Data Mining

Find meaningful data patterns from a large data set. Computationally very expensive process.

Frequency of an itemset

\[ \text{Freq} \{ \text{bread, egg} \} = \text{No. of shopping carts that included bread as well as egg} \]

Support of an itemset = \[
\frac{\text{No. of shopping carts with the itemset}}{\text{Total # of shopping carts}}
\]

Support of \{bread, egg\} = \[
\frac{75}{1000} = \frac{75}{300} = 0.25
\]

Association rule: \{LHS itemset\} \Rightarrow \{RHS itemset\}

i.e., if a customer purchases LHS itemset, s/he is likely to purchase the RHS itemset.

\[
\text{Confidence of a rule} = \frac{\text{Support} \{ \text{LHS U RHS} \}}{\text{Support} \{ \text{LHS} \}}
\]

\[
\text{Confidence} \{ \text{milk} \Rightarrow \text{cheese} \} = \frac{\text{Support} \{ \text{milk, cheese} \}}{\text{Support} \{ \text{milk} \}}
\]

\[
\frac{75}{1000} = \frac{75}{300} = 0.25
\]

If \text{Support (RHS)} is close to the confidence value, then the distribution of RHS among all shopping carts is uniform.

\text{Support (cheese)} = \frac{250}{1000} = 0.25 among all shopping carts.
Interest of a rule:

Deviation of the support \{RHS\} from the confidence of the rule.

\{beer\} \Rightarrow \{chips\}

\text{Confidence} = \frac{\text{Support} \{\text{beer, chips}\}}{\text{Support} \{\text{beer}\}} = \frac{150}{450} = \frac{1}{3} = 0.33

\text{Support} \{\text{RHS}\} = \text{Support} \{\text{chips}\} = \frac{600}{1000} = 0.60

\text{interest} = \left| \frac{\text{Confidence} - \text{Support} \{\text{RHS}\}}{\text{Support} \{\text{RHS}\}} \right|

\{beer\} \Rightarrow \{chips\}

= \left| 0.33 - 0.60 \right| = 0.27
Generation of Frequent Itemsets

Frequency of an itemset must be \( \geq \min \text{Support} \) (or Support for the itemset).

E.g., \( \min \text{Support} = 0.4 \)

\( \Rightarrow \text{Frequent Level-1 Itemsets} \)

\( \downarrow \)

\( \text{Level-2} \)

\( \downarrow \)

\( \text{Level-N} \)

Algorithms

- \text{A-Priori}
- \text{Frequent-Pattern}

\text{Brute Force} [Generate all candidate itemsets for each level.]

\text{Generate} [check for minimum frequency for each candidate itemset]

\text{Frequent-Pattern}

- Compute Level-1 Frequent Itemset
- Build Frequent Pattern Tree (FP Tree)
- Prune FP Tree to generate higher-order frequent itemsets.