Bayesovské sítě – řešené úlohy

Shromáždil: Jiří Kléma, klema@labe.felk.cvut.cz

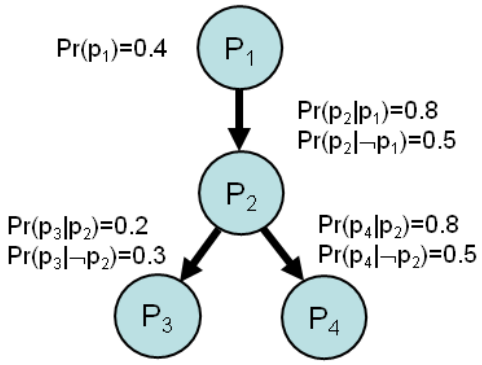
ZS 2011/2012

Cíle materiálu:

Text poskytuje řešené úlohy jako podpůrný výukový materiál ke cvičení předmětu RZN na bayesovské sítě. Materiál doplňuje úlohy uvedené v přednáškách: Grafické pravděpodobnostní modely – úvod, Grafické pravděpodobnostní modely – inference, Grafické pravděpodobnostní modely – učení.

1 Inference

Příklad 1. *Pro síť na obrázku počítejte marginální a podmíněné pravděpodobnosti* $Pr(\neg p_3)$, $Pr(p_2|\neg p_3)$, $Pr(p_1|p_2, \neg p_3)$ *a* $Pr(p_1|\neg p_3, p_4)$. *Použijte metodu* **výčtové inference**.



Výčtová inference pravděpodobnosti vyčísluje vysčítáním přes sdružené pravděpodobnosti atomických událostí. Ty počítá z definice síťového modelu: $Pr(P_1, \ldots, P_n) = Pr(P_1|rodice(P_1)) \times \cdots \times Pr(P_n|rodice(P_n))$. Nevyužívá vztahů podmíněné nezávislosti k dalším zjednodušením. Jde o rutinní, snadno formalizovatelný, ale výpočetně náročný postup. Počet operací roste exponenciálně s počtem proměnných.

$$
\begin{aligned}
Pr(\neg p_3) &= \sum_{P_1,P_2,P_4} Pr(P_1,P_2,\neg p_3,P_4) = \sum_{P_1,P_2,P_4} Pr(P_1)Pr(P_2|P_1)Pr(\neg p_3|P_2)Pr(P_4|P_2) = \\
&= Pr(p_1)Pr(p_2|p_1)Pr(\neg p_3|p_2)Pr(p_4|p_2) + Pr(p_1)Pr(p_2|p_1)Pr(\neg p_3|p_2)Pr(\neg p_4|p_2) + \\
&+ Pr(p_1)Pr(\neg p_2|p_1)Pr(\neg p_3|\neg p_2)Pr(p_4|\neg p_2) + Pr(p_1)Pr(\neg p_2|p_1)Pr(\neg p_3|\neg p_2)Pr(\neg p_4|\neg p_2) + \\
&+ Pr(\neg p_1)Pr(p_2|\neg p_1)Pr(\neg p_3|p_2)Pr(p_4|p_2) + Pr(\neg p_1)Pr(p_2|\neg p_1)Pr(\neg p_3|p_2)Pr(\neg p_4|p_2) + \\
&+ Pr(\neg p_1)Pr(\neg p_2|\neg p_1)Pr(\neg p_3|\neg p_2)Pr(p_4|\neg p_2) + Pr(\neg p_1)Pr(\neg p_2|\neg p_1)Pr(\neg p_3|\neg p_2)Pr(\neg p_4|\neg p_2) = \\
&= .4\times.8\times.8\times.8 + .4\times.8\times.8\times.2 + .4\times.2\times.7\times.5 + .4\times.2\times.7\times.5 + \\
&+ .6\times.5\times.8\times.8 + .6\times.5\times.8\times.2 + .6\times.5\times.7\times.5 + .6\times.5\times.7\times.5 = \\
&= .2048 + .0512 + .028 + .028 + .192 + .048 + .105 + .105 = \mathbf{.762}
\end{aligned}
$$

$$
\begin{aligned}
Pr(p_2|\neg p_3) &= \frac{Pr(p_2,\neg p_3)}{Pr(\neg p_3)} = \frac{.496}{.762} = \mathbf{.6509} \\
Pr(p_2,\neg p_3) &= \sum_{P_1,P_4} Pr(P_1,p_2,\neg p_3,P_4) = \sum_{P_1,P_4} Pr(P_1)Pr(p_2|P_1)Pr(\neg p_3|p_2)Pr(P_4|p_2) = \\
&= Pr(p_1)Pr(p_2|p_1)Pr(\neg p_3|p_2)Pr(p_4|p_2) + Pr(p_1)Pr(p_2|p_1)Pr(\neg p_3|p_2)Pr(\neg p_4|p_2) + \\
&+ Pr(\neg p_1)Pr(p_2|\neg p_1)Pr(\neg p_3|p_2)Pr(p_4|p_2) + Pr(\neg p_1)Pr(p_2|\neg p_1)Pr(\neg p_3|p_2)Pr(\neg p_4|p_2) = \\
&= .4\times.8\times.8\times.8 + .4\times.8\times.8\times.2 + .6\times.5\times.8\times.8 + .6\times.5\times.8\times.2 = \\
&= .2048 + .0512 + .192 + .048 = \mathbf{.496}
\end{aligned}
$$

$$
\begin{aligned}
Pr(p_1|p_2,\neg p_3) &= \frac{Pr(p_1,p_2,\neg p_3)}{Pr(p_2,\neg p_3)} = \frac{.256}{.496} = \mathbf{.5161} \\
Pr(p_1,p_2,\neg p_3) &= \sum_{P_4} Pr(p_1,p_2,\neg p_3,P_4) = \sum_{P_4} Pr(p_1)Pr(p_2|p_1)Pr(\neg p_3|p_2)Pr(P_4|p_2) = \\
&= Pr(p_1)Pr(p_2|p_1)Pr(\neg p_3|p_2)Pr(p_4|p_2) + Pr(p_1)Pr(p_2|p_1)Pr(\neg p_3|p_2)Pr(\neg p_4|p_2) = \\
&= .4\times.8\times.8\times.8 + .4\times.8\times.8\times.2 = .2048 + .0512 = \mathbf{.256} \\
Pr(p_2,\neg p_3) &= Pr(p_1,p_2,\neg p_3) + Pr(\neg p_1,p_2,\neg p_3) = .256 + .24 = \mathbf{.496} \\
Pr(\neg p_1,p_2,\neg p_3) &= \sum_{P_4} Pr(\neg p_1,p_2,\neg p_3,P_4) = \sum_{P_4} Pr(\neg p_1)Pr(p_2|\neg p_1)Pr(\neg p_3|P_2)Pr(P_4|P_2) = \\
&= Pr(\neg p_1)Pr(p_2|\neg p_1)Pr(\neg p_3|p_2)Pr(p_4|p_2) + Pr(\neg p_1)Pr(p_2|\neg p_1)Pr(\neg p_3|p_2)Pr(\neg p_4|p_2) = \\
&= .6\times.5\times.8\times.8 + .6\times.5\times.7\times.2 = .192 + .048 = \mathbf{.24}
\end{aligned}
$$

$$
\begin{aligned}
Pr(p_1|\neg p_3, p_4) &= \frac{Pr(p_1, \neg p_3, p_4)}{Pr(\neg p_3, p_4)} = \frac{.2328}{.5298} = \mathbf{.4394} \\
Pr(p_1, \neg p_3, p_4) &= \sum_{P_2} Pr(p_1, P_2, \neg p_3, p_4) = \sum_{P_2} Pr(p_1)Pr(P_2|p_1)Pr(\neg p_3|P_2)Pr(p_4|P_2) = \\
&= Pr(p_1)Pr(p_2|p_1)Pr(\neg p_3|p_2)Pr(p_4|p_2) + Pr(p_1)Pr(\neg p_2|p_1)Pr(\neg p_3|\neg p_2)Pr(p_4|\neg p_2) = \\
&= .4 \times .8 \times .8 \times .8 + .4 \times .2 \times .7 \times .5 = .2048 + .028 = \mathbf{.2328} \\
Pr(\neg p_3, p_4) &= Pr(p_1, \neg p_3, p_4) + Pr(\neg p_1, \neg p_3, p_4) = .2328 + .297 = \mathbf{.5298} \\
Pr(\neg p_1, \neg p_3, p_4) &= \sum_{P_2} Pr(\neg p_1, P_2, \neg p_3, p_4) = \sum_{P_2} Pr(\neg p_1)Pr(P_2|\neg p_1)Pr(\neg p_3|P_2)Pr(p_4|P_2) = \\
&= Pr(\neg p_1)Pr(p_2|\neg p_1)Pr(\neg p_3|p_2)Pr(p_4|p_2) + Pr(\neg p_1)Pr(\neg p_2|\neg p_1)Pr(\neg p_3|\neg p_2)Pr(p_4|\neg p_2) = \\
&= .6 \times .5 \times .8 \times .8 + .6 \times .5 \times .7 \times .5 = .192 + .105 = \mathbf{.297}
\end{aligned}
$$

Závěr: $Pr(\neg p3) = 0.762$, $Pr(p2|\neg p3) = 0.6509$, $Pr(p_1|p_2, \neg p_3) = 0.5161$, $Pr(p1|\neg p3, p4) = 0.4394$.

Příklad 2. *Pro stejnou síť počítejte stejné marginální a podmíněné pravděpodobnosti. Využijte vlastností orientovaného grafického modelu ke* **zjednodušení** *manuálního výpočtu.*

V případě $Pr(\neg p3)$ (a analogicky i při výpočtu $Pr(p2|\neg p3)$) je P_4 nedotazovaným a nepozorovaným listem. Lze jej vypustit beze změny výsledku.

$$
\begin{aligned}
Pr(\neg p_3) &= \sum_{P_1,P_2} Pr(P_1, P_2, \neg p_3) = \sum_{P_1,P_2} Pr(P_1)Pr(P_2|P_1)Pr(\neg p_3|P_2) = \\
&= Pr(p_1)Pr(p_2|p_1)Pr(\neg p_3|p_2) + Pr(p_1)Pr(\neg p_2|p_1)Pr(\neg p_3|\neg p_2) + \\
&+ Pr(\neg p_1)Pr(p_2|\neg p_1)Pr(\neg p_3|p_2) + Pr(\neg p_1)Pr(\neg p_2|\neg p_1)Pr(\neg p_3|\neg p_2) = \\
&= .4 \times .8 \times .8 + .4 \times .2 \times .7 + .6 \times .5 \times .8 + .6 \times .5 \times .7 = \\
&= .256 + .056 + .24 + .21 = \mathbf{.762}
\end{aligned}
$$

Ke stejnému závěru jako při vypuštění proměnné na základě grafu dojdeme i úpravou výrazu:

$$
\begin{aligned}
Pr(\neg p_3) &= \sum_{P_1,P_2,P_4} Pr(P_1, P_2, \neg p_3, P_4) = \sum_{P_1,P_2,P_4} Pr(P_1)Pr(P_2|P_1)Pr(\neg p_3|P_2)Pr(P_4|P_2) = \\
&= \sum_{P_1,P_2} Pr(P_1)Pr(P_2|P_1)Pr(\neg p_3|P_2) \sum_{P_4} Pr(P_4|P_2) = \sum_{P_1,P_2} Pr(P_1)Pr(P_2|P_1)Pr(\neg p_3|P_2) \times 1
\end{aligned}
$$

Analogické zjednodušení poslouží i k výpočtu $Pr(p2, \neg p3)$ a tím i $Pr(p2|\neg p3)$:

$$
\begin{aligned}
Pr(p_2, \neg p_3) &= \sum_{P_1} Pr(P_1, p_2, \neg p_3) = \sum_{P_1} Pr(P_1)Pr(p_2|P_1)Pr(\neg p_3|p_2) = \\
&= Pr(p_1)Pr(p_2|p_1)Pr(\neg p_3|p_2) + Pr(\neg p_1)Pr(p_2|\neg p_1)Pr(\neg p_3|p_2) = \\
&= .4 \times .8 \times .8 + 6 \times .5 \times .8 = .256 + .24 = \mathbf{.496}
\end{aligned}
$$

V případě $Pr(p_1|p_2, \neg p_3)$ využijeme vztahu $P_1 \perp\!\!\!\perp P_3|P_2$ – P_2 je lineárním spojením mezi P_1 a P_3, pozorováním P_2 se cesta přeruší a uzly P_1 a P_3 jsou d-odděleny. $Pr(p_1|p_2, \neg p_3)$ lze zjednodušit na $Pr(p_1|p_2)$ a využít snadné vyčíslitelnosti této psti (lze si představit, že P_3 i P_4 jsou opět nepozorované a nedotazované listy nebo zjednodušit výraz pro výpočet sdružené psti zanedbáním jednotkového chvostu):

$$\begin{aligned} Pr(p_1|p_2) &= \frac{Pr(p_1, p_2)}{Pr(p_2)} = \frac{.32}{.62} = \mathbf{.5161} \\ Pr(p_1, p_2) &= Pr(p_1)Pr(p_2|p_1) = .4 \times .8 = \mathbf{.32} \\ Pr(p_2) &= Pr(p_1, p_2) + Pr(\neg p_1, p_2) = .32 + 0.3 = \mathbf{.62} \\ Pr(\neg p_1, p_2) &= Pr(\neg p_1)Pr(p_2|\neg p_1) = .6 \times .5 = \mathbf{.3} \end{aligned}$$

V případě $Pr(p_1|\neg p_3, p_4)$ žádné zjednodušení použít nelze.

Závěr: Důsledné využití vlastností grafických modelů a předpočítávání opakujících se výpočtů výrazně zjednodušuje a urychluje výpočet.

Příklad 3. *Pro stejnou síť znovu počítejte podmíněnou pravděpodobnost $Pr(p_1|p_2, \neg p_3)$ pomocí vzorkování. Diskutujte výhody a nevýhody dopředného a Gibbsova vzorkování. V tabulce je k dispozici výstup uniformního náhodného generátoru na intervalu (0,1), použijte jej ke tvorbě vzorků.*

r_1	r_2	r_3	r_4	r_5	r_6	r_7	r_8	r_9	r_{10}
0.2551	*0.5060*	*0.6991*	*0.8909*	*0.9593*	*0.5472*	*0.1386*	*0.1493*	*0.2575*	*0.8407*
r_{11}	r_{12}	r_{13}	r_{14}	r_{15}	r_{16}	r_{17}	r_{18}	r_{19}	r_{20}
0.0827	*0.9060*	*0.7612*	*0.1423*	*0.5888*	*0.6330*	*0.5030*	*0.8003*	*0.0155*	*0.6917*

Nejprve použijeme **dopředné vzorkování**. Uzly uspořádáme topologicky (stačí použít značení proměnných, $P_1 < P_2 < P_3 < P_4$). Vzorky budeme generovat postupně takto:

- $Pr(p_1) > r_1 \rightarrow p_1$,
- $Pr(p_2|p_1) > r_2 \rightarrow p_2$,
- $Pr(p_3|p_2) < r_3 \rightarrow \neg p_3$,
- P_4 je pro danou pravděpodobnost irelevantní,
- $Pr(p_1) < r_4 \rightarrow \neg p_1$,
- $Pr(p_2|\neg p_1) < r_5 \rightarrow \neg p_2$,
- $Pr(p_3|\neg p_2) < r_6 \rightarrow \neg p_3$,
- ...

Použitím prvních 18 náhodných čísel získáme 6 vzorků uvedených v tabulce. V úvahu vezmeme vzorky, které splňují podmínkovou část vyčíslované pravděpodobnosti, pravděpodobnost odhadneme jako podíl z těchto vzorků, u kterých nastal i jev p_1:

	p^1	p^2	p^3	p^4	p^5	p^6
P_1	T	F	T	F	F	F
P_2	**T**	F	**T**	**T**	**T**	F
P_3	**F**	F	**F**	**F**	**F**	F
P_4	?	?	?	?	?	?

$$Pr(p_1|p_2, \neg p_3) \sim \frac{N(p_1, p_2, \neg p_3)}{N(p_2, \neg p_3)} = \frac{2}{4} = 0.5$$

Závěr: I s relativně malým počtem vzorků jsme se přiblížili skutečné hodnotě odhadované pravděpodonosti. Počet vzorků pro spolehlivý odhad by ale musel být řádově vyšší. Nevýhodou dopředného vzorkování je to, že řada vygenerovaných vzorků zůstane nepoužita (viz p^2 a p^6). Zastoupení zbytečně generovaných vzorků by dále narůstalo pro méně časté podmínky uzlů s vysokými topologickými indexy, příkladem je $Pr(p_1|\neg p_3, p_4)$. Pro větší sítě je použití dopředného vzorkování nevhodné.

Gibbsovo vzorkování eliminuje nevýhodu uvedenou výše. Abychom ale mohli generovat vzorky, je třeba vyčíslit pravděpodobnosti $Pr(P_i|MB(P_i))$, kde MB je Markov blanket uzlu P_i, tedy všechny rodiče uzlu P_i, následníci uzlu P_i a jejich rodiče. Vyčíslení se provede pro všechny proměnné, které nejsou dány podmínkou, popřípadě jsou relevantní. V dané úloze bychom museli vyčíslit $Pr(P_1|P_2)$. Pravděpodobnosti $Pr(P_2|P_1, P_3, P_4)$, $Pr(P_3|P_2)$ a $Pr(P_4|P_2)$ není nutné stanovit (P_2 a P_3 jsou pozorované a tudíž zafixované, P_4 je irelevantní), dvě poslední pravděpodobnosti bychom odečetli přímo v síti. Samotné nalezení $Pr(P_1|P_2)$ je ale de facto řešením problému $Pr(p_1|p_2, \neg p_3)$. Plyne z toho, že Gibbsovo vzorkování je vhodné pouze pro větší sítě, kde platí $\forall i = 1 \dots n \; |MB(P_i)| \ll n$.

Závěr: Gibbsovo vzorkování má smysl pro velké sítě, kde $\forall P_i : |MB(P_i)| \ll n$, kde n je celkový počet proměnných.

Příklad 4. *Nechť před budovou fakulty jezdí tři linky tramvaje – 6, 22 a 24. Linka 22 jezdí častěji než 24, ta jezdí častěji než linka 6 (poměr je 5:3:2 a je zachován po celou dobu provozu). Na lince 6 jezdí ve dne krátká souprava v 9 z 10 případů, večer je krátká vždy. Na lince 22 jede krátká souprava jen zřídka večer (pouze v 1 z 10 případů). Na lince 24 se může krátká souprava objevit kdykoli, obvykle ale jezdí souprava dlouhá (8 z 10 případů). Na Albertov jezdí linka 24, linky 6 a 22 jedou na IP Pavlova. Ke změně trasy dojde pouze když tramvaje jednou do vozovny (nechť 24 má vozovnu směrem na IP Pavlova a 6 a 22 směrem na Albertov). Do vozovny jezdí každá desátá tramvaj rovnoměrně po celou dobu provozu. Večer je od 18 do 24 hod, den je od 6 do 18 hod.*

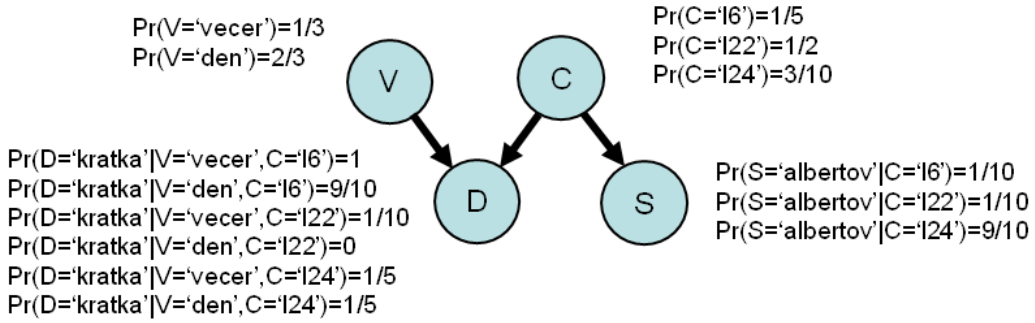
a) Nakreslete **správnou, efektivní a příčinnou** *bayesovskou síť.*

b) Anotujte síť tabulkami podmíněných pstí.

c) Je večer. Do zastávky vjíždí krátká tramvaj. Jaká je pravděpodobnost, že pojede na Albertov?

d) V zastávce na KN čelně pozoruji tramvaj 22. Bude dlouhá nebo krátká?

Ad a) a b)



Které vztahy podmíněné nezávislosti v realitě platí:

- $V \perp\!\!\!\perp C|\emptyset$ – neznám-li délku tramvaje, pak číslo nesouvisí s denní dobou.
- $D \perp\!\!\!\perp S|C$ – znám-li číslo tramvaje, pak délka nesouvisí se směrem.
- $V \perp\!\!\!\perp S|C$ – znám-li číslo tramvaje, pak denní doba nesouvisí se směrem.
- $V \perp\!\!\!\perp S|\emptyset$ – neznám-li délku tramvaje, pak denní doba nesouvisí se směrem.

Ad c) Vyčíslujeme $Pr(S = albertov|V = vecer, D = kratka)$, celá cesta od V k S je otevřena (V je připojen přes pozorovaný konvergentní uzel D, ten přes nepozorovaný divergentní uzel C, platí $S \top\!\!\!\top V|D$, $S \top\!\!\!\top D|\emptyset$). Zjednodušení lze dosáhnout uspořádáním proměnných a vypuštěním S ve jmenovateli:

$$
\begin{aligned}
Pr(S = albertov|V = vecer, D = kratka) &= \frac{Pr(S = albertov, V = vecer, D = kratka)}{Pr(V = vecer, D = kratka)} = \\
&= \frac{\sum_C Pr(V = vecer, C, D = kratka, S = albertov)}{\sum_{C,S} Pr(V = vecer, C, D = kratka, S)} = \\
&= \frac{Pr(V = vecer) \sum_C Pr(C) Pr(D = kratka|V = vecer, C) Pr(S = albertov|C)}{Pr(V = vecer) \sum_C Pr(C) Pr(D = kratka|V = vecer, C) \sum_S Pr(S|C))} = \\
&= \frac{\sum_C Pr(C) Pr(D = kratka|V = vecer, C) Pr(S = albertov|C)}{\sum_C Pr(C) Pr(D = kratka|V = vecer, C)} = \\
&= \frac{\frac{1}{5} \times 1 \times \frac{1}{10} + \frac{1}{2} \times \frac{1}{10} \times \frac{1}{10} + \frac{3}{10} \times \frac{1}{5} \times \frac{9}{10}}{\frac{1}{5} \times 1 + \frac{1}{2} \times \frac{1}{10} + \frac{3}{10} \times \frac{1}{5}} = \mathbf{.2548}
\end{aligned}
$$

Ad d) K vyčíslení $Pr(D = dlouha|C = 22)$ postačí informace uložená v uzlech V a D, provedeme vysčítání proměnné V:

$$
\begin{aligned}
Pr(D = dlouha|C = 22) &= \sum_V Pr(D = dlouha, V|C = 22) = \\
&= \sum_V Pr(V|C = 22) \times Pr(D = dlouha|V, C = 22) = \\
&= \sum_V Pr(V) \times Pr(D = dlouha|V, C = 22) = \\
&= \frac{1}{3} \times \frac{9}{10} + \frac{2}{3} \times 1 = \mathbf{.9667}
\end{aligned}
$$

2 Testy (podmíněné) nezávislosti, volba struktury

Příklad 5. *Na základě níže uvedené frekvenční tabulky rozhodněte o vztahu mezi veličinami A a B.*

	c		$\neg c$	
	b	$\neg b$	b	$\neg b$
a	14	8	25	56
$\neg a$	54	25	7	11

Lze uvažovat vztah nezávislosti ($A \perp\!\!\!\perp B|\emptyset$) a podmíněné nezávislosti ($A \perp\!\!\!\perp B|C$). Budeme demonstrovat tři postupy analýzy možného vztahu. První vychází z definice nezávislosti a je pouze ilustrativní. Další dva vychází z prakticky používaných metod.

Postup 1: Prosté porovnání (podmíněných) pravděpodobností.

Na nezávislost lze usuzovat z níže uvedených vztahů:
$A \perp\!\!\!\perp B|\emptyset \Leftrightarrow Pr(A,B) = Pr(A)Pr(B) \Leftrightarrow Pr(A|B) = Pr(A) \wedge Pr(B|A) = Pr(B)$

Na základě frekvenční tabulky proveďme MLE odhad výše uvedených pstí:
$Pr(a|b) = \frac{39}{100} = 0.39, Pr(a|\neg b) = \frac{64}{100} = 0.64, Pr(a) = \frac{103}{200} = 0.51$
$Pr(b|a) = \frac{39}{103} = 0.38, Pr(b|\neg a) = \frac{61}{97} = 0.63, Pr(b) = \frac{100}{200} = 0.5$

Na podmíněnou nezávislost lze usuzovat z níže uvedených vztahů:
$A \perp\!\!\!\perp B|C \Leftrightarrow Pr(A,B|C) = Pr(A|C)Pr(B|C) \Leftrightarrow Pr(A|B,C) = Pr(A|C) \wedge Pr(B|A,C) = Pr(B|C)$

Na základě frekvenční tabulky proveďme MLE odhad výše uvedených pstí:
$Pr(a|b,c) = \frac{14}{68} = 0.21, Pr(a|\neg b,c) = \frac{8}{33} = 0.24, Pr(a|c) = \frac{22}{101} = 0.22$
$Pr(a|b,\neg c) = \frac{25}{32} = 0.78, Pr(a|\neg b,\neg c) = \frac{56}{67} = 0.84, Pr(a|\neg c) = \frac{81}{99} = 0.82$
$Pr(b|a,c) = \frac{14}{22} = 0.64, Pr(b|\neg a,c) = \frac{54}{79} = 0.68, Pr(b|c) = \frac{68}{101} = 0.67$
$Pr(b|a,\neg c) = \frac{25}{81} = 0.31, Pr(b|\neg a,\neg c) = \frac{7}{18} = 0.39, Pr(b|\neg c) = \frac{32}{99} = 0.32$

Lze vidět, že vztah nezávislosti spíše neplatí, uvedené definiční rovnosti neplatí. Vztah podmíněné nezávislosti spíše platí, protože pravděpodobnosti se ve všech řádcích zhruba rovnají. Potřebujeme ale vědečtější nástroj k potvrzení či vyvrácení obou vztahů, dvě možnosti naznačíme dále.

Postup 2: Statistické testování hypotéz.

Nezávislost lze testovat například Pearsonovým χ^2 testem dobré shody. Pro ověření $A \perp\!\!\!\perp B|\emptyset$ test aplikujeme na kontingenční (frekvenční) tabulku pro veličiny A a B (tabulka vlevo):

O_{AB}	b	$\neg b$	sum
a	39	64	103
$\neg a$	61	36	97
sum	100	100	200

E_{AB}	b	$\neg b$	sum
a	51.5	51.5	103
$\neg a$	48.5	48.5	97
sum	100	100	200

Nulovou hypotézou je nezávislost veličin. Test pracuje s očekávanými četnostmi za předpokladu platnosti nulové hypotézy (tabulka vpravo):

$$E_{AB} = \frac{N_A \times N_B}{N} \quad \rightarrow \quad E_{a\neg b} = \frac{N_a \times N_{\neg b}}{N} = \frac{103 \times 100}{200} = 51.5$$

Tyto očekávané četnosti srovnává s pozorovanými. Testová statistika:

$$\chi^2 = \sum_{A,B} \frac{(O_{AB} - E_{AB})^2}{E_{AB}} = 12.51 \gg \chi^2(\alpha = 0.05, df = 1) = 3.84$$

Nulovou hypotézu lze zamítnout ve prospěch alternativní hypotézy o závislosti veličin na hladině významnosti $\alpha = 0.05$. Frekvenční tabulku s pozorovanou nebo větší odchylkou od očekávaných hodnot lze při použití nezávislého modelu pozorovat v minimálním procentu případů $p = 0.0004 \ll \alpha$. **Veličiny A a B jsou závislé.**

Analogicky lze ověřit i $A \perp\!\!\!\perp B|C$[1]. Testovou statistiku χ^2 testu napočítáme odděleně pro obě kontingenční tabulky pro c a $\neg c$. Celková hodnota statistiky je dána součtem obou dílčích statistik, má dva stupně volnosti.

O_{AB}	c			$\neg c$		
	b	$\neg b$	sum	b	$\neg b$	sum
a	14	8	22	25	56	81
$\neg a$	54	25	79	7	11	18
sum	68	33	101	32	67	99

E_{AB}	c			$\neg c$		
	b	$\neg b$	sum	b	$\neg b$	sum
a	14.8	7.2	22	26.2	54.8	81
$\neg a$	53.2	25.8	79	5.8	12.2	18
sum	68	33	101	32	67	99

Nulovou hypotézou je podmíněná nezávislost veličin, alternativní hypotézou je úplný model se všemi parametry. Testová statistika:

$$\chi^2 = \sum_{A,B|C} \frac{(O_{AB|C} - E_{AB|C})^2}{E_{AB|C}} = 0.175 + 0.435 = 0.61 \ll \chi^2(\alpha = 0.05, df = 2) = 5.99$$

Nulovou hypotézu nelze zamítnout ve prospěch alternativní hypotézy o saturovaném modelu na hladině významnosti $\alpha = 0.05$. Frekvenční tabulku s danou nebo větší odchylkou od očekávaných hodnot lze při použití podmíněně nezávislého modelu pozorovat ve většině případů – $p = 0.74 \gg \alpha$. **Veličiny A a B jsou podmíněně nezávislé za předpokladu znalosti C.** Proměnná C vysvětluje závislost mezi A a B.

Postup 3: Srovnání hodnocení dvou struktur modelu.

Nejprve srovnáme nulový (A a B nezávislé) a alternativní (A a B závislé) model dvou proměnných, viz obrázek níže.

[1]Prakticky se test dobré shody pro testování podmíněné nezávislosti nepoužívá pro malou sílu. Je nahrazen mj. **testem poměrem věrohodností**, který srovnává věrohodnost null modelu (AC,BC) s věrohodností alt modelu (AC,BC,AB). Nulový model předpokládá nulovou interakci mezi A a B, uvažuje pouze vztahy mezi A a C, resp. B a C. Alternativní model bere v úvahu i možný vztah mezi A a B.

Porovnáme hodnoty BIC (a bayesovského) kritéria obou modelů. Vybereme strukturu s vyšším ohodnocením. Pokud chceme pojmout jako statistický test, provedeme test poměrem věrohodností.

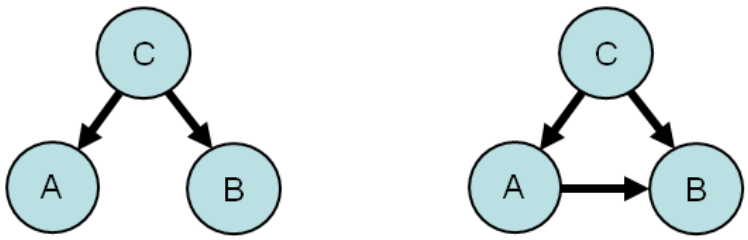
$\ln L_{null} = (39+64)\ln\frac{103}{200}\frac{100}{200} + (61+36)\ln\frac{97}{200}\frac{100}{200} = -277.2$
$\ln L_{alt} = 39\ln\frac{103}{200}\frac{39}{103} + 64\ln\frac{103}{200}\frac{64}{103} + 61\ln\frac{97}{200}\frac{61}{97} + 36\ln\frac{97}{200}\frac{36}{97} = -270.8$
$BIC(null) = -\frac{K}{2}\ln M + \ln L_{null} = -\frac{2}{2}\ln 200 - 277.2 = -282.5$
$BIC(alt) = -\frac{K}{2}\ln M + \ln L_{alt} = -\frac{3}{2}\ln 200 - 270.8 = -278.8$
$BIC(null) < BIC(alt) \Leftrightarrow$ alternativní model je věrohodnější, nulový předpoklad $A \perp\!\!\!\perp B|\emptyset$ neplatí.

Na základě výpočtu bayesovského hodnotícího kritéria (stanoveno v Matlab BNT, fce *score_dags*), lze určit: $\ln Pr(D|null) = -282.9 < \ln Pr(D|alt) = -279.8 \Leftrightarrow$ alternativní model je věrohodnější, nulový předpoklad $A \perp\!\!\!\perp B|\emptyset$ neplatí.

Varianta testu poměrem věrohodností:
$D = -2(\ln L_{null} - \ln_{alt}) = -2(-277.2 + 270.8) = 12.8$
D statistika má χ^2 rozdělení s $3-2=1$ stupněm volnosti. Nulová hypotéza má p=0.0003 a lze ji zamítnout.

Závěr: **veličiny A a B jsou závislé.**

Analogicky srovnáme nulový (A a B podmíněně nezávislé) a alternativní (A a B podmíněně závislé) model tří proměnných, viz obrázek níže.



nulový model alternativní model

Opět porovnáme hodnoty kritérií obou modelů a vybereme strukturu s vyšším ohodnocením.

Na základě výpočtu $\ln L_{null}$ a $\ln L_{alt}$ (stanoveno v Matlab BNT, fce *log_lik_complete*), lze určit:
$BIC(null) = -\frac{K}{2}\ln M + \ln L_{null} = -\frac{5}{2}\ln 200 - 365.1 = -377.9$
$BIC(alt) = -\frac{K}{2}\ln M + \ln L_{alt} = -\frac{7}{2}\ln 200 - 364.3 = -382.9$
$BIC(null) > BIC(alt) \Leftrightarrow$ nulový model je věrohodnější, předpoklad $A \perp\!\!\!\perp B|C$ platí.

Na základě výpočtu bayesovského hodnotícího kritéria (stanoveno v Matlab BNT, fce *score_dags*), lze určit: $\ln Pr(D|null) = -379.4 > \ln Pr(D|alt) = -385.5 \Leftrightarrow$ alternativní model je méně věrohodný, předpoklad $A \perp\!\!\!\perp B|C$ platí.

Varianta testu poměrem věrohodností:
$D = -2(\ln L_{null} - \ln_{alt}) = -2(-365.1 + 364.3) = 1.6$
D statistika má χ^2 rozdělení se $7-5=2$ stupni volnosti. Nulová hypotéza má p=0.45 a nelze ji zamítnout.

Závěr: **veličiny A a B jsou podmíněně nezávislé za předpokladu znalosti C.**

3 Dynamické bayesovské sítě

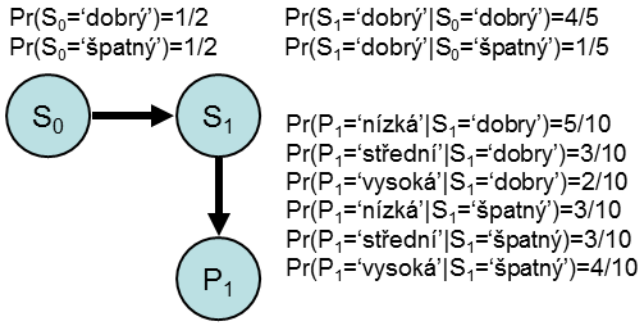
Příklad 6. *U pacienta s nemocí N lékaři každý den měří hodnotu klíčového parametru P, ta může nabývat hodnot {nízká, střední, vysoká}. Hodnota P je ovlivněna pacientovým aktuálním skrytým stavem S. Pro S rozlišujeme hodnoty {dobrý, špatný}. Mezi dvěma následujícími dny dochází ke změně stavu v pětině případů. Je-li pacient v dobrém stavu je hodnota parametru P spíše nízká (z 10 měření je průměrně 5 nízkých, 3 střední a 2 vysoká), je-li pacient ve špatném stavu je hodnota spíše vysoká (z 10 měření jsou průměrně 3 nízká, 3 střední a 4 vysoká). Při příchodu do nemocnice v den 0 byl stav pacienta neznámý, tj.* $Pr(S_0 = dobry) = 0.5$.

a) Nakreslete přechodový a senzorický model dynamické bayesovské sítě modelující zadání,

b) určete pravděpodobnost, že pacient je v dobrém stavu ve dni 2, pokud hodnota P ve dni 1 i 2 byla nízká,

c) lze bez dalších výpočtů rozhodnout o nejpravděpodobnějším průchodu pacienta stavy ve dnech 0, 1 a 2?, zdůvodněte.

ad a) Jde o úlohu **filtrace**. Přechodový model vyjadřuje vztah mezi stavy, senzorický model vztah mezi stavem a senzorickou veličinou. Oba modely jsou na následujícím obrázku:



ad b) vyčíslujeme $Pr(s_2|P_1 = nizka, P_2 = nizka)$ (pracujeme s notací s dobrý stav, $\neg s$ špatný stav):

$$Pr(S_1|P_1 = nizka) = \alpha_1 Pr(P_1 = nizka|S_1) \sum_{S_0 \in \{s_0, \neg s_0\}} Pr(S_1|S_0) Pr(S_0)$$

$$Pr(s_1|P_1 = nizka) = \alpha_1 \times 0.5 \times 0.5 = 0.625$$

$$Pr(\neg s_1|P_1 = nizka) = \alpha_1 \times 0.3 \times 0.5 = 0.375$$

$$Pr(S_2|P_1 = nizka, P_2 = nizka) = \alpha_2 Pr(P_2 = nizka|S_2) \sum_{S_1 \in \{s_1, \neg s_1\}} Pr(S_2|S_1) Pr(S_1)$$

$$Pr(s_2|P_1 = nizka, P_2 = nizka) = \alpha_2 \times 0.5(0.8 \times 0.625 + 0.2 \times 0.375) = \alpha_2 \times 0.2875 = \mathbf{0.6928}$$

$$Pr(\neg s_2|P_1 = nizka, P_2 = nizka) = \alpha_2 \times 0.3(0.2 \times 0.625 + 0.8 \times 0.375) = \alpha_2 \times 0.1275 = 0.3072$$

Problém lze složitěji řešit i klasickou inferencí:

$$Pr(s_2|P_1 = nizka, P_2 = nizka) = \frac{Pr(s_2, P_1 = nizka, P_2 = nizka)}{Pr(P_1 = nizka, P_2 = nizka)} =$$
$$= \frac{\sum_{S_0,S_1} Pr(S_0, S_1, s_2, P_1 = nizka, P_2 = nizka)}{\sum_{S_0,S_1,S_2} Pr(S_0, S_1, S_2, P_1 = nizka, P_2 = nizka)} =$$
$$= \frac{Pr(s_0)Pr(s_1|s_0)Pr(s_2|s_1)Pr(P_1 = nizka|s_1)Pr(P_2 = nizka|s_2) + \ldots}{Pr(s_0)Pr(s_1|s_0)Pr(s_2|s_1)Pr(P_1 = nizka|s_1)Pr(P_2 = nizka|s_2) + \ldots} = \ldots$$

ad c) Nelze, nejpravděpodobnější průchod $Pr(S_{1:2}|P_{1:2})$ je úloha jiná než stanovení psti stavů v jednotlivých dnech. Stavy se navzájem ovlivňují, při filtrování je navíc mezivýpočet ve dni 1 pouze přibližný (nebere v úvahu následná pozorování). Pro daný model by šel použít Viterbiho algoritmus (algoritmus dynamického programování používaný v HMM).