

# Association analysis.

## Basic concepts

Lecture 12

# Classification rules: reminder

Outlook	Temp	Humidity	Windy	Play
Sunny	Hot	High	False	No
Sunny	Hot	High	True	No
Overcast	Hot	High	False	Yes
Rainy	Mild	High	False	Yes
Rainy	Cool	Normal	False	Yes
Rainy	Cool	Normal	True	No
Overcast	Cool	Normal	True	Yes
Sunny	Mild	High	False	No
Sunny	Cool	Normal	False	Yes
Rainy	Mild	Normal	False	Yes
Sunny	Mild	Normal	True	Yes
Overcast	Mild	High	True	Yes
Overcast	Hot	Normal	False	Yes
Rainy	Mild	High	True	No

R1: if humidity=normal and windy=false  
then yes  
R2: if outlook=overcast then yes  
R3: if temp=hot then no  
R4: if outlook=rainy and windy=true then no

- *LHS*: rule *antecedent* : in this case – combination of attribute-values
- *RHS*: rule *consequent*: in this case – class label

# Association rules: no class

Outlook	Temp	Humidity	Windy	Play
Sunny	Hot	High	False	No
Sunny	Hot	High	True	No
Overcast	Hot	High	False	Yes
Rainy	Mild	High	False	Yes
Rainy	Cool	Normal	False	Yes
Rainy	Cool	Normal	True	No
Overcast	Cool	Normal	True	Yes
Sunny	Mild	High	False	No
Sunny	Cool	Normal	False	Yes
Rainy	Mild	Normal	False	Yes
Sunny	Mild	Normal	True	Yes
Overcast	Mild	High	True	Yes
Overcast	Hot	Normal	False	Yes
Rainy	Mild	High	True	No

R1: if temp=cool then humidity=normal

- *LHS*: rule *antecedent* : combination of attribute-values
- *RHS*: rule *consequent*: combination of attribute-values

# Association rules: no class

Outlook	Temp	Humidity	Windy	Play
Sunny	Hot	High	False	No
Sunny	Hot	High	True	No
Overcast	Hot	High	False	Yes
Rainy	Mild	High	False	Yes
Rainy	Cool	Normal	False	Yes
Rainy	Cool	Normal	True	No
Overcast	Cool	Normal	True	Yes
Sunny	Mild	High	False	No
Sunny	Cool	Normal	False	Yes
Rainy	Mild	Normal	False	Yes
Sunny	Mild	Normal	True	Yes
Overcast	Mild	High	True	Yes
Overcast	Hot	Normal	False	Yes
Rainy	Mild	High	True	No

R1: if temp=cool then humidity=normal  
R2: if temp=hot then humidity=high

- *LHS*: rule *antecedent* :  
combination of attribute-values
- *RHS*: rule *consequent*:  
combination of attribute-values

# The goal: discover relationships, not prediction

Outlook	Temp	Humidity	Windy	Play
Sunny	Hot	High	False	No
Sunny	Hot	High	True	No
Overcast	Hot	High	False	Yes
Rainy	Mild	High	False	Yes
Rainy	Cool	Normal	False	Yes
Rainy	Cool	Normal	True	No
Overcast	Cool	Normal	True	Yes
Sunny	Mild	High	False	No
Sunny	Cool	Normal	False	Yes
Rainy	Mild	Normal	False	Yes
Sunny	Mild	Normal	True	Yes
Overcast	Mild	High	True	Yes
Overcast	Hot	Normal	False	Yes
Rainy	Mild	High	True	No

R1: if temp=cool then humidity=normal  
R2: if temp=hot then humidity=high

- The rules – one form of representing relationships between objects which point to their related behavior – appearing in the same observation

# Terminology: market basket

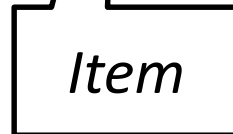
<i>TID</i>	<i>Items</i>
1	Bread, Coke, Milk
2	Beer, Bread
3	Beer, Coke, Diaper, Milk
4	Beer, Bread, Diaper, Milk
5	Coke, Diaper, Milk

Market  
basket

# Terminology: market basket

<i>TID</i>	<i>Items</i>
1	Bread, Coke, Milk
2	Beer, Bread
3	Beer, Coke, Diaper, Milk
4	Beer, Bread, Diaper, Milk
5	Coke, Diaper, Milk

*Item*



# Terminology: market basket

<i>TID</i>	<i>Items</i>
1	Bread, Coke, Milk
2	Beer, Bread
3	Beer, Coke, Diaper, Milk
4	Beer, Bread, Diaper, Milk
5	Coke, Diaper, Milk

Observation - *transaction*



# Terminology: market basket

<i>TID</i>	<i>Items</i>
1	Bread, Coke, Milk
2	Beer, Bread
3	Beer, Coke, Diaper, Milk
4	Beer, Bread, Diaper, Milk
5	Coke, Diaper, Milk

Combination of  $k$  items –  $k$ -itemset

{Coke, Diaper} – 2-itemset

If itemset  $A$  is a subset of items in transaction  $t_i$ ,  
we say  $t_i$  contains  $A$  or *supports*  $A$

# Terminology: support count

<i>TID</i>	<i>Items</i>
1	Bread, Coke, Milk
2	Beer, Bread
3	Beer, <b>Coke, Diaper</b> , Milk
4	Beer, Bread, Diaper, Milk
5	<b>Coke, Diaper</b> , Milk

Number of transactions which contain  
itemset  $A$  – *support count*

support count {Coke, Diaper} = 2

# Terminology: support

<i>TID</i>	<i>Items</i>
1	Bread, Coke, Milk
2	Beer, Bread
3	Beer, Coke, Diaper, Milk
4	Beer, Bread, Diaper, Milk
5	Coke, Diaper, Milk

Fraction of transactions which  
contain itemset  $A$  – *support*

$$\text{support } \{\text{Coke, Diaper}\} = 2/5$$

# Terminology: frequent itemset

<i>TID</i>	<i>Items</i>
1	Bread, Coke, Milk
2	Beer, Bread
3	Beer, Coke, Diaper, Milk
4	Beer, Bread, Diaper, Milk
5	Coke, Diaper, Milk

An itemset whose support is greater than or equal to a *minsup* threshold – *frequent itemset*

For *minsup*=40% frequent itemsets are:

{Coke, Diaper}

{Bread, Coke, Milk}

...

# Association rules

- **Association Rule**

- An implication expression of the form  $X \rightarrow Y$ , where  $X$  and  $Y$  are itemsets
- Example:  
 $\{\text{Milk, Diaper}\} \rightarrow \{\text{Beer}\}$

<i>TID</i>	<i>Items</i>
1	Bread, Milk
2	Bread, Diaper, Beer, Eggs
3	Milk, Diaper, Beer, Coke
4	Bread, Milk, Diaper, Beer
5	Bread, Milk, Diaper, Coke

- **Rule Evaluation Metrics ( $X \rightarrow Y$ )**

- **Support ( $s$ )**
  - Fraction of transactions that contain both  $X$  and  $Y$
- **Confidence ( $c$ )**
  - Measures how often items in  $Y$  appear in transactions that contain  $X$

**Example:**

$\{\text{Milk, Diaper}\} \Rightarrow \text{Beer}$

$$s = \frac{\sigma(\text{Milk, Diaper, Beer})}{|T|} = \frac{2}{5} = 0.4$$

$$c = \frac{\sigma(\text{Milk, Diaper, Beer})}{\sigma(\text{Milk, Diaper})} = \frac{2}{3} = 0.67$$

# Why Use Support and Confidence?

## Support

- A rule that has very low support may occur simply by chance.
- Support is often used to eliminate uninteresting rules.

## Confidence

- Measures the reliability of the inference made by a rule.
- For a rule  $X \rightarrow Y$ , the higher the confidence, the more likely it is for  $Y$  to be present in transactions that contain  $X$ .
- Confidence provides an estimate of the **conditional probability** of  $Y$  given  $X$ .

# Market basket analysis

- Marketing and Sales Promotion:

- Let the rule discovered be

*{Bagels, ... } --> {Potato Chips}*

- Potato Chips as consequent

Can be used to determine what should be done to boost its sales.

- Bagels in the antecedent

Can be used to see which products would be affected if the store discontinues selling bagels.

# Association Rule Mining Task

- Given a set of transactions  $\mathbf{T}$ , the goal of association rule mining is to find all rules having
  - $\text{support} \geq \text{minsup}$  threshold
  - $\text{confidence} \geq \text{minconf}$  threshold
- Brute-force approach:
  - List all possible association rules
  - Compute the support and confidence for each rule
  - Prune rules that fail the *minsup* and *minconf* thresholds

⇒ **Computationally prohibitive!**



# Brute-force approach:

## how many rules $R$

- Suppose there are  $d$  items. We first choose  $k$  of the items to form the left hand side of the rule. There are  $C_{d,k}$  ways for doing this.
- Now, there are  $C_{d-k,i}$  ways to choose the remaining items to form the right hand side of the rule, where  $1 \leq i \leq d-k$ .

We applied:

$$\sum_{i=1}^n \binom{n}{i} = 2^n - 1$$

We also have that :

$$(1+x)^d = \sum_{i=1}^d \binom{d}{i} x^{d-i} + x^d$$

For  $x = 2$

$$3^d = \sum_{i=1}^d \binom{d}{i} 2^{d-i} + 2^d$$

Therefore  $R = 3^d - 2^d - (2^d - 1) = 3^d - 2^{d+1} + 1$

$$R = \sum_{k=1}^d \binom{d}{k} \sum_{i=1}^{d-k} \binom{d-k}{i}$$

$$= \sum_{k=1}^d \binom{d}{k} (2^{d-k} - 1)$$

$$= \sum_{k=1}^d \binom{d}{k} 2^{d-k} - \sum_{k=1}^d \binom{d}{k}$$

$$= \sum_{k=1}^d \binom{d}{k} 2^{d-k} - (2^d - 1)$$

# Brute-force approach

- $R=3^d-2^{d+1}+1$
- For  $d=6$ ,  
 $3^6-2^7+1=602$  possible rules
- However, 80% of the rules are discarded after applying  $minsup=20\%$  and  $minconf=50\%$ , thus making most of the computations become wasted.
- So, it would be useful to prune the rules early without having to compute their support and confidence values.

<i>TID</i>	<i>Items</i>
1	Bread, Milk
2	Bread, Diaper, Beer, Eggs
3	Milk, Diaper, Beer, Coke
4	Bread, Milk, Diaper, Beer
5	Bread, Milk, Diaper, Coke

An initial step toward improving the performance:  
*decouple the support and confidence requirements.*

# Mining Association Rules

<i>TID</i>	<i>Items</i>
1	Bread, Milk
2	Bread, Diaper, Beer, Eggs
3	Milk, Diaper, Beer, Coke
4	Bread, Milk, Diaper, Beer
5	Bread, Milk, Diaper, Coke

## Example of Rules:

$\{\text{Milk, Diaper}\} \rightarrow \{\text{Beer}\}$  ( $s=0.4, c=0.67$ )

$\{\text{Milk, Beer}\} \rightarrow \{\text{Diaper}\}$  ( $s=0.4, c=1.0$ )

$\{\text{Diaper, Beer}\} \rightarrow \{\text{Milk}\}$  ( $s=0.4, c=0.67$ )

$\{\text{Beer}\} \rightarrow \{\text{Milk, Diaper}\}$  ( $s=0.4, c=0.67$ )

$\{\text{Diaper}\} \rightarrow \{\text{Milk, Beer}\}$  ( $s=0.4, c=0.5$ )

$\{\text{Milk}\} \rightarrow \{\text{Diaper, Beer}\}$  ( $s=0.4, c=0.5$ )

## Observations:

- All the above rules are binary partitions of the same itemset:  
 $\{\text{Milk, Diaper, Beer}\}$
- Rules originating from the same itemset have identical support but can have different confidence
- Thus, we may decouple the support and confidence requirements

*If the itemset is infrequent, then all six candidate rules can be pruned immediately without us having to compute their confidence values.*

# Mining Association Rules

- Two-step approach:
  1. **Frequent Itemset Generation**
    - Generate all itemsets whose **support**  $\geq$  **minsup** (these itemsets are called *frequent itemset*)
  2. **Rule Generation**
    - Generate high confidence rules from each frequent itemset, where each rule is a binary partitioning of a frequent itemset (these rules are called *strong rules*)

We focus first on **frequent itemset generation**.