Sane's STATS Academy of Statistics

Association Rules Mining

Apriori Algorithm

June 1, 2022

A Customer’s Basket

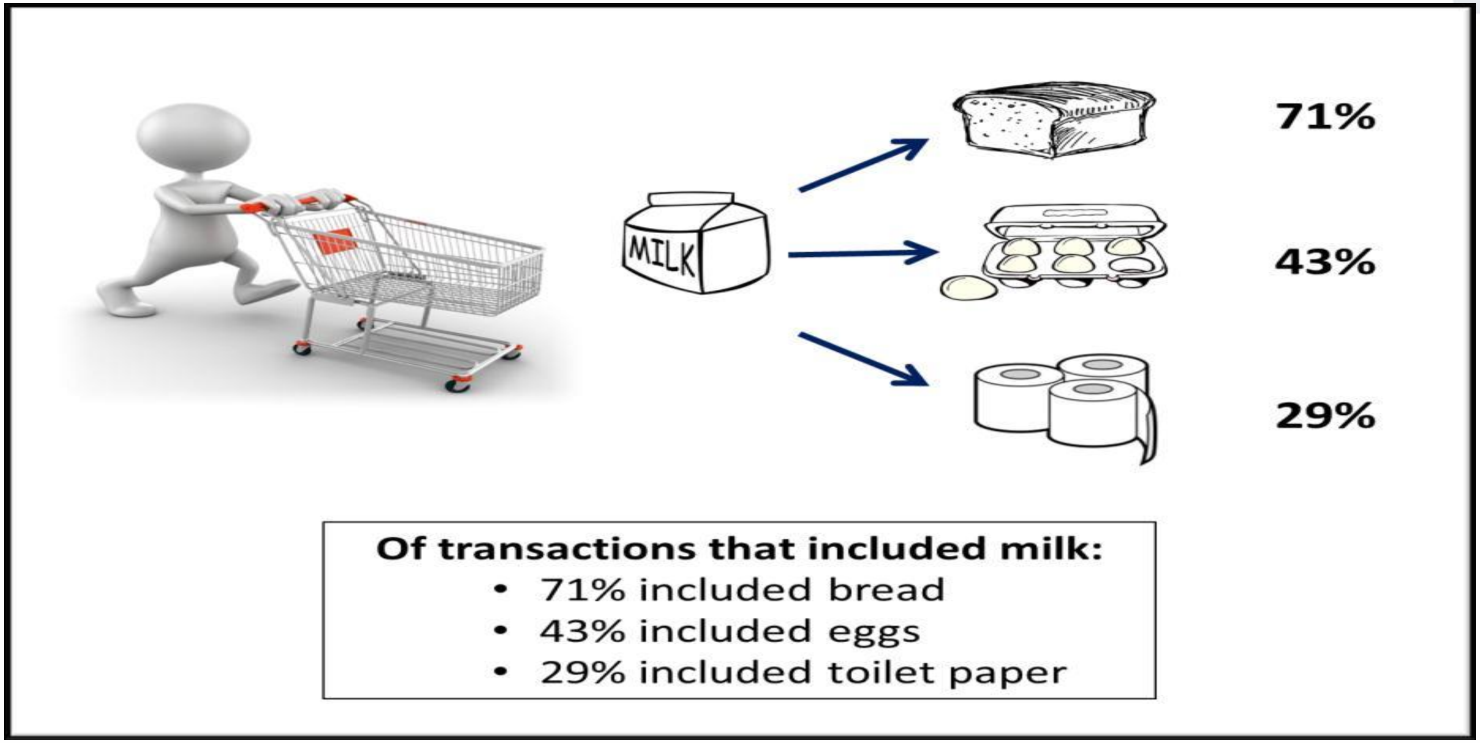
- If a customer buys bananas and she buys apple then she buys a fruit beverage also.
- If its a late noon time and a customer buys coconut biscuits then he also buys chips.

Association Rules

- Association rules provide information of this type in the form of “if–then” statements.
- These rules are computed from the data.

Generating Rules

- Examine all possible rules between items in an if–then format, and select only those that are most likely to be indicators of true dependence.



Of transactions that included milk:

- 71% included bread
- 43% included eggs
- 29% included toilet paper

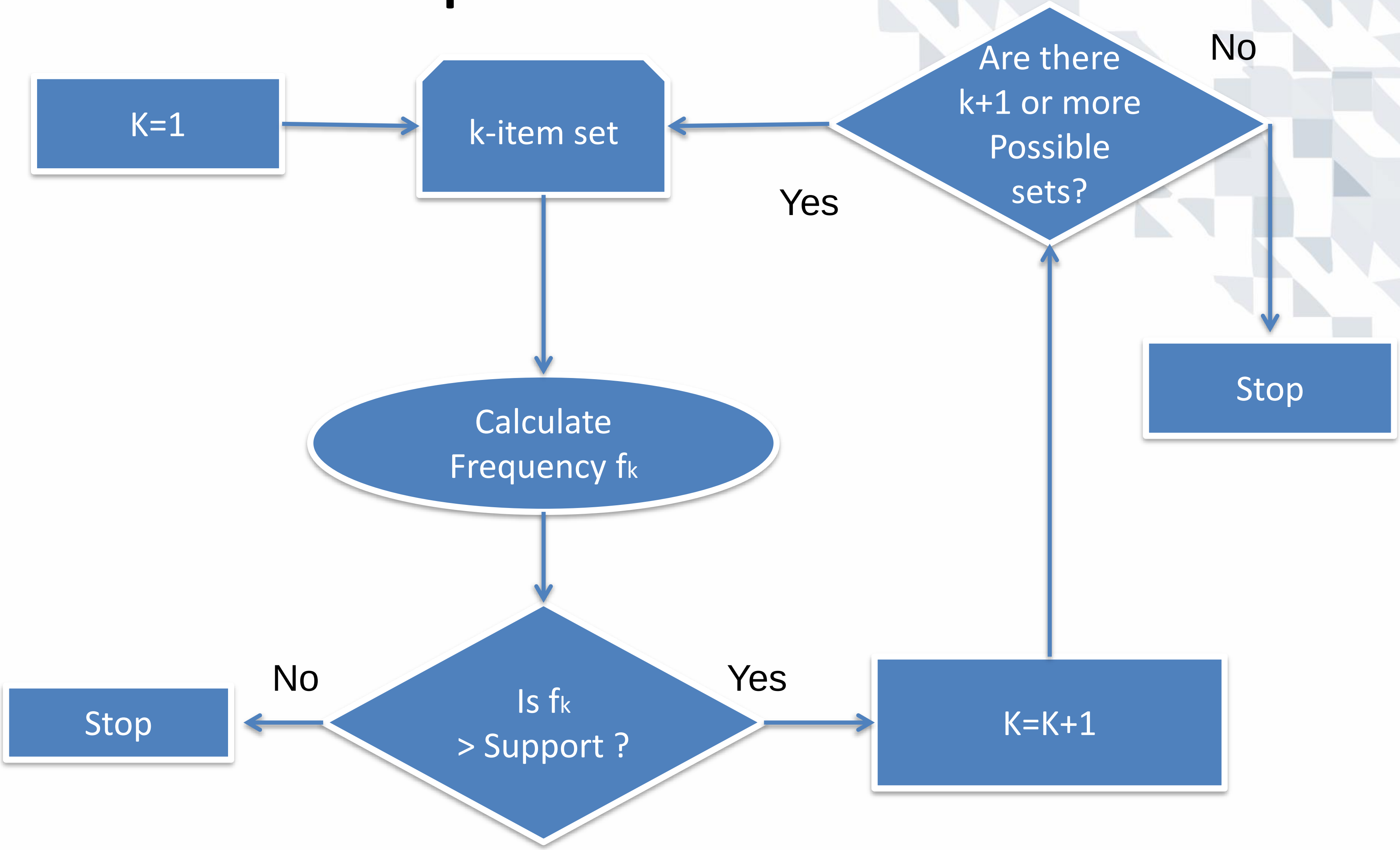
Apriori Algorithm

- Generate frequent item sets with just one item (one-item sets)
- Recursively generate frequent item sets with two items, then with three items, and so on, until we have generated frequent item sets of all sizes.
- Count, for each item, how many transactions in the database include the item.

Apriori Algorithm

- These transaction counts are the supports for the one-item sets.
- We drop one-item sets that have support below the desired minimum support to create a list of the frequent one-item sets.
- To generate frequent two-item sets, we use the frequent one-item sets.

Apriori Flow Chart



K=1

k-item set

Calculate Frequency f_k

Is f_k > Support ?

No

Stop

Yes

K=K+1

Are there k+1 or more Possible sets?

Yes

No

Stop

Strength of Association

- Support
- Confidence
- Lift Ratio

Support

- The support of an item set is the number of transactions that include that item set.
- The support of a rule is the number of transactions that include both the antecedent (if-part) and consequent (then-part) item sets.

Confidence

- Confidence is defined as the measure of trustworthiness associated with each discovered rule

$$Confidence = \frac{Support(if - part\ and\ then - part)}{Support(if - part)}$$

Example

- Consider a database of shopping mall of about 10,00,000 transactions. Out of these transactions, there are 40,000 transactions with purchase of soft toys and hand towel (purchased together) and 24,000 of these transactions include the room freshener purchases.
  - n{Soft Toy, Hand Towel} = 40,000 ,
  - n{Soft Toy, Hand Towel, Room Freshener} = 24,000
- Rule : “If anyone purchases soft toys and hand towel then he/she also purchases room freshener in the same trip” has Support of 24,000 (2.4%) transactions and a confidence of 24,000/40,000 %= 60%.
  - Conf ({Soft Toy, Hand Towel} → {Room Freshener}) = 0.6

Support as Probability

- Support is the (estimated) probability that a transaction selected randomly from the database will contain all items in the if-part and the then-part:
  - P(if-part AND then-part)

Confidence as Probability

- Confidence is the (estimated) conditional probability that a transaction selected randomly will include all the items in the consequent given that the transaction includes all the items in the antecedent.

$$Confidence = \text{P(then−part|if−part)} = \frac{P(if - part\ AND\ then - part)}{P(if - part)}$$

Possible Independence

- If if-part and then-part are independent then the support would be

$$P(if - part\ AND\ then - part) = P(if - part) * P(then - part)$$

- Based on this, the benchmark confidence is defined as

$$P(then - part|if - part) = \frac{P(if - part\ AND\ then - part)}{P(if - part)}$$

$$= \frac{\mathrm{P(if{-}part)} * P(then{-}part)}{P(if{-}part)}$$

$$= P(then - part)$$

Benchmark Confidence

- Benchmark Confidence can be estimated from the data as,

$$Benchmark\ Confidence = \frac{No.of\ transactions\ with\ then - part}{No.of\ transactions\ in\ the\ database}$$

- In shopping mall example if 300,000 transactions are of then-part (room freshener purchases) then benchmark confidence can be calculated as

$$Benchmark\ Confidence = \frac{300000}{1000000} = 0.3$$

Lift Ratio

- The lift ratio is the confidence of the rule divided by the confidence, assuming independence of consequent from antecedent.

$$Lift\ Ratio = \frac{Confidence}{Benchmark\ Confidence}$$

- A lift ratio greater than 1.0 suggests that there is some usefulness to the rule.

- In shopping mall example if 300,000 transactions are of then-part (room freshener purchases) then lift ratio of the said transaction can be calculated as

$$Lift\ Ratio = \frac{Confidence}{Benchmark\ Confidence} = \frac{0.6}{0.3} = 2$$

Interpreting the Results

- The support for the rule indicates its impact in terms of overall size as proportion of transactions getting affected.
- If only a small number of transactions are affected, the rule may be of little use.
- The lift ratio indicates how efficient the rule is in finding consequents, compared to random selection.