Exercise 2 - Association Rule Mining

Rommel Bartolome

March 13, 2019

Data Loading and Preliminaries

```
library(arules)
library(arulesViz)
library(tidyverse)
load("marketing_sparse.Rdata")
m <- marketing %>% apriori(parameter =
                             list(support =0.07, confidence = 0.75, minlen = 2))
```

```
## Apriori
##
## Parameter specification:
##  confidence minval smax arem  aval originalSupport maxtime support minlen
##        0.75    0.1    1 none FALSE            TRUE       5    0.07      2
##  maxlen target   ext
##      10  rules FALSE
##
## Algorithmic control:
##  filter tree heap memopt load sort verbose
##     0.1 TRUE TRUE  FALSE TRUE    2    TRUE
##
## Absolute minimum support count: 481
##
## set item appearances ...[0 item(s)] done [0.00s].
## set transactions ...[84 item(s), 6876 transaction(s)] done [0.00s].
## sorting and recoding items ... [58 item(s)] done [0.00s].
## creating transaction tree ... done [0.00s].
## checking subsets of size 1 2 3 4 5 6 7 done [0.03s].
## writing ... [3800 rule(s)] done [0.00s].
## creating S4 object  ... done [0.00s].
```

1) What are the top 5 association rules in terms of lift? In terms of confidence?

In terms of lift, we check the top 10 first:

```
inspect(sort(m, by = 'lift')[1:10])
```

```
##      lhs                                   rhs         support
## [1]  {age.1}                            => {educ.2}    0.07111693
## [2]  {marital.5,age.1,hh_stat.3}        => {income.1}  0.07737056
## [3]  {marital.5,age.1,dual.1,hh_stat.3} => {income.1}  0.07737056
## [4]  {age.1,dual.1,hh_stat.3}           => {income.1}  0.07780686
## [5]  {age.1,hh_stat.3}                  => {income.1}  0.07795230
## [6]  {marital.5,age.1}                  => {income.1}  0.07897033
## [7]  {marital.5,age.1,dual.1}           => {income.1}  0.07897033
## [8]  {income.1,marital.5,age.1}         => {hh_stat.3} 0.07737056
## [9]  {income.1,marital.5,age.1,dual.1}  => {hh_stat.3} 0.07737056
## [10] {income.1,age.1,dual.1}            => {hh_stat.3} 0.07780686
##      confidence lift     count
## [1]  0.7557960  6.603371 489
## [2]  0.9001692  4.931923 532
## [3]  0.9001692  4.931923 532
## [4]  0.8991597  4.926392 535
## [5]  0.8918469  4.886326 536
## [6]  0.8729904  4.783013 543
## [7]  0.8729904  4.783013 543
## [8]  0.9797422  4.777807 532
## [9]  0.9797422  4.777807 532
## [10] 0.9762774  4.760910 535
```

Checking the data, it appears that dual.1 and marital.5 are the same. If you answer dual.1 as YES, you should answer marital.5 as YES too, and vice versa. This is also the case of the difference in the LHS and RHS. Thus, we remove identical rules. The top 5 association rules in terms of lift are:

1. Being aged 14-17 AND being in Grades 9-11
2. Being single, aged 14-17 and living with parents/family AND having a personal income of less than $10,000
3. Being aged 14-17 and living with parents/family AND having a personal income of less than $10,000
4. Being single and aged 14-17 AND having a personal income of less than $10,000
5. Having a personal income of less than $10,000, being single and aged 14-17, AND living with parents/family

In terms of confidence, we do the same. However, since we are talking about confidence we should be wary of the direction of the LHS and RHS:

```
inspect(sort(m, by = 'confidence')[1:5])
```

```
##     lhs                        rhs      support    confidence lift
## [1] {marital.3}             => {dual.1} 0.09787667 1          1.671366
## [2] {marital.5}             => {dual.1} 0.40910413 1          1.671366
## [3] {marital.5,age.1}       => {dual.1} 0.09045957 1          1.671366
## [4] {marital.3,num_child.0} => {dual.1} 0.07184410 1          1.671366
## [5] {marital.3,lang.1}      => {dual.1} 0.09235020 1          1.671366
##     count
## [1]  673
## [2] 2813
## [3]  622
## [4]  494
## [5]  635
```

One may also do the analysis made above, removing the possible duplicates.

2) What makes houseowners different from renters?

We first check the difference of homeowners from renters based on lift:

```
homeowners_bylift <- m %>% subset(rhs %in% "hh_stat.1") %>%
  sort(by="lift", decreasing = F)
renters_bylift <- m %>% subset(rhs %in% "hh_stat.2") %>%
  sort(by="lift", decreasing = F)
inspect(homeowners_bylift[1:5])
```

```
##     lhs            rhs            support confidence     lift count
## [1] {sex.1,
##      marital.1,
##      yrsbay.5}  => {hh_stat.1} 0.07969750  0.7537827 2.005809   548
## [2] {marital.1,
##      yrsbay.5,
##      dual.2,
##      ethnic.7,
##      lang.1}    => {hh_stat.1} 0.08086097  0.7554348 2.010205   556
## [3] {marital.1,
##      dual.3}    => {hh_stat.1} 0.10587551  0.7559709 2.011632   728
## [4] {sex.1,
##      marital.1,
##      yrsbay.5,
##      lang.1}    => {hh_stat.1} 0.07431646  0.7570370 2.014469   511
## [5] {marital.1,
##      age.4,
##      lang.1}    => {hh_stat.1} 0.07388016  0.7570790 2.014580   508
```

```
inspect(renters_bylift[1:5])
```

```
##     lhs                              rhs         support    confidence
## [1] {marital.5,age.3}             => {hh_stat.2} 0.07300756 0.7606061
## [2] {marital.5,age.3,dual.1}      => {hh_stat.2} 0.07300756 0.7606061
## [3] {age.3,dual.1}                => {hh_stat.2} 0.11125654 0.7611940
## [4] {age.3,dual.1,lang.1}         => {hh_stat.2} 0.10413031 0.7665953
## [5] {age.3,dual.1,num_child.0} => {hh_stat.2} 0.09482257 0.7752675
##     lift     count
## [1] 1.814687 502
## [2] 1.814687 502
## [3] 1.816090 765
## [4] 1.828976 716
## [5] 1.849667 652
```

Here, we see that if you are a single male, living in the bay area for at least 5 years, you are likely a homeowner. If you are single and aged 25-34, you are likely a renter.

3) Provide 3 association rules that you deem to be actionable. Briefly explain the insights that you have obtained from them.

1. We check income:

```
income <- m %>% subset(rhs %pin% "income") %>%
  sort(by="confidence", decreasing = F)
inspect(income[1:1])
```

```
##     lhs        rhs        support    confidence lift    count
## [1] {age.1} => {income.1} 0.08042467 0.8547141  4.68288 553
```

Here, we see than if you are young (aged 14-17), you are likely to have an income of less than $10,000.

2. We check sex:

```
sex <- m %>% subset(rhs %pin% "sex") %>% sort(by="confidence", decreasing = F)
inspect(sex[1:1])
```

```
##     lhs           rhs     support    confidence lift     count
## [1] {occup.5} => {sex.2} 0.07038976 0.9603175  1.733563 484
```

Here we see that if you are homemaker, you are likely a female.

3. We check occupation:

```r
occup <- m %>% subset(rhs %pin% "occup") %>% sort(by="confidence", decreasing = F)
inspect(occup[1:1])
```

```
##     lhs                                      rhs       support
## [1] {income.1,marital.5,hh_stat.3,lang.1} => {occup.6} 0.0718441
##     confidence lift     count
## [1] 0.7670807  4.675928 494
```

Here we see that if you are aged 14-17, single, living with your parents/family and speaks english, you are likely a student (HS/College).