Data Mining for Business Analytics
MSBA 511

Association Rules

What are Association Rules?

Identifies patterns and relationships between items or events to understand "what goes with what".

Originated from the study of customer transaction databases to determine ***associations*** among items purchased hence the term "**Market Basket Analysis**".

In this basket, the shopper has added milk, bread, cheese, eggs, carrots and broccoli.

- Is milk typically purchased with bread?
- Is cheese typically purchased when milk and eggs are purchased together?
- What product is the most likely to be added next based on the current basket?

Rule Format

Given a set of transactions, find **rules** that predict the occurrence of an item based on the occurrences of other items in the database.

- Implication means **co-occurrence**, **not causality**
  - IF {set of items} $\Rightarrow$ THEN {set of items}
    - Example: If {bread} $\Rightarrow$ then {milk}
  - "IF" part: **Antecedent** or Body of the rule
  - "THEN" part: **Consequent** or Head of the rule
  - "Item set" = the items comprising the antecedent or consequent
  - Antecedent and consequent are ***disjoint*** (no items in common)

Many Rules are Possible

Consider the example to the right:

tran_id	items
1	Bread, Milk
2	Bread, Diapers, Beer, Eggs
3	Milk, Diapers, Beer, Soda
4	Bread, Milk, Diapers, Beer
5	Bread, Milk, Diapers, Soda

Transaction 2 supports several rules:

- If bread, then diapers
- If beer, then diapers
- If bread and beer, then eggs
- + many more….

Ideally, we want to create all possible combinations of items

Problem: computation time grows exponentially as # of items increases

Solution: consider only “frequent itemsets”

➢ Criterion for “frequent”: **support**

Support

The **support** of a rule is:

$$\frac{\#\ of\ transactions\ with\ both\ the\ antecedent\ and\ consequent\ itemsets}{\#\ of\ transactions}$$

tran_id	items
1	Bread, Milk
2	Bread, Diapers, Beer, Eggs
3	Milk, Diapers, Beer, Soda
4	Bread, Milk, Diapers, Beer
5	Bread, Milk, Diapers, Soda

- Support for {beer} ⇒ {diapers} is 3/5
  - ✓ 60% of transactions include this pair of items

- Support quantifies the significance of the **co-occurrence** of the items involved in a rule.
- In practice, we only care about itemsets with strong support based on subjective measurement.

Let's Practice: Support

$$\frac{\#\ of\ transactions\ with\ both\ the\ antecedent\ and\ consequent\ itemsets}{\#\ of\ transactions}$$

tran_id	items
1	Pizza, Salad, Soda
2	Burger, Soda
3	Pizza, Garlic Bread, Soda
4	Burger, Fries, Water
5	Burger, Fries, Ice Cream
6	Pizza, Soda
7	Burger, Fries, Soda
8	Soup, Salad, Water
9	Pizza, Fries, Soda
10	Pizza, Salad, Soda

1. What is the support of {**Pizza**}? **50%**
2. What is the support of {**Soda**}? **70%**
3. What is the support of {**Burger, Fries**}? **30%**

Assume a minimum support requirement of 50%, are there any other itemsets of size 2 that we care about?

Confidence

The **confidence** of a rule measures the strength of association:

$$\frac{\#\ of\ transactions\ with\ both\ the\ antecedent\ and\ consequent\ itemsets}{\#\ of\ transactions\ with\ antecedent\ itemset}$$

tran_id	items
1	Bread, Milk
2	Bread, Diapers, Beer, Eggs
3	Milk, Diapers, Beer, Soda
4	Bread, Milk, Diapers, Beer
5	Bread, Milk, Diapers, Soda

- Confidence for {**Bread**} ⇒ {**Diapers**} **3/4**
  - ✓ Conditional that the basket contains bread, there is a 75% change that the same basket also has diapers

Let's Practice: Confidence

$$\frac{\#\ of\ transactions\ with\ both\ the\ antecedent\ and\ consequent\ itemsets}{\#\ of\ transactions\ with\ antecedent\ itemset}$$

tran_id	items
1	Pizza, Salad, Soda
2	Burger, Soda
3	Pizza, Garlic Bread, Soda
4	Burger, Fries, Water
5	Burger, Fries, Ice Cream
6	Pizza, Soda
7	Burger, Fries, Soda
8	Soup, Salad, Water
9	Pizza, Fries, Soda
10	Pizza, Salad, Soda

1. What is the confidence for {**Pizza**} $\Rightarrow$ {**Soda**}?
   100%
2. What is the confidence for {**Soda**} $\Rightarrow$ {**Pizza**}?
   71.43%

Is it relevant to consider rules with the antecedent and consequent switched in association rules?

Generating Association Rules

The standard approach for generating association rules is the Apriori algorithm developed by Agrawal et all (1993).

Generate all association rules that meet the following:

1. support greater than a user-specified support threshold referred to as ***min_sup*** (minimum support)
2. confidence greater than a user-specified confidence threshold ***min_conf*** (minimum confidence)

Why do we need both thresholds?

In the prior transaction data, the rule: {**Salad, Water**} $\Rightarrow$ {**Soup**} has confidence of 100%, but there is only 1 transaction so in theory, the association rule has low impact.

Valid Association Rules Phases

Identifying valid association rules can be decomposed into two phases:

1. Find all sets of items (***itemsets***) with support above a minimum support threshold (***min_sup)***
   - itemsets with support ≥ ***min_sup*** are considered **frequent** itemsets.

2. From each frequent itemset, generate rules that use items from that frequent itemset.
   - Given a frequent itemset Y, and X, a subset of Y
     - Calculate the confidence of the rule $X \Rightarrow (Y \mid X)$ and compare to the minimum confidence threshold ***(min_conf)***
   - If confidence >=c ≥ ***min_conf***, $X \Rightarrow (Y \mid X)$ is a valid association rule.

Step 1: Finding Frequent Itemsets

1. Start by finding all itemsets of size 1 that are frequent.
2. Expand these by counting the frequency of all itemsets of size 2 that include frequent itemsets of size 1.
3. Next, we take itemsets of size 2 that are frequent, and try to expand them, and continue expanding this way until we cannot expand further.

Requirement: *Minimum support*: 50%

tran_id	items
1	Bread, Milk
2	Bread, Diapers, Beer, Eggs
3	Milk, Diapers, Beer, Soda
4	Bread, Milk, Diapers, Beer
5	Bread, Milk, Diapers, Soda

Frequent Pattern	Support
{Bread}	80%
{Milk}	80%
{Diapers}	80%
{Beer}	60%
{Milk, Bread}	60%
{Diapers, Bread}	60%
{Diapers, Milk}	60%
{Diapers, Beer}	60%

Step 2: Selecting Strong Rules

1. Generate rules that use items from that frequent itemset.
2. Calculate confidence for the rule and compare to ***min_conf***.

Requirement: *Minimum confidence*: 80%

Frequent Pattern	Support
{Bread}	80%
{Milk}	80%
{Diapers}	80%
{Beer}	60%
{Milk, Bread}	60%
{Diapers, Bread}	60%
{Diapers, Milk}	60%
{Diapers, Beer}	60%

Rule	Support	Confidence
{**Beer**} ⇒ {**Diapers**}	60%	1.00
{**Diapers**} ⇒ {**Beer**}	60%	0.75
{**Milk**} ⇒ {**Bread**}	60%	0.75
{**Bread**} ⇒ {**Milk**}	60%	0.75
{**Bread**} ⇒ {**Diapers**}	60%	0.75
{**Diapers**} ⇒ {**Bread**}	60%	0.75
{**Milk**} ⇒ {**Diapers**}	60%	0.75
{**Diapers**} ⇒ {**Bread**}	60%	0.75

Every transaction when Beer was purchased, Diapers was also purchased!

Application Across Industries

Generating association rules has applicability across many different industries and not just limited to shopping basket analysis.

- **Healthcare**: Identifying symptom-diagnosis correlations (e.g., fever and cough linked to flu).
- **Entertainment**: Recommending playlists based on commonly grouped songs.
- **Telecommunications**: Bundling services based on customer usage patterns (e.g., internet + streaming).
- **Education**: Identifying courses students frequently enroll in together to optimize scheduling.

Another Look at Confidence

Consider the below purchase matrix for customers and the rule {**Tea**} $\Rightarrow$ {**Coffee**}

	Coffee	**NOT** Coffee	Total
Tea	15	5	20
NOT Tea	75	5	80
Total	90	10	100

What is the confidence for this rule? **15 / 20 = 75%**

But support for Coffee is very high! **90 / 100 = 90%**

So, given that tea has been bought, the probability of buying coffee has dropped. **Although confidence is high, rule is misleading!**

In fact, the confidence for {**NOT Tea**} $\Rightarrow$ {**Coffee**} is higher!

75 / 80 = 93.75%

Another Performance Measure: Lift

The **lift** of a rule measures how much more likely the consequent is, given the antecedent:

$$\frac{confidence\ of\ rule}{consequent\ support}$$

Referred to as benchmark confidence

	Coffee	**NOT** Coffee	Total
Tea	15	5	20
NOT Tea	75	5	80
Total	90	10	100

Confidence is 75% and Support of Coffee is 90%

What is the lift for this rule? **0.75 / 0.9 = 0.833 < 1**

A lift ratio greater than 1.0 suggests the rule is useful in finding consequent itemsets.

Lift Example

Consequent Support:

$$\frac{\#\ of\ transactions\ with\ consequent\ itemsets}{\#\ of\ transactions}$$

Lift:

$$\frac{confidence\ of\ rule}{consequent\ support}$$

tran_id	items
1	Pizza, Salad, Soda
2	Burger, Soda
3	Pizza, Garlic Bread, Soda
4	Burger, Fries, Water
5	Burger, Fries, Ice Cream
6	Pizza, Soda
7	Burger, Fries, Soda
8	Soup, Salad, Water
9	Pizza, Fries, Soda
10	Pizza, Salad, Soda

1. What is the lift for {**Pizza**} $\Rightarrow$ {**Soda**} given confidence of 100%?
   1 / .7 = 1.43
2. What is the lift for {**Soda**} $\Rightarrow$ {**Pizza**} given confidence of 71.43%?
   .7143 / .5 = 1.43

Do you think it is possible to have a different lift ratio when the itemsets are the same in the association rule?

Let's Practice: Finding Frequent Itemsets

tran_id	items
1	Laptop, Mouse, Keyboard
2	Laptop, Mouse, Monitor
3	Desk, Chair, Lamp
4	Laptop, Desk, Mouse
5	Mouse, Keyboard, Desk

Requirement:

- *Minimum support*: 60%
- *Minimum confidence*: 80%

- Find all 1-item itemsets that meet the minimum support
- What are the 2-item itemsets that you need to investigate?
- Find all 2-item itemsets that meet the minimum support
- Do you need to investigate any 3-item itemset?

Let's Practice: Selecting Strong Rules

tran_id	items
1	Laptop, Mouse, Keyboard
2	Laptop, Mouse, Monitor
3	Desk, Chair, Lamp
4	Laptop, Desk, Mouse
5	Mouse, Keyboard, Desk

Requirement:

- *Minimum support*: 40%
- *Minimum confidence*: 80%
- Recall that we already find the following frequent Itemsets:
  - {Mouse} (sup = 0.8),
  - {Laptop} (sup = 0.6),
  - {Desk} (sup = 0.6)
  - {Laptop, Mouse} (support = 0.6)

- For each multi-item itemset, list all possible association rules and calculate confidence and lift.
- Identify all strong association rules.

Summary

- Association rules (or *affinity analysis,* or *market basket analysis*) produce rules on associations between items from a database of transactions.
- Widely used in **recommender systems**
- Most popular method is **Apriori algorithm**
- To reduce computation, we consider only “frequent” item sets (**support**).
- Performance of rules is measured by **confidence** and **lift**
- Can produce a profusion of rules; review is required to identify useful rules and to reduce redundancy.