Mining Association Rule

Week 6

Nurul Afiqah Burhanuddin
nurul.afiqah@ukm.edu.my

Room 2119
Department of Mathematical Science

Outline

1. Market Basket Analysis

2. Efficient Mining of Association Rule

Market Basket Analysis

- Analyzes customer buying habits by finding associations between the different items that customers place in their "shopping baskets".
- Helps retailers develop marketing strategies by gaining insight into which items are frequently purchased together by customers. For e.g., if customers are buying milk, how likely are they to also buy bread on the same trip.

Transaction	Items
1	Bread, Milk
2	Bread, Diaper, Beer, Eggs
3	Milk, Diaper, Beer, Coke
4	Bread, Milk, Diaper, Beer
5	Bread, Milk, Diaper, Coke

Example of association rule: $\{\text{Diaper}\} \Rightarrow \{\text{Beer}\}$

$\{\text{Milk, Bread}\} \Rightarrow \{\text{Eggs,Coke}\}$

$\{\text{Beer, Bread}\} \Rightarrow \{\text{Milk}\}$

Definition

- Let dataset T consists of N transactions (i.e., baskets) such that $T = \{t_1, t_2, \ldots, t_N\}$ over a set of I of d items with $t_i \subseteq I$ for $1 \leq i \leq N$.
- An **itemset** is a subset $X \subseteq I$ and its support w.r.t. T, denoted by $supp(X)$, is defined as the fraction of transactions of T that contain X,

  $$supp(X) = \frac{N_X}{N}$$

  An itemset that contains k items is called a k-itemset. For e.g.: {Eggs,Coke} is a 2-itemset. An itemset X is called frequent if $supp(X)$ is greater than some user-defined threshold, α.

Definition

- An **association rule** is a rule: $A \Rightarrow B$ satisfying:
  - $A, B \subset I$
  - $A, B \neq \emptyset$
  - A and B are disjoint itemsets
- The strength of an association rule can be measured in terms of its support, confidence, and lift.
  - support: determines how often a rule is applicable to a given data set.

$$supp(A \Rightarrow B) = \Pr(A \cap B) = \frac{N_{A \cap B}}{N}$$

  - confidence: determines how frequently items in B appear in transactions that contain A.

$$conf(A \Rightarrow B) = \Pr(B|A) = \frac{\Pr(A \cap B)}{\Pr(A)} = \frac{N_{A \cap B}/N}{N_A/N}$$

  - lift: assesses the degree to which the occurrence of A increases the occurrence of the B.

$$lift(A \Rightarrow B) = \frac{conf(A \Rightarrow B)}{\Pr(B)} = \frac{\Pr(A \cap B)}{\Pr(A)\Pr(B)} = \frac{N_{A \cap B}/N}{(N_A/N)(N_B/N)}$$

Definition

- A rule that has very low support might occur simply by chance. From a business perspective a low support rule is unlikely to be interesting because it might not be profitable to promote items that customers seldom buy together.
- Confidence measures the reliability of the inference made by a rule. For a given rule $A \Rightarrow B$, the higher the confidence, the more likely it is for B to be present in transactions that contain A.
- Lift is a correlation measure between A and B. Notice that $\Pr(A) \times \Pr(B) = \Pr(A \cap B)$ if A and B are independent.
  - $lift(A \Rightarrow B) < 1$: the occurrence of A is negatively correlated with the occurrence of B, meaning that the occurrence of one likely leads to the absence of the other one.
  - $lift(A \Rightarrow B) > 1$: then A and B are positively correlated, meaning that the occurrence of one implies the occurrence of the other.
  - $lift(A \Rightarrow B) = 1$: A and B are independent and there is no correlation between them.

Example

Example:

Transaction	Items
1	Bread, Milk
2	Bread, Diaper, Beer, Eggs
3	Milk, Diaper, Beer, Coke
4	Bread, Milk, Diaper, Beer
5	Bread, Milk, Diaper, Coke

- $supp(\{\text{Milk,Diaper}\} \Rightarrow \{\text{Beer}\}) = 0.4000$
- $conf(\{\text{Milk,Diaper}\} \Rightarrow \{\text{Beer}\}) = 0.6667$
- $lift(\{\text{Milk,Diaper}\} \Rightarrow \{\text{Beer}\}) = 1.1111$

Example

Example:

Transaction	Items
1	Bread, Milk
2	Bread, Diaper, Beer, Eggs
3	Milk, Diaper, Beer, Coke
4	Bread, Milk, Diaper, Beer
5	Bread, Milk, Diaper, Coke

- $supp(\{\text{Milk,Diaper}\} \Rightarrow \{\text{Beer}\}) = 0.4000$
- $supp(\{\text{Milk,Beer}\} \Rightarrow \{\text{Diaper}\}) = 0.4000$
- $supp(\{\text{Diaper,Beer}\} \Rightarrow \{\text{Milk}\}) = 0.4000$
- $supp(\{\text{Beer}\} \Rightarrow \{\text{Milk,Diaper}\}) = 0.4000$
- $supp(\{\text{Diaper}\} \Rightarrow \{\text{Milk,Beer}\}) = 0.4000$
- $supp(\{\text{Milk}\} \Rightarrow \{\text{Diaper,Beer}\}) = 0.4000$

Example

Example:

Transcation	Items
1	Bread, Milk
2	Bread, Diaper, Beer, Eggs
3	Milk, Diaper, Beer, Coke
4	Bread, Milk, Diaper, Beer
5	Bread, Milk, Diaper, Coke

- $conf(\{\text{Milk,Diaper}\} \Rightarrow \{\text{Beer}\}) = 0.6667$
- $conf(\{\text{Milk,Beer}\} \Rightarrow \{\text{Diaper}\}) = 1.0000$
- $conf(\{\text{Diaper,Beer}\} \Rightarrow \{\text{Milk}\}) = 0.6667$
- $conf(\{\text{Beer}\} \Rightarrow \{\text{Milk,Diaper}\}) = 0.6667$
- $conf(\{\text{Diaper}\} \Rightarrow \{\text{Milk,Beer}\}) = 0.5000$
- $conf(\{\text{Milk}\} \Rightarrow \{\text{Diaper,Beer}\}) = 0.5000$

Example

Example:

- $\{\text{Milk,Diaper}\} \Rightarrow \{\text{Beer}\}$
- $\{\text{Milk,Beer}\} \Rightarrow \{\text{Diaper}\}$
- $\{\text{Diaper,Beer}\} \Rightarrow \{\text{Milk}\}$
- $\{\text{Beer}\} \Rightarrow \{\text{Milk,Diaper}\}$
- $\{\text{Diaper}\} \Rightarrow \{\text{Milk,Beer}\}$
- $\{\text{Milk}\} \Rightarrow \{\text{Diaper,Beer}\}$

All the above rules are subsets of the same itemset:

$$A \cap B = \{\text{Milk,Diaper,Beer}\}$$

Rules originating from the same itemset have identical support but can have different confidence.

Efficient Mining of Association Rule

- Given a set of transactions, the goal of association rule mining is to find all rules $A \Rightarrow B$ having:

  $$supp(A \Rightarrow B) \geq \alpha$$
  $$conf(A \Rightarrow B) \geq \gamma$$

  α is the minimum support threshold.
  γ is the minimum confidence threshold.
- Some approaches in mining association rules:
  - Brute-force approach
  - Apriori algorithm

Efficient Mining of Association Rule

Brute-force approach

- To compute the support and confidence for every possible rule.
- Assuming that neither the LHS nor the RHS of the rule is an empty set, the total number of possible rules, R, extracted from a dataset that contains d items is

  $$\begin{aligned} R &= \sum_{k=1}^{d-1} \binom{d}{k} (2^{d-k} - 1) \\ &= 3^d - 2^{d+1} + 1 \end{aligned}$$

  For k items, $1 \leq k < d$, there are $\binom{d}{k}$ possible itemsets of size k on the LHS. For each of these possible itemsets, we can form a rule with $2^{d-k} - 1$ distinct non-empty itemsets on the RHS that disjoint from the LHS.
- This approach is prohibitively expensive because there are exponentially many rules that can be extracted from a dataset.

Efficient Mining of Association Rule

Apriori algorithm

- Exploits anti-monotone property of the $support$:

  $$\text{For every } A, B \subseteq I,\ A \subseteq B \Rightarrow supp(B) \le supp(A)$$

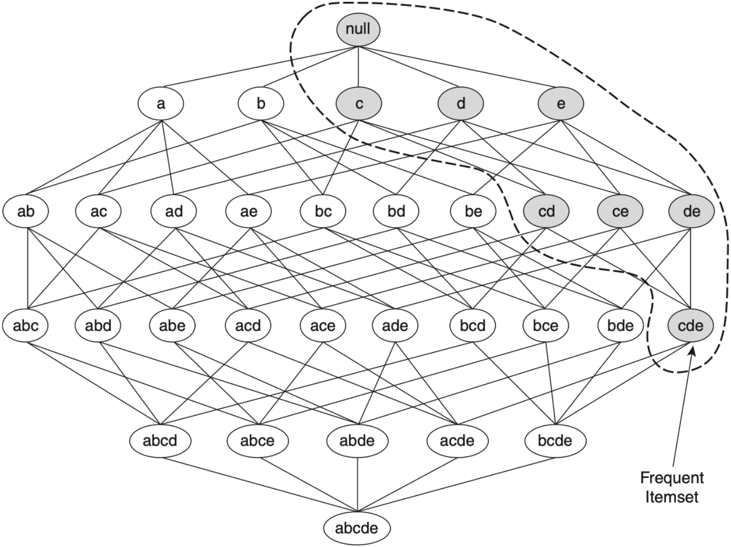
  Consequence:
  - If an itemset A is frequent, then all of its subsets, $W \subseteq A$, must also be frequent.
  - If an itemset A is infrequent, then all of its supersets, $W \supseteq A$, must be infrequent.
- Two phases:

  I: Determine the set F of all frequent itemsets with $supp \ge \alpha$

  II: For each itemset $W \in F$, generate all rules $A \Rightarrow B$, with $A \cup B = W$ and $conf(A \Rightarrow B) \ge \gamma$
- The strategy of trimming the exponential search space k based on the $support$ is known as support-based pruning.
- The use of $support$-based pruning systematically controls the exponential growth of candidate itemsets.

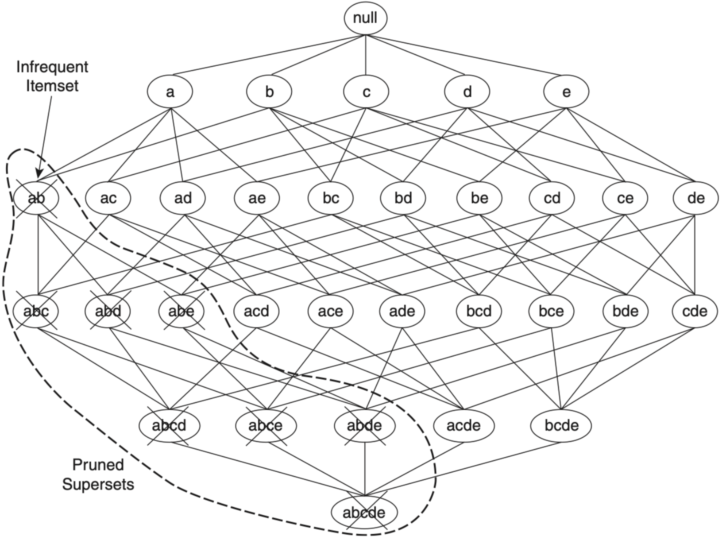
Efficient Mining of Association Rule

Apriori algorithm



Efficient Mining of Association Rule

Apriori algorithm



Efficient Mining of Association Rule

Apriori algorithm

Example 1: Find association rule with $\alpha = 0.4$ and $\gamma = 0.7$.

Transaction	Items
1	Eggs, Flour, Milk
2	Beer, Eggs, Flour
3	Eggs, Milk
4	Eggs, Flour, Milk
5	Milk

Efficient Mining of Association Rule

Apriori algorithm

Example 1: Find association rule with $\alpha = 0.4$ and $\gamma = 0.7$.
Phase I: Determine the set F of all frequent itemsets with $supp \geq 0.4$

1-itemset	$supp$
Eggs	0.80
Flour	0.60
Milk	0.80
Beer	0.20

2-itemset	$supp$
Eggs, Flour	0.60
Eggs, Milk	0.60
Flour, Milk	0.40

3-itemset	$supp$
Eggs, Flour, Milk	0.40

Efficient Mining of Association Rule

Apriori algorithm

Example 1: Find association rule with $\alpha = 0.4$ and $\gamma = 0.7$.
Phase II: For each itemset $W \in F$, generate all rules $A \Rightarrow B$, with $A \cup B = W$ and $conf(A \Rightarrow B) \geq 0.7$.
F={{Eggs, Flour}, {Eggs, Milk}, {Flour, Milk}, {Eggs, Flour,Milk}}

rule	$conf$
{Eggs}⇒{Flour}	0.75
{Flour}⇒{Eggs}	1.00
{Eggs}⇒{Milk}	0.75
{Milk}⇒{Eggs}	0.75
{Flour}⇒{Milk}	0.67
{Milk}⇒{Flour}	0.50

rule	$conf$
{Eggs,Flour}⇒{Milk}	0.67
{Eggs,Milk}⇒{Flour}	0.67
{Flour,Milk}⇒{Eggs}	1.00
{Milk}⇒{Eggs,Flour}	0.50
{Flour}⇒{Eggs,Milk}	0.67
{Eggs}⇒{Flour,Milk}	0.50

Efficient Mining of Association Rule

Apriori algorithm

Example 1: Find association rule with $\alpha = 0.4$ and $\gamma = 0.7$.
Phase II: For each itemset $W \in F$, generate all rules $A \Rightarrow B$, with $A \cup B = W$ and $conf(A \Rightarrow B) \geq 0.7$.
Based on $\gamma = 0.7$, only rule exceeding these thresholds will be retained.

rule	$supp$	$conf$	$lift$
{Flour}⇒{Eggs}	0.60	1.00	1.25
{Eggs}⇒{Flour}	0.60	0.75	1.25
{Milk}⇒{Eggs}	0.60	0.75	0.94
{Eggs}⇒{Milk}	0.60	0.75	0.94
{Flour,Milk}⇒{Eggs}	0.40	1.00	1.25

Efficient Mining of Association Rule

Apriori algorithm

Example 2: Find association rule with $\alpha = 0.6$ and $\gamma = 0.8$.

Transaction	Items
1	Bread, Milk
2	Bread, Diaper, Beer, Eggs
3	Milk, Diaper, Beer, Coke
4	Bread, Milk, Diaper, Beer
5	Bread, Milk, Diaper, Coke

Efficient Mining of Association Rule

Apriori algorithm

Example 2: Find association rule with $\alpha = 0.6$ and $\gamma = 0.8$.

Phase I: Determine the set F of all frequent itemsets with $supp \geq 0.6$

1-itemset	$supp$
Bread	0.80
Milk	0.80
Diaper	0.80
Beer	0.60
Eggs	0.20
Coke	0.40

2-itemset	$supp$
Bread, Milk	0.60
Bread, Diaper	0.60
Bread, Beer	0.40
Milk, Diaper	0.60
Milk, Beer	0.40
Diaper, Beer	0.60

3-itemset	$supp$
Bread, Milk, Diaper	0.40

Efficient Mining of Association Rule

Apriori algorithm

Example 2: Find association rule with $\alpha = 0.6$ and $\gamma = 0.8$.
Phase II: For each itemset $W \in F$, generate all rules $A \Rightarrow B$, with $A \cup B = W$ and $conf(A \Rightarrow B) \geq 0.8$.
F={{Bread, Milk}, {Bread, Diaper}, {Milk, Diaper}, {Diaper, Beer}}

rule	$conf$
{Bread}⇒{Milk}	0.75
{Milk}⇒{Bread}	0.75
{Bread}⇒{Diaper}	0.75
{Diaper}⇒{Bread}	0.75
{Milk}⇒{Diaper}	0.75
{Diaper}⇒{Milk}	0.75
{Diaper}⇒{Beer}	0.75
{Beer}⇒{Diaper}	1.00

Efficient Mining of Association Rule

Apriori algorithm

Example 2: Find association rule with $\alpha = 0.6$ and $\gamma = 0.8$.
Phase II: For each itemset $W \in F$, generate all rules $A \Rightarrow B$, with $A \cup B = W$ and $conf(A \Rightarrow B) \geq 0.8$.
Based on $\gamma = 0.8$, only rule exceeding these thresholds will be retained.

rule	$supp$	$conf$	$lift$
{Beer}⇒{Diaper}	0.6	1.00	1.25

Thank you!