MIS772
Predictive Analytics

Association rule mining

Refer to your textbook by Vijay Kotu and Bala Deshpande, *Data Science: Concepts and Practice*, 2nd ed, Elsevier, 2018.

AACSB ACCREDITED

EFMD EQUIS ACCREDITED

Deakin University CRICOS Provider Code: 00113B

Association rule mining

- Understanding association rules
  - Evaluation metrics
    - Support
    - Confidence
    - Lift
- Example rule generation algorithm
  - Apriori algorithm

DEAKIN BUSINESS SCHOOL

Question:

- A person 35 years of age, shopping at around 6.00pm on a Friday, has just purchased a pack of nappies on the way home. What would you think will be the most likely items bought next?
  - 1: A pack of plastic bags
  - 2: A DVD of a newly released movie
  - 3: A soft toy
  - 4: A bottle of milk
  - 5: Pair of sunglasses

Image source: https://www.maleskin.co.uk/male-skin-care-assistant

Association rules

- What are they?
  - Are a measure of how strongly two (or more) items co-occur
  - Find patterns in the data
  - Are rules extracted from large amounts of data
  - {Item A} -> {Item B}: if A is in the item set, then B will most likely be there too
  - {Item A and Item B} -> {Item C and Item D and Item E}
    - If a shopper buys milk, then they will most likely buy bread too
    - If a football team is awarded a penalty, then they will most likely score a goal
    - If a customer buys one product per quarter, then they will most likely not churn for a year

Refer KD, Chapter 6

Association rule generic form:

{Antecedent(s)} → {Consequent(s)}
e.g., if {A,B} Then {C}

Association rules

- Containers
  - Frequent item sets reside in…
    - Baskets of occurrence (e.g., one transaction, one episode of care, one online session, etc.)
    - Windows of time (e.g., one day, one quarter [of a game], etc.)
  - Data may need to be pre-processed to…
    - Create containers
    - Find co-occurrences in those containers

Association rules

- Pre-processing
  - Example…field hockey, finding containers

can be within a time window of 15 seconds

Time	Location	Event
7:05:05 PM	first quarter - own side	passed the ball
7:05:08 PM	first quarter - own side	lost the ball
7:05:12 PM	second quarter - own side	intercepted the ball
7:05:14 PM	midfield	passed the ball
7:05:18 PM	midfield	passed the ball
7:05:20 PM	second quarter - own side	passed the ball
7:05:22 PM	second quarter - opponent's side	passed the ball
7:05:25 PM	first quarter - opponent's side	shot on goal - returned
7:05:27 PM	first quarter - opponent's side	intercepted the ball
7:05:29 PM	first quarter - opponent's side	passed the ball
7:05:35 PM	first quarter - opponent's side	passed the ball
7:05:40 PM	second quarter - opponent's side	passed the ball
7:05:52 PM	first quarter - opponent's side	shot on goal - scored
7:06:40 PM	second quarter - own side	passed the ball
...	...	...

Sample data sorted by time

can be within a specific location

Time	Location	Event
7:05:12 PM	second quarter - own side	intercepted the ball
7:05:20 PM	second quarter - own side	passed the ball
7:06:40 PM	second quarter - own side	passed the ball
7:05:22 PM	second quarter - opponent's side	passed the ball
7:05:40 PM	second quarter - opponent's side	passed the ball
7:05:14 PM	midfield	passed the ball
7:05:18 PM	midfield	passed the ball
7:05:05 PM	first quarter - own side	passed the ball
7:05:08 PM	first quarter - own side	lost the ball
7:05:25 PM	first quarter - opponent's side	shot on goal - returned
7:05:27 PM	first quarter - opponent's side	intercepted the ball
7:05:29 PM	first quarter - opponent's side	passed the ball
7:05:35 PM	first quarter - opponent's side	passed the ball
7:05:52 PM	first quarter - opponent's side	shot on goal - scored
...	...	...

Sample data sorted by locations

Association rules

- Pre-processing
  - Example…media website, transforming data

Session ID	List of media categories accessed
1	{News, Finance}
2	{News, Finance}
3	{Sports, Finance, News}
4	{Arts}
5	{Sports, News, Finance}
6	{News, Arts, Entertainment}

Sample data set
Source: Page 197, KD Ch6

clickstream converted to binary codes
(items=visits to specific categories)

Session ID	News	Finance	Entertainment	Sports	Arts
1	1	1	0	0	0
2	1	1	0	0	0
3	1	1	0	1	0
4	0	0	0	0	1
5	1	1	0	1	0
6	1	0	1	0	1

Sample data set
Source: Page 198, KD Ch6

Association rules

- Pre-processing
  - Example…media website, transforming data (cont.)
  - Which rules are likely to be valid?
    - {News} -> {Entertainment}
    - {News} -> {Sports}
    - {Finance} -> {Arts}
    - {Finance} -> {News}
    - {News, Finance} -> {Sports}
    - {News, Finance} -> {Arts}

Session ID	News	Finance	Entertainment	Sports	Arts
1	1	1	0	0	0
2	1	1	0	0	0
3	1	1	0	1	0
4	0	0	0	0	1
5	1	1	0	1	0
6	1	0	1	0	1

Sample data set
Source: KD Page 198, Ch6

Question

Given the very large number of possible permutations between items, how do we know when to keep or not to keep a rule(s)?

Association rules

Session ID	News	Finance	Entertainment	Sports	Arts
1	1	1	0	0	0
2	1	1	0	0	0
3	1	1	0	1	0
4	0	0	0	0	1
5	1	1	0	1	0
6	1	0	1	0	1

- Evaluation metrics
  - Support
    - Is the relative frequency of occurrence of an item set in the container set.
      - (i.e. Fraction of total items that contain a specific occurrence)
    - **Filters out rules that are not worth considering further.**
    - Support({News})=5/6=0.83
    - Support({News, Finance})=4/6=0.67
    - {News} -> {Sports}: Support({News, Sports})=2/6=0.33
    - {News, Finance} -> {Arts}: Support({News, Finance, Arts})=0/6=0

Association rules

Association rule generic form:

{Antecedent(s)} → {Consequent(s)}

- Evaluation metrics
  - Confidence
    - Measures the likelihood of occurrence of the right-side of the rule (i.e., consequent) out of all the items in the container that contain the left-side of the rule (i.e., antecedent). This is the *reliability* of the rule.

$$Confidence(X \rightarrow Y) = \frac{Support(X \cup Y)}{Support(X)}$$

$$Confidence(\{News\} \rightarrow \{Finance\}) = \frac{Support(\{News, Finance\})}{Support(\{News\})} = \frac{4/6}{5/6} = 0.8$$

$$Confidence(\{News, Finance\} \rightarrow \{Sports\}) = \frac{Support(\{News, Finance, Sports\})}{Support(\{News, Finance\})} = \frac{2/6}{4/6} = 0.5$$

Important note: the use of ∪ in these formulas in the textbook differs from the meaning commonly used in mathematical set theory (where ∪ indicates a union between two sets). In the textbook formulas, the symbol is used to indicate ***intersection*** (i.e., A ∪ B in the formulas refer to instances where A and B co-occurs).

Association rules

- Evaluation metrics
  - Lift
    - Is similar to confidence; however, it **considers the support of the right-side of the rule too**.
    - **Values closer to 1 indicate non-useful rules, larger lift values indicate more significant rules.**

$$Lift(X \rightarrow Y) = \frac{Support(X \cup Y)}{Support(X) \times Support(Y)}$$

$$Lift(\{News, Finance\} \rightarrow \{Sports\}) = \frac{Support(\{News, Finance, Sports\})}{Support(\{News, Finance\}) \times Support(Sports)} = \frac{2/6}{4/6 \times 2/6} = 1.5$$

Session ID	News	Finance	Entertainment	Sports	Arts
1	1	1	0	0	0
2	1	1	0	0	0
3	1	1	0	1	0
4	0	0	0	0	1
5	1	1	0	1	0
6	1	0	1	0	1

Quiz!

- In a set of 10,000 transactions…
  - an analysis shows that 6,000 of customer transactions include computer games, while 7,500 include videos, and 4,000 include both computer games and videos. What is the confidence of the rule {computer games} -> {videos}?
    - A: 0.40
    - B: 0.89
    - C: 0.76
    - D: 0.67

Step-by-step calculation..

- $Confidence(X \rightarrow Y) = \dfrac{Support(X \cup Y)}{Support(X)}$
- $Confidence\ (games\ \rightarrow videos) = \dfrac{Support\ (games \cup videos)}{Support\ (games)}$
- Support {games, videos} =
  - 4000/10000 = 0.4
  - Why?
    - *of the 10,000 transactions "4,000 include both computer games and videos"*
- Support {games} =
  - 6000/10000 = 0.6
  - Why?
    - *of the 10,000 transactions "6,000 of customer transactions include computer games"*
- Therefore:
  - $Confidence\ (games\ \rightarrow videos) = \dfrac{0.4}{0.6} = 0.67$ (i.e., option D)

Rule generation process

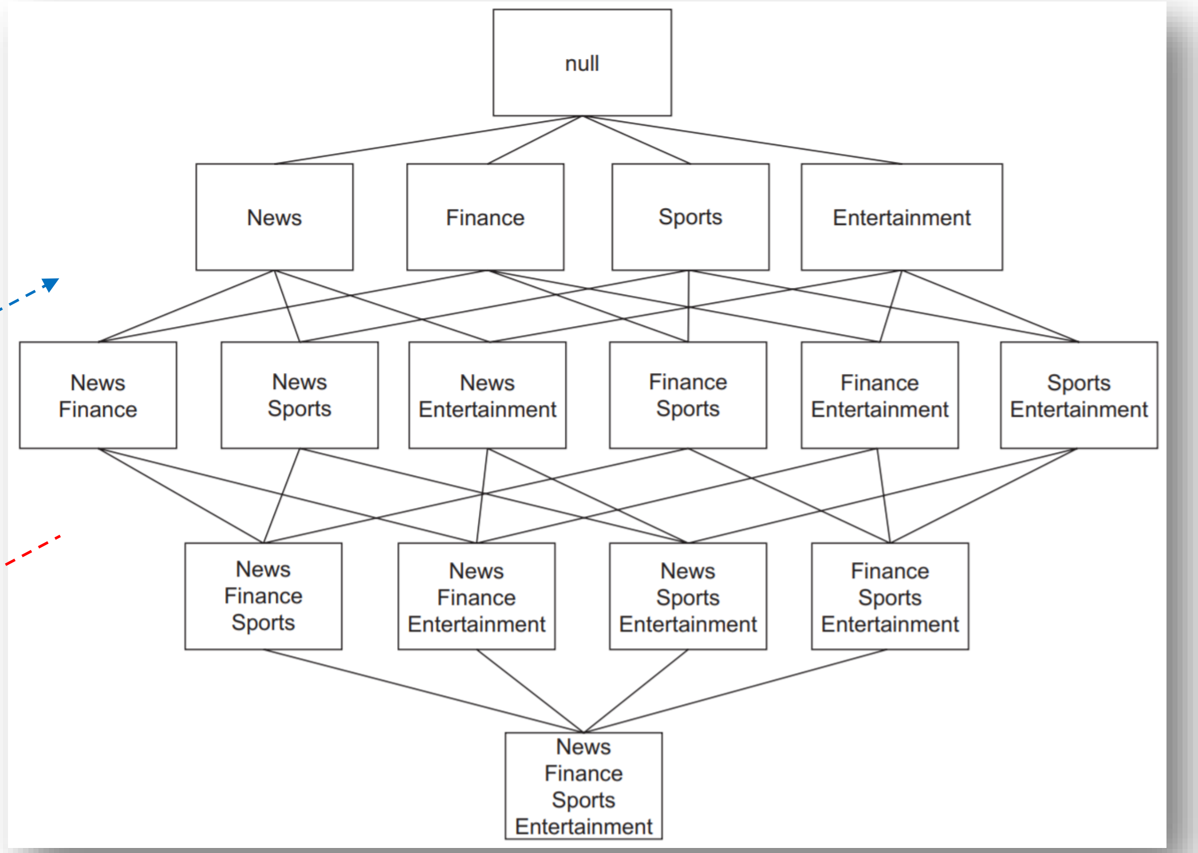
- Two main steps
  - Step 1: Finding all frequent item sets
    - Look at all possible combinations of items
    - There will be $2^n - 1$ item sets in a set of n items
    - Filtering non-important items out (using support)
  - Step 2: Generating/extracting rules from frequent item sets
    - Look at all possible rules
    - For a dataset with n items, there will be $3^n - 2^{n+1} + 1$ rules
    - Filter out rules that are not significant (using confidence or lift)

Rule generation process

- Example…
  - Media website
    - Items: News, Finance, Sports, Entertainment

all possible item sets in a lattice form,
to be used to find frequent item sets

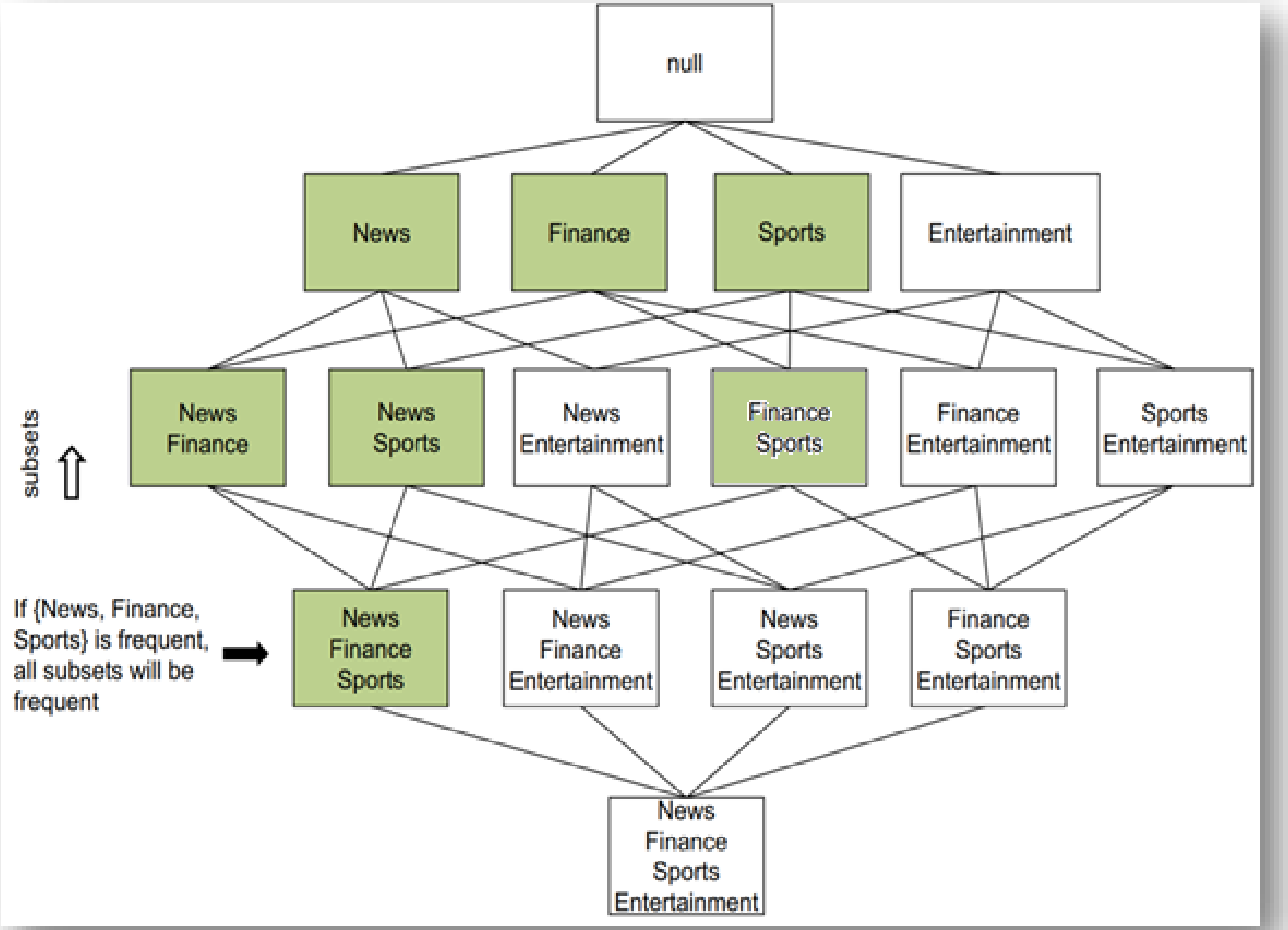
which item set is **frequent**?



Source: KD Page 201, Ch6

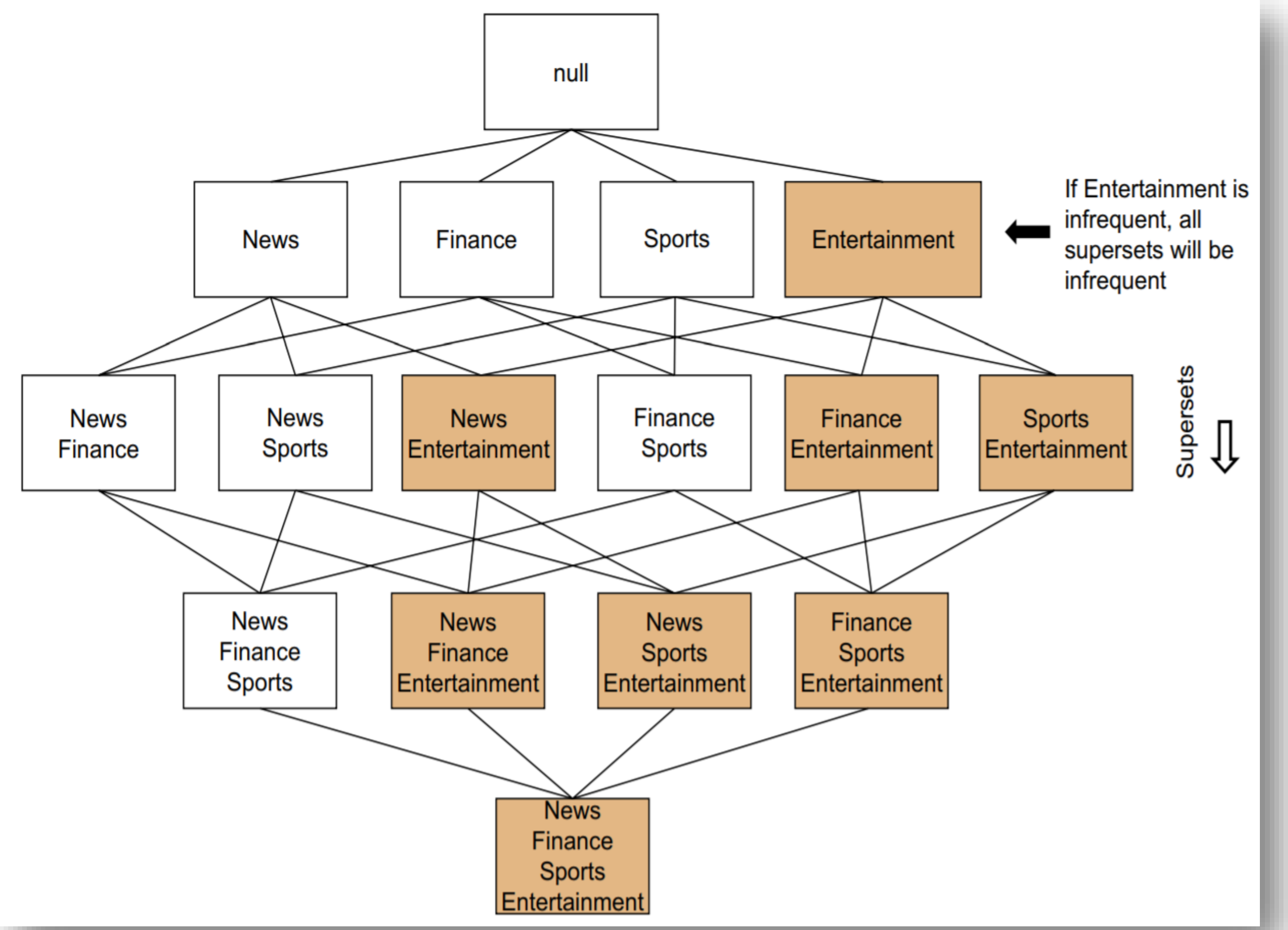
Rule generation process

- Apriori algorithm
  - To find frequent item sets **more efficiently**
  - Makes use of support of item sets
    - Item sets with a support of larger than a threshold are frequent
    - Rule 1…
      - If an item set is frequent, then all its **subsets** are frequent



Rule generation process

- Apriori algorithm



- Rule 2…
  - If an item set is infrequent, then all its **supersets** are infrequent

Rule generation process

- Generating/extracting rules
  - Generate all rules for each frequent item set with n items
  - Makes use of confidence or lift of rules to filter out non-significant rules
  - In the previous example…
    - For the item set {News, Sports, Finance}, there will be the following rules/confidence values:
    - {News, Sports} -> {Finance}: confidence=1.0
    - {News, Finance} -> {Sports}: confidence=0.5
    - {Sports, Finance} -> {News}: confidence=1.0
    - {News} -> {Sports, Finance}: confidence=0.4
    - {Sports} -> {News, Finance}: confidence=1.0
    - {Finance} -> {News, Sports}: confidence=0.5

all rules with a confidence $\geq$ a threshold will be kept as output

Rule generation process

- Frequent pattern (FP)-growth algorithm
  - Another algorithm for finding frequent item sets
  - Extra reading…
  - Details are not examinable
  - Reference: Pages 206-210, KD Ch6

  - FP-growth algorithm…
    - works on the basis of compressing item sets into compressed tree structures called FP-Trees
    - is often more efficient than the Apriori algorithm

Sample exam question

You are given a data set including 1,000 shopping transactions. In this data set, there are transactions that include items as listed in the table below:

a) Given the transaction set, will you say the association rule $\{Milk\} \rightarrow \{Beer\}$ represents a correct and likely association? Justify your answer.

b) Given $\mathrm{Confidence}(\mathrm{X} \rightarrow \mathrm{Y}) = \frac{\mathrm{Support}(\mathrm{X} \cup \mathrm{Y})}{\mathrm{Support}(\mathrm{X})}$, calculate the Confidence of the association rule $\{Clock\} \rightarrow \{Towel\}$ in the transaction set, and

c) Briefly explain the main shortcoming/s of Confidence as related to this case. What other association rule analysis evaluation metric do you suggest to be used to address the shortcoming/s of Confidence? Justify your answer.

6 + 8 + 6 = 20 Marks

Items	Frequency of occurrence in transactions
Milk and DVD	**650**
Milk and Beer	**20**
Bread and Beer	**35**
Towel and Milk and DVD	**15**
Clock and Towel	**575**
Clock	**620**

Summary / Review

- What are frequent item sets?
- Give the large number of possible permutations between items, how do we know when to keep or not to keep a rule(s)?
- What is the shortcoming of support?
- What is the shortcoming of confidence?
- How should interpret lift values?
- Describe the apriori algorithm.
- What are frequent item subsets?
- What are frequent item supersets?
- How could you apply the insights from association rules to inform which items to stock and where to place them on supermarket shelves (e.g. in the cleaning products aisle)?