NANYANG TECHNOLOGICAL UNIVERSITY
SINGAPORE

Association Rules

SESSION 3 IN-CLASS SLIDES

Intended Learning Outcomes

Identify aspects of business problems that cause standard analytics models to become useless or less effective.

Apply advanced techniques to overcome or mitigate the weaknesses of standard analytics models.

Evaluate performance of the advanced predictive techniques.

Explain the workings and results of the advanced predictive techniques in the context of the business problem to client/employer.

Propose business solutions/recommendations based on the advanced predictive techniques.

Quiz

Ungraded. Check your understanding of this Session Content.
Use your real name (not nickname) in the quiz.

Class Activity 1

Association Rules Concepts

Part is Pre-class learning activities.

Est. Duration: 15 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. [Pre-class]
2. Individually write down your answers to the questions in the next 2 slides. [Pre-class]
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 may 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?

Questions for Class Activity 1

4. 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.

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

6. 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?

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

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?

Answers to Questions for Class Activity 1

4. 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.

Answers to Questions for Class Activity 1

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

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

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)?
-
-

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

Activity 2

milk.csv

Part is Pre-class learning activities.

Est. Duration: 20 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. [Pre-class]
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 what type of data format?

Class Activity 3

Groceries Example

Est: 30 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?

Summary

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

Reminder

Please complete the Pre-Class Learning Activities before next class.

Reflection on your Learning

Go	NTULearn Class Site > Journal
Post	Read the instructions and post entry on this week's learning. • Reply on the 3 questions as stated in the Journal Instructions.