James Machado, Jonny Shakerchi, and Justin Kaplan

BUS256 Marketing Analytics

Spring 2017

Case: Market Basket (Affinity) Analysis

Introduction: We have now read Miller's Chapter 9 as well as two articles applying Association Rules to "market basket" studies. In these studies, the common question is "what goes with what?" The goal is typically to discern probabilistic rules that can guide marketers to understand purchasing patterns.

The article by Brijs presents a specific algorithm that he entitles the PROFSET model with the goal of selecting "the most profitable set of products in terms of cross-selling opportunities between the delegates of each category." Instead of PROFSET we will use the modeling approach that Miller teaches in Chapter 9, but modify the objectives of Miller's analysis to instead emulate Brijs' approach. Note that our dataset does not include price or profit data, so we cannot directly estimate monetary value. We will instead think in terms of likelihoods that customers who purchase itemset "A" might be encouraged to add item (or itemset) "B" to their purchase.

Objectives:

- Gain insight into Association Rules logic and methods by modifying existing code for a new purpose
- Gain experience using the GitHub environment for version control
- Learn to *interpret* the results of an affinity analysis for business value

Summary: Miller's Chapter 9 uses a dataset ("Groceries") that comes with the arules package. In his example, he focuses on finding rules related to purchases of **vegetables**. As suggested at the end of his code (page 155), we'll shift focus to a different product category, specifically **dairy produce**. Moreover, as you work track the modifications in your R script using Git.

NOTE: Because Groceries is 'built-in' with arules, you won't need to save or read in the dataset.

Tasks and Questions:

1. **PRE-WORK**: Create a new project enabled for version control
    a. On GitHub, create a new repository for this assignment; initialize it with a README that indicates the purpose of the assignment.
    b. Copy the URL for your repository and paste it into your paper. I'll want to visit it later ☺

    https://github.com/jamesjmachado/BrijMBProject.git

    c. In Rstudio, create a new Project that links to the GitHub repo.
    d. Open Miller 09.R within your new project

The next few steps ask you to explore the data by making selected modifications to code before you finally run the entire script. Following each step, Commit and Push your modified script to

GitHub.

2. (10 pts) Modify Miller's script to focus to identify which items are categorized as "dairy produce". Report the specific items that make up the category.

	labels	level2	level1
25	whole milk	dairy produce	fresh products
26	butter	dairy produce	fresh products
27	curd	dairy produce	fresh products
28	dessert	dairy produce	fresh products
29	buttermilk	dairy produce	fresh products
30	yogurt	dairy produce	fresh products
31	whipped/sour cream	dairy produce	fresh products
32	beverages	dairy produce	fresh products

3. (10) Miller decides to use a support cutoff of 0.025, which yields 344 rules. Modify the code to create approximately 200 rules. Report on the support cutoff that accomplishes this and also the number of resulting rules.

```
second.rules <- apriori(groceries,
                        parameter = list(support = 0.03435, confidence = 0.05))
print(summary(second.rules))  # yields 201 rules
rule length distribution (lhs + rhs):sizes
  1   2   3   4
 21 106  66   8
```

4. (10) Working from the same cutoff as you found in step 2, suppose we tighten the confidence level to 0.01 rather than 0.05. Describe the impact of these changes on the set of rules.

```
second.rules <- apriori(groceries,
                        parameter = list(support = 0.03435, confidence = 0.01))
print(summary(second.rules))  # yields 206 rules
rule length distribution (lhs + rhs):sizes
```

1 2 3 4

26 106 66 8

5. (25) Now, using the cutoff from step 2 and a confidence level of 0.05, modify the script one last time to generate Miller's analysis, but for dairy produce rather than vegetables. This should create 5 pdfs (rename the last one from "farmer rules" to "dairy rules").

- Complete: in R script and located in attached .pdf files.

6. (25) In your document, write a short commentary and explanation for each of the five graphs. Write as if you are presenting this to a merchandising team in charge of store layout as well as promotions for various food items.

The "Graph for 10 Rules" shows a map of the top 10 dairy rules we created. As you can see, "dairy produce" is in the middle of the graph with 10 arrows pointing at it. The arrows come from the top 10 basket sets that most probably lead a person to buying a dairy product. The basket sets are represented by the circles in the graph. The bigger the circle, the stronger the support level is in that basket, ranging from 3.4% to 7.9%. The color of the circles indicates the lift level; the darker the circle, the higher the lift. The circles are made up of the products that belong in that basket. The biggest basket with the highest lift is made up with fruit and vegetables, with the basket containing fruit and bread and baked goods coming next.

The next two graphs (Market Basket Initial Support, and Market Basket Final Support) show which items have support greater than 3.345%. The difference between these two graphs is how products are categorized. The "initial support" chart shows individual products, while the "final support" chart shows products by level 2 groupings. This tells us what the most common items are in market baskets.

The "Grouped Matrix for 201 Rules" shows the relationship between the left-handed side and right-handed side in predicting which items cause people to buy another certain product. To simplify, if i buy product A (the left-hand side,) what is the probability I will buy product B (the right-hand side). The size of the circles shows how supportive the products are, like in the "Graph for 10 Rules," the bigger the circle, the more likely these items are bought together. Also, the darker the circle, the higher the lift is, meaning how effective the support and confidence levels we set are.

Finally, the "Scatter Plot for 201 Rules" shows the relationship between the confidence and support levels we set for our rules. The color of the points shows how strong the lift level is, an indicator of how effective the relationship between the confidence and support levels are. Apparently, the lower the support level and the higher the confidence level, the better lift we get on predicting which baskets are sold together. This is intuitive, because we know if we want effective rules, we need a lower support level and a higher confidence level.

7. (10) Make specific recommendations aimed at cross-selling to customers who purchase dairy produce – what other items might we encourage them to buy, and how might we accomplish that based on the intelligence revealed in your analysis.

Based on the data we obtained from question 2 in which we selected rules with dairy products in consequent (right-hand-side), item subsets, I would recommend cross-selling the following items to customers who purchase dairy products:

	lhs		rhs	support	confidence	lift
[1]	{bread and backed goods,fruit,sausage}	=>	{dairy produce}	0.03060498	0.7984085	1.802237
[2]	{bread and backed goods,fruit,vegetables}	=>	{dairy produce}	0.04077275	0.7956349	1.795976
[3]	{cheese,fruit}	=>	{dairy produce}	0.03965430	0.7707510	1.739806
[4]	{cheese,vegetables}	=>	{dairy produce}	0.04219624	0.7628676	1.722011
[5]	{fruit,non-alc. drinks,vegetables}	=>	{dairy produce}	0.03304525	0.7575758	1.710066
[6]	{bread and backed goods,sausage,vegetables}	=>	{dairy produce}	0.03284189	0.7494200	1.691656
[7]	{vegetables,vinegar/oils}	=>	{dairy produce}	0.03141840	0.7481840	1.688866
[8]	{bread and backed goods,fruit,non-alc. drinks}	=>	{dairy produce}	0.03528216	0.7430407	1.677256
[9]	{frozen foods,fruit}	=>	{dairy produce}	0.03070666	0.7401961	1.670835
[10]	{fruit,vegetables}	=>	{dairy produce}	0.07869853	0.7350427	1.659203

All of these products have high amounts of lift values which illustrate that these items have a strong likelihood of being purchased with dairy products. To improve the cross-selling of these items, we recommend multiple approaches. One approach would to create a promotion leaflet that pairs items next to each other and offers promotions when pairing the items. Additionally, the grocery store could engage in inbound marketing tactics, such as featuring the recipes online that contain both products (i.e. similar to tablespoon.com). Another inbound tactic could be partnering with social media influencers (i.e. Delish or Spoon U) to have them post recipes using the products together. A third way to cross-sell would be to put the items in locations next to each other such as putting dairy products near cheese or eggs. Lastly, grocery stores could create rewards programs that allow them to track customer purchases. This would allow them to isolate which items individuals prefer to buy alongside dairy products. In turn, the grocery story could create individualized promotions or recipes geared toward cross-selling those specific items.

8. (10) (reflection—no wrong answers here!) In a corporate setting, version control is quite common and serves important functions. Comment on your impressions of this approach. Did it facilitate or impede your workflow? How might it be helpful going forward?

In terms of Version Control and PROFSET, it is mostly positive. To have commits so that you can keep track of what's done, what has been done, is incredibly helpful, especially with large data sets. For the most part, it facilitated our workflow. The only negative we really encountered was figuring out or losing track of our changes. When we would look or use another branch, it was hard to get back to our commit.

As for PROFSET itself, since we're looking at frequent sets and not individual products, plus the cross-selling opportunities, we could see this being very useful in finding out patterns, especially in the case of Basket Market where you could see the associations, that can help retailers and individual products try to match for example, put one product next to each other for convenience since they are shown matching together. The problem is because the PROFSET, especially in question 1, is random; it can lead to some false positives. For example, when we used Dairy Produce, there were some

categories that were games/books/hobbies. These are not Dairy Products. So, it seems we'll get a mostly accurate account. But not a comprehensive account and sometimes there can be too many categories, which again, based on Brijs article, gives us numbers that would underestimate the market. Meaning, if we have individual products, this would not be the best way to determine what they are good for. Also, as it's mentioned in the Brijs article, this method should only be used for 7 different products, not necessarily the 10 to 15 in the Market Basket situation.

Version Control and PROFSET can be useful in the corporate world to help with generalized results. However, if you want individual product results, than this is not necessarily the best way to go about it.