MA5810: Introduction to Data Mining

Week 6; Collaborate Session 1: Assosciation Rule Mining

Martha Cooper, PhD

JCU Masters of Data Science

2021-08-10

Housekeeping

- Collaborate 1 = Tuesday 6.45-8pm (Martha)
- Collaborate 2 = Thursday 6.45-8pm (Martina)

For my Collaborate Sessions, you can get the slides & R code for each week here:

https://github.com/MarthaCooper/MA5810

Subject: MA5810 Intro to Data Mining

MA5810 Learning Outcomes

1. Overview of Data Mining and Examples
2. Unsupervised data mining methods e.g. clustering and outlier detection;
3. Unsupervised and supervised techniques for dimensionality reduction (Today = PCA);
4. Supervised data mining methods for pattern classification;
5. Apply these concepts to real data sets using R (Today).

Today's Goals

- Understand the background behind association rule mining using the Apriori algorithm
- Understand the pros and cons of the Apriori algorithm
- Apply frequent pattern mining using the Apriori algorithm to real datasets using R

Association Rules

Frequent pattern mining

- Finding strong associations between variables or sequences of values, which manifest themselves as recurring patterns in observed data.

Association rule mining

- Aims to detect interesting rules of association between the values of a collection of variables using some measure of interesting-ness.

Market Basket Analysis

- The most well-known application is basket or transaction analysis.

Finding association rules

- Database of transactions
  - e.g. "Basket" of items purchased in a supermarket

```
##   Transaction_ID                Items
## 1              1       Cookies, Juice
## 2              2  Bread, Butter, Milk
## 3              3 Cookies, Milk, Juice
## 4              4          Juice, Milk
```

- The goal is to find interesting rules *e.g.*
  - $A_i, A_j, A_k \Rightarrow A_m$
  - When items A_i, A_j, A_k are purchased together, it is likely that A_m will be purchased as well.
- Utility for marketing, sales, bioinformatics, meteorology, finance etc...

Defining itemsets and rules

- Itemset: Let T be the collection of items in a database of transactions. Any subset of items $I \subseteq T$ is referred to as an itemset *e.g.* entire transaction or part of a transaction.
  - $T = \{Bread, Butter, Cookies, Juice, Milk\}$,
  - $I = \{Milk\}$, $I = \{Bread, Juice\}$, $I = \{Bread, Butter, Cookies\}$, etc...
- k-Itemset: An itemset containing k items, $I = \{i_1, \ldots, i_k\}$.
- Association Rule An association rule, R, is a rule of the form $I \Rightarrow J$ where $I \subseteq T$ and $J \subseteq T$ are disjoint itemsets, that is, $I \cap J = \emptyset$. Think *IF* I *THEN* J .For example, given an itemset {A,B,C}, two possible rules are:
  - $\{A\} \Rightarrow \{B, C\}$
  - $\{A, B\} \Rightarrow \{C\}$
  - Note that $I \Rightarrow J$ does not imply causation.
  - Left = *Antecedent*; Right = *Consequent*

Defining how "interesting" a rule is

- Support of an itemset: The support of itemset I, $Sup(I)$, is the number of transactions that contain I out of the total number of transactions. *Probability that a transaction contains* I.
- Support of a rule: The support of the rule $Sup(I \Rightarrow J)$ is the fraction of transactions that contain all items involved in the rule. *Probability that a transaction that contains* I *and* J.
- Confidence of a rule: The fraction of all transactions that contain all items in the defined rule, $(I \Rightarrow J)$, out of all transactions that contain the antecedent itemset, I. *Conditional probability that a transaction that contains* I *also contains* J.

Defining how "interesting" a rule is

- $R : I \Rightarrow J$
- $Sup = \frac{frq(I,J)}{N}$
- $Conf = \frac{frq(I,J)}{freq(I)}$

Apriori Algorithm

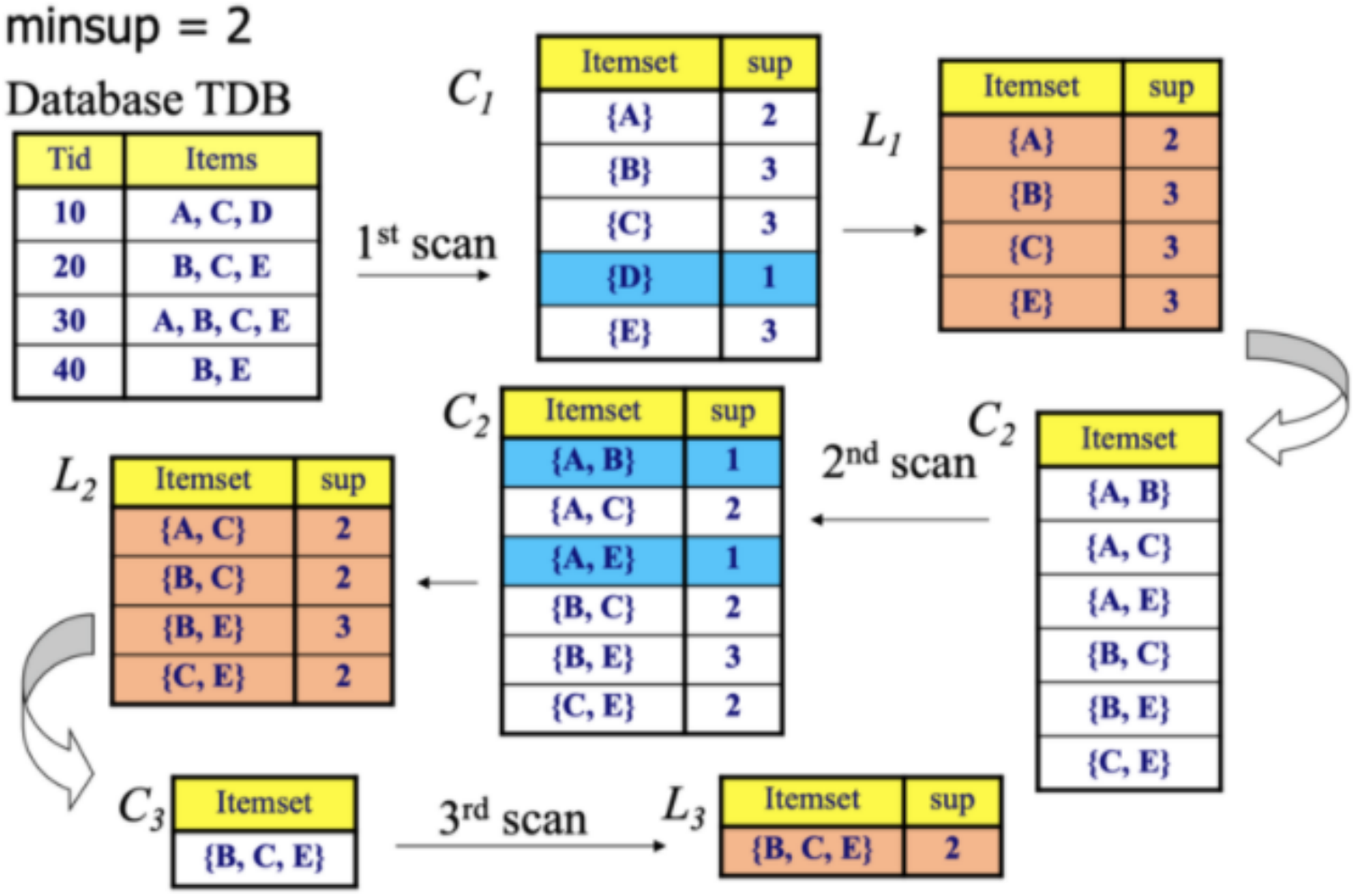
- Uses frequent item sets to generate association rules.
- Principal: If an itemset is frequent, then any of its subsets are also frequent. Conversely, if an itemset is infrequent then any of its supersets are also infrequent.
- How do we define frequent?
    - We define minimum threshold for support.
    - We define itemsets as frequent if they have a support value above that minimum threshold.

Apriori Algorithm Intuition

1. Define thresholds for Support (*e.g.* 0.5) and Confidence (*e.g.* 1)
2. Create frequency table of all itemsets of size 1 ($k = 1$). Remove itemsets that are *below the support threshold* (Pruning).
3. Create a frequency table for all itemsets of size 2 ($k = 2$). Remove itemsets that are *below the support threshold*.
4. Create a frequency table for all itemsets of size 3 ($k = 3$). Remove itemsets that are *below the support threshold*.
5. Repeat until no frequent or candidate itemsets can be generated.
6. Association rules are generated from the frequent itemsets in steps 2-4. Optional filter to remove rules that are below the confidence threshold.

See Chapter 6 in Introduction to Data Mining by Tan, Steinbach, & Kumar for more detail and extensions.

Apriori Algorithm Intuition



Apriori Algorithm Pro's and Con's

Pros and Cons

Pros

- Easy to implement and easy to understand
- Scale-able

Cons

- Computationally expensive (but extensions exist)

Apriori in R

```
Tr_list <- list( #list of transactions
          c("BISCUITS", "MILK","ORANGEJUICE", "TOOTHPASTE", "MOUTHWASH"
          c("BREAD", "BUTTER", "JAM", "MILK", "TOOTHPASTE", "MOUTHWASH"
          c("JAM", "ORANGEJUICE", "TOOTHPASTE", "MOUTHWASH"),
          c("TOOTHPASTE", "JAM", "MOUTHWASH"),
          c("MILK", "BREAD", "BISCUITS", "TOOTHPASTE"),
          c("MOUTHWASH", "TOOTHPASTE", "TOOTHBRUSH", "BREAD"),
          c("MOUTHWASH", "TOOTHPASTE", "TOOTHBRUSH", "TOILETPAPER"),
          c("BREAD", "BUTTER", "JAM", "MILK", "TOOTHPASTE"),
          c("BREAD", "BUTTER", "JAM", "MILK"),
          c("BISCUITS", "JAM", "ORANGEJUICE", "TOOTHPASTE","MOUTHWASH"
          c("BREAD", "BUTTER", "BISCUITS"),
          c("BREAD", "BISCUITS", "JAM", "MILK"),
          c("BREAD", "BUTTER", "MILK", "TOOTHPASTE", "MOUTHWASH"),
          c("BREAD", "JAM", "ORANGEJUICE", "MILK"),
          c("BUTTER", "JAM", "TOOTHPASTE", "MOUTHWASH"),
          c("BREAD", "BUTTER", "DIAPERS"),
          c("MOUTHWASH", "TOOTHPASTE"),
          c("BREAD", "BUTTER", "BISCUITS","TOOTHPASTE"),
          c("MOUTHWASH", "MILK", "BREAD", "TOOTHPASTE"),
          c("MOUTHWASH", "TOOTHPASTE", "TOOTHBRUSH", "TOILETPAPER"),
          c("TOOTHPASTE", "JAM", "MOUTHWASH"),
          c("MILK", "BREAD", "BISCUITS", "TOOTHPASTE"),
```

Apriori in R

```
library(arules) #for association rule mining
Tr_object <- as(Tr_list, "transactions") #create object of class transac
summary(Tr_object) #summarise
inspect(Tr_object) #see all transactions

#extract frequent with a minimum support of 0.4
FI <- apriori(Tr_object, parameter = list(support = 0.4,
                                          target = "frequent itemsets"))
inspect(sort(FI), by = "support")

#extract rules with a minimum support of 0.4 and confidence of 1
AR <- apriori(Tr_object, parameter = list(support = 0.4,
                                          confidence = 1,
                                          target = "rules",
                                          minlen=2))
inspect(sort(AR), by = "support")

library(arulesViz) #for visualisation
plot(AR) #plot
```

Extra reading

- Chapter 6 in Introduction to Data Mining by Tan, Steinbach, & Kumar.

References

- MA5810 Week 6 Topic 1
- Chapter 6 in Introduction to Data Mining by Tan, Steinbach, & Kumar.

Slides

- xaringhan, xaringanthemer, remark.js, knitr, R Markdown