MIS2502: Data Analytics
Association Rule Mining
Association Rule Mining

Find out which items predict the occurrence of other items

Also known as “affinity analysis” or “market basket” analysis

Uses

• What products are bought together?
• Amazon’s recommendation engine
• Telephone calling patterns
Examples of Association Rule Mining

• Market basket analysis/affinity analysis
  – What products are bought together?
  – Where to place items on grocery store shelves?

• Amazon’s recommendation engine
  – “People who bought this product also bought…”

• Telephone calling patterns
  – Who do a set of people tend to call most often?

• Social network analysis
  – Determine who you “may know”
### Market-Basket Transactions

<table>
<thead>
<tr>
<th>Basket</th>
<th>Items</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Bread, Milk</td>
</tr>
<tr>
<td>2</td>
<td>Bread, Diapers, Beer, Eggs</td>
</tr>
<tr>
<td>3</td>
<td>Milk, Diapers, Beer, Coke</td>
</tr>
<tr>
<td>4</td>
<td>Bread, Milk, Diapers, Beer</td>
</tr>
<tr>
<td>5</td>
<td>Bread, Milk, Diapers, Coke</td>
</tr>
</tbody>
</table>

#### Association Rules from these transactions

\[ X \rightarrow Y \]  
(antecedent → consequent)  
(aka LHS → RHS)

- \( \{\text{Diapers}\} \rightarrow \{\text{Beer}\} \)
- \( \{\text{Milk, Bread}\} \rightarrow \{\text{Diapers}\} \)
- \( \{\text{Beer, Bread}\} \rightarrow \{\text{Milk}\} \)
- \( \{\text{Bread}\} \rightarrow \{\text{Milk, Diapers}\} \)
Core idea: The itemset

**Itemset**
A group of items of interest
{Milk, Diapers, Beer}

**Association rules** express relationships between itemsets

\[ X \rightarrow Y \]
{Milk, Diapers} \rightarrow {Beer}

“when you have milk and diapers, you also have beer”
Support

- **Support count ($\sigma$)**
  - In how many baskets does the itemset appear?
  - $\sigma\{\text{Milk, Diapers, Beer}\} = 2$ (i.e., in baskets 3 and 4)

- **Support ($s$)**
  - Fraction of transactions that contain all items in $X \rightarrow Y$
  - $s\{\{\text{Milk, Diapers, Beer}\}\} = 2/5 = 0.4$

- You can calculate support for both $X$ and $Y$ separately
  - Support for $X = 3/5 = 0.6$
  - Support for $Y = 3/5 = 0.6$
Confidence

- **Confidence** is the strength of the association
  - Measures how often items in Y appear in transactions that contain X

\[ c = \frac{\sigma(X \rightarrow Y)}{\sigma(X)} = \frac{\sigma(\text{Milk, Diapers, Beer})}{\sigma(\text{Milk, Diapers})} = \frac{2}{3} = 0.67 \]

This says 67% of the times when you have milk and diapers in the itemset you also have beer!

<table>
<thead>
<tr>
<th>Basket</th>
<th>Items</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Bread, Milk</td>
</tr>
<tr>
<td>2</td>
<td>Bread, Diapers, Beer, Eggs</td>
</tr>
<tr>
<td>3</td>
<td>Milk, Diapers, Beer, Coke</td>
</tr>
<tr>
<td>4</td>
<td>Bread, Milk, Diapers, Beer</td>
</tr>
<tr>
<td>5</td>
<td>Bread, Milk, Diapers, Coke</td>
</tr>
</tbody>
</table>

\( c \) must be between 0 and 1
1 is a complete association
0 is no association
Calculating and Interpreting Confidence

<table>
<thead>
<tr>
<th>Association Rule (a→b)</th>
<th>Confidence (a→b)</th>
<th>What it means</th>
</tr>
</thead>
<tbody>
<tr>
<td>{Milk,Diapers} → {Beer}</td>
<td>2/3 = 0.67</td>
<td>• 2 baskets have milk, diapers, beer</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• 3 baskets have milk and diapers</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• So, 67% of the baskets with milk and diapers also have beer</td>
</tr>
<tr>
<td>{Milk,Beer} → {Diapers}</td>
<td>2/2 = 1.0</td>
<td>• 2 baskets have milk, diapers, beer</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• 2 baskets have milk and beer</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• So, 100% of the baskets with milk and beer also have diapers</td>
</tr>
<tr>
<td>{Milk} → {Diapers,Beer}</td>
<td>2/4 = 0.5</td>
<td>• 2 baskets have milk, diapers, beer</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• 4 baskets have milk</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• So, 50% of the baskets with milk also have diapers and beer</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Basket</th>
<th>Items</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Bread, Milk</td>
</tr>
<tr>
<td>2</td>
<td>Bread, Diapers, Beer, Eggs</td>
</tr>
<tr>
<td>3</td>
<td>Milk, Diapers, Beer, Coke</td>
</tr>
<tr>
<td>4</td>
<td>Bread, Milk, Diapers, Beer</td>
</tr>
<tr>
<td>5</td>
<td>Bread, Milk, Diapers, Coke</td>
</tr>
</tbody>
</table>
But don’t blindly follow the numbers

i.e., high confidence suggests a strong association...

- But this can be deceptive
- Consider \{Bread\} → \{Diapers\}
  - Support for the total itemset is 0.6 (3/5)
  - And confidence is 0.75 (3/4) – pretty high
  - But is this just because both are frequently occurring items (s=0.8)?
- You’d almost *expect* them to show up in the same baskets by chance
Lift

Takes into account how co-occurrence differs from what is expected by chance – i.e., if items were selected independently from one another

\[
Lift = \frac{s(X \rightarrow Y)}{s(X) \times s(Y)}
\]
Lift Example

- What’s the lift for the rule: \{Milk, Diapers\} → \{Beer\}

- So \( X = \{\text{Milk, Diapers}\} \)
  \( Y = \{\text{Beer}\} \)

\[
s(\{\text{Milk, Diapers, Beer}\}) = 2/5 = 0.4 \\
s(\{\text{Milk, Diapers}\}) = 3/5 = 0.6 \\
s(\{\text{Beer}\}) = 3/5 = 0.6
\]

So

\[
Lift = \frac{s(X \rightarrow Y)}{s(X) \times s(Y)}
\]

\[
Lift = \frac{0.4}{0.6 \times 0.6} = \frac{0.4}{0.36} = 1.11
\]

When Lift > 1, the occurrence of \( X \rightarrow Y \) together is more likely than what you would expect by chance.

<table>
<thead>
<tr>
<th>Basket</th>
<th>Items</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Bread, Milk</td>
</tr>
<tr>
<td>2</td>
<td>Bread, Diapers, Beer, Eggs</td>
</tr>
<tr>
<td>3</td>
<td>Milk, Diapers, Beer, Coke</td>
</tr>
<tr>
<td>4</td>
<td>Bread, Milk, Diapers, Beer</td>
</tr>
<tr>
<td>5</td>
<td>Bread, Milk, Diapers, Coke</td>
</tr>
</tbody>
</table>
Another example

**What is the effect of Netflix on Cable TV?**

<table>
<thead>
<tr>
<th></th>
<th>Netflix</th>
<th>Cable TV</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>No</td>
<td>200</td>
<td>3800</td>
</tr>
<tr>
<td>Yes</td>
<td>8000</td>
<td>1000</td>
</tr>
</tbody>
</table>

Total = 200 + 3800 + 8000 + 1000 = 13000

People with **both services** = 1000/13000 → 7%
People with **Cable TV** = (8000+1000)/13000 → 69%
People with **Netflix** = (3800+1000)/13000 → 37%

\[
Lift = \frac{0.07}{0.69 \times 0.37} = \frac{0.07}{0.24} = 0.25
\]

Having one negatively affects the purchase of the other (lift closer to 0 than 1)
Selecting the rules

• We know how to calculate the measures for each rule
  – Support
  – Confidence
  – Lift

• Then we set up **thresholds** for the minimum rule strength we want to accept

The steps

• List all possible association rules
• Compute the support and confidence for each rule
• Drop rules that don’t make the thresholds
• Use lift to further check the association
Once you are confident in a rule, take action

\{\text{Milk, Diapers}\} \rightarrow \{\text{Beer}\}

Possible Marketing Actions

- Create “New Parent Coping Kits” of beer, milk, and diapers
- What are some others?
Summary

• Support, confidence, and lift
  – Explain what each means
    • Can you have high confidence and low lift?
  – How to compute

• In-Class Exercises:
  – Part 1: Computing Confidence, Support, and Lift
  – Part 2: Association Rule Mining Using R