Data Mining Algorithms

CS102
Fall 2017
Big Data Tools and Techniques

- **Basic Data Manipulation and Analysis**
  Performing well-defined computations or asking well-defined questions ("queries")

- **Data Mining**
  Looking for patterns in data

- **Machine Learning**
  Using data to make inferences or predictions

- **Data Visualization**
  Graphical depiction of data

- **Data Collection and Preparation**
Looking for patterns in data

Similar to unsupervised machine learning
  • Popularity predates popularity of machine learning
  • “Data mining” often associated with specific data types and patterns

We will focus on “market-basket” data
  • Widely applicable (despite the name)

And two types of data mining patterns
  • Frequent item-sets
  • Association rules
Other Data and Patterns

Other types of data
  • Graphs/networks
  • Streams
  • Text ("text mining")

Other patterns
  • Similar items
  • Structural patterns in large graphs/networks
  • Clusters, anomalies

Specific techniques for each one
(In)Famous Early Success Stories

- Victoria’s Secret
- Walmart
- Beer & Diapers
Market-Basket Data

Originated with retail data
- Each shopper buys “market basket” of groceries
- Mine data for patterns in buying habits

General definition
- Domain of items
- Transaction - one or more items occurring together
- Dataset - set of transactions (usually large)
# Market-Basket Examples

<table>
<thead>
<tr>
<th>Items</th>
<th>Transaction</th>
</tr>
</thead>
<tbody>
<tr>
<td>Groceries</td>
<td>Grocery cart</td>
</tr>
<tr>
<td>Online goods</td>
<td>Virtual shopping cart</td>
</tr>
<tr>
<td>University courses</td>
<td>Student transcript</td>
</tr>
<tr>
<td>University students</td>
<td>Party</td>
</tr>
<tr>
<td>Movies</td>
<td>Person</td>
</tr>
<tr>
<td>Symptoms</td>
<td>Patient</td>
</tr>
<tr>
<td>Menu items</td>
<td>Restaurant customer</td>
</tr>
<tr>
<td>Words</td>
<td>Document</td>
</tr>
</tbody>
</table>
Data Mining Algorithms

**Frequent Item-Sets** - sets of items that occur frequently together in transactions

- Groceries bought together
- Courses taken by same students
- Students going to parties together
- Movies watched by same people

**Association Rules** - When certain items occur together, another item frequently occurs with them

- Shoppers who buy phone + charger also buy case
- Students who take Databases also take Machine Learning
- Diners who order curry and rice also order bread
Frequent Item-Sets

Sets of items that occur frequently together in transactions

- How large is a “set”?
- What does “frequently” mean?
Frequent Item-Sets

Sets of items that occur frequently together in transactions

- How large is a “set”? Usually specify a minimum \textit{min-set-size}. Possibly also a maximum \textit{max-set-size}.

- What does “frequently” mean? Notion of support.
Support for a set of items $S$ in a dataset of transactions is the fraction of the transactions containing $S$:

\[
\frac{\text{# of transactions containing } S}{\text{total # of transactions}}
\]

Specify \textit{support-threshold} for frequent item-sets

Only return sets where \textit{support} > \textit{support-threshold}
Your Turn

Transactions:

- T1: milk, eggs, juice
- T2: milk, juice, cookies
- T3: eggs, chips
- T4: milk, eggs
- T5: milk, juice, cookies, chips

What are the frequent item-sets if:

- \( \text{min-set-size} = 2 \) (no \( \text{max-set-size} \))
- \( \text{support-threshold} = 0.3 \)

Support: \[
\frac{\# \text{ of transactions containing } S}{\text{total \# of transactions}}
\]
Computing Frequent Item-Sets

“Apriori” algorithm

Efficiency relies on the following property:

If $S$ is a frequent item-set satisfying support-threshold $t$, then every subset of $S$ is also a frequent item-set satisfying support-threshold $t$.

Or the contrapositive:

If $S$ is not a frequent item-set satisfying support-threshold $t$, then no superset of $S$ can be a frequent item-set satisfying support-threshold $t$. 
Association Rules

When a set of items $S$ occurs together, another item $i$ frequently occurs with them

$S \rightarrow i$

- How large is a “set”?
- What does “occurs together” mean?
- What does “frequently occurs with them” mean?
Association Rules

When a set of items $S$ occurs together, another item $i$ frequently occurs with them

$S \rightarrow i$

- How large is a “set”? Usually specify a minimum $\text{min-set-size}$ for $S$
  Possibly also a maximum $\text{max-set-size}$ for $S$

- What does “occurs together” mean?

- What does “frequently occurs with them” mean?
Association Rules

When a set of items $S$ occurs together, another item $i$ frequently occurs with them

$S \rightarrow i$

- **How large is a “set”?**
  - Usually specify a minimum $\text{min-set-size}$ for $S$
  - Possibly also a maximum $\text{max-set-size}$ for $S$

- **What does “occurs together” mean?**
  - Notion of support

- **What does “frequently occurs with them” mean?**
  - Notion of confidence
Support and Confidence

Support for association rule $S \rightarrow i$ in a dataset of transactions is fraction of transactions containing $S$:

\[
\frac{\text{# of transactions containing } S}{\text{total # of transactions}}
\]

Confidence for association rule $S \rightarrow i$ in a dataset of transactions is the fraction of transactions containing $S$ that also contain $i$:

\[
\frac{\text{# of transactions containing } S \text{ and } i}{\text{# of transactions containing } S}
\]
Support and Confidence

Specify *support*-threshold and *confidence*-threshold for association rules

Only return rules where:

\[
\text{support} > \text{support}-\text{threshold} \quad \text{and} \quad \text{confidence} > \text{confidence}-\text{threshold}
\]
Your Turn

Transactions:

T1: milk, eggs, juice
T2: milk, juice, cookies
T3: eggs, chips
T4: milk, eggs
T5: milk, juice, cookies, chips

What are the association rules $S \rightarrow i$ if:

- $\text{min-set-size} = 1$ (no $\text{max-set-size}$)
- $\text{support-threshold} = 0.5$
- $\text{confidence-threshold} = 0.5$

Support:

\[
\frac{\text{# of transactions containing } S}{\text{total # of transactions}}
\]

Confidence:

\[
\frac{\text{# of transactions containing } S \text{ and } i}{\text{# of transactions containing } S}
\]
Computing Association Rules

1. Use frequent item-sets to find left-hand sides $S$ satisfying support threshold
2. Then extend to find right-hand sides $S \rightarrow i$ satisfying confidence threshold

NOT a property:

If $S \rightarrow i$ is an association rule satisfying support-threshold $t$ and confidence-threshold $c$, and $S' \subseteq S$, then $S' \rightarrow i$ is an association rule satisfying support-threshold $t$ and confidence-threshold $c$. Why Not?
Association Rules: Lift

Association rule $S \rightarrow i$ might have high confidence because item $i$ appears frequently, not because it’s associated with $S$.

Lift for association rule $S \rightarrow i$ in a dataset of transactions is the fraction of transactions containing $S$ that also contain $i$, divided by the overall frequency of $i$:

$$\frac{\text{#trans containing } S \text{ and } i}{\text{#trans containing } S} \div \frac{\text{#trans containing } i}{\text{total #trans}}$$
Lift: Examples

Transactions:
- T1: milk, eggs, juice
- T2: milk, juice, cookies
- T3: eggs, chips
- T4: milk, eggs
- T5: milk, juice, cookies, chips

juice → cookies  \[ \text{Lift} = \frac{2/3}{2/5} = \frac{10}{6} = 1.67 \]

eggs → milk  \[ \text{Lift} = \frac{2/3}{4/5} = \frac{10}{12} = 0.83 \]

Lift: \[ \frac{\#\text{trans containing } S \text{ and } i}{\#\text{trans containing } S} \div \frac{\#\text{trans containing } i}{\text{total #trans}} \]
Data Mining Algorithms

CS102
Fall 2017