17 Jan 2019

Frequent Itemsets – Market Basket Analysis

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

$T = \{t_1, t_2, \ldots, t_M\}$

Each $t_i \subseteq I$ is an item set

Given $X \subseteq I$, support of X is fraction of T containing X

$$sup(x) = \frac{|\{t_i | X \subseteq t_i\}|}{M}$$

Frequent : $Sup(X) \geq$ threshold

min-sup

Examples

1. Retail - shopping baskets
2. Text - items are words, baskets are documents
3. Subjects in which schoolchildren fail

Frequent itemsets - possibilities are combinatorially large, but sparsity helps

Apriori Principle

If X is frequent, every $Y \subseteq X$ must also be frequent.

Corollary

If Y is not frequent, no $X \supseteq Y$ is frequent — Prune space for country

Level by level calculation

F_1 – frequent sets of size 1 – count

C_2 – candidate-frequent sets of size 2

↓ $F_1 \times F_1 \setminus Id \setminus$ symmetric pairs

count → F_2

C_3 – all $\{x,y,z\}$ s.t. $\{x,y\}, \{y,z\}$ $\{x,z\}$ in F_2

$F_{k-1} \rightarrow C_k$

Enumerate all k-subsets of I

Check if each k-1-subset is in F_{k-1}

Combinatorially infeasible

lexicographically order each subset

Assume I is ordered

$i_1 < i_2 < \cdots < i_M$

Given $\{i_1, i_2, \ldots, i_{k-2}, i_{k-1}\} \in F_{k-1}$

$\{i_1, i_2, \ldots, i_{k-2}, i'_{k-1}\}$

↓

$\{i_1, i_2, \ldots, i_{k-2}, i_{k-1}, i'_{k-1}\}$

Option 1 Use this more permissive list of candidates $C'_k \supseteq C_k$

Check this

Suppose $\{j_1, j_2, \ldots, j_k\}$ genuinely in C_k

- Every $k-1$ subset is in F_{k-1}

$\{j_1, \ldots, j_{k-2}, j_{k-1}\} \in F_{k-1}$

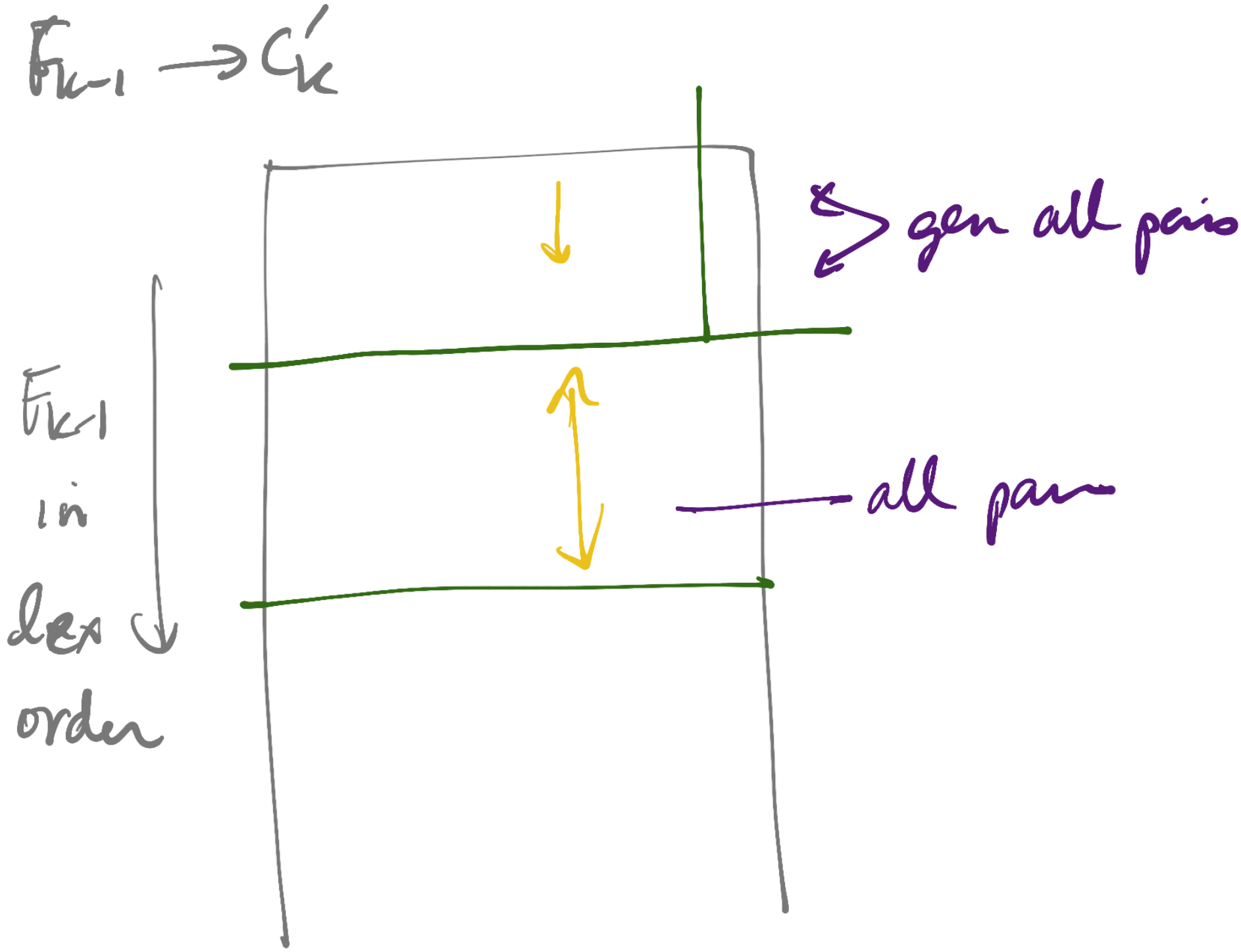
$\{j_1, \ldots j_{k-2}, j_k\} \in F_{k-1}$

→ generates this candidate in C_k

Option 2 Prune C_k' to C_k

$\{f_{j_1}, f_{j_2}, \ldots, f_{j_3}\} \in C_k'$

Extra work, exact C_k


$F_{k-1} \to C'_k$
gen all pairs
F_{k-1} in lex order
all pairs

Termination

1. All t_i have size $\leq S$

   level S is largest possible

2. Some F_{k-1} generates empty C_k

   or F_k is empty

3. Fix some small k that is useful

This problem has a correct answer

Uses of frequent itemsets

Association Rules

$$X \rightarrow Y \quad , \quad X, Y \subseteq I$$
$$X \cap Y = \phi$$

"People who bought X also bought Y"

Two parameters to make a rule worth considering

- Support — how often is it seen?

$sup(X \to Y)$ is just $sup(X \cup Y)$

min-sup

- Confidence $= \dfrac{sup(X \cup Y)}{sup(X)}$

min-conf

From frequent itemsets (A-priori)

to association rules

$F = F_1 \cup F_2 \cup \ldots \cup F_s$ — candidates for $X \cup Y$ for a rule $X \to Y$

Given $Z \in F$

does it break up as $X \cup Y$

s.t. $X \to Y$ has min-conf?

Namely, consider every partition of Z as $A, Z \setminus A$

Given A, $B = Z \setminus A$

$$conf(A \to B) = \frac{sup(A \cup B = Z)}{sup(A)}$$

both are in F_i, F_j for some i, j

Counts are available already

A priori again !

$A \to \{a, b\}$ is a rule with

$conf > min\text{-}conf$

What about $A \cup \{a\} \to \{b\}$?

$$\frac{sup(A \cup \{a, b\})}{sup(A)} \overset{=}{\longleftrightarrow} \frac{sup((A \cup \{a\}) \cup \{b\})}{sup(A \cup \{a\})}$$

$$\frac{}{} \quad \geq \quad \frac{}{}$$

As a fraction $LHS \leq RHS$

<u>In general</u> $A \to B \Rightarrow A \cup \{a\} \to B \setminus \{a\}$

is "good" is also "good"

Conversely

If $A \to B$ is not good

$A \setminus \{a\} \to B \cup \{a\}$ is also not good

Decomposition of $X \subseteq F$ as A, B

level 1 $X \setminus \{a\} \to \{a\}$

level 2 $X \setminus \{a, b\} \to \{a, b\}$ if both

$X \setminus \{a\} \to \{a\}$ &

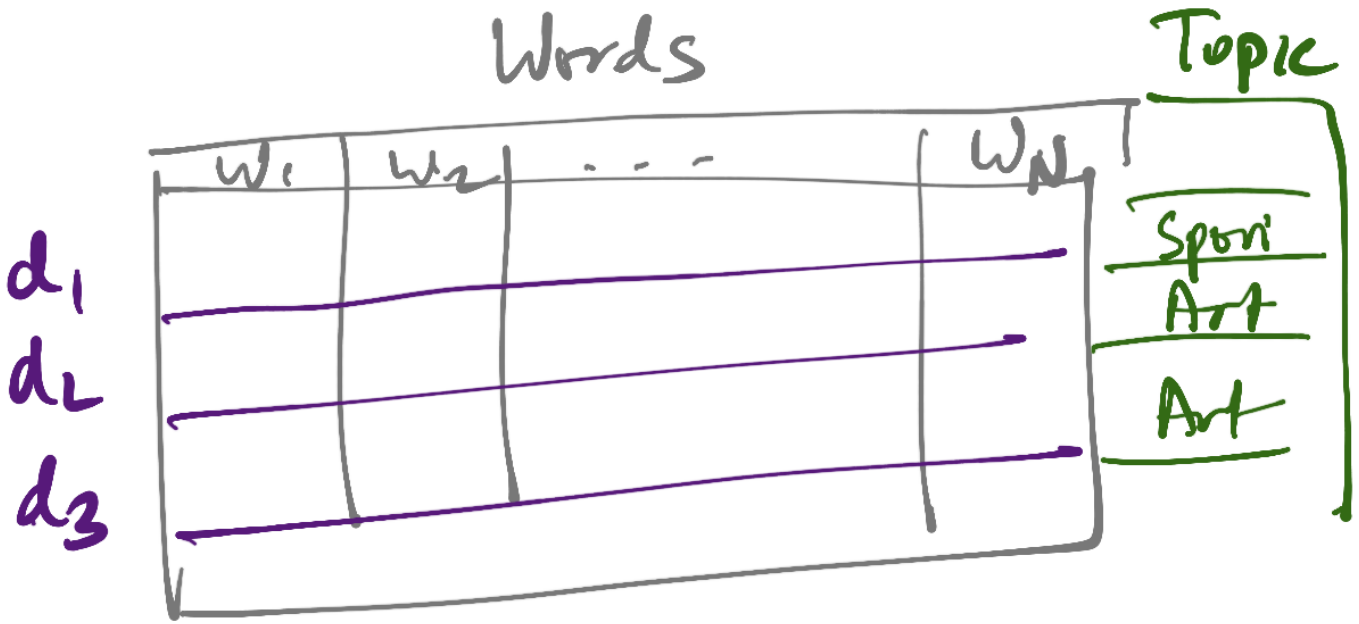
$X \setminus \{b\} \to \{b\}$ are

validated in level 1

Special Case

Basket as rows in a table

A column that is a category



Rules of the form: $X \subseteq$ Words $\rightarrow$ Topic

Association rules $\Rightarrow$ Topic classifier

Association rules can be used for classification = "supervised learning"

One problem to be addressed –
items have different "natural" frequency

Split I into disjoint buckets
– Will only find relationships within a category

"Manually" assign min-sup(i) separately for each $i \in I$

$X \subseteq I \quad \{i_1, i_2, \ldots, i_k\}$

When is X frequent? Each i_j has different min-sup()

Natural choice: $\text{min-sup}(X) = \min_j \text{min-sup}(i_j)$

Lose a-priori property!

$\{i_1, i_2, i_3\}$ ⟶ min-sup is 0.1

0.3 0.1 0.4

$\{i_1, i_3\}$ occurs with sup 0.25

not freq — min-sup is <u>0.3</u>

Problem only occurs when smallest min-sup item is dropped

Modified level wise calculation

Earlier - each $X \subseteq I$ ordered acc to I

Instead order by min-sup

$\{i_1, i_2, \ldots, i_k\}$

$$\text{min-sup}(i_1) \leq \text{min-sup}(i_2) \leq \ldots \leq \text{min-sup}(i_k)$$

F_1 ✓

$C_2 = F_1 \times F_1$?

$(i, j) \in C_2$

↓

should meet minsup(i)

$$C_2 = \{(j,k) \mid j \in F_1$$

$$k > j, \text{ k-count} \geq \text{min-sup}\}$$

Finish next time