**In-Class Activity #13.2: Association Rule Mining Using R**

(Due Friday, 4/28/2017, 9:00 am)

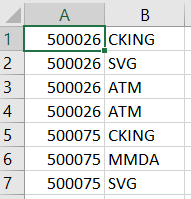
What to submit: a single word/pdf file with answers for the questions in Part 5.

You’ll need two files to do this exercise: **aRules.r** (the R script file) and **Bank.csv** (the data file[[1]](#footnote-1)), both available on the course site. The data file contains 32,366 rows of bank customer data covering 7,991 customers and the financial services they use.

Download both files and save them to the folder where you keep your R files. Also make sure you are connected to the Internet when you do this exercise!

**Part 1: Look at the Data File**

1. Open the Bank.csv data file. If it warns you, that’s ok. Just click “Yes” or “OK.”
2. You’ll see something like this:



Look at the contents of the file. Each row represents a customer/product pair. The first value is the customer number (ID) and the second value is the product/service name. We can see from the excerpt that customer 500026 uses CHCKG, SVG, and ATM services (ATM may be in there twice because they have two ATM cards, but that won’t matter for our analysis).

We will use this data set to predict which bank services customers are likely to own, given the other services they own.

The input file for an Association Rule analysis follows this general format. The first value is the “basket” and the second value is the “event.” If multiple events occur within a basket, then the basket number is repeated for each event. Both the basket and event are discrete values representing a particular basket and a particular event.

For the bank data set, here is the complete list of products and services they offer:

|  |  |  |  |
| --- | --- | --- | --- |
| **ITEM** | **Description** | **ITEM** | **Description** |
| ATM | ATM card | IRA | Individual retirement account |
| AUTO | Auto loan | MMDA | Money market deposit account |
| CCRD | Credit card | MTG | Mortgage |
| CD | Certificate of deposit | PLOAN | Personal loan |
| CKCRD | Check card | SVG | Savings account |
| CKING | Checking account | TRUST | Trust account |
| HMEQLC | Home equity loan |  |  |

1. Close the Bank.csv file. If it asks you to save the file, choose “Don’t Save”.

**Part 2: Explore the aRules.r Script**

1. Open the aRules.r file. This contains the R script that performs the association mining analysis.
2. Look at lines 8 through 20. These contain the parameters for the decision tree model. Here’s a rundown:

|  |  |  |
| --- | --- | --- |
| INPUT\_FILENAME | Bank.csv | The data is contained in Bank.csv |
| OUTPUT\_FILENAME | ARulesOutput.csv | The association rules and their statistics (lift, confidence, and support) are output to ARulesOutput.csv. |
| SUPPORT\_THRESH | 0.01 | If the support for a rule is below 0.01, it won’t appear in the final list. |
| CONF\_THRESH | 0.01 | If the confidence for a rule is below 0.01, it won’t appear in the final list. |

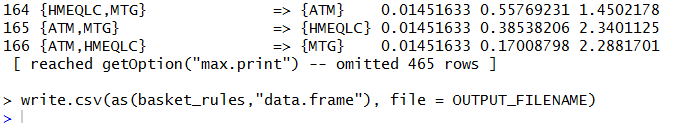
1. Look at lines 27 and 28. These install (when needed) the arules package, which is needed for computing the association rules.
2. Now let’s look at the the apriori() function that computes the association rules. Scroll down to line 50:



You can see a few things at work:

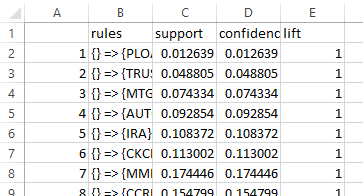
* The apriori() function is used to compute the association rules (and the results are stored in basket\_rules).
* txn is the transaction data read from the data file.
* Our SUPPORT\_THRESH and CONF\_THRESH parameters from above are used here.

**Part 3: Execute the aRules.r Script**

1. Select Session/Set Working Directory/To Source File Location to change the working directory to the location of your R script.
2. Select Code/Run Region/Run All. It could take a few seconds to run since the first time it has to install some extra modules to do the analysis. Be patient!
3. You’ll see a lot of action in the Console window at the bottom left side of the screen, ending with this:   
     
   

**Part 4: Importing the generated rules into Microsoft Excel for analysis**

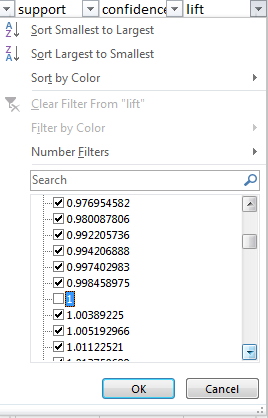
Working with the association rules in RStudio can be cumbersome and aren’t a very good way to look at all the rules at once. Fortunately, our aRules.r script generates an output file listing every rule. The only restriction is that it won’t include rules with support and confidence levels below the thresholds we set using SUPPORT\_THRESH and CONF\_THRESH.

1. Close the aRules.r script. If it asks you to save the file, click “Save.”
2. Open ARulesOutput.csv in Microsoft Excel by double clicking on the file icon. You’ll see your rules in a spreadsheet:  
     
   

The Column A lists the rule numbers. Column B presents the association rules. Columns C through E contain the support, confidence and lift values of each association rule.

1. Choose File/Save As… and save the file to your working directory as BankRules.xlsx (you can change the file format under “Save as type:”

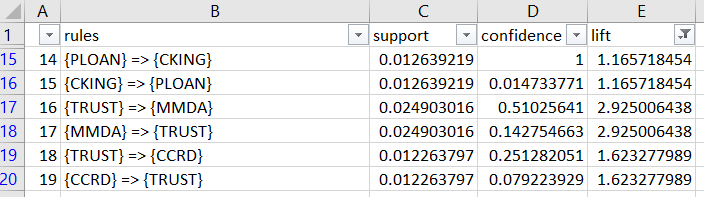
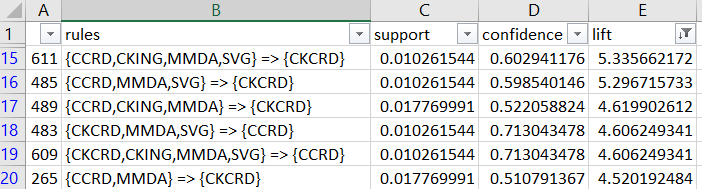
So now let’s work with the rules in Excel to make them easier to interpret.

1. First, we want to eliminate all rules with a lift of exactly 1. Those are also our rules with no antecedent (left hand side), like the ones listed above.

Select columns A through E, go to the Data tab and select Filter.

Select the down arrow next to the **lift** column header and scroll down until you see the value “1”. Uncheck the box.

Then click OK.

1. You’ve now removed the first 13 rules:  
     
   
2. Since we really care about the highest lift rules, click again on the icon next to the lift column header and select “Sort Largest to Smallest.”
3. You’ll now see the list sorted by lift values:  
     
   
4. Save your workbook.
5. Now that’s look at rule 611 (in row 15):

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | rules | support | confidence | lift |
| 611 | {CCRD,CKING,MMDA,SVG} => {CKCRD} | 0.010262 | 0.602941 | 5.335662 |

We see that this is the rule with the highest lift is the first one {CCRD, CKING, MMDA, SVG} => {CKCRD}. It doesn’t occur that often (support = 0.01), but the relationship is fairly strong (confidence = 0.603) and the rule has high predictive power (lift = 5.336>1). The high lift indicates that if a person has a credit card, checking account, money market account, and a savings account makes it much more likely than chance (lift>1) that they will also get a check card.

1. Now scroll down to look at rule 72 (in the row 547):

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | rules | support | confidence | lift |
| 72 | {IRA} => {CKING} | 0.088224 | 0.814088 | 0.948997 |

This indicates the rule is about whether people who have an IRA also have a checking account. We see that it doesn’t happen that often (support = 0.088), when it happens the relationship is fairly strong (confidence = 0.814), and that buying an IRA is **not** very predictive of whether someone will have a checking account (lift = 0.949≈1, about the same as random chance).

**Part 5: Try it yourself**

Looking at your Excel worksheet with the imported rules and answer the following questions:  
  
a) How many rules are there with a lift value between 4 and 5?  
  
b) What service are customers with an automobile loan (AUTO) and a checking account (CKING) most likely to also have (or be interested in having)? Explain your answer.  
 **(HINT: Sort the rules in alphabetical order to make those rules easier to find.)**

c) Find rule 138: {SVG,TRUST} => {CD}). Explain in business terms what it means and what you, as a bank manager, should do with that information.

1. Adapted from SAS Enterprise Miner sample data set. [↑](#footnote-ref-1)