Vytěžování dat, přednáška 3: Grafické pravděpodobnostní modely

Filip Železný

Evropský sociální fond
Praha & EU: Investujeme do vaší budoucnosti

Fakulta elektrotechnická, ČVUT

Mnoharozměrné rozdělení

- Minulá přednáška: odhad rozdělení pro data s jedním příznakem $\vec{x} = x$ resp. dvěma příznaky $\vec{x} = (x_1, x_2)$
  - jednorozměrné resp. dvourozměrné rozdělení $P(\vec{x})$
- V této přednášce: odhadujeme rozdělení pro více příznaků (rozměrů)
  - $P(\vec{x}) = P(x_1, x_2, \ldots x_n)$
- Příklad:

	Věk	**Pohlaví**	**Kuřák**	**Rakovina**
$\vec{x}_1$:	56	muž	$+$	$+$
$\vec{x}_2$:	32	žena	$-$	$-$
$\vec{x}_3$:	48	žena	$+$	$+$
$\vec{x}_4$:	60	muž	$+$	$+$

Značení náhodných veličin

Věk	Pohlaví	Kuřák	Rakovina
56	muž	+	+
32	žena	−	−
48	žena	+	+
60	muž	+	+

- Vektorové/maticové značení

$$\vec{x}_3 = (48, \text{žena}, -, -)$$
$$x_{3,2} = \text{žena}$$

- Pomocí prvních písmen příznaků (= náhodných veličin)

$$(V_3, P_3, K_3, R_3) = (48, \text{žena}, -, -)$$
$$P_3 = \text{žena}$$

- Obor hodnot

$$X = \mathcal{V} \times \mathcal{P} \times \mathcal{K} \times \mathcal{R} = \{1, 2, \ldots 100\} \times \{\text{muž}, \text{žena}\} \times \{+, \text{-}\} \times \{+, \text{-}\}$$

Odvozování ze sdruženého rozdělení

Známe-li sdružené rozdělení všech příznaků, můžeme odvodit rozdělení (sdružené i podmíněné) přes kteroukoliv podmnožinu příznaků. Např.

- pravděpodobnost, že osoba je muž - kuřák

$$P_{P,K}(\text{muž}, +) = \sum_{v \in \mathcal{V}} \sum_{r \in \mathcal{R}} P_{V,P,K,R}(v, \text{muž}, +, r)$$

- pravděpodobnost, že kouřící muž má rakovinu

$$\begin{aligned} P_{R|P,K}(+|\text{muž}, +) &= \frac{P_{R,P,K}(+, \text{muž}, +)}{P_{P,K}(\text{muž}, +)} \\ &= \frac{\sum_{v \in \mathcal{V}} P_{V,P,K,R}(v, \text{muž}, +, +)}{\sum_{v \in \mathcal{V}} \sum_{r \in \mathcal{R}} P_{V,P,K,R}(v, \text{muž}, +, r)} \end{aligned}$$

Reprezentace sdruženého rozdělení

- **Parametrická**: např. *mnoharozměrné normální rozdělení* s parametry: vektor $\vec{\mu}$ středních hodnot a matice Σ tzv. kovariancí.
  (Mimo rozsah tohoto předmětu)
- **Neparametrická**: jedno číslo $\in [0; 1]$ pro každou kombinaci hodnot příznaků. V našem příkladě tedy:

$$|\mathcal{V}| \cdot |\mathcal{P}| \cdot |\mathcal{K}| \cdot |\mathcal{R}| = 100 \cdot 2 \cdot 2 \cdot 2 = 800$$

(Odpovídá 4-rozměrné kontingenční tabulce.)

Ve skutečnosti stačí 799 čísel. Proč?

$$\sum_{v \in \mathcal{V}} \sum_{p \in \mathcal{P}} \sum_{k \in \mathcal{K}} \sum_{r \in \mathcal{R}} P_{V,P,K,R}(v, p, k, r) = 1$$

tedy jednu pravděpodobnost dopočítáme z ostatních.

Kombinatorická exploze

Problémy s neparametrickým sdruženým rozdělelním:

- **Paměťová náročnost.** I kdyby všechny příznaky byly pouze binární, potřebujeme pro reprezentaci sdruženého rozdělení $2^n - 1$ čísel, kde n je počet příznaků. Exponenciální nárůst!
  - Např. pro $n = 40$, jedno číslo - float - 4 bajty $\Rightarrow$ potřebujeme přes 4 TB
- **Datová náročnost.** Pro odhad každého čísla (pravděpodobnosti) z relativní četnosti, např.

$$P_{V,P,K,R}(30, \text{muž}, \text{-}, \text{-}) \approx \frac{\text{počet 30-letých zdravých nekuřáků v datech}}{\text{počet dat}}$$

roste potřebný počet dat také exponenciálně. (Pro danou spolehlivost odhadu)

Jak z toho ven?

Využití nezávislosti

- Kdyby byly všechny příznaky navzájem nezávislé, tak

$$P_{V,P,K,R} = P_V \cdot P_P \cdot P_K \cdot P_R$$

a stačí tak znát jen 4 marginální rozdělení, zde tedy

$$(100 - 1) + (2 - 1) + (2 - 1) + (2 - 1) = 102$$

čísel (místo původních 799).

- Obecně pro binární příznaky: n čísel místo $2^n - 1$ čísel.
- Většinou ale všechny příznaky navzájem nezávislé nejsou!

Využití nezávislosti (pokr.)

- I nezávislost jedné veličiny na ostatních znamená značné ulehčení:

Věk	Pohlaví	Kuřák	Rakovina	Měsíc narození
56	muž	+	+	7
32	žena	−	−	2
48	žena	+	+	12
60	muž	+	+	6

$$\underbrace{P_{V,P,K,R,M}}_{100\cdot 2\cdot 2\cdot 2\cdot 12-1=9599 \text{ čísel}} = \underbrace{P_{V,P,K,R} \cdot P_M}_{800-1+12-1=810 \text{ čísel}}$$

- Ale ani nezávislost jediné veličiny nelze obvykle předpokládat.

Podmíněná nezávislost

Pozorování: výskyt rakoviny R je závislý na pohlaví P, tj.

$$P_{R|P} \neq P_R, \text{ ekvivalentně: } P_{P|R} \neq P_P$$

ale pouze proto, že muži častěji kouří. Tedy jakmile víme, zda osoba kouří, na pohlaví už nezáleží

$$P_{R|P,K} = P_{R|K}, \text{ ekvivalentně: } P_{P|R,K} = P_{P|K}$$

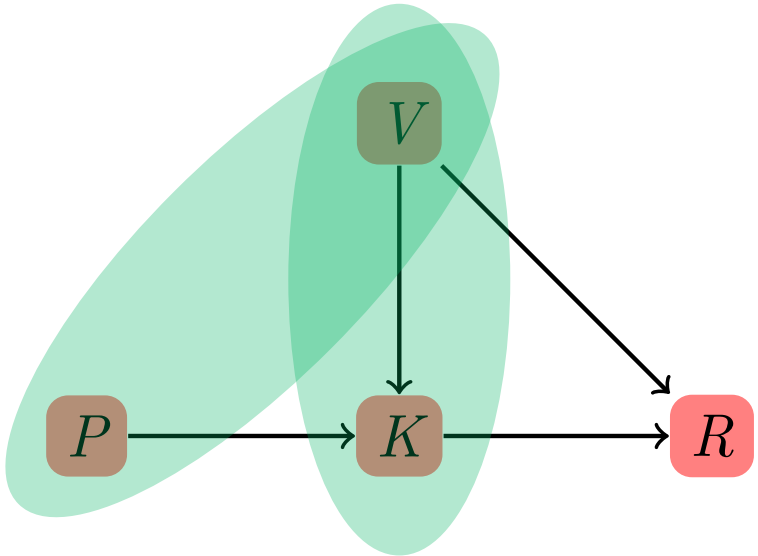
R a P jsou tedy podmíněně nezávislé, přičemž podmínkou je K.

Totéž jinými slovy:

- V celé populaci mají častěji rakovinu muži: $P_{R|P}(+|\text{muž}) > P_R(+)$
- U kuřáků už na pohlaví nezáleží: $P_{R|P,K}(+|\text{muž}, +) = P_{R|K}(+|+)$
- Totéž u nekuřáků: $P_{R|P,K}(+|\text{muž}, -) = P_{R|K}(+|-)$

Grafické znázornění podmíněných nezávislostí

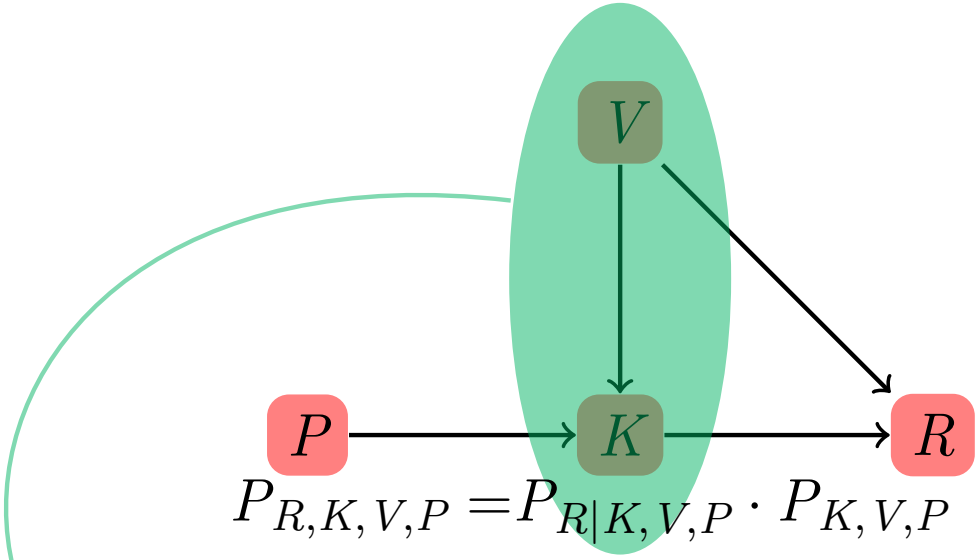
Pro dané rodiče uzlu je uzel podmíněně nezávislý na všech uzlech, které nejsou jeho potomky.



- Orientovaný graf bez cyklů

Pro dané rodiče uzlu je uzel podmíněně nezávislý na všech uzlech, které nejsou jeho potomky. Kouření (K) závisí na všech ostatních příznacích. Výskyt rakoviny (R) závisí na kouření (K) a věku (V), ale pro dané K a V nezávisí na pohlaví P.

Výpočet sdruženého rozdělení



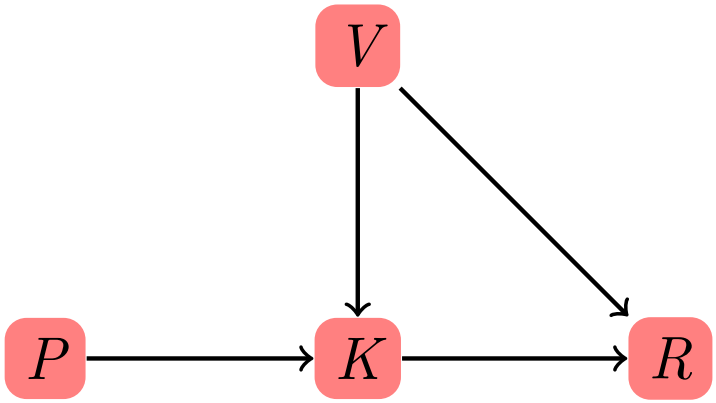
$$P_{R,K,V,P} = P_{R|K,V,P} \cdot P_{K,V,P}$$

$$= P_{R|K,V} \cdot P_{K,V,P}$$

$$= P_{R|K,V} \cdot P_{K|V,P} \cdot P_{V,P}$$

$$= P_{R|K,V} \cdot P_{K|V,P} \cdot P_V \cdot P_P \quad \text{(nezávislost } V \text{ a } P\text{)}$$

Výpočet sdruženého rozdělení (pokr.)



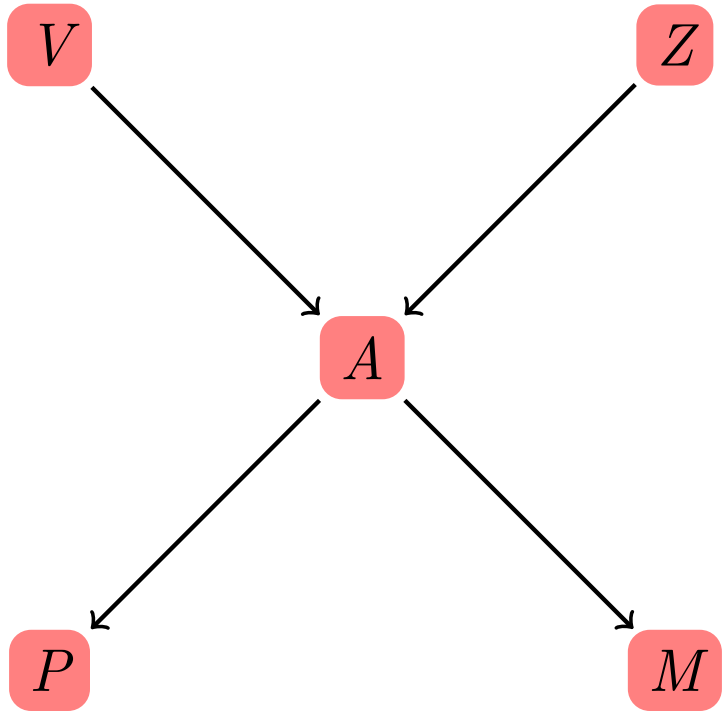
Obecně pro příznaky $X = X_1 \times X_2 \times \ldots \times X_n$:

$$P_X = \Pi_{i=1}^{n} P_{X_i|\text{rodiče}(X_i)}$$

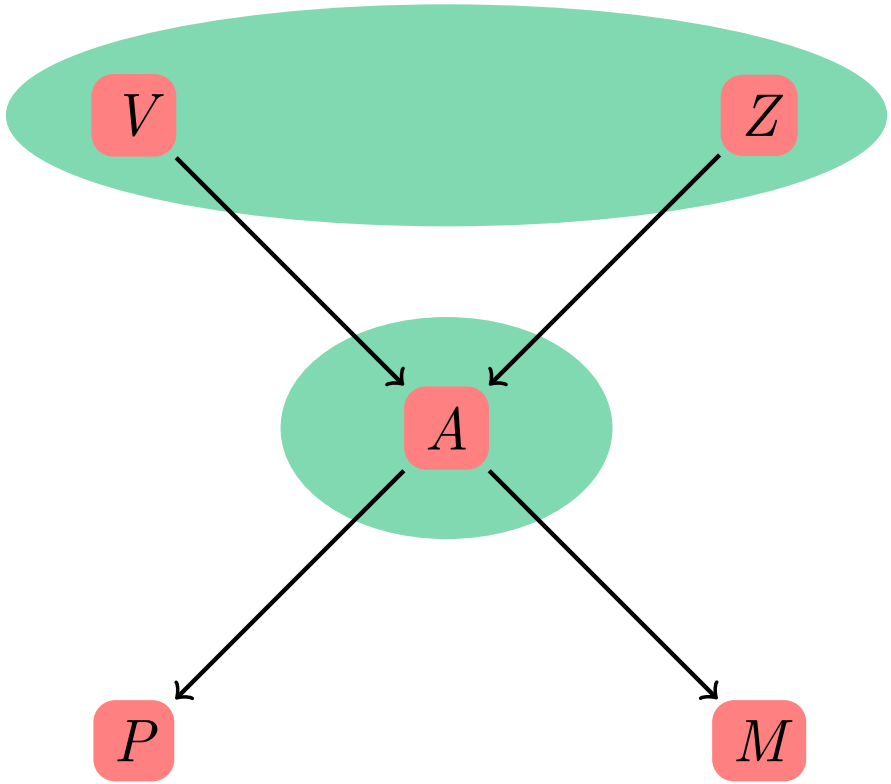
rodiče(X_i): v nezávislostním grafu, např. rodiče$(R) = \{K, V\}$

Příklad s binárními příznaky

V: vloupání do domu Z: zemětřesení A: ozval se alarm
P: volá soused Pepa M: volá sousedka Marie

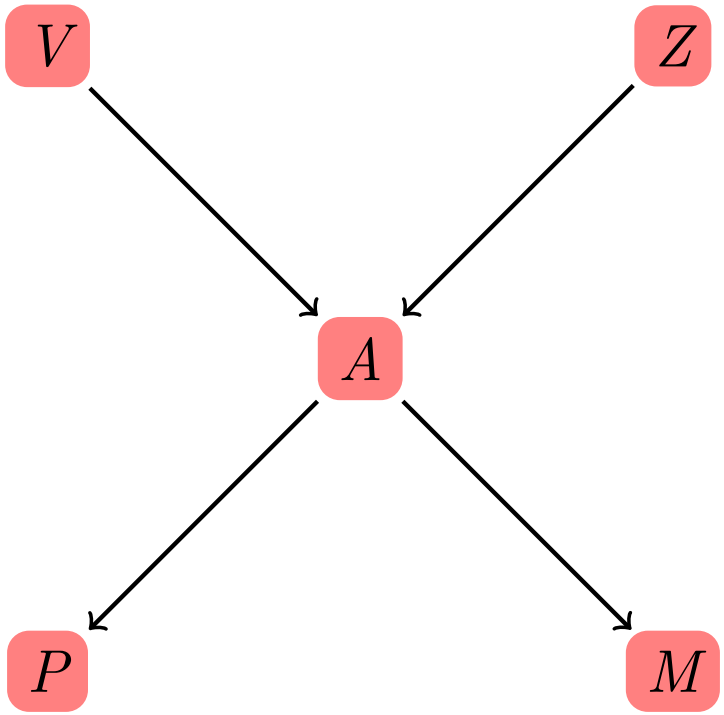


Podmíněné nezávislosti



V nezávisí na Z. Z nezávisí na V.

Výpočet sdruženého rozdělení



$$P_X = \Pi_{i=1}^{n} P_{X_i|\text{rodiče}(X_i)}$$

$$P_{V,Z,A,P,M} = P_{P|A} \cdot P_{M|A} \cdot P_{A|V,Z} \cdot P_V \cdot P_Z$$

Tabulky podmíněných pravděpodobností

$$P_{V,Z,A,P,M} = P_{P|A} \cdot P_{M|A} \cdot P_{A|V,Z} \cdot P_V \cdot P_Z$$

$P_{V,Z,A,P,M}$ jsme dekomponovali na rozdělení $P_{P|A}$, $P_{M|A}$, $P_{A|V,Z}$, P_V, P_Z. Každé z nich popíšeme *tabulkou podmíněných pravděpodobností* (TPP).

$P_{P\|A}(+\|a)$	a
0.90	$+$
0.05	$-$

$P_{M\|A}(+\|a)$	a
0.70	$+$
0.01	$-$

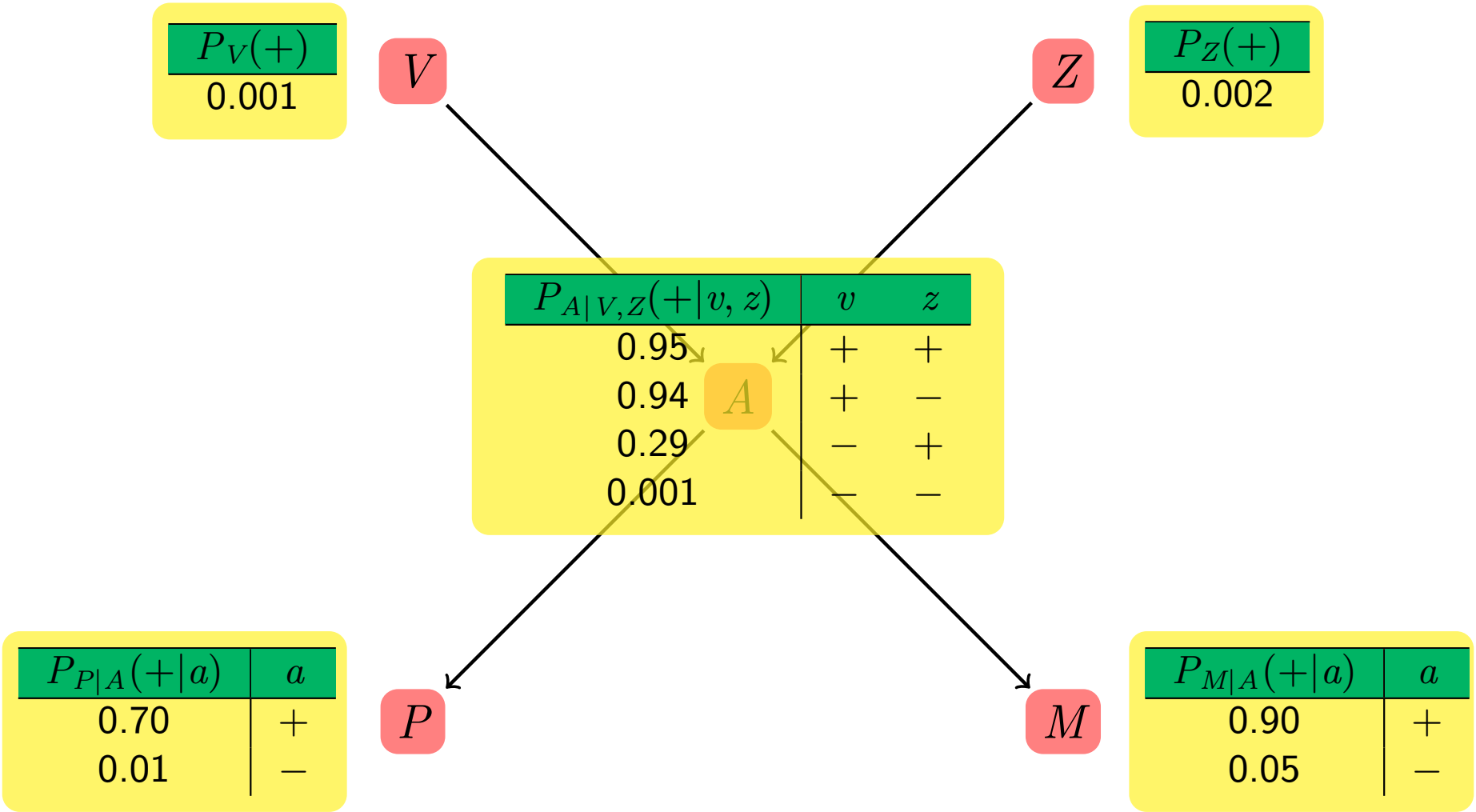
$P_{A\|V,Z}(+\|v,z)$	v	z
0.95	$+$	$+$
0.94	$+$	$-$
0.29	$-$	$+$
0.001	$-$	$-$

$P_V(+)$
0.001

$P_Z(+)$
0.002

10 čísel (místo $2^5 = 32$)

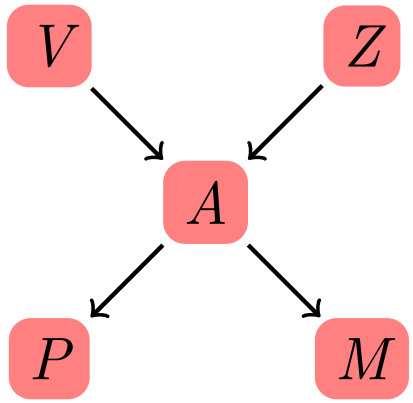
Bayesovská síť



Graf + TPP = Bayesovská síť

Příčinné vztahy v BS

- Hrany v tomto grafu odpovídají příčinným (kauzálním) vztahům mezi uzly.
- Příčinnost = vodítko pro návrh grafu BS
- Graf BS ale obecně *nemusí* odpovídat příčinnosti!



Sestavení grafu BS

Algoritmus pro sestavení grafu BS bez znalosti příčinných vztahů

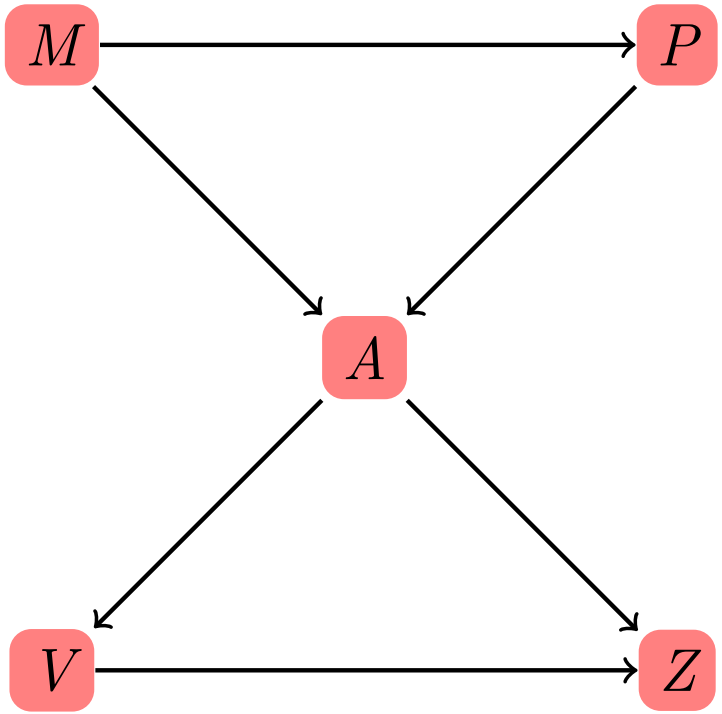
1. Zvol pořadí příznaků $X_1, X_2, \ldots X_n$
   /* šťastná volba $\Rightarrow$ kompaktní síť */
2. Pro $i = 1$ až n:
   přidej X_i jako uzel do grafu
   vyber co nejmenší množinu rodičů z $X_1, \ldots X_{i-1}$ tak, že

$$P_{X_i|\text{rodiče}(X_i)} = P_{X_i|X_1,\ldots X_{i-1}}$$

   vyveď hrany z rodičů do X_i

Příklad sestavení grafu BS

Bez znalosti příčinných vztahů volíme např. pořadí M, P, A, V, Z



žádné rodiče $P_{P|M} = P_P$?

Ne, M musí být rodičem. $P_{A|P,M} = P_{A|P}$ nebo $P_{A|P,M} = P_{A|M}$ nebo $P_{A|P,M} = P_A$?

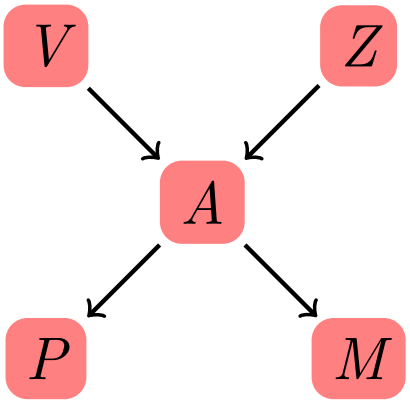
Ne, M i P musí být rodiči. $P_{V|A,P,M} = P_V$?

Ne. $P_{V|A,P,M} = P_{V|A}$? Ano. $P_{Z|V,A,P,M} = P_Z$?

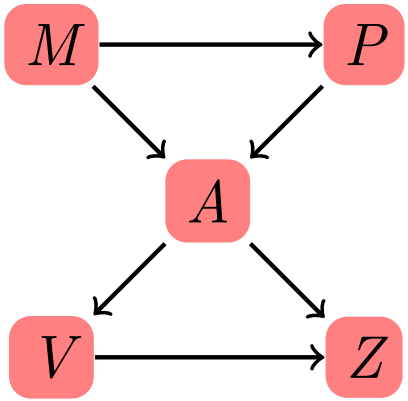
Ne. $P_{Z|V,A,P,M} = P_{Z|A}$?

Ekvivalentní BS

Dvě BS. Různé grafy, různé TPP. Reprezentují totéž sdružené rozdělení.



- Graf sestaven na základě příčinných vztahů.
- TPP vyžadují $1 + 1 + 4 + 2 + 2 = 10$ čísel.



- Graf sestaven obecným algoritmem.
- TPP vyžadují $1 + 2 + 4 + 2 + 4 = 13$ čísel.

Odvozování z BS

Příklad: Volá Pepa, Marie nevolá, nevíme, zda zvonil alarm, zemětřesení není. Jaká je pravděpodobnost vloupání?

$$P_{V|Z,P,M}(+|-,+,-) = \frac{P_{V,Z,P,M}(+,-,+,-)}{P_{Z,P,M}(-,+,-)}$$

$$= \alpha P_{V,Z,P,M}(+,-,+,-) = \alpha \sum_{a\in\{+,-\}} P_{V,Z,A,P,M}(+,-,a,+,-)$$

dosadíme dle $P_{V,Z,A,P,M} = P_{P|A}P_{M|A}P_{A|V,Z}P_V P_Z$

$$= \alpha P_V(+)P_Z(-) \sum_{a\in\{+,-\}} P_{P|A}(+|a)P_{M|A}(-|a)P_{A|V,Z}(a|+,-)$$

$$= \alpha \cdot 0.001 \cdot 0.998 \cdot (0.90 \cdot 0.30 \cdot 0.94 + 0.05 \cdot 0.99 \cdot 0.06)$$

$$\approx \alpha \cdot 0.00025$$

Odvozování z BS (pokr.)

Analogicky

$$
\begin{aligned}
&P_{V|Z,P,M}(-|-,+,-) \\
&= \alpha P_V(-) P_Z(-) \sum_{a \in \{+,-\}} P_{P|A}(+|a) P_{M|A}(-|a) P_{A|V,Z}(a|-,-) \\
&= \alpha \cdot 0.999 \cdot 0.998 \cdot (0.90 \cdot 0.30 \cdot 0.001 + 0.05 \cdot 0.99 \cdot 0.999) \\
&\approx \alpha \cdot 0.00520
\end{aligned}
$$

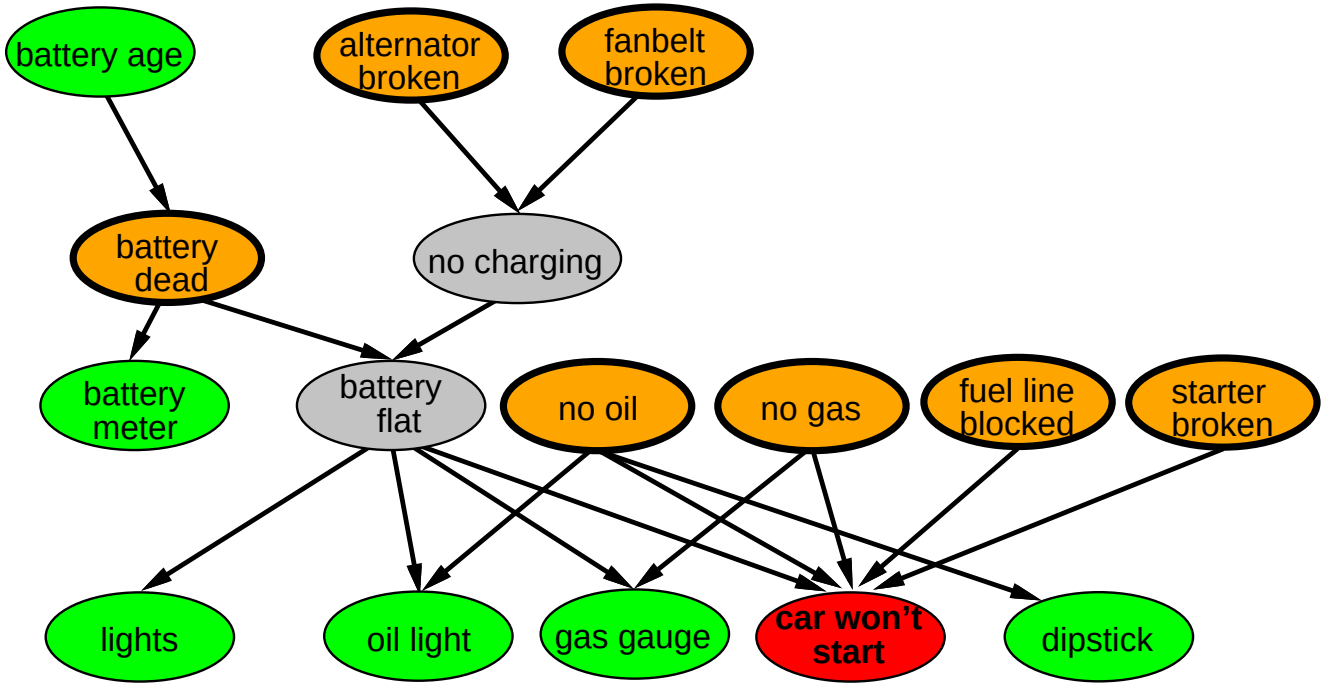
Protože $P_{V|Z,P,M}(+|-,+,-) + P_{V|Z,P,M}(-|-,+,-) = 1$, máme:

$$
P_{V|Z,P,M}(+|-,+,-) = \frac{\cancel{\alpha} \cdot 0.00025}{\cancel{\alpha} \cdot 0.00025 + \cancel{\alpha} \cdot 0.00520} \approx 0.046
$$

Apriorní pravděpodobnost vloupání je 0.001, ale volá-li soused Pepa a není zemětřesení, vzroste na 0.046.

Příklad využití BS: Diagnóza poruchy auta [Russel, Norvig]

červený uzel: počáteční příznak, zelené: testovatelné příznaky, oranžové: 'opravitelné' příznaky, šedivé: skryté příznaky - zjednodušují graf, snižují potřebný počet parametrů.



Příklad využití BS: Pojištění auta [Russel, Norvig]