Data Set'. It includes a 'Description' section stating: 'The Groceries data set contains 1 month (30 days) of real-world point-of-sale transaction data from a typical local grocery outlet. 9835 transactions and the Items are aggregated to 169 categories. If you use this data set in your paper, please refer to the paper in the references section.' There are also sections for 'Usage' and 'Format'. The 'Usage' section shows the command `data(Groceries)`. The 'Format' section states 'Object of class transactions'.

● 연관규칙 - 데이터 설명(Groceries)

✓ Groceries("arules" package에 탑재되어 있는 데이터)

```
> summary(Groceries)
transactions as itemMatrix in sparse format with
9835 rows (elements/itemsets/transactions) and
169 columns (items) and a density of 0.02609146
```

9835거래건수
169항목

```
most frequent items:
  whole milk other vegetables    rolls/buns      soda      yogurt    (Other)
    2513         1903         1809      1715      1372    34055

element (itemset/transaction) length distribution:
sizes
  1    2    3    4    5    6    7    8    9   10   11   12   13   14   15   16   17   18   19   20   21
2159 1643 1299 1005  855  645  545  438  350  246  182  117  78  77  55  46  29  14  14  9  11
 22  23  24  26  27  28  29  32
  4   6   1   1   1   1   3   1

  Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 1.000  2.000   3.000   4.409  6.000  32.000

includes extended item information - examples:
  labels level2      level1
1 frankfurter sausage meat and sausage
2  sausage sausage meat and sausage
3  liver loaf sausage meat and sausage
```

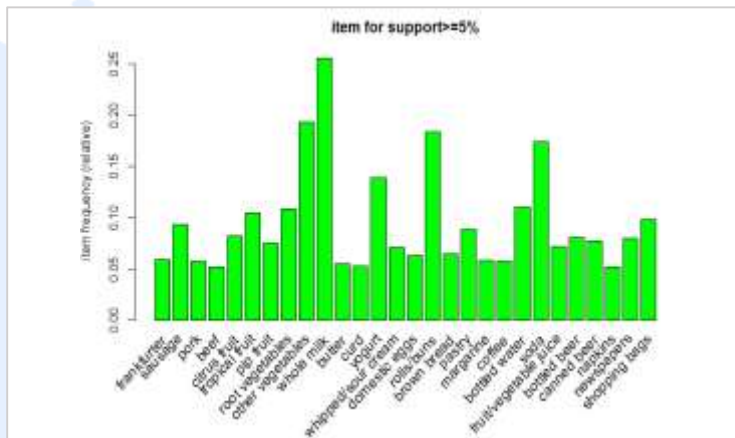
가장 많이
거래된 항목

● 연관규칙 – visualization(지지도)

☑ 그래프로 표현한 연관규칙(지지도)

지지도 5%이상의 item 막대 그래프

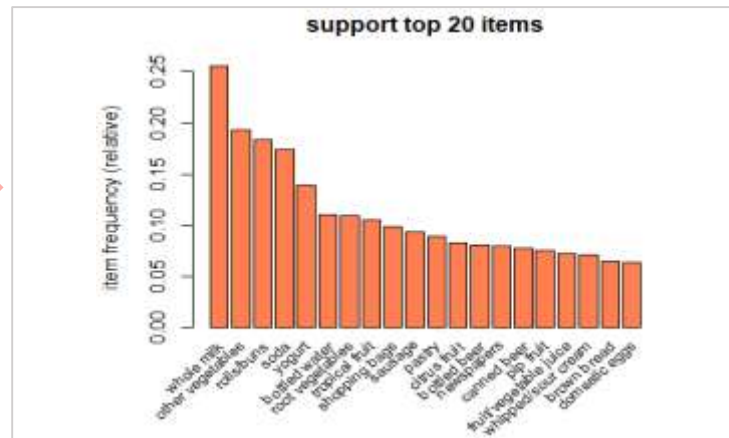
```
itemFrequencyPlot(Groceries,support
=0.05,main="items for support>= 5%",
col="green")
```



상위에서
하위 정렬

지지도 상위 20개 막대 그래프

```
itemFrequencyPlot(Groceries,topN=20
,main="support top 20 items",
col="coral")
```



● 연관규칙 분석결과 – Groceries 데이터

☑ 연관규칙 분석

```
# Association rule with support>5%, confidence>20% in minimum  
Grocery_rule<-apriori(data=Groceries,  
  parameter = list(support=0.05,  
                    confidence = 0.20,  
                    minlen = 2))  
Grocery_rule
```

✧ support, confidence와 length는 minimum 값
으로 너무 높게 잡으면 연관규칙 도출이 어려움

```
> Grocery_rule<-apriori(data=Groceries,  
+                         parameter = list(support=0.05,  
+                                           confidence = 0.20,  
+                                           minlen = 2))  
Apriori  
  
Parameter specification:  
confidence minval smax arem avar originalSupport maxtime  
0.2 0.1 1 none FALSE TRUE 5  
support minlen maxlen target ext  
0.05 2 10 rules TRUE  
  
Algorithmic control:  
filter tree heap memopt load sort verbose  
0.1 TRUE TRUE FALSE TRUE 2 TRUE  
  
Absolute minimum support count: 491  
  
set item appearances ... [0 item(s)] done [0.00s].  
set transactions ... [169 item(s), 9835 transaction(s)] done  
[0.00s].  
sorting and recoding items ... [28 item(s)] done [0.00s].  
creating transaction tree ... done [0.00s].  
checking subsets of size 1 2 done [0.00s].  
writing ... [6 rule(s)] done [0.00s].  
creating S4 object ... done [0.00s].  
> Grocery_rule  
set of 6 rules
```


● 연관규칙 분석결과 – Groceries 데이터

☑ 연관규칙 조회 및 평가

```
#analyzing result
summary(Grocery_rule)
inspect(Grocery_rule)
```

```
> inspect(Grocery_rule)
```

	lhs	rhs	support	confidence	coverage	lift	count
[1]	{yogurt}	=> {whole milk}	0.05602440	0.4016035	0.1395018	1.571735	551
[2]	{whole milk}	=> {yogurt}	0.05602440	0.2192598	0.2555160	1.571735	551
[3]	{rolls/buns}	=> {whole milk}	0.05663447	0.3079049	0.1839349	1.205032	557
[4]	{whole milk}	=> {rolls/buns}	0.05663447	0.2216474	0.2555160	1.205032	557
[5]	{other vegetables}	=> {whole milk}	0.07483477	0.3867578	0.1934926	1.513634	736
[6]	{whole milk}	=> {other vegetables}	0.07483477	0.2928770	0.2555160	1.513634	736

```
> summary(Grocery_rule)
set of 6 rules

rule length distribution (lhs + rhs): sizes
  2  6
  2  2  2  2  2  2

summary of quality measures:
  support confidence coverage lift count
Min. :0.05602 Min. :0.2193 Min. :0.1395 Min. :1.205 Min. :551
1st Qu.:0.05618 1st Qu.:0.2395 1st Qu.:0.1863 1st Qu.:1.282 1st Qu.:557
Median :0.05663 Median :0.3004 Median :0.2245 Median :1.514 Median :557
Mean :0.06250 Mean :0.3050 Mean :0.2139 Mean :1.430 Mean :614
3rd Qu.:0.07028 3rd Qu.:0.3670 3rd Qu.:0.2555 3rd Qu.:1.557 3rd Qu.:691
Max. :0.07483 Max. :0.4016 Max. :0.2555 Max. :1.572 Max. :736

mining info:
  data ntransactions support confidence
Groceries 9835 0.05 0.2
```

- ▶ 6개의 rule이 item 2개로 구성되어 있음
- ▶ 향상도 최소값이 1보다 큰 것을 알 수 있음
- ▶ 요쿠르트와 우유를 동시에 구매할 확률(지지도:5.6%), 요쿠르트를 구매한 조건에서 우유도 구매할 확률(신뢰도 40%)

● 연관규칙 분석결과 – Groceries 데이터

☑ 연관규칙-향상도(Lift)순서로 정렬

```
# sorting by Lift  
inspect(sort(Grocery_rule,by="lift"))  
# inspect(sort(Grocery_rule, by="support"))
```



```
> inspect(sort(Grocery_rule,by="lift"))
```

	lhs	rhs	support	confidence	coverage	lift
[1]	{yogurt}	=> {whole milk}	0.05602440	0.4016035	0.1395018	1.571735
[2]	{whole milk}	=> {yogurt}	0.05602440	0.2192598	0.2555160	1.571735
[3]	{other vegetables}	=> {whole milk}	0.07483477	0.3867578	0.1934926	1.513634
[4]	{whole milk}	=> {other vegetables}	0.07483477	0.2928770	0.2555160	1.513634
[5]	{rolls/buns}	=> {whole milk}	0.05663447	0.3079049	0.1839349	1.205032
[6]	{whole milk}	=> {rolls/buns}	0.05663447	0.2216474	0.2555160	1.205032

➤ sort() 함수를 통해 분석가가 보고자 하는 기준으로 정렬하는 것도 가능


● 연관규칙 분석결과 – Groceries 데이터

☑ 연관규칙-품목별 연관성 탐색

- sort() 함수를 통해 분석가가 보고자 하는 기준으로 정렬
- subset() 함수를 통해 원하는 item이 포함된 연관규칙만 추출
- %in%, %pin% 을 이용해 다양한 조건의 규칙 도출

yogurt가 들어있는 연관규칙

```
rule_interest3<-subset(Grocery_rule, items %in% c("yogurt"))  
inspect(rule_interest3)
```



```
> inspect(rule_interest3)
```

	lhs	rhs	support	confidence	coverage	lift	count
[1]	{yogurt}	=> {whole milk}	0.0560244	0.4016035	0.1395018	1.571735	551
[2]	{whole milk}	=> {yogurt}	0.0560244	0.2192598	0.2555160	1.571735	551

● 연관규칙 분석결과 – Groceries 데이터

☑ 연관규칙-품목별 연관성 탐색

➤ (other)라는 품목이 들어있고 & 신뢰도>25% 규칙

```
rule_interest5<-subset(Grocery_rule, items %pin% c("other") & confidence>0.25)  
inspect(rule_interest5)
```



```
> rule_interest5<-subset(Grocery_rule, items %pin% c("other") & confidence>0.25)  
> inspect(rule_interest5)
```

	lhs	rhs	support	confidence	coverage	lift	count
[1]	{other vegetables}	=> {whole milk}	0.07483477	0.3867578	0.1934926	1.513634	736
[2]	{whole milk}	=> {other vegetables}	0.07483477	0.2928770	0.2555160	1.513634	736

● 연관규칙 분석결과 저장

☑ 연관규칙결과를 data.frame으로 저장

```
# save as dataframe  
Grocery_rule_df<-as(Grocery_rule,"data.frame")  
Grocery_rule_df
```



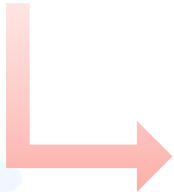
```
> Grocery_rule_df
```

	rules	support	confidence	coverage	lift	count
1	{yogurt} => {whole milk}	0.05602440	0.4016035	0.1395018	1.571735	551
2	{whole milk} => {yogurt}	0.05602440	0.2192598	0.2555160	1.571735	551
3	{rolls/buns} => {whole milk}	0.05663447	0.3079049	0.1839349	1.205032	557
4	{whole milk} => {rolls/buns}	0.05663447	0.2216474	0.2555160	1.205032	557
5	{other vegetables} => {whole milk}	0.07483477	0.3867578	0.1934926	1.513634	736
6	{whole milk} => {other vegetables}	0.07483477	0.2928770	0.2555160	1.513634	736

● 연관규칙 분석결과 저장

☑ 연관규칙결과 저장

```
#saving results as csv file  
write(Grocery_rule, file="Grocery_rule.csv",  
      sep=";",  
      quote=TRUE,  
      row.names=FALSE)|
```



	A	B	C	D	E	F
1	rules	support	confidence	coverage	lift	count
2	{yogurt} => {whole milk}	0.056024	0.401603	0.139502	1.571735	551
3	{whole milk} => {yogurt}	0.056024	0.21926	0.255516	1.571735	551
4	{rolls/buns} => {whole milk}	0.056634	0.307905	0.183935	1.205032	557
5	{whole milk} => {rolls/buns}	0.056634	0.221647	0.255516	1.205032	557
6	{other vegetables} => {whole milk}	0.074835	0.386758	0.193493	1.513634	736
7	{whole milk} => {other vegetables}	0.074835	0.292877	0.255516	1.513634	736