NANYANG TECHNOLOGICAL UNIVERSITY
SINGAPORE

# Association Rules

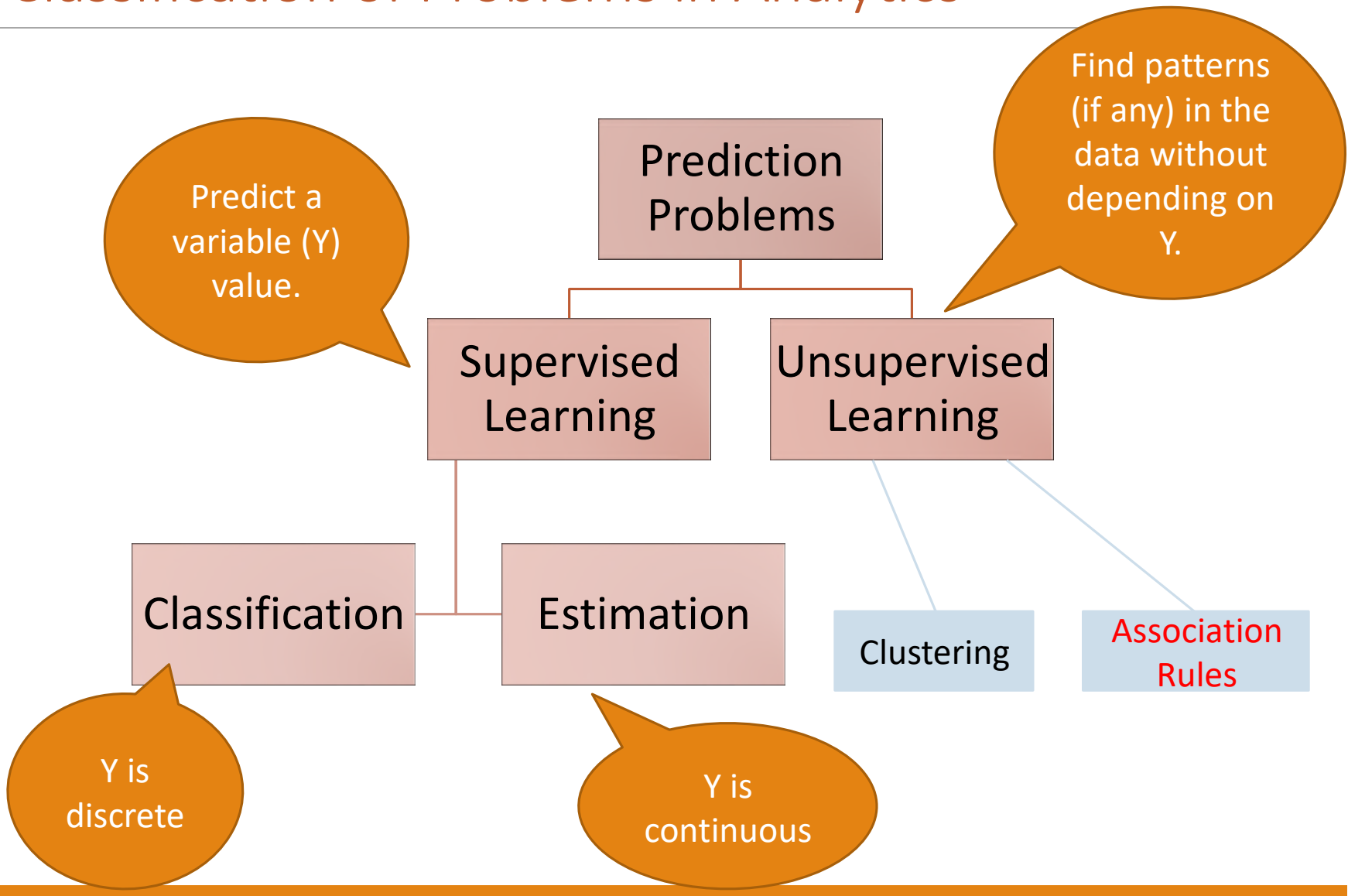
BC2407 ANALYTICS II SESSION 3

BASED ON CHEW C.H. TEXTBOOK AAD VOLUME 2.

# Applications of Association Rules

- Originally started to analyze retail transaction (i.e. market basket analysis) and to:
  - Recommend products to Customer.
  - Re-design store shelfs to optimize retail sales.
- Amazon.com
- Netflix
- Spotify
- Basic Analytics in most Recommendation System

# Classification of Problems in Analytics



# Data Format in Association Rules

- Transaction Record (aka receipt):
  - A record of all the purchases made in one transaction. E.g. Receipt from NTUC supermarket, Invoice for mobile phone contract, Invoice of software purchase, etc...
- Wide Data Format:
  - Each row is a transaction record.
  - Each column represents a unique item.
  - Each row then records a sequence of 0 or 1 (the item in that column was purchased).
- Long Data Format:
  - Each row contains only the transaction ID and one item.

# Example of Wide Data Format

**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 |

- Each row (transaction ID) shows the items purchased in that single transaction (aka receipt).
- If many possible items, dataset is sparse.
  - i.e. many zeros or NAs.

# Example of Long Data Format

| | A | B |
|---|---|---|
| 1 | ID | Item |
| 2 | 1 | Milk |
| 3 | 4 | Milk |
| 4 | 1 | Bread |
| 5 | 4 | Bread |
| 6 | 5 | Bread |
| 7 | 2 | Butter |
| 8 | 4 | Butter |
| 9 | 3 | Beer |
| 10 | 3 | Diapers |

What is the difference between wide format vs long format?

# Concepts in Association Rules

- Itemset
  - A set of items in a transaction. An itemset of size 3 means there are 3 items in the set.
  - In general, it can be of any sized, unless specified otherwise.
- Association Rule
  - Form: X → Y
  - X associated to Y.
  - X is the “antecedent” itemset, Y is the “consequent” itemset.
    - There might be more than one item in X or Y.

# 3 Key Concepts in Association Rules

- Many textbooks and websites, even Wikipedia, defines and explains the 3 key concepts (Support, Confidence, Lift) in terms of Set.
- The 3 key concepts can be easier to understand and appreciate if we express them in terms of Probability instead of Sets [see next slide].
- Example:
  - Set A = {Bread, Peanut Butter}
  - P(A) = 0.9
  - Which of the two above provide more useful info?

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

# 3 Key Concepts in Association Rules X → Y

- Define the following in terms of Probability:
  - Supp (X) ≡ P(contains X)

*Support indicates how frequently certain combinations of items in both antecedent and consequent occur together in the data.*

  - Supp (X and Y) [aka Rule Support] ≡ P(contains X and Y)
  - Conf (X → Y) ≡ P(contains Y | contains X)

*Confidence is a measure of how often the consequent is true, given that the antecedent is also true.*

  - Lift (X → Y) ≡ $\frac{P(contains\ Y | contains\ X)}{P(contains\ Y)}$

*Lift is a measure of how predictive the rule is, compared to random association. It is defined as the ratio of the observed confidence to expected confidence.*

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.

# A Simple Numerical Example

- Given the rule R1:
  - {milk, bread} → Butter
- Calculate:
  - Supp ({milk})
  - Supp ({milk, bread})
  - Supp ({milk, bread, butter})
  - Confidence of the rule R1.
  - Lift of the rule R1.

**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 |

# 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 |

Assoc 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

*= P(contains{butter}|containts{milk and bread}} divide P(contains butter)*

*0.5/P(butter)*

# Goals in Association Rule Mining

- Find potentially useful rules automatically from the data:
  - Rules with high enough Confidence
  - Rules with high enough support
  - Rules with high enough Lift

# Class Activity 1

Association Rules Concepts

Est. Duration: 30 mins

1. Complete Pre-class reading: 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. At the table, each student to take the lead in explaining their answer to at least one question. Do the rest understand? Other teammates at the same table to ask questions for clarification.

*Notes:*

- *All answers are either direct or can be inferred from the review paper and your previous study of statistics.*
- *There are many ways to express your answers in English. Keep it simple and straight to the point.*

# Questions for Class Activity 1

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

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

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?

4. Many application of association rules require both Confidence and Support. Explain in your own words how both measures can be used to define good association rules.

# Questions for Class Activity 1

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

6. Is Confidence or Lift a symmetric concept? Explain. Implications?

7. Provide another potential application of association rules beyond groceries and insurance claims.

# 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. Explain in your own words how both measures can be used to define good association rules.

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 generating association rules. (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.

# Answers to Questions for Class Activity 1

6. Is Confidence or Lift a symmetric concept? Explain. Implications?

(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.

# Example: Recommend Burger?

| Transaction ID | Items Purchased |
|---|---|
| 1 | Burger, Fries |
| 2 | Burger |
| 3 | Burger |
| 4 | Burger |

- Conf(Burger → Fries) = ¼
- Conf(Fries → Burger) = P(Burger | Fries) = 1/1 = 100%.
- Hence, recommend burger to all who purchased Fries (w/o burger)?
- Ans: No. Lift(Fries → Burger) = P(Burger | Fries) / P(Burger) = 1.
- i.e. Fries did not boost the sales of Burger.

# 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.

# Association Rules using Python

- Python packages with example:
  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 R. Perhaps you can find an alternative easier-to-use Python package for Association Rules. The R packages arules & arulesViz are also available in Anaconda at* https://anaconda.org/conda-forge/r-arules

# Class Activity 2

milk.csv

Est. Duration: 30 mins

Using the Wikipedia 5 transactions example (milk.csv), produce association rules with min support = 0.4 and min Conf = 0.3.

Questions:

1. Create the wide data format and long data format.
2. A default parameter setting may generate empty set in either the LHS or RHS. What setting can we use to prevent this?
3. Compare the results using (a) wide data format and (b) long data format.
   a. What's the cause of the difference?
   b. When should you use which type of format?

# Class Activity 3

Groceries Example

Est: 20 mins

- Run the Rscript groceries.R line by line and view the outputs.
- Based on the R reference link provided.
- Answer the questions in the Rscript.

# 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.

# Summary

- Association Analysis:
  - To identify potentially useful association rules fast and automatically.
  - 3 Key Measures:
    - Support
    - Confidence
    - Lift