David Eccles School of Business

THE UNIVERSITY OF UTAH

Association Rule Mining

Data Mining for Business Intelligence

AGENDA

- Motivating examples and main approach
- Set basics and important metrics
  - Support, confidence and lift
- Apriori property and Apriori algorithm
- Data preparation decisions
- Extension of association rules

SET AND NOTATION

***Set** is a collection of objects*

E.g., set A = {3,5} and set B= {1,3,5}

***Elements** of a set are the objects belonging to (∈) it.*

E.g., 3∈{3,5}, 3 ∈ {1,3,5}, 3 ∈ A and 3 ∈ B

*Set X is **a subset of** set Y if every element in X belongs to Y, denoted as X⊆ Y.*

E.g., A⊆B or {3,5}⊆{1,3,5}

SET PROPERTIES, UNION AND SIZE

Two properties of set

Uniqueness of elements

E.g., set A = {3,5} and set B= {1,3,5}

{3,3,5} reduced to A
{1,3,3,5} reduced to B

Order of elements

Apriori algorithm uses ordered sets

E.g., {3,1,5} = {1,3,5}

Set union: *X U Y includes unique elements of X and Y*

E.g., {3,5,7} U {1,3,5} = {1,3,5,7}

Size of a set

The number of elements in a set.

E.g., size of {1,3,5} = 3

DEFINITIONS AND FIRST METRIC: SUPPORT

***Itemset X**: A set of items.*

E.g., {eggs, milk}

***K-itemset**: An itemset of size K.*

First Metric: Support **support(X) = freq(X)/D**

***Support (X)**: The ratio of the # of transactions purchasing X to (÷) the total # of transactions in the DB (also probability (X))*

EXAMPLE: SUPPORT

An example in the “single” or “long” file format

TID	CID	Item	Price	Date
101	201	Computer	1500	1/4/99
101	201	MS Office	300	1/4/99
101	201	MCSE Book	100	1/4/99
102	201	Hard disk	500	1/8/99
102	201	MCSE Book	100	1/8/99
103	202	Computer	1500	1/21/99
103	202	Hard disk	500	1/2199
103	202	MCSE Book	100	1/2199

support ({Computer}) = 2/3 support ({Hard disk}) = 2/3 support ({MS Office}) = 1/3 support ({MCSE Book}) = 3/3

support ({Computer, Hard disk}) = 1/3 support ({Computer, MS Office}) = 1/3 support ({Hard disk, MS Office}) = 0

EXAMPLE: SUPPORT

An example in the “wide” file format

Date	TID	CID	Computer	MS Office	MCSE Book	Hard disk
1/4/1999	101	201	1	1	1	0
1/8/1999	102	201	0	0	1	1
1/21/1999	103	202	1	0	1	1

support ({Computer}) = 2/3 support ({Hard disk}) = 2/3 support ({MS Office}) = 1/3 support ({MCSE Book}) = 3/3

support ({Computer, Hard disk}) = 1/3 support ({Computer, MS Office}) = 1/3 support ({Hard disk, MS Office}) = 0

ANOTHER EXAMPLE

An example in the “transaction” or “wide” file format

date	trans	cust id	egg	milk	bread	soda
11/1/2010	1	1	1	1	0	0
11/1/2010	2	2	1	1	1	0
11/1/2010	3	3	0	0	1	0
11/2/2010	4	3	1	0	1	1
11/2/2010	5	4	0	1	1	0

SUPPORT FOR ITEMSET

1-itemset

trans id	egg	milk	bread	soda
1	1	1	0	0
2	1	1	1	0
3	0	0	1	0
4	1	0	1	1
5	0	1	1	0
Support	0.6	0.6	0.8	0.2

3 out of 5 transactions contain egg or milk

SUPPORT FOR ITEMSET

2-itemset

trans id	egg bread	egg soda	milk bread	milk soda	bread soda
1	0	0	0	0	0
2	1	0	1	0	0
3	0	0	0	0	0
4	1	1	0	0	1
5	0	0	1	0	0
Support	0.4	0.2	0.4	0	0.2

SUPPORT FOR ITEMSET

3-itemset

4-itemset

trans id	egg milk bread	egg milk soda	egg bread soda	milk bread soda	egg milk bread soda
1	0	0	0	0	0
2	1	0	0	0	0
3	0	0	0	0	0
4	0	0	1	0	0
5	0	0	0	0	0
Support	0.2	0	0.2	0	0

SUPPORT FOR ITEMSET

1-itemset | 2-itemset | 3-itemset | 4-itemset

trans id	egg	milk	bread	soda	egg milk	egg bread	egg soda	milk bread	milk soda	bread soda	egg milk bread	egg milk soda	egg bread soda	milk bread soda	egg milk bread soda
1	1	1	0	0	1	0	0	0	0	0	0	0	0	0	0
2	1	1	1	0	1	1	0	1	0	0	1	0	0	0	0
3	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
4	1	0	1	1	0	1	1	0	0	1	0	0	1	0	0
5	0	1	1	0	0	0	0	1	0	0	0	0	0	0	0
Support	0.6	0.6	0.8	0.2	0.4	0.4	0.2	0.4	0	0.2	0.2	0	0.2	0	0

3 out of 5 transactions contain egg or milk

2ND METRIC - CONFIDENCE

If two itemsets X and Y co-exist in a transaction DB,

For association rule (R):

***Support (R)**:*

The ratio of the # of transactions purchasing

Both X and Y to (÷) the total # of transactions in the DB

(also probability (R))

***Confidence (R)**:*

The ratio of the support of transactions purchasing both X and Y to (÷) the support of transactions purchasing X only.

Hence confidence(X -> Y) = support(X U Y) /support(X).

(also P(Y|X))

2ND METRIC - CONFIDENCE

If two itemsets X and Y co-exist in a transaction DB,

For association rule (R):

***Confidence (R)**:*

The ratio of the support of transactions purchasing both X and Y to (÷) the support of transactions purchasing X only.
Hence confidence(X -> Y) = support(X U Y) /support(X).

(also P(Y|X))

Example: Every transaction that has X also has Y.
Assume the support(X) = .6

support(X U Y) = .6

.6/.6 = 1.0

The probability of observing Y given we saw X is 1.0.

2ND METRIC - CONFIDENCE

If two itemsets X and Y co-exist in a transaction DB,

For association rule (R):

***Confidence (R)**:*

The ratio of the support of transactions purchasing both X and Y to (÷) the support of transactions purchasing X only.
Hence confidence(X -> Y) = support(X U Y) /support(X).

(also P(Y|X))

Example: No transaction that has X also has Y. Assume the support(X) = .6

support(X U Y) = .0

.0/.6 = 0.0

The probability of observing Y given we saw X is 0.

CONFIDENCE FOR ASSOCIATION RULES

egg -> milk	conf(egg->milk) = supp(egg U milk)/supp(egg)	.4/.6 = .67
milk -> egg	conf(milk->egg) = supp(milk U egg)/supp(milk)	.4/.6 = .67
egg -> bread	conf(egg->bread) = supp(egg U bread)/supp(egg)	.4/.6 = .67
bread -> egg	conf(bread->egg) = supp(bread U egg)/supp(bread)	.4/.8 = .5
egg -> soda	conf(egg->soda) = supp(egg U soda)/supp(egg)	.2/.6 = .33
milk-> bread	conf(milk->bread) = supp(milk U bread)/supp(milk)	.4/.6 = .67
bread -> soda	conf(bread->soda) = supp(bread U soda)/supp(bread)	.2/.8 = .25
egg -> milk, bread	conf(egg->milk,bread) = supp(egg U milk, bread) /supp(egg)	.2/.6 = .33
egg -> bread, soda	conf(egg->bread,soda) = supp(egg U bread,soda)/supp(egg)	.2/.6 = .33

INTERESTINGNESS LEVELS AND PATTERNS

- Minimum Support (S)
  - Large (frequent) itemsets:
    - Support(large itemsets) >= S
- Minimum Confidence (C)
  - Association rule R: X -> Y
  - X U Y is a large itemset
  - Confidence(R) >= C
- S and C: Decided by an analyst and vary by application

THE THIRD METRIC: LIFT

- Lift $(X \rightarrow Y)$=confidence$(X \rightarrow Y)$/support(Y)
  = P(Y|X)/P(Y)=P(X U Y)/P(X)/P(Y) = P(X U Y)/[P(X)*P(Y)]

- Lift is the ratio of the probability of X and Y occurring together to the probability that X and Y occurring independently.
  - If Lift $(X \rightarrow Y)=1$ then X and Y are independent
  - If Lift $(X \rightarrow Y)< 1$, then X and Y are negatively correlated
  - If Lift $(X \rightarrow Y)>1$, then X and Y are positively correlated

- Interesting association rules tend to have lift greater than 1 (or less than 1 in some cases).

LIFT EXAMPLES: A-> B Independence.

Lift $(X \rightarrow Y)$=confidence$(X \rightarrow Y)$/support(Y)

= P(Y|X)/P(Y)=P(X U Y)/P(X)/P(Y) = P(X U Y)/[P(X)*P(Y)]

Also if A and B are independent P(A and B) = P(A)*P(B)

A and B are independent			
Event =>	**A**	**B**	**A and B**
1	1	1	1
2		1	
3	1	1	1
4	1	1	1
5	1		
6	1	1	1
7		1	
8		1	
9		1	
10			
P(Event)	**0.5**	**0.8**	**0.4**
Lift(A -> B)	**P(B|A)**	**P(A|B)**	**P(A)*P(B)**
1	**0.8**	**0.5**	**0.4**

Lift $(A \rightarrow B)$=confidence$(A \rightarrow B)$/support(B)

= P(A U B)/[P(A)*P(B)]

=.4/.4 = 1.0

LIFT EXAMPLES: A-> B Negative Correlation.

Lift $(X \rightarrow Y)$=confidence$(X \rightarrow Y)$/support(Y)

= P(Y|X)/P(Y)=P(X U Y)/P(X)/P(Y) = P(X U Y)/[P(X)*P(Y)]

Also if A and B are independent P(A and B) = P(A)*P(B)

A and B are negatively correlated			
Event =>	**A**	**B**	**A and B**
1	1	1	1
2	1	1	1
3	1	1	1
4		1	
5	1		
6		1	
7		1	
8		1	
9		1	
10	1		
P(Event)	**0.5**	**0.8**	**0.3**
Lift(A -> B)	**P(B|A)**	**P(A|B)**	**P(A)*P(B)**
0.75	**0.6**	**0.375**	**0.4**

Lift $(A \rightarrow B)$=confidence$(A \rightarrow B)$/support(B)

= P(A U B)/[P(A)*P(B)]

=.3/.4 = 0.75

LIFT EXAMPLES: A-> B Positive Correlation.

Lift (X → Y)=confidence(X → Y)/support(Y)

= P(Y|X)/P(Y)=P(X U Y)/P(X)/P(Y) = P(X U Y)/[P(X)*P(Y)]

Also if A and B are independent P(A and B) = P(A)*P(B)

A and B are positively correlated			
Event =>	**A**	**B**	**A and B**
1	1	1	1
2	1	1	1
3	1	1	1
4	1	1	1
5			
6			
7		1	
8		1	
9		1	
10	1	1	1
P(Event)	**0.5**	**0.8**	**0.5**
Lift(A -> B)	**P(B|A)**	**P(A|B)**	**P(A)*P(B)**
1.25	**1**	**0.625**	**0.4**

Lift (A → B)=confidence(A → B)/support(B)

= P(A U B)/[P(A)*P(B)]

=.5/.4 = 1.25

David Eccles
School of Business
THE UNIVERSITY OF UTAH