Association Rule Mining

Key Terms: **A→B (If A, then B)**

Itemset: A collection of items. `{Milk, Bread, Butter}`

Support: $Support(A) = \frac{Transactions\ contain\ A}{Total\ Transactions}$

Confidence: $Confidence(A \rightarrow B) = \frac{Support(A \cup B)}{Support(A)}$

Sequential: ordered sets of items. <(A, B),(C),(A,D)>

Apriori Algorithm

Identify frequent itemset:

1. Start with 1-itemsets (individual items).
2. Calculate support and filter out items below `min_sup`
3. Use freq. 1-itemsets to generate candidate 2-itemsets, then 3-itemsets...

Generate Association rules:

1. For each freq. itemset, generate rules and calculate confidence
2. Keep rules with conf. above `min_conf`.

$$Interest(A \rightarrow B) = \frac{P(A \wedge B)}{P(A)P(B)} = \frac{P(A \wedge B)}{P(A)} \times \frac{1}{P(B)} = conf(A \rightarrow B) \times \frac{1}{P(B)}$$

AprioriAll Algorithm

Data Transformation: Convert the dataset into customer sequences.

Customer	Transaction Sequence
David	⟨(A,B),(B),(C)⟩
John	⟨(A),(B),(C,D)⟩

Sequence Generation:

1. Identify frequent **1-sequences**.
2. Use frequent **k-sequences** to generate candidate (k+1) sequences.

Support Calculation:

1. Count the support for each candidate sequence.
2. Retain sequences meeting the `min_sup` threshold.

Rule Generation:

1. Generate sequential rules of the form A→B.
2. Calculate confidence for each rule and filter by **`min_conf`**.

Sequence Phase:

1-sequence	Count	2-sequence	Count	3-sequence	Count	4-sequence	Count
<1>	5	~~<1 1>~~	~~1~~	**<1 2 3>**	**3**	**<1 2 3 2>**	**2**
<2>	4	**<1 2>**	**3**	**<1 3 2>**	**2**	~~<1 2 2 3>~~	~~0~~
<3>	4	~~<2 1>~~	~~0~~	**<1 2 2>**	**2**	~~<1 2 3 3>~~	~~0~~
<4>	3	**<1 3>**	**3**	~~<1 3 3>~~	~~0~~	~~<1 3 2 2>~~	~~0~~

1. A database has four transactions. Let min_sup=60% and min_conf=80%.

TID	Date	Items_bought
100	10/15/99	{K,A,D,B}
200	10/15/99	{D,A,C,E,B}
300	10/19/99	{C,A,B,E}
400	10/22/99	{B,A,D}

a) Find all frequent itemsets using Aprior algorithm.

b) List all of the strong association rules (with support s and confidence c) matching the following metarule (form), where X is a variable representing customers, and $item_i$ denotes variables representing items (e.g., "A", "B", etc.):

$\forall x \in$ transaction, buys$(X, item_1) \wedge$ buys$(X, item_2) \Rightarrow$ buys$(X, item_3)$ $[s,c]$

Suggested Answer: (Please try it first!)

a) min _sup=60% (i.e., ≥ 3 transactions)

1-itemset	Count	2-itemset	Count	3-itemset	Count
A	4	**A-B**	4	**A-B-D**	3
B	4	**A-D**	3		
~~C~~	~~2~~	**B-D**	3		
D	3				
~~E~~	~~2~~				
~~K~~	~~1~~				

The frequent 2-itemsets and 3-itemsets are bolded.

b)

Rule	Confidence
A,B⇒D	3/4
A,D⇒B	3/3
B,D⇒A	3/3

All except the first one are strong rules for *min_conf*=80%.

Sort Phase:

Customer ID	Customer Sequence
David	<(30 50),(50),(70)>
John	<(10 30),(50),(40 60 70),(50 90)>
Peter	<(30 50 70)>
Aaron	<(30),(30 50),(70),(50)>
Leon	<(30)>

Litemset Phase: min_sup = 25% (i.e. >= 2 customers)

Freq. Itemsets	Mapped to
(30)	1
(50)	2
(70)	3
(30 50)	4

Transformation Phase

Customer ID	Customer Sequence	Transformed Sequence	Mapping
David	<(30 50),(50),(70)>	<{(30),(50),(30 50)},{(50)},{(70)}>	<{1,2,4}{2}{3}>
John	<(10 30),(50),(40 60 70),(50 90)>	<{(30)},{(50)},{(70)},{(50)}>	<{1}{2}{3}{2}>
Peter	<(30 50 70)>	<{(30) (50) (70) (30 50)}>	<{1,2,3,4}>
Aaron	<(30),(30 50),(70),(50)>	<{(30)},{(30),(50),(30 50)},{(70)},{(50)}>	<{1}{1,2,4}{3}{2}>
Leon	<(30)>	<(30)>	<{1}>

(b) *Answer*

Rule	Support	Confidence	Strong or not?
1 → 2 3 2	2 (40%)	2/5 (40%)	No
1 2 → 3 2	2 (40%)	2/3 (~66.7%)	Yes
1 2 3 → 2	2 (40%)	2/3 (~66.7%)	Yes