NOVA IMS Information Management School

ITEMSETS AND ASSOCIATION RULES FOR MARKET BASKET ANALYSIS: BRIEF REVIEW

Business Cases with Data Science

SUMMARY

Itemsets and association rules

1. Introduction
2. Market basket analysis: Apriori algorithm
3. Demo

1. INTRODUCTION

FREQUENT PATTERNS

(AKA. “ASSOCIATION RULES”)

Patterns that occur frequently in data, which are usually divided into:

- Frequent itemsets
- Frequent subsequences
- Frequent substructures

FREQUENT ITEMSETS

Identifying recurring relationships in a dataset, like the co-occurrence of two or mode objects of interest is also known as “Market basket analysis”. For example, identifying set of items that often appear together in a transactional dataset (e.g., in a restaurant: steak and red wine)

> *“Customers purchase a wide variety of items together. For example, the model shows that a **high percentage of customers purchase bread and milk in** the same transaction, so if a customer purchases bread they are highly likely to purchase milk, and vice versa. Milk, however, is purchased so frequently that many products exhibit the same kind of behavior: **Whether a customer purchases cereal, chicken, or Limburger cheese, there is still a high likelihood that milk is in the same basket**.*
>
> *However, the converse is not true: **If one purchases milk, only rarely is Limburger cheese in that basket**, in large part because **very few customers purchase Limburger cheese**. Customers who purchases **high-end, specialty crackers, however, purchase Limburger cheese much more frequently than the rate of purchasing Limburger cheese alone**, which is low for this store. In addition, the **purchasers of Limburger cheese spent more on average than the average customer**, a statistic that was computed after the model was computed and wasn't a part of the association rules model. **As a result of the analysis, the decision-makers kept the high-end cracker on the shelf even though its sales were not strong ...**”*[2]

ASSOCIATION ALGORITHMS

The purpose of finding items associations is to enumerate interesting interactions between items. Due the numerous possible combinations, a brute-force approach is not a solution to identify interactions. Therefore, several algorithms are used to the effect, namely:

- Apriori
- Eclat
- FP-growth

FREQUENT SUBSEQUENCES

Allows the discovery of patterns across time or positions in a dataset. For example, the sequential order of purchasing history (e.g., airline ticket, followed by hotel, then by transfer), or travelers' trajectories. Algorithms to explore sequences include:

GSP

Spade

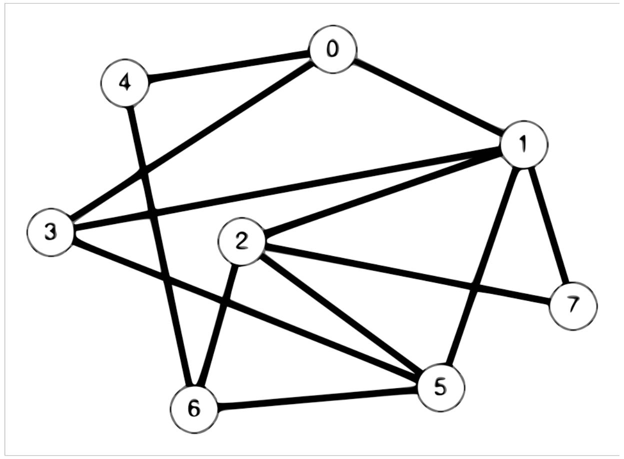
PrefixScan

FREQUENT SUBSTRUCTURES

Combination of frequent itemsets and frequent subsequences (e.g., graphs or trees) that as the objective of finding interesting subgraphs in data.

Graphs - set of interconnected objects

- Objects: nodes or vertices
- Links between objects: edges



COMMON MINED GRAPH PATTERNS

Frequent subgraphs: (using for example the Gspan algorithm)

Closed graphs

Coherent graphs

Dense graphs

2.

MARKET BASKET ANALYSIS: APRIORI ALGORITHM

APPLICATIONS

Cross-selling: understand customer behavior to identify cross-selling opportunities

Product placement: identify complementary products (e.g., pen and paper) and substitute products (e.g., tea and coffee) and place them near by

Affinity promotion: promote products/services bought in association with other items

Customer behavior: understand customer purchasing behavior towards a product or service

Fraud detection: identify unusual patterns that may indicate fraud (by building a classification model)

TERMINOLOGY (1/6)

Conditions (aka as “if-then”): logical constructs from categorical variables that evaluate to true or false, for example:

Product = “wine of brand x”
Age = range [45, 65]
Red wine = 1

TERMINOLOGY (2/6)

Itemsets/association rules (“if-then”) are made of two parts:

- **Antecedent(s)/Left-Hand-Sides (LHS)**: the part ”being compared” (“before” the ”then”). For example, in the rule “if steak then red wine”, steak is the LHS
- **Consequent/Output/Right-Hand-Sides (RHS)**: the part ”being compared” (“after” the ”then”). In the previous example, red wine is the RHS

TERMINOLOGY (3/6)

Support: fraction of the rule(s) that occurs in all observations [0, 1]

$$support(A \Longrightarrow \mathrm{B}) = \mathrm{P}(\mathrm{A} \cup B) = \frac{\#\ transactions\ where\ rule(s)\ is\ present}{total\ \#\ transactions}$$

Steak=1 $= \frac{steak{=}1}{50} = \frac{10}{50} = 20\%$

Red wine=1 $= \frac{red\ wine{=}1}{50} = \frac{8}{50} = 16\%$

Both rules $= \frac{steak{=}1\ and\ red\ wine{=}1}{50} = \frac{6}{50} = 12\%$

Transaction	Steak	Red wine
1	1	
2	1	
3	1	
4	1	
5	1	1
6	1	1
7	1	1
8	1	1
9	1	1
10	1	1
11		1
12		1
…		
50		

TERMINOLOGY (4/6)

Confidence: probability of a rule being correct for a new observation [0,1]

$$confidence(A \Longrightarrow \mathrm{B}) = \mathrm{P}(\mathrm{B}|A)$$

$$= \frac{support(A \cup B)}{support(A)} = \frac{\#\ transactions\ A\ and\ B}{\#\ transactions\ A}$$

$$confidence\ (if\ steak\ then\ red\ wine) = \frac{0.12}{0.2} = 60\%$$

$$confidence\ (if\ red\ wine\ then\ steak) = \frac{0.12}{0.16} = 75\%$$

Transaction	Steak	Red wine
1	1	
2	1	
3	1	
4	1	
5	1	1
6	1	1
7	1	1
8	1	1
9	1	1
10	1	1
11		1
12		1
…		
50		

TERMINOLOGY (5/6)

Lift: ratio by which the confidence of a rule exceeds the expected confidence. In other words, how many times is more likely RHS to occur when LHS is true, compared to when RHS occurs on its own [0,∞]

$$lift(A \Longrightarrow \mathrm{B}) = \frac{support(A \cup B)}{support(A) \times support(B)}$$

$$lift\ (if\ steak\ then\ red\ wine) = \frac{0.12}{0.20 \times 0.16} = 3.75$$

=1: A and B are independent

>1: complementary effects between A and B

<1: substitution effects between A and B

Transaction	Steak	Red wine
1	1	
2	1	
3	1	
4	1	
5	1	1
6	1	1
7	1	1
8	1	1
9	1	1
10	1	1
11		1
12		1
…		
50		

TERMINOLOGY (6/6)

Other measures:

- Leverage
- Conviction

More information at http://rasbt.github.io/mlxtend/user_guide/frequent_patterns/association_rules/

PARAMETERS SETTINGS

Due to the high number of rules that can be found, usually, parameters are defined to speedup algorithms and to reduce the number of rules:

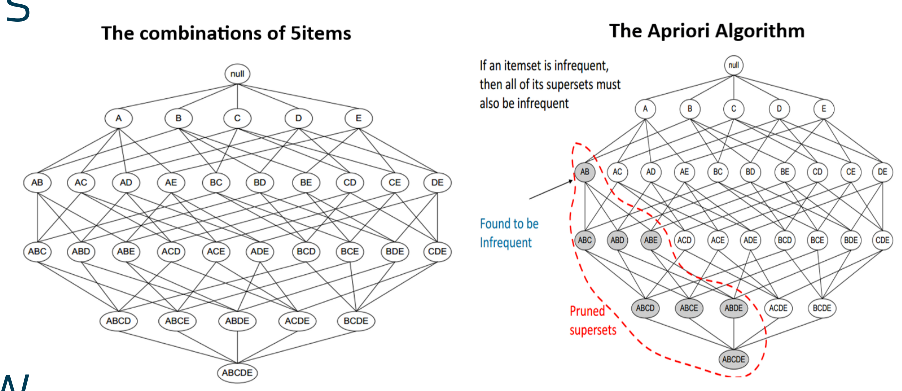
- Minimum support
- Minimum confidence
- Maximum itemsets length (#antecedents + 1 for the consequent)
- Maximum number of rules

PROBLEMS WITH ASSOCIATION RULES

Redundant rules: rules that include other rules

Too many rules: too many rules makes the process finding interesting patterns difficult. Redundant rules should be removed, or increase parameters thresholds

Too few rules: if minimum support or confidence thresholds are too high, few rules may be found


The combinations of 5items
null
A B C D E
AB AC AD AE BC BD BE CD CE DE
ABC ABD ABE ACD ACE ADE BCD BCE BDE CDE
ABCD ABCE ABDE ACDE BCDE
ABCDE
The Apriori Algorithm
If an itemset is infrequent, then all of its supersets must also be infrequent
Found to be Infrequent
Pruned supersets

[https://towardsdatascience.com/market-basket-analysis-978ac064d8c6]

TYPICAL APPROACH TO DEVELOP MODELS

1. Setup association rules to have one consequent (one RHS): the predictive model target variable
2. Set the parameters to include only two or three antecedents (LHS)
3. After building the rules, sort them by confidence
4. Identify top rules and build new columns (dummy variables) to indicate this combination. Repeat the process

3.
DEMO

RESTAURANT ASIA: EXAMPLE

1. Copy from the files folder the dataset "AsianRestaurant_Cyprus_2018_partial.txt"
2. Copy and open the file "MarketBasketAnalysis.py" (in Visual Code)
3. Follow the presentation of the file

REFERENCES

[1] Abbott, D. (2014). *Applied Predictive Analytics: Principles and Techniques for the Professional Data Analyst*. Indianapolis, IN: Wiley

[2] Han, J., Kamber, M., Pei, J. (2012). *Data Mining: Concepts and Techniques (Third edition)*. Waltham, MA: Elsevier