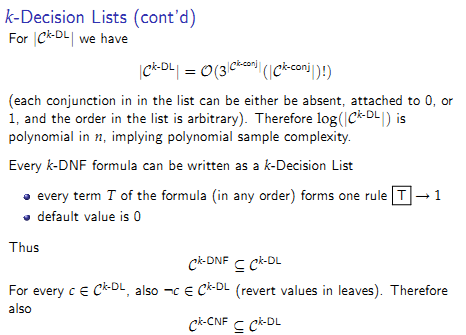
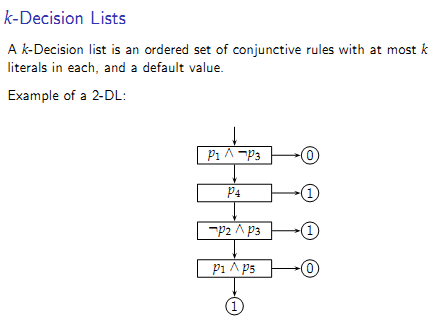
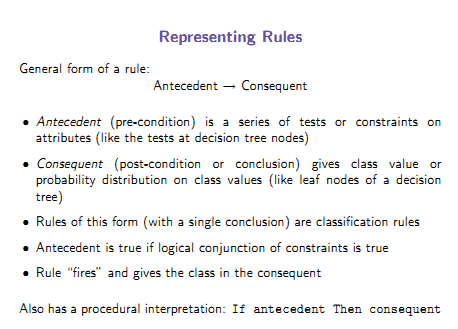
# 13. Učení rozhodovacích pravidel a stromů: pokrývací a rozdělovací strategie, algoritmy AQ a C4.5. Asociační pravidla: algoritmus Apriori.

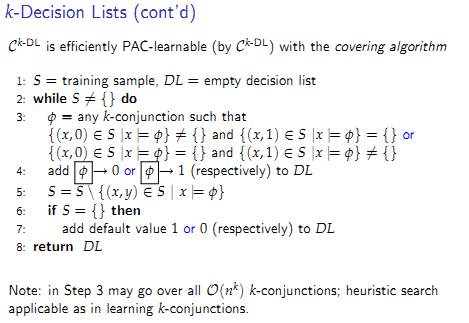


**(23)**

kazdou k-DNF lze zapsat jako k-decision list

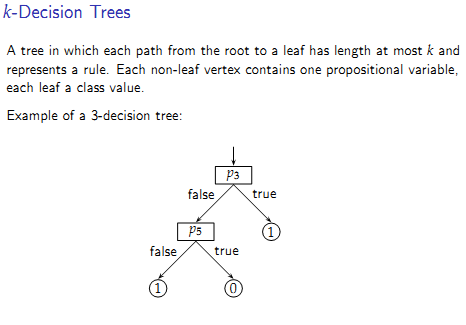
kazdy Term predstavuje jedno pravislo

T -> 1, jinak je 0



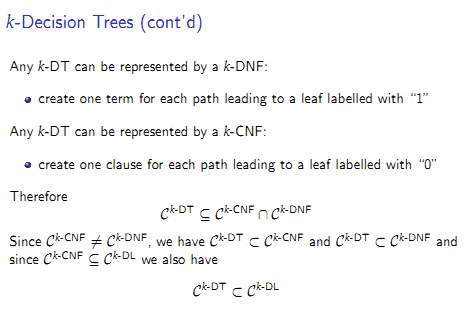
**(24)**

tady je asi jak se ten DL tvori, nejdriv je DL prazdny

nevim, ale udelal bych to jak to bylo nekde predtim napsane, ze Term je jedno rozhodnuti

**(25)** **Decision Trees**

takze na zaklade podminky se to vetvi a **listy** jsou vyhodnoceni **delka** nejvyse k



**(26)**

k-DT lze taky udelat z k-DNF

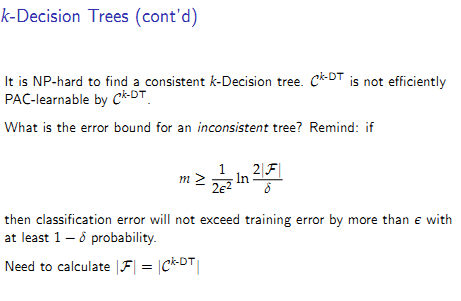
tady je jeden Term jeden path z korene k listu

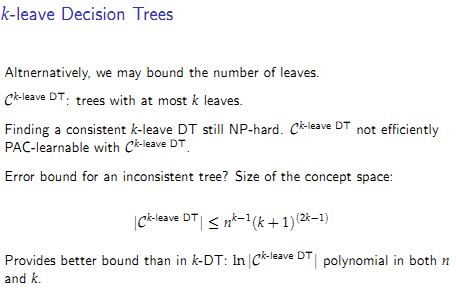
a taky ho lze udelat z CNF

s tim ze, u DNF listy oznacujeme 1 a u CNF 0

takze treba DNF bych tvoril asi takto:

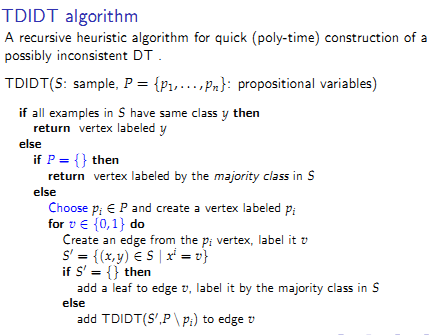
vzal bych termy a udelal paths a k listum bych napsal 1, pak bych to doplnil, aby to bylo binarni a ke zbylym listum doplnil nulu



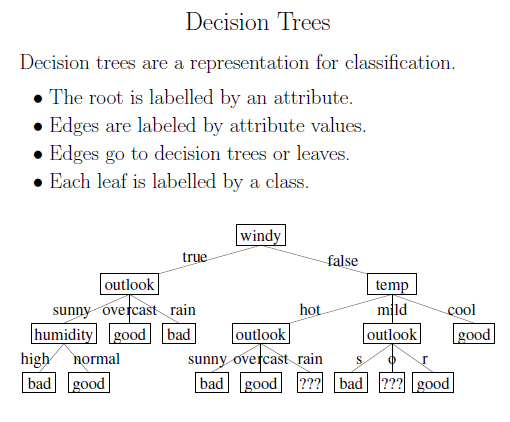


**(29)**

takze muzeme jeste omezit pocet listu, ale je to stejne Nphard



**(30)**

****

alg. pro konstrukci v poly-time, ale nemusi byt konzistentni

to je co konzistentní?

ze ma nulovou chybu na trenovacich datech nejspis

if all examples jsou jablka, return uzel s oznacenim jablko

pokud nejsou vsechno jablka, tak:

kdyz P={}, tak return uzel s labelem nejcastejsiho ovoce

else

vyber pi a vytvor uzel s labelem pi

v tom alg. je modre choose pi na str. 30

tak k tomu asi potrebujem nejakou heuristiku

takze to vybirame maximalizaci zmeny entropie asi

**entropie** je informace, kterou **poskytuje zdroj**

treba budes hazet kostkou tak ti bude davat nejaky informace, nejaky cisla

kdybys mel kostku, ktera by mela na vsech stranach cislo 1 a na jedne strane cislo 2

tak ti bude poskytovat mene informaci, nizsi entropie

protoze i kdybys ji nehodil, tak si muzes byt jisty, ze nejspis padne 1

pokud bych vedel jak ten zdroj vypada ...

kdyby kostka mela vsude jen 1 na vsech stranach, tak entropie bude 0

pze vim co padne ...

cili kostka je dokonale symetricka, usporadana, a zadne informace ti neposkytuje

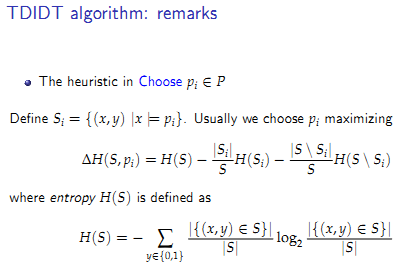
kde je chaos je vysoka entropie otázka předběhla odpoved :-)

a kdy je entroipie maximální?

a pod **heuristikou** si mohu představit zase nějaky alg

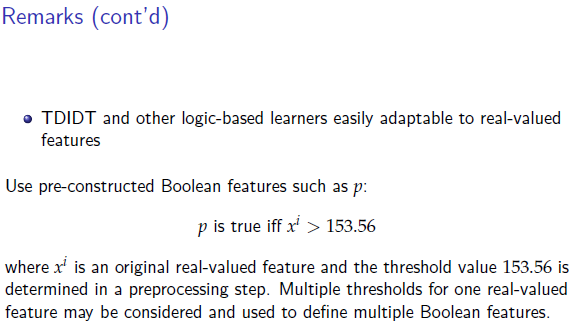
to je nejaky alg. ktery ti radi, co **mas udelat**.. zjednodusene receno

treba chces tahnout figurkou v sachach, tak mas funkci heuristika, ktera ti rekne, kterou figurkou mas tahnout



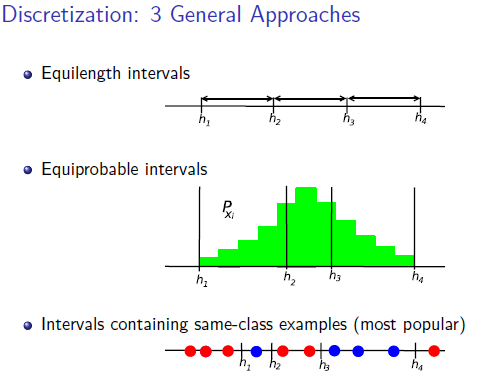
**(32)**

nevim jestli je to dulezite, ale mozna je dobry si uvedomit, ze Decision lists jsou vlastne Decision trees, resp. decision list lze prevest na decision tree, obracene to nemusi jit



**(33)**

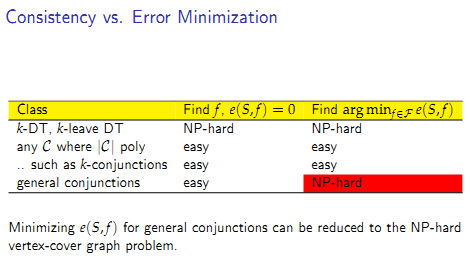
takze real values resime tak, ze delame podminky vetsi mensi, muzeme delat i vice podminek - multiple thresholds



**(34)**

takze intervaly jsou stejne

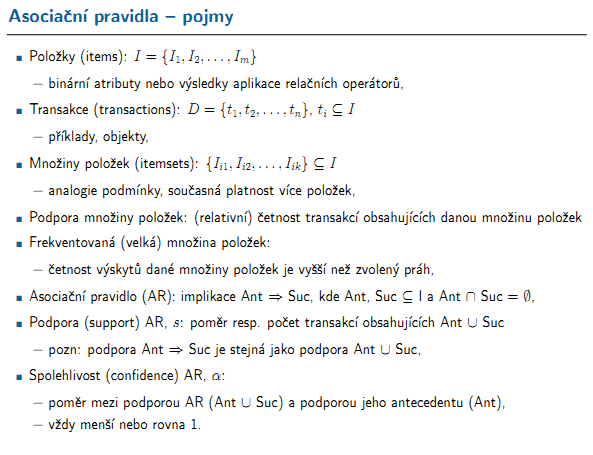
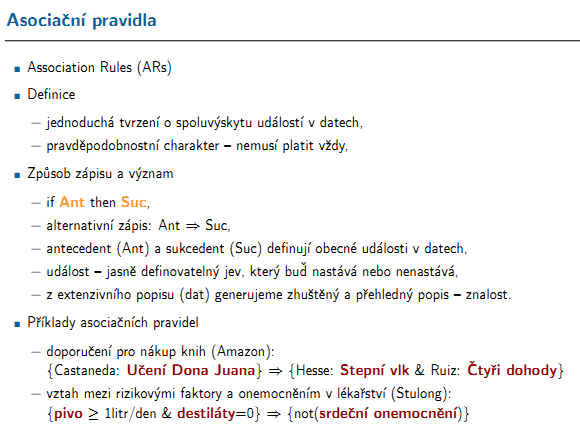
pak intervaly obsahuji stejne mnozstvi zelene, Equiprobable znamená stejně pravdepodobný ...   
a intervaly co obsahuji stejne tridy ... ty jsou most polular ...



**(36)**

takze tady je asi srovna ni nalezeni konzistentniho f a nalezeni nejlepsiho f, ktere nemusi byt konzistentni

## Asociacni pravidla



**(5)**

Ant a Suc (předchůdce a následník)

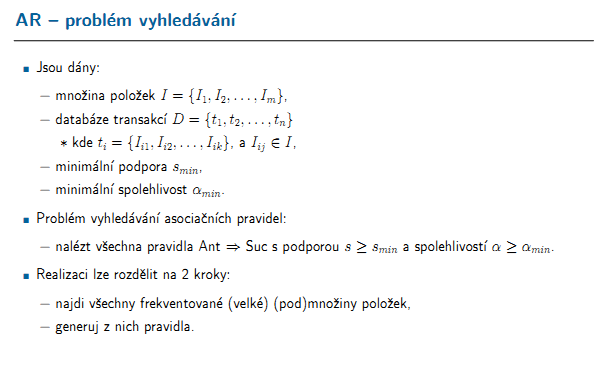
mame eshop, neco prodavame a chceme zjistit nejakou podporu tvrzeni, ze kdyz si nekdo kupuje Pocitac, tak si k nemu prikoupi i Monitor, **takze Pocitac je Ant a Monitor je Suc,** takze mame **transakce** - seznam nakoupenych polozek a je tam tedy ANO/NE, nikoliv pocet

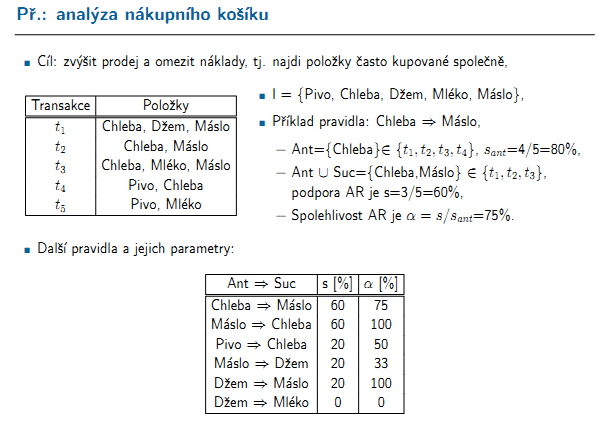
**podpora** je pocet koupenych Pocitacu nebo monitoru, **relativni** je pak, kdyz vydelim **poctem vsech transakci**

**and je prunik**!!! **or je sjednoceni**!!! do pruniku patri ty prvky, ktere jsou v prvni mnozine **A ZAROVEN** ve druhe

do sjednoceni patri ty prvky, ktere jsou v prvni mnozine **NEBO** ve druhe, kdyz si to predstavis s temi mnozinami vybarvenymi, tak to funguje

**spolehlivost** - pomer mezi podporou **Ant U Suc a Ant**

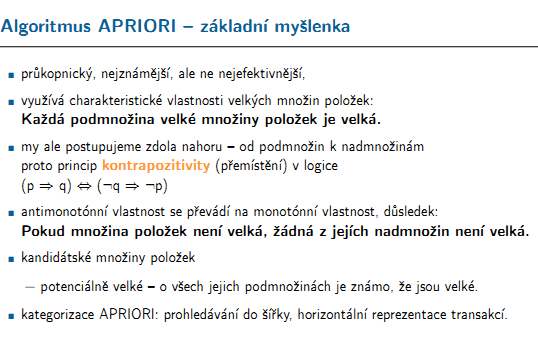
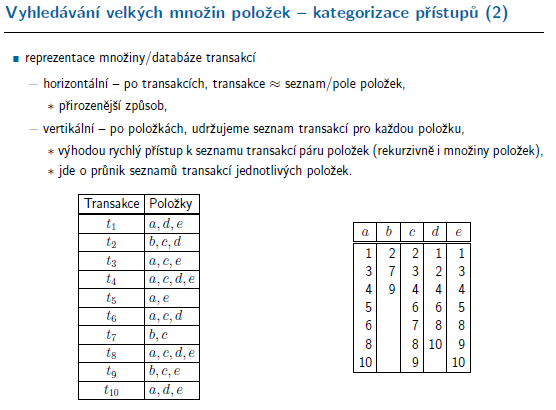
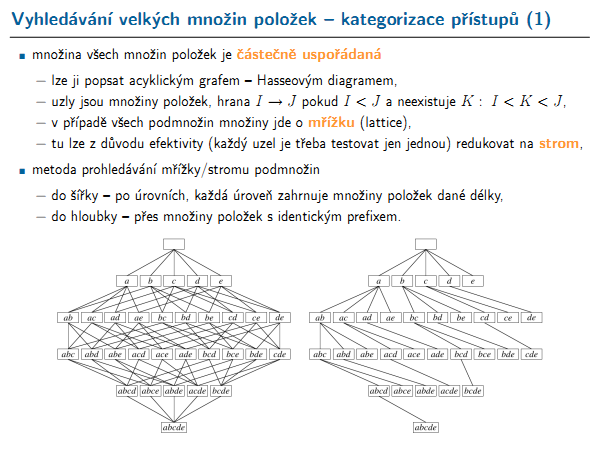


**(6 AR)** no tak potrebujeme proste dobra pravidla, kde plati ze ty hodnoty (podpora a spolehlivost) jsou dostatecne velke, pravidla se hledaji tak, ze se nejprve **najdou caste polozky** a pak se z toho **vytvori pravidla**

**(7)**

takze mame **pravidlo** Chleb -> Maslo, podpora chleba je 4/5, protoze je ve 4 transakcich z 5, jasný ... jedná se spíš o relativní podporu, takze oni hledaji AND (sjednoceni), je to tedy 3/5, spolehlivost taky jasna (60/80=75%)

prostě tam, kde je chleba i maslo, coz je ve trech ... a lomeno všechny transakce

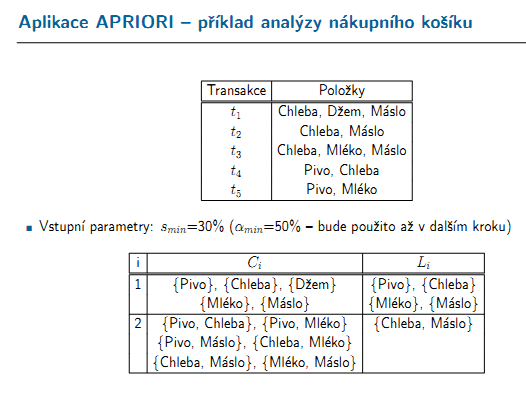
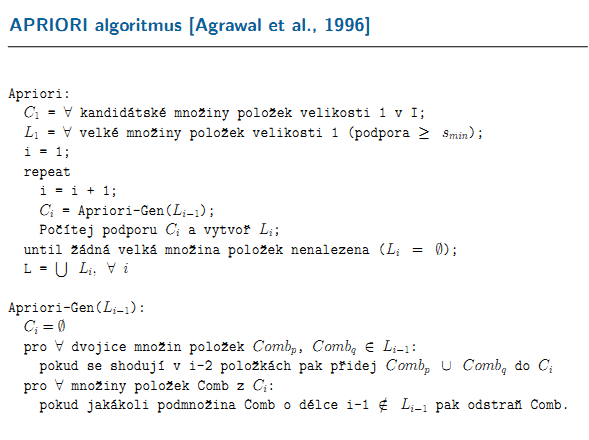
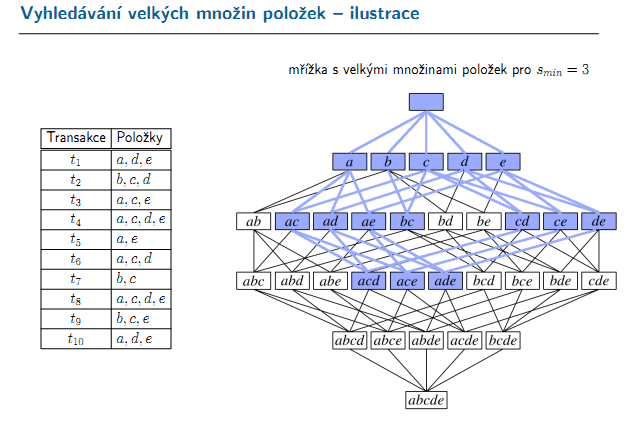


**(11)**

**kotrapozitivita** ... napil jsem se, tak pujdu na WC :D nepujdu na WC, tak jsem nepil ...

**APRIORI**

jako u pismenek je to jasne, kdyz ab nema velkou podporu, tak kdyz tam pridam dalsi pismenko a budu mit abc, tak to taky nemuze mit velkou podporu, to je jasny, protoze pridanim pismenka se **podpora jenom zmensuje**

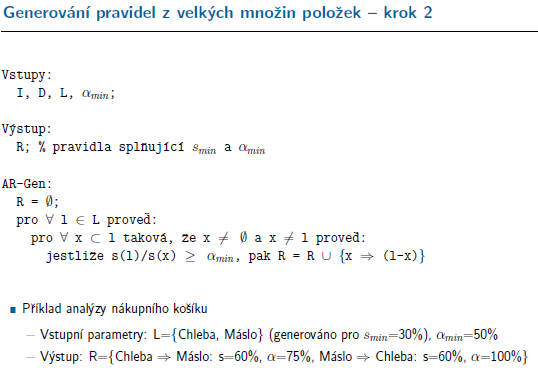
**podmnozina se mysli asi ab a nadmnozina abc**

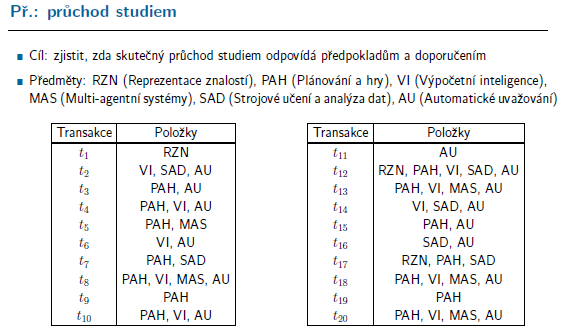
**(14)**

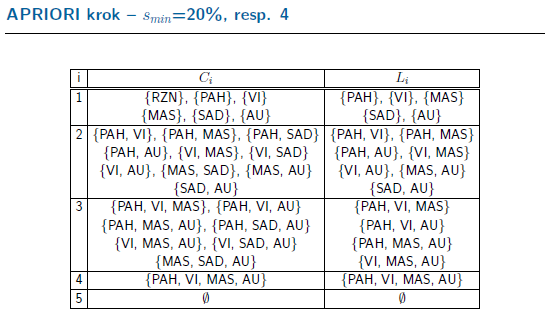
v L neni dzem, musí mít vyšší podporu než 30% a to nemá, vyskytuje se jen v jednom záznamu ...

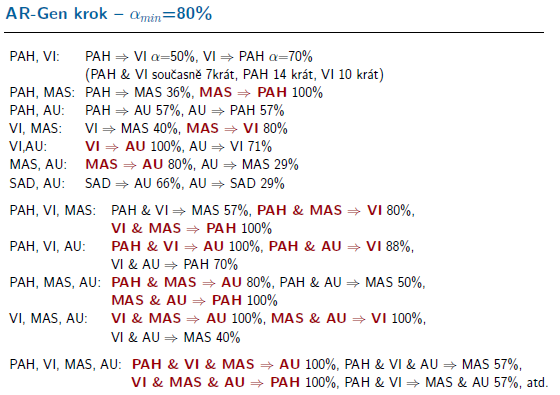
není, protože se vyskytuje jen v prvním záznamu ... pak už ne a to je 1 lomeno 5 = 20%, což je míň než smin = 30

jj, v C je, pak se zkontroluje podpora 1/5, neprojde tak neni v L a do C se zaradi vsechny kombinace z L









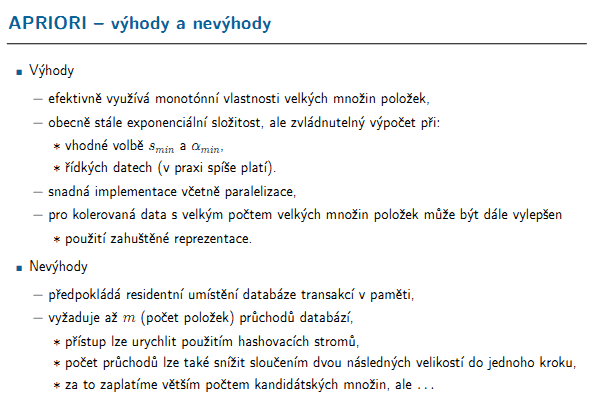
**(22)**

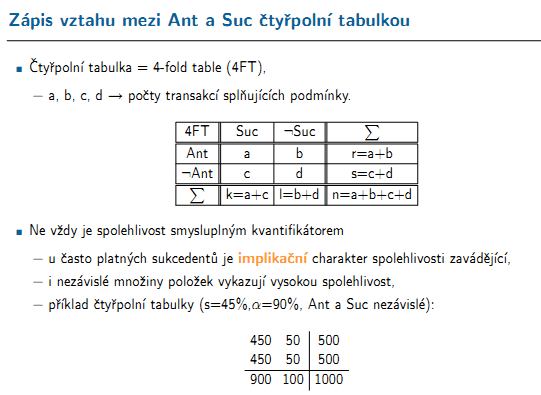
ten **support s** se používá k **prořezávání** a co ten confidence? rekl bych, ze se to vyuzije az v te druhe fazi pri **generaci pravidel**, pravidla – implikace, jj na 21 jsme zjistili, ze nejcastejsi kombinace je PAH, VI, MAS, AU

kouká se do toho L že až? jj, do toho posledniho L, tam jsou prorezane a ted na 22 se hledaji ta pravidla mezi nimi

**spolehlivost** - pomer mezi podporou **Ant U Suc a Ant**, 50% bych rekl jako 7/14 = PAH & VI soucasne 7 krat, a PAH 14 krát

pojdme rozebrat PAH and VI implikuje MAS prosimte ... takze bych to pocital jako:   
sup(PAH & VI & MAS) / sup(PAH & VI) **4 lomeno 7 = 0,5714**, kde **MAS je Suc,** ostatni je **Ant**



**(23)** jak lze snížit počet průchodu? jakože nebudeš rozšiřovat o jedno písmenko, ale rovnou o dvě?

**(24) Čtyřpolní tabulka**

proste pocty kdy to plati, nechápu moc tu tabulku ... jednou je tam suc a potom je tam negace suc, vsechny 4 moznosti, plati neplati suc a ant a napises v kolika pripadech to bylo splneno, a ty čísla v té tabulce, to se zjišťuje od kud? přímo z té tabulky transakcí? z transakcni db, takže kdyby jsi to měl sestavit, potřebuješ tabulku a pravidla ... jo

potrebuju vedet co je ant a co je suc

**45% je podpora** a to je 450/1000, takze **a/n**

**spolehlivost** na slajdech taky maji alfa = 90%, coz by odpovidalo: 450/500, takze spravne je **a/r**

