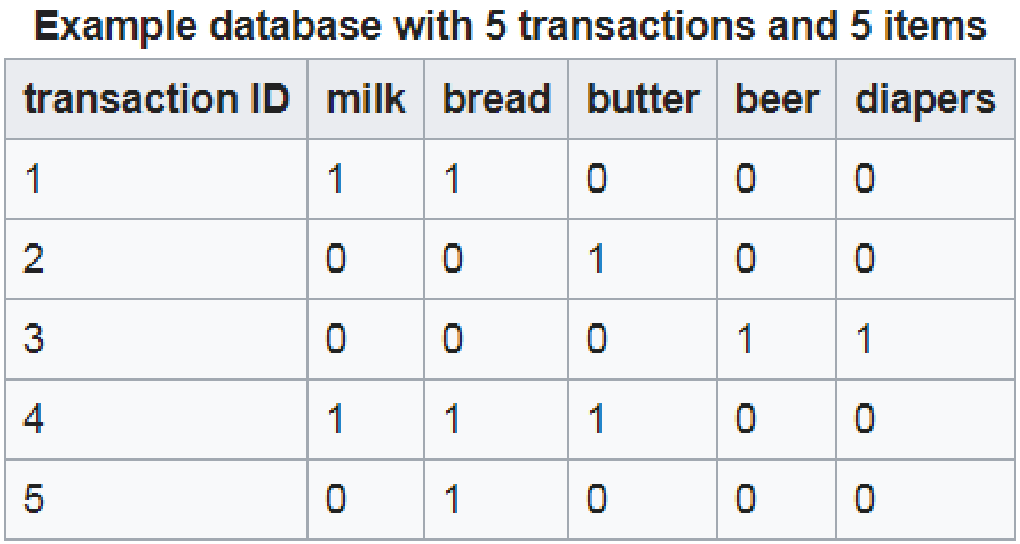
3 Key Concepts in Association Rules X => Y

- Define the following in terms of Probability:
  - Supp (X) ≡ P(contains X)
  - Supp (X and Y) [aka Rule Support] ≡ P(contains X and Y)
  - Conf (X => Y) ≡ P(contains Y|contains X)
  - Lift (X => Y) ≡ $\frac{P(contains\ Y|contains\ X)}{P(contains\ Y)}$

Using the keyword “contains” avoids the awkward definition of P(Y | X) = P(X intersect Y)/P(X). In probability, X and Y are events but in Association Rules, X and Y are itemsets that typically has no items in common and thus no intersection. X and Y in itemsets actually mean items in X together with items in Y, and hence is a union of two itemsets and not the intersection.

Source: Chew C.H. (2020) AI, Analytics and Data Science Vol. 2.

A Simple Numerical Example (Answers)



Example database with 5 transactions and 5 items

transaction ID	milk	bread	butter	beer	diapers
1	1	1	0	0	0
2	0	0	1	0	0
3	0	0	0	1	1
4	1	1	1	0	0
5	0	1	0	0	0

Rule: {milk, bread} -> Butter

- Supp ({milk}) = 2/5
- Supp ({milk, bread}) = 2/5
- Supp ({milk, bread, butter}) = 1/5 = 0.2
- Confidence of the rule = ½ = 0.5
- Lift of the rule = ½ / 2/5 = 1.25

Class Activity 1

Association Rules

Est. Duration: 45 mins

1. Read the Review Paper: Lucas Lau and Arun Tripathi (2001) Mine Your Business — A Novel Application of Association Rules for Insurance Claims Analytics. Casualty Actuarial Society E-Forum, Winter 2011.
2. Individually write down your answers to the questions in the next 2 slides.
3. Volunteer to answer to gain extra participation points.
   - Rubics: Critical Thinking, Oral Communication.

Notes:

- *All answers are either direct or can be inferred from the review paper and your previous study of statistics.*
- *There are multiple ways to express your answers in English. Keep it simple and straight to the point.*
- *Instructor's answers will be posted by end of the week into Main Site.*

Answers to Questions for Class Activity 1

1. Explain in your own words, the meaning of the concept "Confidence", and why is this measure useful?

Association Rule, X -> Y is actually a probabilistic concept as only a certain percentage of transactions will contain items in Y, if those transactions already contain items in X. Not every transaction has both X and Y. To precisely capture this probabilistic concept, we use the condition probability P(Y | X). This measures the probability that Y occurs, given that X already occurred, and can be interpreted as the confidence of Y occurring, in the presence of X. It is a measure of the "strength" or "predictability" of the rule.

2. Explain in your own words, the meaning of the concept "Support", and why is this measure useful?

Support is the frequency of occurrence of the items and is a measure of prevalence or popularity of the items. It is a measure of the "applicability" of the rule. It tells you how often you will be able to apply the association rule, whereas confidence tells you the strength of the rule in situations where it is applicable. Thus, they complements each other.

Answers to Questions for Class Activity 1

3. Explain in your own words, the meaning of the concept "Lift", and why is this measure useful? Is Lift still necessary if we have a rule that has high confidence and high support?

Association Rule: X -> Y

Lift measures how useful the rule is, in the context of the existing situation. Even if a rule has high confidence and high support, it may not be useful if P(Y) is already high. Example: Conf(X→Y) = 99% but P(Y) = 99.99%

i.e. The additional information about X has no/minimal impact on Y if P(Y) is already very high i.e. If Y is already popular, why do you need X to boost Y?

If Lift = 1, then X has no impact on Y. [X and Y are independent.]
If Lift > 1, then X can help to boost Y.
If Lift < 1, then X will reduce Y.

Answers to Questions for Class Activity 1

4. Many application of association rules require both Confidence and Support. (a) Explain in your own words how to use both of the measures at the same time to define good association rules. (b) Why can't we just choose and use only one of the two?

Support and Confidence (and Lift) reveals different information and thus, complements each other in understanding the situation. Support measures the applicability of the rule, whereas Confidence measures the strength of the rule in situations where it is applicable.

The real business question is how high is high enough for support and confidence. This depends on the business. If it is a high volume, low profit margin business (e.g. supermarket) then it needs rules with higher support levels, compared to low volume, high profit margin business (e.g. Diamond stores).

Answers to Questions for Class Activity 1

5. The Apriori algorithm is the standard method for eventually calculating Confidence and Support. (a) Explain in your own words, in a few sentences, how it works. (b) Is this sufficient to compute Confidence? Explain.

(a) It computes the support of all 1-itemsets that meets the minimum support criteria, then use it to compute the support of all 2-itemsets, and so on. By exploiting the fact that adding more items will either reduce the support or stay constant, but never increase support, it smartly reduces the amount of items to consider. The benefit is speed and scale, especially if there are many products to consider. It can assemble different product combinations (that are likely to result in good rules) and calculate their supports in order generate a lot of "good" rules, for a lot of products, very fast.

(b) It is sufficient as confidence can be expressed as the support of m items divided by the support of n items, and Apriori already has all these supports computed.

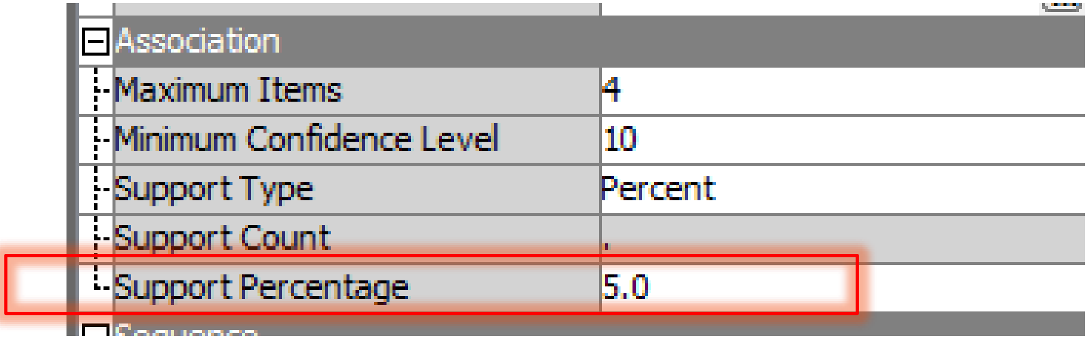
Class Activity 2

SAS EM Practice

Est. Duration: 30 mins

- Follow and practice SAS EM Data Mining Case Study textbook chapter 5. [3 pages only].

- Check the results and compare against the Association node property setting. Explain why the minimum support in the Rules table results is lower than 5% as specified in the Association node setting.



Association	
Maximum Items	4
Minimum Confidence Level	10
Support Type	Percent
Support Count	.
Support Percentage	5.0

Support in Rules Table refers to Rule Supp, not just Antecedent Supp.

- Try different settings for Minimum Confidence Level and Maximum Items and watch the results. See the impact of those settings.

Class Activity 3

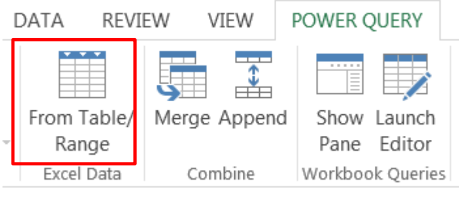
SAS EM Practice

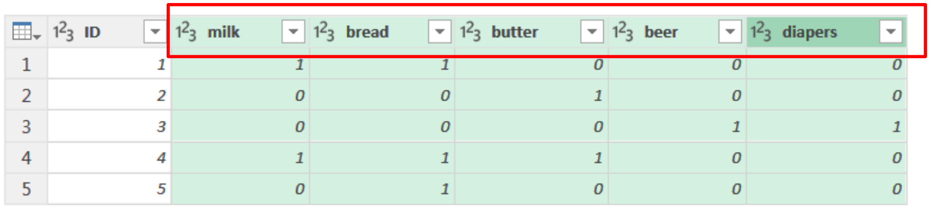
Est. Duration: 30 mins

- Using the Wikipedia 5 transactions example (see milk.csv), apply SAS EM to produce association rules.
- You will need to prepare the data in the correct format before importing into SAS EM.
  - Hint: The correct data format can be viewed in Class Activity 2
- Questions:
  1. How would you verify the results of the calculations done in previous slide against the software outputs.
  2. Verify if the Support(%) in SAS software output refers to Support(X) or Rule Support(X and Y)?
  3. Is Expected Confidence(%) referring to P(X), P(Y), or other probabilities? How would you verify from the SAS EM output?

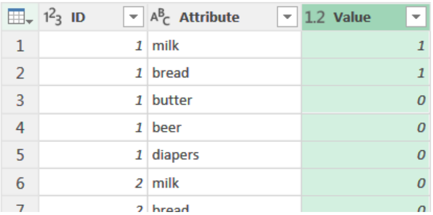
Ans to Class Activity 3

- 2 ways to prepare data in long format:
  - Excel 2016 onwards: Unpivot columns via Get & Transform > From Table > Power Query
  - Before Excel 2016:
    1. Download & install free Power Query from Microsoft
    2. Select Data in table > Power Query > From Table/Range
    3. Select all columns excluding ID columns > Right-click > Unpivot Columns
    4. Filter to remove value = 0.
    5. Click Close & Load

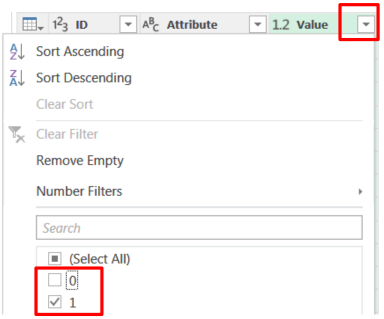


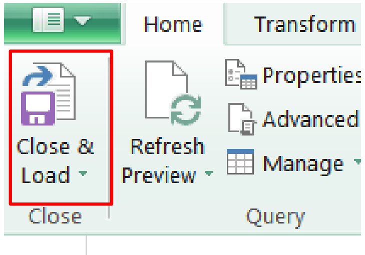


	ID	milk	bread	butter	beer	diapers
1	1	1	1	0	0	0
2	2	0	0	1	0	0
3	3	0	0	0	1	1
4	4	1	1	1	0	0
5	5	0	1	0	0	0



	ID	Attribute	Value
1	1	milk	1
2	1	bread	1
3	1	butter	0
4	1	beer	0
5	1	diapers	0
6	2	milk	0
7	2	bread	0

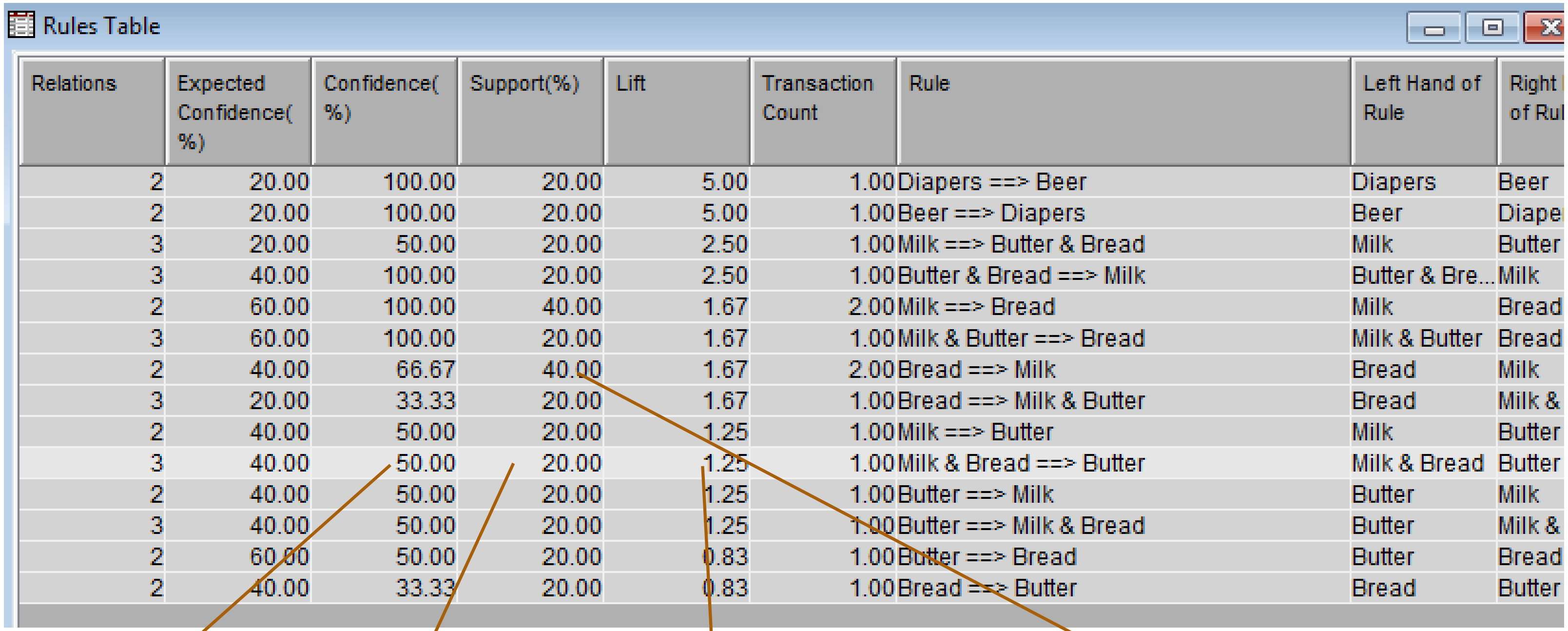




- R: melt() from reshape2 package.

Ans to Class Activity 3

For the rule: {milk, bread} -> Butter



Rules Table

Relations	Expected Confidence(%)	Confidence(%)	Support(%)	Lift	Transaction Count	Rule	Left Hand of Rule	Right of Rul
2	20.00	100.00	20.00	5.00	1.00	Diapers ==> Beer	Diapers	Beer
2	20.00	100.00	20.00	5.00	1.00	Beer ==> Diapers	Beer	Diape
3	20.00	50.00	20.00	2.50	1.00	Milk ==> Butter & Bread	Milk	Butter
3	40.00	100.00	20.00	2.50	1.00	Butter & Bread ==> Milk	Butter & Bre...	Milk
2	60.00	100.00	40.00	1.67	2.00	Milk ==> Bread	Milk	Bread
3	60.00	100.00	20.00	1.67	1.00	Milk & Butter ==> Bread	Milk & Butter	Bread
2	40.00	66.67	40.00	1.67	2.00	Bread ==> Milk	Bread	Milk
3	20.00	33.33	20.00	1.67	1.00	Bread ==> Milk & Butter	Bread	Milk &
2	40.00	50.00	20.00	1.25	1.00	Milk ==> Butter	Milk	Butter
3	40.00	50.00	20.00	1.25	1.00	Milk & Bread ==> Butter	Milk & Bread	Butter
2	40.00	50.00	20.00	1.25	1.00	Butter ==> Milk	Butter	Milk
3	40.00	50.00	20.00	1.25	1.00	Butter ==> Milk & Bread	Butter	Milk &
2	60.00	50.00	20.00	0.83	1.00	Butter ==> Bread	Butter	Bread
2	40.00	33.33	20.00	0.83	1.00	Bread ==> Butter	Bread	Butter

Conf = 0.5

Rule Supp = 0.2

Lift = 1.25

Supp({milk, bread}) = 0.4

Ans to Class Activity 3

EM Help > Explore > Association Node

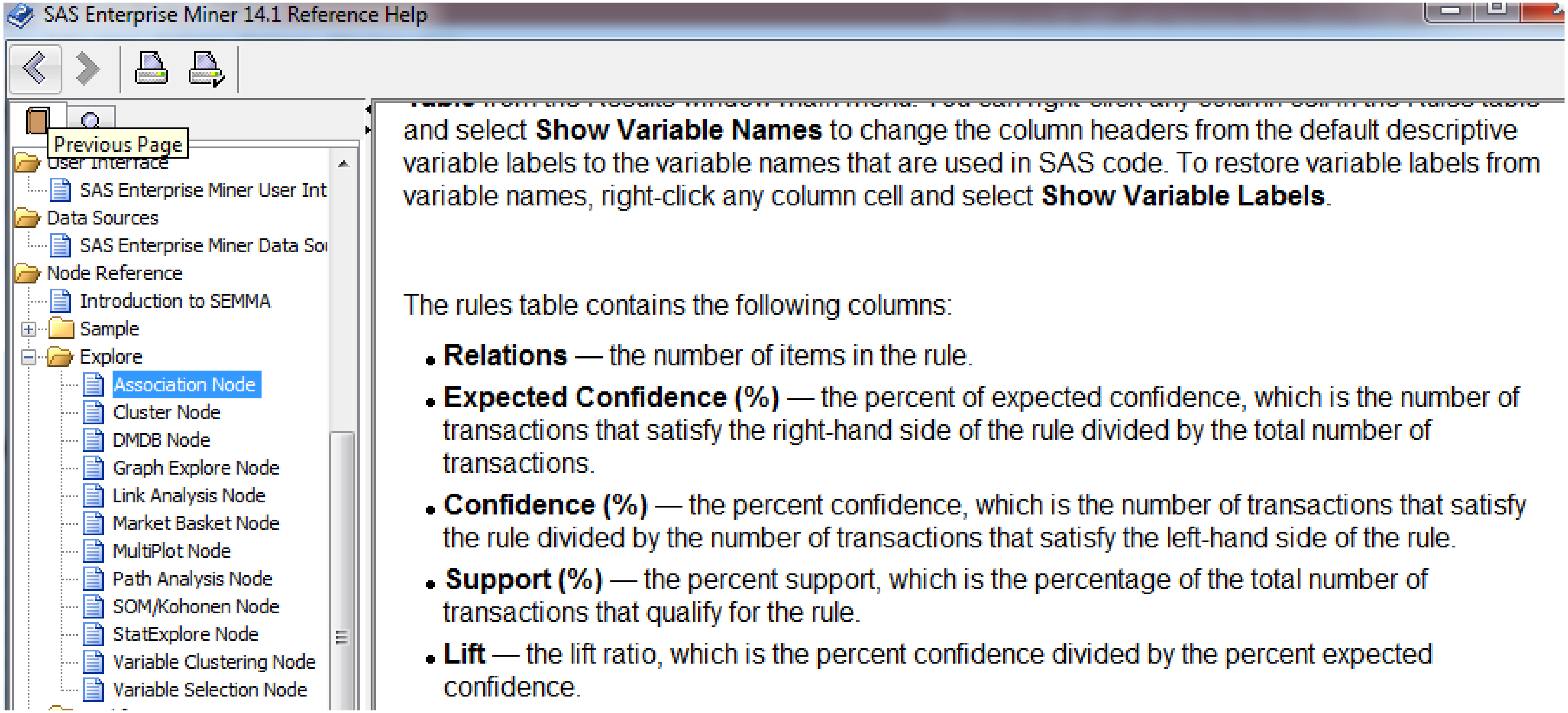


Table from the Results window main menu. You can right-click any column cell in the Rules table and select **Show Variable Names** to change the column headers from the default descriptive variable labels to the variable names that are used in SAS code. To restore variable labels from variable names, right-click any column cell and select **Show Variable Labels**.

The rules table contains the following columns:

- **Relations** — the number of items in the rule.
- **Expected Confidence (%)** — the percent of expected confidence, which is the number of transactions that satisfy the right-hand side of the rule divided by the total number of transactions.
- **Confidence (%)** — the percent confidence, which is the number of transactions that satisfy the rule divided by the number of transactions that satisfy the left-hand side of the rule.
- **Support (%)** — the percent support, which is the percentage of the total number of transactions that qualify for the rule.
- **Lift** — the lift ratio, which is the percent confidence divided by the percent expected confidence.

Expected Conf ≡ P(Y)

]

Reflection

- The Supp, Conf and Lift are easy to calculate.
- What do you think is the main problem/difficulty with doing Association Analysis?

The Support, Confidence and Lift are easy to compute **if the rule is given**. The problem is that we are not given all the “good” rules, and need to generate potentially good rules for a list of (many) items, fast. Imagine NTUC store with 10,000 unique items in the store.

Apriori algorithm is a smarter way than brute force to quickly generate “potentially good” rules, fast, even for many items. i.e. speed and scale. It avoids wasting time considering “useless” rules.

Association Rules using R

- R Packages:
  - arules
  - arulesViz
- Excellent Tutorial with dataset: https://towardsdatascience.com/association-rule-mining-in-r-ddf2d044ae50
- Explanation about the Transaction class [data format] required in arules
  - https://www.jdatalab.com/data_science_and_data_mining/2018/10/10/association-rule-transactions-class.html
- How to convert from 5 different data formats into transactions data format for arules to work
  - https://rdrr.io/cran/arules/man/transactions-class.html
- Two PDF vignettes in subfolder R.
- Please see demo Rscripts arules-milk.R and meltmilk.R newly uploaded in NTUlearn.

Association Rules using Python

- Two Alternative Python Packages:
  1. mlxtend
     - http://rasbt.github.io/mlxtend/user_guide/frequent_patterns/apriori/
  2. apyori
     - https://stackabuse.com/association-rule-mining-via-apriori-algorithm-in-python/

Comment: These packages are not as easy to use as SAS EM or R. Perhaps you can find an alternative easier-to-use Python package for Association Rules. BTW R packages arules & arulesViz are also available in Anaconda at https://anaconda.org/conda-forge/r-arules