## Hledání asociací

*„ 90 % spotřebitelů, kteří nakupují zboží A a B, kupují i zboží C a D"*

### Asociační pravidla

$\mathcal{I} = \{i_1, i_2, ..., i_m\}$ množina literálů, položek
$D$ množina transakcí, transakce $T \subseteq \mathcal{I}$ *transakce* např. nákup

$\mathcal{I} = \{i_1, i_2, ..., i_m\}$ množina atributů nad doménou $\{0, 1\}$
$A \subset \mathcal{I}, B \subset \mathcal{I}, A \cap B = \emptyset$ tj. na pravé straně může být víc položek

$$\boxed{A \Rightarrow B[support, confidence]}$$

$support(podpora)$ $s\%$ transakcí v $D$ obsahuje $A$ a současně $B$
$confidence(konfidence, spolehlivost)$ $c\%$ transakcí v $D$, které obsahují $A$, obsahují i $B$

**Úloha:** Najdi všechna asociační pravidla, kde

$support \geq minsup \wedge confidence \geq minconf$

## Základní algoritmus

*naivní algoritmus* $\mathcal{O}(exp(m))$ $m$ ... počet položek

1. Najdi všechny *velké* množiny položek $support \geq minsup$
2. $ABCD, AB$ velké množiny
   $AB \Rightarrow CD[support(ABCD), \frac{support(ABCD)}{support(AB)}]$

## Algoritmus pro nalezení velkých množin

1. Z kandidátů $L_{k-1}$ generuj kandidáty $C_k$ délky $k$

   (a) Pokud se v $L_{k-1}$ vyskytují dvě $(k-1)$-tice, které se liší jen v 1 položce, vytvoř z nich $k$-tici a přidej do $C_k$

   (b) Zruš z $C_k$ ty prvky, jejichž některá $(k-1)$-tice není v $L_{k-1}$

   Příklad

   $L_3 = \{\{ABC\}\{ABD\}\{ACD\}\{ACE\}\{BCD\}\}$
   $C_4 = \{\{ABCD\}\{ACDE\}\}$
   $L_4 = \{\{ABCD\}\}$ protože $\{ADE\} \notin L_3$

2. Z $C_k$ vyber ty, pro něž $support(C_k) > minsup$

prohledávání do šířky

## Generování asociačních pravidel

$l$ ... velká množina, $a \subset l$

$(l-a) \Rightarrow a[supp(l), \frac{supp(l)}{supp(l-a)}]$ , $\quad \frac{supp(l)}{supp(l-a)} \geq minconf$

Příklad

$l = \{ABCD\}$

$a = \{CD\}$ : $AB \Rightarrow CD[support(ABCD), \frac{support(ABCD)}{support(AB)}]$

$a' = \{D\}$ : $ABC \Rightarrow D[support(ABCD), \frac{support(ABCD)}{support(ABC)}]$

1. Generuj množinu pravidel $H_1$ s 1 položkou v konsekventu
2. Z pravidel $H_{k-1}$ tvoř $H_k$

## Velikost učící množiny

**Otevřené problémy**

- $is - a$ hierarchie
- neuvažovala se velikost položek
- pravidla s danými položkami v předpokladu/důsledku