Data Mining Algorithms

CS102
Spring 2020

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 build models and make predictions
- Data Visualization
  Graphical depiction of data
- Data Collection and Preparation

Data Mining

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

- Networks/graphs
- Streams
- Text ("text mining")

Specific techniques for each one

Other patterns

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

(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

Items	Transaction

Market-Basket Examples

Items	Transaction
Groceries	

Market-Basket Examples

Items	Transaction
Groceries	Grocery cart

Market-Basket Examples

Items	Transaction
Groceries	Grocery cart
Online goods	

Market-Basket Examples

Items	Transaction
Groceries	Grocery cart
Online goods	Virtual shopping cart

Market-Basket Examples

Items	Transaction
Groceries	Grocery cart
Online goods	Virtual shopping cart
University courses	

Market-Basket Examples

Items	Transaction
Groceries	Grocery cart
Online goods	Virtual shopping cart
University courses	Student transcript

Market-Basket Examples

Items	Transaction
Groceries	Grocery cart
Online goods	Virtual shopping cart
University courses	Student transcript
University students	

Market-Basket Examples

Items	Transaction
Groceries	Grocery cart
Online goods	Virtual shopping cart
University courses	Student transcript
University students	Party

Market-Basket Examples

Items	Transaction
Groceries	Grocery cart
Online goods	Virtual shopping cart
University courses	Student transcript
University students	Party
Movies	

Market-Basket Examples

Items	Transaction
Groceries	Grocery cart
Online goods	Virtual shopping cart
University courses	Student transcript
University students	Party
Movies	Person

Market-Basket Examples

Items	Transaction
Groceries	Grocery cart
Online goods	Virtual shopping cart
University courses	Student transcript
University students	Party
Movies	Person
Symptoms	

Market-Basket Examples

Items	Transaction
Groceries	Grocery cart
Online goods	Virtual shopping cart
University courses	Student transcript
University students	Party
Movies	Person
Symptoms	Patient

Market-Basket Examples

Items	Transaction
Groceries	Grocery cart
Online goods	Virtual shopping cart
University courses	Student transcript
University students	Party
Movies	Person
Symptoms	Patient
Menu items	

Market-Basket Examples

Items	Transaction
Groceries	Grocery cart
Online goods	Virtual shopping cart
University courses	Student transcript
University students	Party
Movies	Person
Symptoms	Patient
Menu items	Restaurant customer

Market-Basket Examples

Items	Transaction
Groceries	Grocery cart
Online goods	Virtual shopping cart
University courses	Student transcript
University students	Party
Movies	Person
Symptoms	Patient
Menu items	Restaurant customer
Words	

Market-Basket Examples

Items	Transaction
Groceries	Grocery cart
Online goods	Virtual shopping cart
University courses	Student transcript
University students	Party
Movies	Person
Symptoms	Patient
Menu items	Restaurant customer
Words	Document

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 *min-set-size*
  - Possibly also a maximum *max-set-size*
- What does "frequently" mean?
  - Notion of support

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 *support-threshold* for frequent item-sets

Only return sets where

support > *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:

- *min-set-size* = 2 (no *max-set-size*)
- *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 inverse:

> 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 *min-set-size* for S
  - Possibly also a maximum *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 *min-set-size* for S
  - Possibly also a maximum *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:

support > *support-threshold* and

confidence > *confidence-threshold*

Your Turn

Transactions:

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

Support:

$$\frac{\text{\# of transactions containing } S}{\text{total \# of transactions}}$$

Reminder: support and confidence must be > threshold, not ≥

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

- *min-set-size* = 1 (no *max-set-size*)
- *support-threshold* = 0.5
- *confidence-threshold* = 0.5

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:

Why Not?

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 an association rule satisfying support-threshold t and confidence-threshold c.

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 :

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

Lift: Examples

Transactions:

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

Lift = 1: no association
Lift > 1: association
Lift < 1: anti-association

juice → cookies Lift = (2/3) ÷ (2/5) = 10/6 = 1.67

eggs → milk Lift = (2/3) ÷ (4/5) = 10/12 = 0.83

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

Data Mining Algorithms

CS102
Spring 2020