# Association rule mining

# Association Rule Mining

- Given a set of transactions, find rules that will predict the occurrence of an item based on the occurrences of other items in the transaction databases, relational databases, and other information repositories..
- Proposed by Agrawal et al. in 1993.
- Assume all data are categorical.

## Market-Basket transactions

| *TID* | *Items* |
|---|---|
| **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 Association Rules

$\{Diaper\} \rightarrow \{Beer\}$,
$\{Milk, Bread\} \rightarrow \{Eggs, Coke\}$,
$\{Beer, Bread\} \rightarrow \{Milk\}$,

Implication means co-occurrence, not causality!

# Applications

- **Market Basket Analysis:** given a database of customer transactions, where each transaction is a set of items the goal is to find groups of items which are frequently purchased together.
- **Telecommunication** (each customer is a transaction containing the set of phone calls)
- **Credit Cards/ Banking Services** (each card/account is a transaction containing the set of customer's payments)
- **Medical Treatments** (each patient is represented as a transaction containing the ordered set of diseases)
- **Basketball-Game Analysis** (each game is represented as a transaction containing the ordered set of ball passes)

# Association Rule: Basic Concepts

- Given: (1) database of transactions, (2) each transaction is a list of items (purchased by a customer in a visit)
- Find: all rules that correlate the presence of one set of items with that of another set of items
  - E.g., *98% of people who purchase tires and auto accessories also get automotive services done*

# Association Rule Problem

- Given a database of transactions:

  Market basket example:

  Basket 1: {bread, cheese, milk}

  Basket 2: {apple, eggs, salt, yogurt}

  …

  Basket n: {biscuit, eggs, milk}

  Definitions:

  - An *item*: an article in a basket, or an attribute-value pair
  - A *transaction*: items purchased in a basket; it may have TID (transaction ID)
  - A *transactional dataset*: A set of transactions

- Find all the association rules:

| $X \Rightarrow Y$ | $s$ | $\alpha$ |
|---|---|---|
| Bread ⇒ PeanutButter | 60% | 75% |
| PeanutButter ⇒ Bread | 60% | 100% |
| Beer ⇒ Bread | 20% | 50% |
| PeanutButter ⇒ Jelly | 20% | 33.3% |
| Jelly ⇒ PeanutButter | 20% | 100% |
| Jelly ⇒ Milk | 0% | 0% |

# Association Rule— Example 1

| *TID* | *Items* |
|---|---|
| 1 | Bread, Milk |
| 2 | Bread, Diaper, Beer, Eggs |
| 3 | Milk, Diaper, Beer, Coke |
| 4 | Bread, Milk, Diaper, Beer |
| 5 | Bread, Milk, Diaper, Coke |

Example:

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

$$S = \frac{N(\text{Milk, Diaper, Beer})}{\text{N}} = \frac{2}{5} = 0.4$$

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

# Goal and key features

- Given a set of transactions T, the goal of association rule mining is to find all rules having
  - support $\geq$ *minsup* threshold
  - confidence $\geq$ *minconf* threshold

# Association rule — Exercise

## Transaction List

| | | | | |
|---|---|---|---|---|
| 1 | Milk | Egg | Bread | Butter |
| 2 | Milk | Butter | Egg | Ketchup |
| 3 | Bread | Butter | Ketchup | |
| 4 | Milk | Bread | Butter | |
| 5 | Bread | Butter | Cookies | |
| 6 | Milk | Bread | Butter | Cookies |
| 7 | Milk | Cookies | | |
| 8 | Milk | Bread | Butter | |
| 9 | Bread | Butter | Egg | Cookies |
| 10 | Milk | Butter | Bread | |
| 11 | Milk | Bread | Butter | |
| 12 | Milk | Bread | Cookies | Ketchup |

# Association rule — Exercise

- Here are a dozen sales transactions.
- The objective is to use this transaction data to find affinities between products, that is, which products sell together often.
- The support level will be set at 33%; confidence level will be set at 50%

**Transactions List**

| | | | | |
|---|---|---|---|---|
| 1 | Milk | Egg | Bread | Butter |
| 2 | Milk | Butter | Egg | Ketchup |
| 3 | Bread | Butter | Ketchup | |
| 4 | Milk | Bread | Butter | |
| 5 | Bread | Butter | Cookies | |
| 6 | Milk | Bread | Butter | Cookies |
| 7 | Milk | Cookies | | |
| 8 | Milk | Bread | Butter | |
| 9 | Bread | Butter | Egg | Cookies |
| 10 | Milk | Butter | Bread | |
| 11 | Milk | Bread | Butter | |
| 12 | Milk | Bread | Cookies | Ketchup |

| 1-item Sets | Frequency |
|---|---|
| Milk | 9 |
| Bread | 10 |
| Butter | 10 |
| Egg | 3 |
| Ketchup | 3 |
| Cookies | 5 |

| Frequent 1-item Sets | Frequency |
|---|---|
| Milk | 9 |
| Bread | 10 |
| Butter | 10 |
| Cookies | 5 |

## Transactions List

| | | | | |
|---|---|---|---|---|
| 1 | Milk | Egg | Bread | Butter |
| 2 | Milk | Butter | Egg | Ketchup |
| 3 | Bread | Butter | Ketchup | |
| 4 | Milk | Bread | Butter | |
| 5 | Bread | Butter | Cookies | |
| 6 | Milk | Bread | Butter | Cookies |
| 7 | Milk | Cookies | | |
| 8 | Milk | Bread | Butter | |
| 9 | Bread | Butter | Egg | Cookies |
| 10 | Milk | Butter | Bread | |
| 11 | Milk | Bread | Butter | |
| 12 | Milk | Bread | Cookies | Ketchup |

| 2-item Sets | Frequency |
|---|---|
| Milk, Bread | 7 |
| Milk, Butter | 7 |
| Milk, Cookies | 3 |
| Bread, Butter | 9 |
| Butter, Cookies | 3 |
| Bread, Cookies | 4 |

| Frequent 2-item Sets | Frequency |
|---|---|
| Milk, Bread | 7 |
| Milk, Butter | 7 |
| Bread, Butter | 9 |
| Bread, Cookies | 4 |

## Transactions List

| | | | | |
|---|---|---|---|---|
| 1 | Milk | Egg | Bread | Butter |
| 2 | Milk | Butter | Egg | Ketchup |
| 3 | Bread | Butter | Ketchup | |
| 4 | Milk | Bread | Butter | |
| 5 | Bread | Butter | Cookies | |
| 6 | Milk | Bread | Butter | Cookies |
| 7 | Milk | Cookies | | |
| 8 | Milk | Bread | Butter | |
| 9 | Bread | Butter | Egg | Cookies |
| 10 | Milk | Butter | Bread | |
| 11 | Milk | Bread | Butter | |
| 12 | Milk | Bread | Cookies | Ketchup |

Milk, Bread, Butter, Cookies

| **3-item Sets** | **Frequency** |
|---|---|
| Milk, Bread, Butter | 6 |
| Milk, Bread, Cookies | 1 |
| Bread, Butter, Cookies | 3 |
| Milk, Butter, Cookies | 2 |

| Frequent 3-item Sets | Frequency |
|---|---|
| Milk, Bread, Butter | 6 |

# Association rule mining- Subset creation

- Frequent 3-Item Set = I => {Milk, Bread, Butter}
- Non-Empty subset are
  - {{Milk}, {Bread}, {Butter}, {Milk, Bread}, {Milk, Butter}, {Bread, Butter}}
- How to form Association Rule…?
  - For every non-empty subset S of I, the association rule is,
    - **S →(I-S)**
    - **If support(I) / support(S) >= min_confidence**

# Association rule mining- Rule creation

- Non-Empty subset are
  - {{Milk}, {Bread}, {Butter}, {Milk, Bread}, {Milk, Butter}, {Bread, Butter}}
  - Min_Support = 30% and Min_Confidence = 60%

- Rule 1: {Milk} → {Bread, Butter} {S=50%, C=66.67%}
  - Support = 6/12 = 50%
  - Confidence = Support (Milk, Bread, Butter)/Support(Milk) = $\frac{6/12}{9/12}$ = 6/9 = 66.67% > 60%
  - Valid

- Rule 2: {Bread} → {Milk, Butter} {S=50%, C=60%}
  - Support = 6/12 = 50%
  - Confidence = Support (Milk, Bread, Butter)/Support(Bread) = 6/10 = 60% >= 60%
  - Valid

# Association rule mining- Rule creation

- Non-Empty subset are
  - {{Milk}, {Bread}, {Butter}, {Milk, Bread}, {Milk, Butter}, {Bread, Butter}}
  - Min_Support = 30% and Min_Confidence = 60%

- Rule 3: {Butter} → {Milk, Bread} {S=50%, C=60%}
  - Support = 6/12 = 50%
  - Confidence = Support (Milk, Bread, Butter)/Support(Butter) = 6/10 = 60%>=60
  - Valid

- Rule 4: {Milk, Bread} → {Butter} {S=50%, C=85.7%}
  - Support = 6/12 = 50%
  - Confidence = Support (Milk, Bread, Butter)/Support(Milk, Bread) = 6/7 = 85.7% > 60%
  - Valid

# Association rule mining- Rule creation

- Non-Empty subset are
  - {{Milk}, {Bread}, {Butter}, {Milk, Bread}, {Milk, Butter}, {Bread, Butter}}
  - Min_Support = 30% and Min_Confidence = 60%

- Rule 5: {Milk, Butter} → {Bread} {S=50%, C=85.7%}
  - Support = 6/12 = 50%
  - Confidence = Support (Milk, Bread, Butter)/Support(Milk, Butter) = 6/7 = 85.7% >= 60%
  - Valid

- Rule 6: {Bread, Butter} → {Milk} {S=50%, C=66.67%}
  - Support = 6/12 = 50%
  - Confidence = Support (Milk, Bread, Butter)/Support(Bread, Butter) = 6/9 = 66.67%>=60
  - Valid

# Association rule mining

- Coverage: Measure of how often a rule can be applied
  - Coverage(X=>Y) = Supp(X)
- Lift: A metric that measures the strength of an association between two items in a dataset
  - Lift(X=>Y) = Conf(X=>Y)/Supp(Y)
  - Lift value near 1 indicates X and Y almost often appear together as expected, greater than 1 means they appear together more than expected and less than 1 means they appear less than expected.
  - Greater lift values indicate stronger association.

# References


- Han, Jiawei, Jian Pei, and Micheline Kamber. *Data mining: concepts and techniques*. Elsevier, 2011.

- T. M. Mitchell, *Machine Learning*. McGraw-Hill Science, 1997