# Section 3
# Apriori Algorithm

# 1. The Role of Apriori Algorithm

# The Usefulness of Apriori Algorithm

- We can quantify associations between items by using metrics.

- However, as the number of items increase, the number of rules increases exponentially.

We can use apriori algorithm for pruning.

# Combination

From 3 different balls: 1 2 3

- Pick 1 ball: 1 2 3 — 3 patterns
- Pick 2 balls: 1 2 1 3 2 3 — 3 patterns
- Pick 3 balls: 1 2 3 — 1 pattern

➡ N of combination = 3 + 3 + 1 = 7

# The Number of Combination

The number of combinations of k objects from n objects

$$ {}_nC_k = \frac{n!}{k!(n-k)!} $$

The number of combination: k=1, 2, …, n

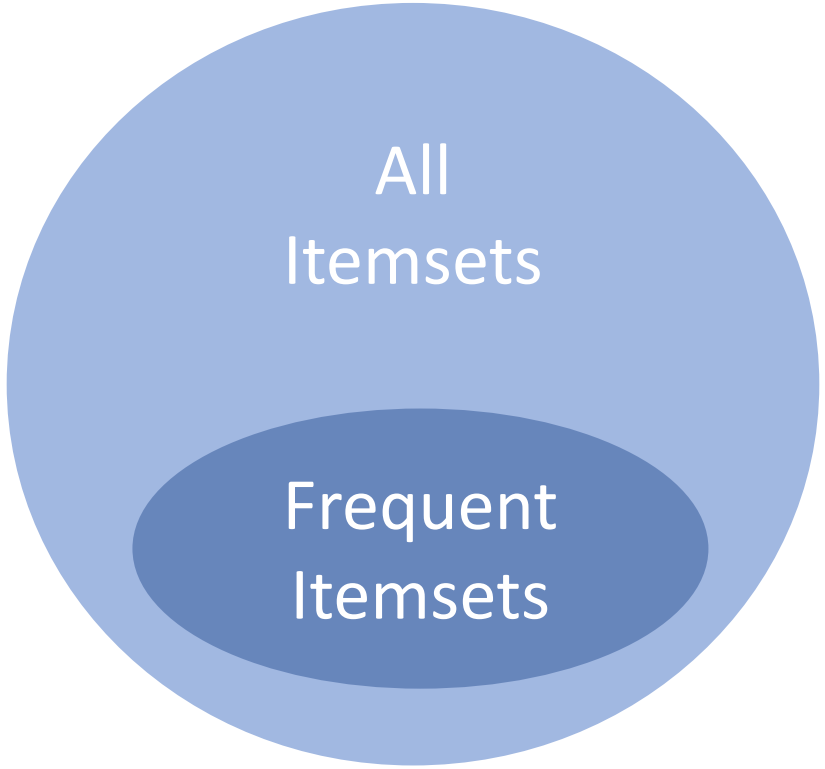
- Pick k from 3: 7 patterns
- Pick k from 4: 15 patterns
- Pick k from 10: 1013 patterns

# 2. Apriori Principle

# Frequent Itemset

**Definition:** An itemset whose support value is greater than minsup

- Minsup is an arbitorary number.
- The first step of pruning is to identify frequent itemsets.



# Apriori Principle

"All subsets of a frequent itemset must be frequent."

Frequent : Support is greater than the threshold.

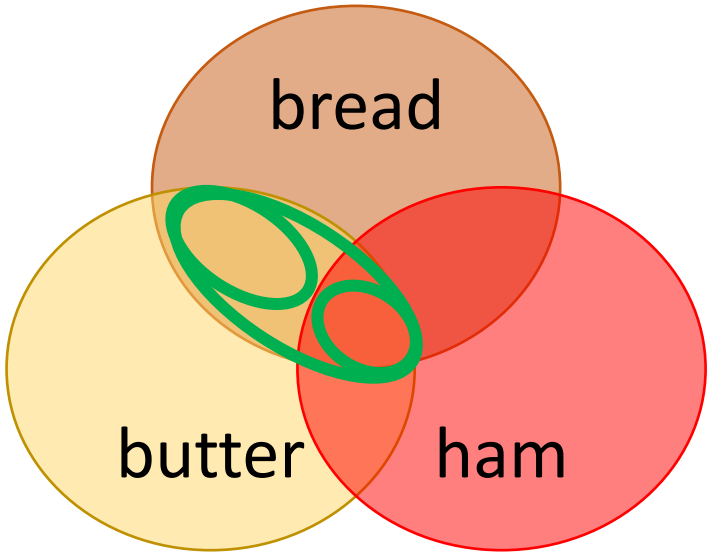
- Keep frequent itemsets for further analysis.
- Prune the itemsets that are found to be not frequent.

# Support and Apriori Principle

- {bread, butter, ham} must contain {bread, butter}.

  → If {bread, butter, ham} is frequent, {bread, butter} must be frequent.

- {bread, butter, ham} cannot be more frequent than {bread, butter}.

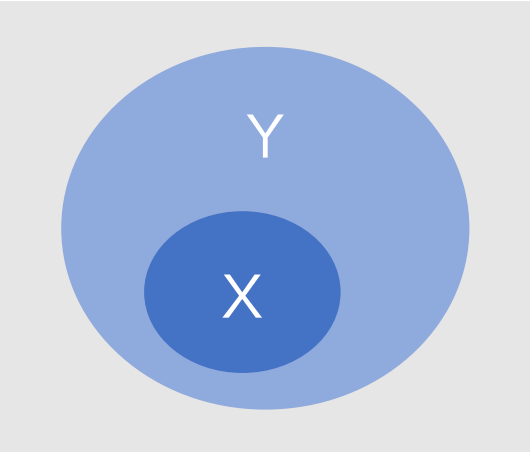



→ $Support(itemset) \leq Support(subset)$

# Anti-Monotone Property and Apriori Principle

- Anti-monotone property of support.

$$\underbrace{\forall X, Y}_{\text{For all X and Y}}: \ \underbrace{(X \subseteq Y)}_{\text{If X is a subset of Y}} \Rightarrow \underbrace{s(X) \geq s(Y)}_{\text{Support (X) ≥ Support (Y)}}$$




- Apriori principle is a result of anti-monotone property of support.

# 3. Apriori Algorithm: Phase 1
# —Find Frequent Itemsets—

# Steps for Finding Frequent Itemsets

| | |
|---|---|
| 1 | Prepare data and set minsup |
| 2 | Create a list of frequent itemsets (support ≥ minsup) of length 1 |
| 3 | Create a list of itemsets of length 2 by combining the frequent itemsets of length 1 |
| 4 | Prune itemsets whose support is less than minsup |
| 5 | Create a list of itemsets of length 3 from the pruned list |
| 6 | Prune itemsets whose support is less than minsup |

- In the following, lengthen the itemsets and check whether “support ≥ minsup.”
- Stop the process when you cannot create a list of frequent itemset.

# Step 1. Prepare data and set minsup

- We generate association rule by using frequent itemsets.
- Frequent itemset: Support ≥ minsup
- The threshold of support (minsup) is an arbitrary number.

Here, $minsup = 40\%$

| ID | Item |
|---|---|
| T1 | bread, ham, cheese |
| T2 | bread, butter, ham, milk |
| T3 | bread, ham, milk |
| T4 | butter, ham, milk |
| T5 | butter, milk |

# Step 2. Create a list of frequent itemsets of length 1

| ID | Item |
|---|---|
| T1 | bread, ham, cheese |
| T2 | bread, butter, ham, milk |
| T3 | bread, ham, milk |
| T4 | butter, ham, milk |
| T5 | butter, milk |

| Itemset | Support |
|---|---|
| bread | 60% |
| butter | 60% |
| cheese | 20% |
| ham | 80% |
| milk | 80% |

| Frequent Itemset |
|---|
| bread |
| butter |
| ham |
| milk |

Prune itemsets if $Support < 40\%$

# Step 3. Create a list of itemsets of length 2

| Frequent Itemset (length 1) |
| --- |
| bread |
| butter |
| ham |
| milk |

List of frequent itemsets created in Step 2

| Itemset (length 2) |
| --- |
| bread, butter |
| bread, ham |
| bread, milk |
| butter, ham |
| butter, milk |
| ham, milk |

Create a list of itemsets of length 2 by combining the frequent itemsets of length 1

# Step 4. Prune itemsets (Support < minsup)

| ID | Item |
|---|---|
| T1 | bread, ham, cheese |
| T2 | bread, butter, ham, milk |
| T3 | bread, ham, milk |
| T4 | butter, ham, milk |
| T5 | butter, milk |

| Itemsets (length 2) | Support |
|---|---|
| bread, butter | 20% |
| bread, ham | 60% |
| bread, milk | 40% |
| butter, ham | 40% |
| butter, milk | 60% |
| ham, milk | 60% |

| Frequent Itemsets (Length 2) |
|---|
| bread, ham |
| bread, milk |
| butter, ham |
| butter, milk |
| ham, milk |

Prune itemsets if $Support < 40\%$

# Step 5. Create a list of itemsets of length 3

| ID | Items |
|---|---|
| T1 | bread, ham, cheese |
| T2 | bread, butter, ham, milk |
| T3 | bread, ham, milk |
| T4 | butter, ham, milk |
| T5 | butter, milk |

| Frequent Itemsets (Length 2) |
|---|
| bread, ham |
| bread, milk |
| butter, ham |
| butter, milk |
| ham, milk |

| Itemsets (Length 3) |
|---|
| bread, butter, ham |
| bread, butter, milk |
| bread, ham, milk |
| butter, ham, milk |

$$ {}_nC_k = \frac{n!}{k!\,(n-k)!} $$

$$ {}_4C_3 = \frac{4!}{3!\,(4-3)!} = \frac{4!}{3!} = 4 $$

# Step 6. Prune itemsets (Support < minsup)

| ID | Items |
|---|---|
| T1 | bread, ham, cheese |
| T2 | bread, butter, ham, milk |
| T3 | bread, ham, milk |
| T4 | butter, ham, milk |
| T5 | butter, milk |

| Itemsets (Length 3) |
|---|
| bread, butter, ham |
| bread, butter, milk |
| bread, ham, milk |
| butter, ham, milk |

| Frequent Itemsets (Length 3) | Support |
|---|---|
| bread, ham, milk | 40% |
| butter, ham, milk | 40% |

minsup = 40%

Apriori Principle

| Itemsets | Support |
|---|---|
| bread, butter | 20% |

# Step 7. Create a list of itemsets of length 4, and check Support

| ID | Item |
|---|---|
| T1 | bread, ham, cheese |
| T2 | bread, butter, ham, milk |
| T3 | bread, ham, milk |
| T4 | butter, ham, milk |
| T5 | butter, milk |

| Frequent Itemset (Length 3) |
|---|
| bread, ham, milk |
| butter, ham, milk |

| Itemset (Length 4) | Support |
|---|---|
| bread, butter, ham, milk | 20% |

Prune itemsets: Support < minsup (40%)

There are no frequent itemsets of length 4.

Terminate the process.

# 4. Apriori Algorithm: Phase 2
# —Association Rule Selection—

# Rule Selection by Confidence

- Support tends to be high when itemsets is purchased very frequently.

  → We use Confidence as a metric

$$Confidence(X \rightarrow Y) = \frac{freq(X, Y)}{freq(X)}$$

- Selection criterion: $minconf = 0.6$

# Step 1. Generate rules from frequent itemsets

| Frequent Itemset (Length 3) |
| --- |
| bread, ham, milk |
| butter, ham, milk |

| Rule |
| --- |
| bread → ham, milk |
| ham → bread, milk |
| milk → bread, ham |
| bread, ham → milk |
| bread, milk → ham |
| ham, milk → bread |

# Step 2. Select rules: Confidence ≥ minconf

| Rule | Formula of Confidence | Confidence |
|---|---|---|
| bread → ham, milk | Support(bread, ham, milk) / Support(bread) | 67% |
| ham → bread, milk | Support(bread, ham, milk) / Support(ham) | 50% |
| milk → bread, ham | Support(bread, ham, milk) / Support(milk) | 50% |
| bread, ham → milk | Support(bread, ham, milk) / Support(bread, ham) | 67% |
| bread, milk → ham | Support(bread, ham, milk) / Support(bread, milk) | 100% |
| ham, milk → bread | Support(bread, ham, milk) / Support(ham, milk) | 67% |

However, Confidence depends on the frequency of the consequent.

# Recap: Lift

$$Lift(X \rightarrow Y) = \frac{freq(X,Y)}{freq(X)} \cdot \frac{1}{Support(Y)}$$

$$= Confidence(X \rightarrow Y) \cdot \frac{1}{Support(Y)}$$

$$= \frac{Confidence(X \rightarrow Y)}{Support(Y)}$$

# Recap: Meaning of Lift

$$Lift(X \rightarrow Y) = \frac{Confidence(X \rightarrow Y)}{Support(Y)} = \frac{P(Y|X)}{P(Y)}$$

$$Lift(X \rightarrow Y) > 1 \Rightarrow P(Y|X) > P(Y)$$

The occurrence of X increased the probability of occurrence of Y

# Step 3. Select rules: Lift > 1.0

| Rule | Confidence | Support(consequent) | Lift |
|---|---|---|---|
| bread→ham, milk | 67% | 60% | 67 / 60 = 1.1 |
| bread, ham→milk | 67% | 80% | 67 / 80 = 0.8 |
| bread, milk→ham | 100% | 80% | 100 / 80 = 1.3 |
| ham, milk→bread | 67% | 60% | 67 / 60 = 1.1 |

$$Lift(X \rightarrow Y) = \frac{Confidence(X \rightarrow Y)}{Support(Y)}$$

# Similar but Different!

All of these rules consist of bread, ham, and milk.

- Bread → ham, milk
- Ham → bread, milk
- Bread, ham → milk

| Rule | Confidence |
|---|---|
| bread → ham, milk | 67% |
| ham → bread, milk | 50% |

| Rule | Lift |
|---|---|
| bread → ham, milk | 1.1 |
| bread, ham → milk | 0.8 |

# 6. Summary

## Summary

- The threshold (minsup, minconf) are arbitrary values.

  The result depends on what value we choose.

- Rules consisting of the same items are not the same.

  $\{X \rightarrow Y\}$ is not same as $\{Y \rightarrow X\}$