Association Rules

- Unsupervised learning but complementary to data exploration in clustering.
- The goal is to find “weak implications” in the data that have “non-negligible coverage”
- Useful in marketing, in understanding application data, as feature generator for supervised learning.

Data Model

- Following the market-basket application
- We assume a table where
  - Each row is a “transaction”
  - Each column is an “item”
- Table entries are in \{0,1\} i.e., discrete
- A transaction can be seen to represent the corresponding set of items

Text from paper by [AIS93] that introduced the topic

Association Rules

- Find all rules that have “Diet Coke” as consequent. These rules may help plan what the store should do to boost the sale of Diet Coke.
- Find all rules that have “bagels” in the antecedent. These rules may help determine what products may be impacted if the store discontinues selling bagels.
- Find all rules that have “sausage” in the antecedent and “mustard” in the consequent. This query can be phrased alternatively as a request for the additional items that have to be sold together with sausage in order to make it highly likely that mustard will also be sold.
Association Rules

- What are useful rules?
  - At least ...% coverage: support
  \[ \text{support}(X) = \frac{\# \text{transactions including } X}{\# \text{transactions}} \]
  - At least ...% predictive: confidence
  \[ \text{confidence}(X \Rightarrow Y) = \frac{\text{support}(X \cup Y)}{\text{support}(X)} \]

- Applications and data characteristics from some early papers:

<table>
<thead>
<tr>
<th>Date</th>
<th>#transact</th>
<th>#items</th>
<th>transactions</th>
<th>Avg size</th>
</tr>
</thead>
<tbody>
<tr>
<td>Market</td>
<td>50K</td>
<td>13K</td>
<td>1-100</td>
<td>10</td>
</tr>
<tr>
<td>Basket</td>
<td>50K</td>
<td>500</td>
<td>1-267</td>
<td>2.5</td>
</tr>
<tr>
<td>Web-Clicks</td>
<td>100K</td>
<td>2000</td>
<td>70</td>
<td>70</td>
</tr>
</tbody>
</table>

- In more demanding applications data does not fit in memory

Association Rules

- More data characteristics

- Too many sets and rules ...

- Huge data: technology challenge making use of memory hierarchy
- Huge data: algorithmic challenge to process it efficiently
- Huge output: conceptual challenge to identify “most interesting” rules

- A concrete task:
  - find all rules with frequency at least \( f \)
  - confidence at least \( c \)
- How can we do this?
Association Rules

• A concrete task 1:
  • find all rules with frequency at least \( f \) and confidence at least \( c \).
  • How can we do this?
  • If \((X \rightarrow Y)\) satisfies conditions then \((X+Y)\) must also have frequency at least \( f \).
• A concrete task 2:
  • find all sets \( Z \) with frequency at least \( f \).

Frequent Set Mining

• find all sets \( Z \) with frequency at least \( f \)
  • How can we do this?
  • Main insight: anti-monotonicity
    if set \( Z \) is frequent then all its subsets are also frequent
  • Algorithmic ideas?

Lattice Structure of Freq Sets

• Notice the notions of
  - positive border
  - negative border
  • that are implicit in the monotonicity property and in the view via the lattice
  • The borders capture all the frequent sets. Some algorithms attempt to find these directly.

Level-wise (Apriori) Algorithm

• level=1
  • \( \text{cands}[1] = \) Sets with single items
  • While \( \text{cands}[\text{level}] \) not empty
    1. \( \text{Calc support for \text{cands}[\text{level}] \)  
    2. \( \text{freq}[\text{level}] = \text{cands}[\text{level}] \text{ with high support} \)
    3. \( \text{cands}[\text{level}+1] = \text{generate from \text{freq}[\text{level}]} \)
    4. \( \text{level} = \text{level}+1 \)
Calculating support for candidates

- Basic implementation of step 1
- For each row $R$
  - For each candidate $X$
    - if $X$ subset of $R$ then: $\text{count}[X]+=1$
- One pass over database
- Improve run time via trie data structure that captures set of candidates

Generating candidates

- Monotonicity can be used to generate and prune potential candidates
- Use trie or lexicographical ordering to identify potential candidates
- Prune via subset relation
- Prune via upper/lower bounds on frequency (we skip details of this idea)

Vertical View

- View each item as a set of transactions
- This directly captures support
- Support of a set is the intersection of support of parents
- This incurs large space cost for storing support
- DFS recursive exploration avoids this and leads to efficient algorithm (we skip the details of this)

Which Rules are Interesting?

- Confidence can often be misleading
  \[ \text{confidence}(X \Rightarrow Y) = \frac{\text{support}(X \cup Y)}{\text{support}(X)} \]
  - If $p(B)$ is large
    - $p(B|A)=p(B)$ i.e., independent
  - $\text{Confidence}(A \Rightarrow B)$ is still large

Which Rules are Interesting?

- Lift measures dist from independence
  \[ \text{Lift}(X \Rightarrow Y) = \frac{\text{support}(X \cup Y)}{\text{support}(X) \times \text{support}(Y)} \]
- Conviction aims at "implication"
  \[ \text{Conviction}(X \Rightarrow Y) = \frac{\text{freq}(X)(1-\text{freq}(Y))}{\text{freq}(X)-\text{freq}(X \cup Y)} \]
- Interpret as inverse of $\text{Lift}(X \text{ and Not } Y)$

Which Rules are Interesting?

<table>
<thead>
<tr>
<th>Implication rule</th>
<th>Conviction</th>
</tr>
</thead>
<tbody>
<tr>
<td>Any two year olds don't work</td>
<td>2.91</td>
</tr>
<tr>
<td>Unemployed people don't earn income from work</td>
<td>2.41</td>
</tr>
<tr>
<td>Men don't give birth</td>
<td>2.04</td>
</tr>
<tr>
<td>People who are not in the military and are not looking for work and had work this year (1990, the year of the census) currently have civilian employment</td>
<td>2.25</td>
</tr>
<tr>
<td>People who are not in the military and who worked last week are not limited in their work by a disability</td>
<td>2.11</td>
</tr>
<tr>
<td>Less of household do not have personal care limitations</td>
<td>2.88</td>
</tr>
<tr>
<td>People not in the US without personal care limitations are not in the military</td>
<td>2.78</td>
</tr>
<tr>
<td>African American women are not in the military</td>
<td>2.81</td>
</tr>
<tr>
<td>African American women are not in the military</td>
<td>2.43</td>
</tr>
<tr>
<td>African American women are not in the military</td>
<td>2.36</td>
</tr>
<tr>
<td>Table 3: Sample Implication Rules from Census Data</td>
<td></td>
</tr>
</tbody>
</table>
• One way to generate enriched features for supervised learning is to generate frequent sets (because they occur and thus have a chance of making a difference)
• Very successful in frequent sub-graph mining, which extends the topic of this lecture to graphs, and its application to classifying molecules

Summary
• Association rules: a novel form of data exploration with different goals from previous supervised and unsupervised learning
• Algorithmic/computational challenge
• Frequent set mining as an important subtask
• Property: Anti-monotonicity
• Level-wise algorithm
• Many alternative algorithms