DMML, 9 Jan 2020

Market Basket Analysis

Items $I = \{i_1, i_2, \ldots, i_N\}$

Transactions $T = \{t_1, t_2, \ldots, t_M\}$, each $t_i \subseteq I$

Identify "association rules" $X \to Y$, $X \cap Y = \phi$

$X, Y \subseteq I$

$X \subseteq I$ → support of X is the number of transactions containing X

$$sup(X) = \left| \{ t_i \in T \mid X \subseteq t_i \} \right|$$

Significance of a rule $X \to Y$

Confidence : $\frac{sup(X \cup Y)}{sup(X)}$ Not $X \cap Y$

Two thresholds

- minimum support — what is $sup(X \to Y)$?
  $= sup(X \cup Y) \geq mm\text{-}sup$
- minimum confidence $conf(X \to Y) \geq mm\text{-}conf$

Note: There is a single "correct" answer once thresholds are fixed

How to find frequent itemsets?

Find X s.t. $sup(X) \geq$ min-sup

If we do this, for each X that is frequent,

for each partition $Y \rightarrow Z$ of X ($Y \cup Z = X$, $Y \cap Z = \phi$)

check if $conf(Y \rightarrow Z) \geq$ min-conf

Use a priori principle

A Priori

If X is frequent, ~~any~~ every $Y \subseteq X$ is also frequent

If some subset of X is not frequent, X cannot be frequent

Level wise search

- Enumerate frequent items $= F_1$
- Next is F_2 – frequent pairs

$F_2 \subseteq F_1 \times F_1$

(Naive) A Priori

Frequent sets of size j, F_j

→ Candidate frequent sets of size $j+1$, C_{j+1}

→ Frequent sets of size $j+1$, F_{j+1}

$F_1 \longrightarrow C_2 = (F_1 \times F_1) \setminus \text{Identity} \xrightarrow{\text{count}} F_2$

$F_2 \longrightarrow C_3 = \{(x,y,z) \mid (x,y), (x,z), (y,z) \in F_2\}$

$\longrightarrow F_3$

$F_j \longrightarrow C_{j+1} = \{(x_1, \ldots, x_{j+1}) \mid$ Every j-size subset is in $F_j\}$

$\rightarrow F_{j+1}$

Enumerate all $j+1$ size subset of I & then prune

$\binom{N}{j+1}$

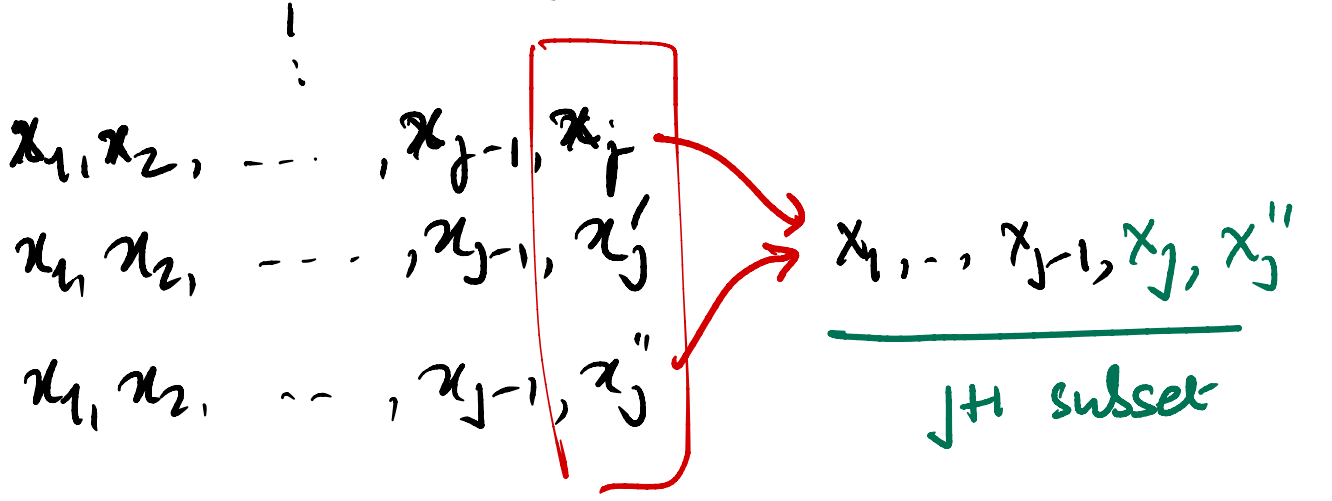
Can we fix this ?

Idea: Can we find a $\hat{C}_{j+1} \supseteq C_{j+1}$ more efficiently s.t. $\hat{C}_{j+1}$ is also "small"

Assume I is ordered $i_1 < i_2 < \cdots < i_N$

Any $X \subseteq I$ is listed in ascending order

F_j can be ordered lexicographically (dictionary order)



$\hat{C}_{j+1}$ is constructed by this fusion operation

Is $\hat{C}_{j+1} \supseteq C_{j+1}$?

Suppose $\{y_1, y_2, \ldots, y_{j+1}\} \in C_{j+1}$ $\quad y_1 < y_2 < \ldots < y_{j+1}$

Every j-subset must be in F_j

$y_1, \ldots, y_{j-1}, y_j$ is a j-subset

$y_1, \ldots, y_{j-1}, y_{j+1}$ also

$\geq$ are in F_j

$\downarrow$ since

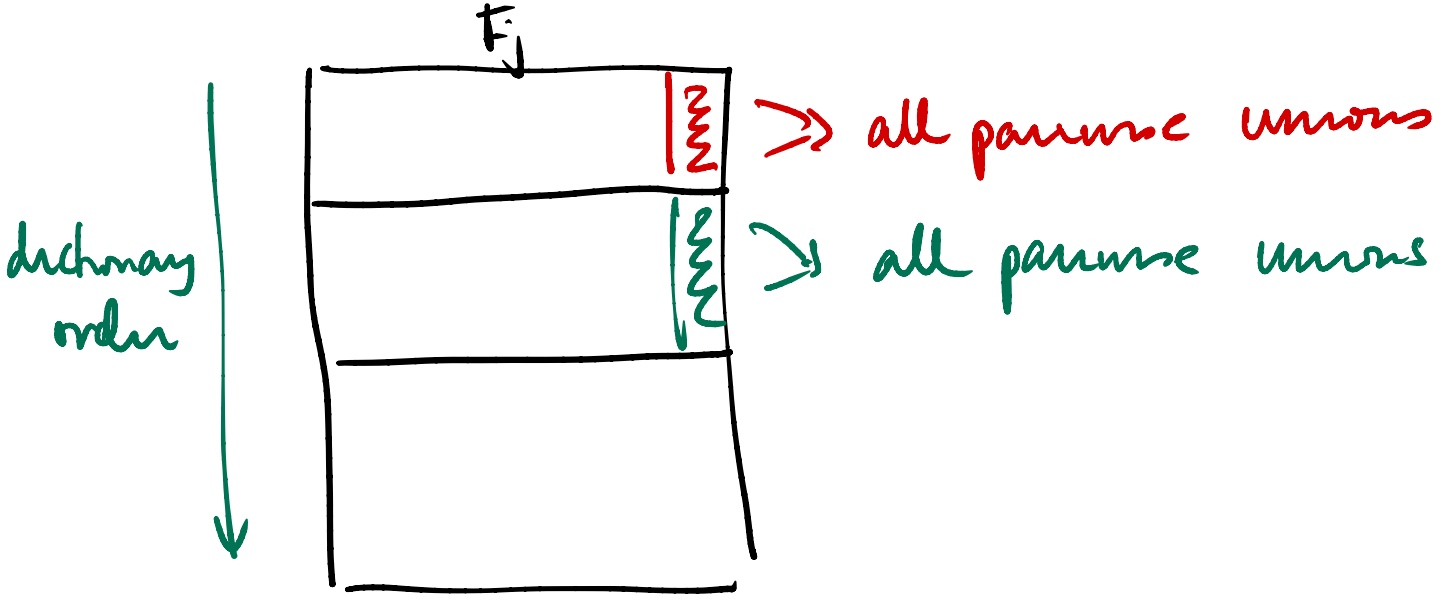
this set would be in $\hat{C}_{j+1}$

$\{x_1, x_2, \ldots, x_{j-1}, x_j, x_j''\} \in \hat{C}_{j+1}$

NOT in C_{j+1}

↓

may not be in F_j



Level by level

$F_1 \rightarrow C_1 \rightarrow F_2 \rightarrow C_2 \rightarrow \cdots$

When to stop ?

① $j >$ max transaction size

② $\hat{C}_{j+1}$ is empty

As j increases, support must decrease

Part 1 Find frequent itemsets

Part 2 Find rules

For each $X \in F_1 \cup F_2 \cup \dots \cup F_g = F$

Check if $Y \to Z$ meet min-conf, $Y \cup Z = X$, $Y \cap Z = \phi$

$\Downarrow$

$\frac{\sup(Y \cup Z)}{\sup(Y)} \in F$ $\in F$ by a-priori

values have already been computed

Bottleneck: All possible decompositions of X

A priori again!

$X = \{a, b, c, d\}$

$Y_1 = \{a, b, c\}$ $Z_1 = \{d\}$

$Y_2 = \{a, b\}$ $Z_2 = \{c, d\}$

$Y_1 \to Z_1$ $\dfrac{\sup(Y_1 \cup Z_1)}{\sup(Y_1)} = \dfrac{\sup(X)}{\sup(Y_1)}$

$Y_2 \subseteq Y_1$

$\sup(Y_2) \geq \sup(Y_1)$

$Y_2 \to Z_2$ $\dfrac{\sup(Y_2 \cup Z_2)}{\sup(Y_2)} = \dfrac{\sup(X)}{\sup(Y_2)}$

$conf(Y_1 \to Z_1) \geq conf(Y_2 \to Z_2)$

If R_1 $X \setminus \{x_1, x_2\} \rightarrow \{x_1, x_2\}$ meets min-conf

then R_2 $X \setminus \{x_1\} \rightarrow \{x_1\}$ also meet- min-conf

R_3 $X \setminus \{x_2\} \rightarrow \{x_2\}$

$\therefore$ if R_2 or R_3 fails min-conf, R_1 must fail min-conf

$conf(R_1) \leq conf(R_2)$

$conf(R_1) \leq conf(R_3)$

Check all rules of the form

$$X \setminus \{x\} \to \{x\}$$

Work up to larger RHS using a-priori

Going back to stopping

Large rules are difficult to interpret