UNIVERZITA PALACKÉHO V OLOMOUCI
PŘÍRODOVĚDECKÁ FAKULTA

BAKALÁŘSKÁ PRÁCE

Grafické modely s R

Katedra matematické analýzy a aplikací matematiky
Vedoucí práce: **prof. RNDr. Karel Hron, Ph.D.**
Vypracoval: **Ondřej Uličný**
Studijní program: B0541A170017 Aplikovaná matematika
Studijní obor: Aplikovaná matematika - specializace Matematika v ekonomické praxi
Forma studia: prezenční
Rok odevzdání: 2024

BIBLIOGRAFICKÁ IDENTIFIKACE

Autor: Ondřej Uličný

Název práce: Grafické modely s R

Typ práce: Bakalářská práce

Pracoviště: Katedra matematické analýzy a aplikací matematiky

Vedoucí práce: prof. RNDr. Karel Hron, Ph.D.

Rok obhajoby práce: 2024

Abstrakt: Tato bakalářská práce se zabývá studiem bayesovských sítí, což jsou statistické modely pro reprezentaci a analýzu nejistoty pomocí grafů. První část práce se zaměřuje na teoretický přehled bayesovských sítí, včetně pravděpodobnostní teorie, struktury sítí a souvislosti mezi nimi. Druhá část práce představuje implementaci a aplikaci bayesovských sítí v programovacím jazyce R. Konkrétně jsou ukázány postupy pro vytváření, analýzu a vizualizaci bayesovských sítí pomocí dostupných knihoven a nástrojů v R. Praktické příklady jsou použity k ilustraci efektivity a flexibilitě těchto metod při modelování a řešení reálných problémů. Výsledkem práce je porozumění bayesovským sítím a schopnost praktického využití těchto metod pomocí softwaru R.

Klíčová slova: bayesovská síť, podmíněná nezávislost, orientovaný acyklický graf

Počet stran: 44

Počet příloh: 1

Jazyk: český

BIBLIOGRAPHICAL IDENTIFICATION

Author: Ondřej Uličný

Title: Graphical models with R

Type of thesis: Bachelor's

Department: Department of Mathematical Analysis and Applications of Mathematics

Supervisor: prof. RNDr. Karel Hron, Ph.D.

The year of presentation: 2024

Abstract: This bachelor thesis deals with the study of Bayesian networks, which are statistical models for representing and analyzing uncertainty using graphs. The first part of the thesis focuses on a theoretical overview of Bayesian networks, including probability theory, the structure of networks and the connections between them. The second part of the thesis presents the implementation and application of Bayesian networks in the programming language R. Specifically, procedures for creating, analyzing, and visualizing Bayesian networks are shown using the available packages and tools in R. Practical examples are used to illustrate the effectiveness and flexibility of these methods in modeling and solving real-world problems. The result of this work is an understanding of Bayesian networks and the ability to use these methods in practice using the R software.

Key words: Bayesian network, conditional independence, directed acyclic graph

Number of pages: 44

Number of appendices: 1

Language: Czech

Prohlášení

Prohlašuji, že jsem bakalářskou práci vypracoval sám pod vedením pana prof. RNDr. Karla Hrona, Ph.D a všechny použité zdroje jsem uvedl v seznamu literatury.

V Olomouci 19. 4. 2024 .....

podpis

Obsah

Poděkování

Děkuji prof. RNDr. Karlu Hronovi, Ph.D. za cenné rady, připomínky a trpělivost při tvorbě této práce, a své rodině a přátelům za podporu a ochotu pomoct.

Úvod

Grafické modely představují efektivní nástroj pro modelování a analýzu nejistoty v různých oblastech, včetně umělé inteligence, strojového učení, medicíny, ekonomie a dalších. Mezi nejpopulárnější typy grafických modelů patří bayesovské sítě, které můžeme dále dělit například na diskrétní či gaussovské. Jejich popularita a využití stále roste díky schopnosti efektivně modelovat komplexní systémy a provádět inferenci na základě dostupných dat.

Tato bakalářská práce se zaměřuje na poskytnutí základního přehledu diskrétních bayesovských sítí, jak teoreticky, tak prakticky s využitím programovacího jazyka R. Úvodní část práce uvádí čtenáře do problematiky bayesovských sítí, kdy jsou na vlastních praktických příkladech vysvětleny jejich principy a vlastnosti.

V druhé části práce se přesouváme k praktickému využití bayesovských sítí pomocí softwaru R. Jsou zde představeny konkrétní postupy pro vytváření, analýzu a vizualizaci bayesovských sítí.

Cílem této práce je názorně představit teorii bayesovských sítí na konkrétních příkladech a jak s těmito sítěmi pracovat v softwaru R.

Kapitola 1

Základní pojmy

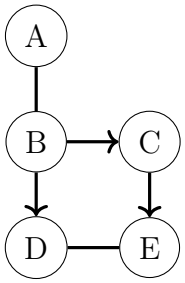
1.1 Teorie grafů

Abychom se mohli začít zabývat problematikou bayesovských sítí, je třeba si zavést několik základních pojmů z teorie grafů a uvést některé vlastnosti grafů, které budeme využívat v následujících částech. Čerpáno ze zdroje ([1], kapitola 2.2).

Graf G je datová struktura určená množinou vrcholů (uzlů) $\mathbf{X}$ a množinou hran $\mathbf{E}$. Dvojice vrcholů $X_i, X_j \in \mathbf{X}$ může být spojena *orientovanou* $X_i \to X_j$ nebo *neorientovanou hranou* X_i — X_j. Grafy dělíme podle typu hran na *orientované*, *neorientované* a *smíšené*.

Pokud v grafu G existuje orientovaná hrana $X_i \to X_j \in \mathbf{E}$, pak X_i nazýváme *rodičem* X_j a X_j *přímým potomkem* X_i. Množinu všech rodičů, resp. přímých potomků, vrcholu X_i značíme Pa_{X_i} (z angl. parents), resp. Ch_{X_i} (z angl. children). Jestliže v G je neorientovaná hrana X_i — $X_j \in \mathbf{E}$, pak X_i nazýváme *sousedem* X_j a opačně.

Cesta v grafu G je posloupnost $X_1, \ldots, X_k$, taková, že $\forall i = 1, \ldots, k-1$; existuje buď $X_i \to X_{i+1}$ nebo X_i — X_{i+1}. Cesta se nazývá orientovaná,



Obrázek 1.1: Příklad smíšeného grafu

jestliže alespoň jedna její hrana je orientovaná.

Trasou v grafu G rozumíme posloupnost $X_1, \ldots, X_k$, takovou, že $\forall i = 1, \ldots, k-1$, je mezi uzly X_i a X_{i+1} orientovaná hrana vycházející buď z uzlu X_i nebo X_{i+1}. Značíme $X_i \rightleftharpoons X_{i+1}$. Na obrázku 1.1 je posloupnost uzlů A, B, C, E, D cestou, tudíž i trasou, zatímco posloupnost A, B, D, E, C je pouze trasou.

Řekneme, že vrchol X je *předchůdcem* vrcholu Y a Y je *potomkem* X, jestliže existuje orientovaná cesta $X_1 \ldots X_k$, kde $X_1 = X$ a $Y = X_k$.

Kružnice či *cyklus* je orientovaná cesta, pro kterou platí, že počáteční vrchol je zároveň i vrcholem posledním ($X_1 = X_k$). Graf nazýváme *acyklický*, pokud neobsahuje žádné kružnice.

Právě acyklické, konkrétně *orientované acyklické grafy* (zkráceně DAG - directed acyclic graph), nám budou reprezentovat strukturu bayesovských sítí a grafy neorientované potom vedou k sítím markovským, kterými se však v této práci nezabýváme.

1.2 Podmíněná nezávislost

V základním kurzu pravděpodobnosti jsme se již seznámili s podmíněnou pravděpodobností a nezávislostí náhodných veličin. Nás nyní bude zajímat koncept podmíněné nezávislosti, který je mnohem častěji se vyskytující než

samotná nezávislost. Na následujícím příkladu si ukážeme rozdíl mezi oběma druhy.

Student střední školy čeká na výsledky přijímacího řízení na stejný obor ze dvou vysokých škol, České vysoké učení technické v Praze ($CVUT$ - binární náhodná veličina, přijetí na ČVUT) a Univerzitu Palackého v Olomouci (UP - přijetí na UP). Předpokládejme, že obtížnost přijímaček je na obě školy stejná. Pokud tedy student obdrží potvrzení o přijetí na UP, zvýší se tím pravděpodobnost přijetí na ČVUT. To znamená, že jevy $CVUT$ a UP **nejsou** nezávislé.

Nyní uvažujme situaci, že se školy budou rozhodovat o přijetí studentů na základě jejich studijních průměrů ze střední školy (AVG - z angl. average). V tomto případě zpráva o výsledku řízení z UP studentovi nijak nepomůže k lepšímu úsudku ohledně výsledku z druhé školy. Pravděpodobnost přijetí na ČVUT je stejná nehledě na znalost výsledku z UP, $\mathsf{P}(CVUT \mid UP, AVG) = \mathsf{P}(CVUT \mid AVG)$.

Řekneme, že náhodné veličiny $CVUT$ a UP jsou *podmíněně nezávislé vzhledem* k AVG, značíme $(CVUT \perp UP \mid AVG)$.

Kapitola 2

Bayesovské sítě

Bayesovské sítě jsou jedním z typu grafických modelů, statistických modelů, které efektivně reprezentují pravděpodobnostní vztahy mezi sledovanými proměnnými. Dalším typem grafických modelů jsou například markovské sítě. V této práci se však budeme zabývat pouze sítěmi bayesovskými, konkrétně diskrétními (obsahují pouze diskrétní proměnné). Diskrétní bayesovské sítě totiž slouží jako pevný základ pro pokročilejší studium dalších typů grafických modelů a mají široké spektrum praktických aplikací v oblastech jako je například strojové učení či diagnostika. Teorii k této kapitole je čerpána zejména ze zdroje [1], kapitoly 3.1 a 3.2.

Bayesovská síť je charakterizována rozdělením pravděpodobnosti a strukturou sítě.

Strukturou bayesovské sítě G je orientovaný acyklický graf (DAG), jehož uzly představují náhodné veličiny $X_1, \ldots, X_n$. Nechť Pa_{X_i} označuje rodiče uzlu X_i v G a $NonDescendants_{X_i}$ označuje uzly v grafu, které nejsou potomky X_i. Pak G reprezentuje množinu předpokladů podmíněné nezávislosti, nazývaných lokální nezávislosti, označovaných $I(G)$:

$$\forall X_i : (X_i \perp NonDescendants_{X_i} \mid Pa_{X_i}). \tag{2.1}$$

Jinak řečeno, každý uzel/proměnná X_i je podmíněně nezávislý se svými *nepotomky* (volný překlad *NonDescendants*) vzhledem ke svým rodičům.

Na následujícím ilustrativním příkladu si objasníme, proč tomu tak je. Jeho zadání je inspirováno mojí seminární prací ze střední školy. Nejednalo se ovšem o práci matematickou, ale chemickou.

Předpokládejme, že jsme provedli dotazníkové šetření v několika ordinacích praktických lékařů. Lékař někdy není schopen z příznaků pacienta jasně určit, zda dotyčný trpí virovým či bakteriálním onemocněním. Ke správné diagnóze pomáhá test, jehož výstupem je hladina C-reaktivního proteinu (CRP) v krvi pacienta. Pokud je hladina nízká, jedná se převážně o virové onemocnění, naopak vysoká hladina indikuje onemocnění bakteriální. Hraniční hodnoty, které oddělují napadení virem od bakterie jsou cca 30-40 mg/l. Pokud se hladina CRP nachází v této šedé zóně 30-40 mg/l, lékař si není stoprocentně jistý příčinou potíží. Nasadí tak dle svého úsudku nejvhodnější lék a sleduje, jestli zabere nebo ne. Cílem výzkumu je vytvořit model, podle kterého by se lékař mohl rozhodovat s větší jistotou v případech, kdy se hladina CRP pohybuje v šedé zóně. V dotazníku jsme se ptali na několik níže vypsaných příznaků, proměnných (v závorkách jsou uvedeny číselné hodnoty pro kategorie jednotlivých proměnných):

- K - typ kašle (0 - žádný, 1 - vlhký, 2 - suchý)
- R - rýma (0 - ne, 1 - ano)
- S - sípání (0 - ne, 1 - ano)
- CRP - hodnota CRP v šedé zóně (0 - ne, 1 - ano)

- L - nasazení správných léků (0 - ne, 1 - ano)

Poznámka: Pro zjednodušení zápisu budeme pro pravděpodobnosti realizací namísto $p(K = 0) = 0.3$ používat značení $p(k^0) = 0.3$.

2.1 Pravděpodobnostní část

Pokud bychom chtěli nyní vytvořit sdruženou pravděpodobnostní funkci tohoto modelu, je zapotřebí $(3 \cdot 2 \cdot 2 \cdot 2 \cdot 2) - 1 = 48 - 1 = 47$ parametrů, respektive pravděpodobností (pro každou možnou kombinaci hodnot proměnných bez jedné, která bude dopočet do 1). Když si uvědomíme, že se jedná o velmi jednoduchý model s 5 náhodnými veličinami, z toho 4 binární a 1 terciární, a potřebujeme znát 47 různých hodnot, je jasné, že tento přístup není optimální.

Abychom mohli pracovat s menším počtem parametrů, je nutné znát, jak na sobě závisí či nezávisí jednotlivé proměnné. Tuto informaci získáme například od experta v daném odvětví. V našem případě od lékaře, který nám sdělil následující:

- typ kašle a rýma na ničem nezávisí - označme schematicky $(K \perp R)$
- hodnota CRP závisí na kašli a rýmě - $(CRP \mid K, R)$
- sípání závisí na kašli - $(S \mid K)$
- nasazení správných léků závisí na hodnotě CRP - $(L \mid CRP)$

Takto tedy vypadá závislostní struktura našeho modelu.

2.1.1 Faktorizace pravděpodobnostní funkce

Pojďme se nyní zaměřit na část našeho modelu, konkrétně na proměnné kašel a sípání. Hodnoty jejich sdružené pravděpodobnostní funkci jsou uvedeny v tabulce 2.1.

K	S	$\mathsf{P}(K, S)$
k^0	s^0	0.12
k^0	s^1	0.08
k^1	s^0	0.1
k^1	s^1	0.2
k^2	s^0	0.15
k^2	s^1	0.35

Tabulka 2.1: Sdružená distribuce pro proměnné K, S

My však máme k dispozici informaci o podmíněnosti sípaní na kašli a můžeme tedy využít definice podmíněné pravděpodobnosti a sdruženou pravděpodobnostní funkci zapsat jako

$$\mathsf{P}(K, S) = \mathsf{P}(K)\,\mathsf{P}(S \mid K). \tag{2.2}$$

Můžeme ji tak reprezentovat pomocí tabulek 2.2 a 2.3. V první je rozdělení pravděpodobnosti veličiny K, což můžeme v tomto kontextu vnímat jako *apriorní rozdělení*. V druhé tabulce je *podmíněné rozdělení pravděpodobnosti* (označované jako CPD, z angl. *conditional probability distribution*) veličiny S vzhledem ke K.

k^0	k^1	k^2
0.2	0.3	0.5

Tabulka 2.2: $\mathsf{P}(K)$

	s^0	s^1
k^0	0.6	0.4
k^1	$\frac{1}{3}$	$\frac{2}{3}$
k^2	0.3	0.7

Tabulka 2.3: CPD pro $(S \mid K)$

Díky této faktorizaci se můžeme jednoduše tázat na pravděpodobnosti různých kombinací našich proměnných. Například by nás mohlo zajímat, jaká je pravděpodobnost, že pacient bude být suchý kašel, ale nebude sípat. Dosadíme příslušné hodnoty z tabulek do vzorce (2.2) a spočítáme,
$p(k^2, s^0) = p(k^2)\, p(s^0 \mid k^2) = 0.5 \cdot 0.3 = 0.15.$

Nyní přidáme do modelu proměnnou CRP, u níž víme, že nabývá dvou hodnot (crp^0, crp^1). V tomto případě je sdružená pravděpodobnostní funkce tvořena třemi náhodnými veličinami a je charakterizována 11 neznámými parametry. Pojďme se pokusit o její faktorizaci jako v případě kašle a sípání. Závislost S a K zůstává stejná. O proměnné CRP máme informaci, že také závisí na K. Je tedy jasné, že proměnná K podmiňuje obě proměnné. Rozeberme si jednotlivé vztahy mezi veličinami:

- Kašel je úzce spjat jak se sípáním, tak s hladinou CRP.
- Sípání a CRP nejsou nezávislé, a to z jednoduchého důvodu. Předpokládejme, že pacient sípe. Tím se zvyšuje i pravděpodobnost na přítomnost kašle, jenž ovlivňuje koncentraci CRP v krvi. My však předpokládáme, že v tomto modelu platí podmíněná nezávislost mezi sípáním a hladinou CRP. Jestliže nám je známa informace o kašli pacienta (např. žádný nemá), tak skutečnost, že sípe, už nijak neovlivní pravděpodobnost, v jaké zóně se nachází hodnota CRP,

  $$p(crp \mid k^0, s^1) = p(crp \mid k^0).$$

  Obecně zapsáno

  $$\mathsf{P} \models (CRP \perp S \mid K). \tag{2.3}$$

Toto značení znamená, že v rámci rozdělení pravděpodobnosti P platí podmíněná nezávislost mezi proměnnými CRP a S vzhledem k proměnné K.

Nutno podotknout, že toto tvrzení o podmíněné nezávislosti platí pouze a jen tehdy, pokud kašel je jediným zprostředkovatelem, který zapříčiňuje jakýkoli vztah mezi sípáním a hodnotou CRP.

Nyní upravíme sdruženou pravděpodobnostní funkci za pomoci definice podmíněné pravděpodobnosti stejně jako v případě kašle a sípání:

$\mathsf{P}(K, S, CRP) = \mathsf{P}(K)\,\mathsf{P}(S, CRP \mid K)$.

Zde využijeme předpoklad podmíněné nezávislosti ze vztahu (2.3), který implikuje následující:

$\mathsf{P}(S, CRP \mid K) = \mathsf{P}(S \mid K)\,\mathsf{P}(CRP \mid K)$.

Dosadíme do předchozího vztahu a dostaneme:

$$\mathsf{P}(K, S, CRP) = \mathsf{P}(K)\,\mathsf{P}(S \mid K)\,\mathsf{P}(CRP \mid K). \tag{2.4}$$

Dokázali jsme tedy napsat sdruženou pravděpodobnostní funkci jako součin tří rozdělení pravděpodobnosti. Tato faktorizace zajistí snížení počtu parametrů potřebných pro charakterizaci pravděpodobnostní funkce. Konkrétně z původních $(2 \cdot 2 \cdot 3) - 1 = 11$, na 8 parametrů. Proč na 8? Jelikož kašel je terciární proměnná, potřebujeme pro $\mathsf{P}(K)$ znát 2 parametry. V případě $\mathsf{P}(S \mid K)$ je třeba znát 3 parametry. Rozdělíme si tuto situaci na tři části - $\mathsf{P}(S \mid k^0)$, $\mathsf{P}(S \mid k^1)$ a $\mathsf{P}(S \mid k^2)$. Jelikož sípání je binární proměnná, tak k parametrizaci každé z částí je třeba 1 parametr, dohromady tedy 3. Situace ohledně $\mathsf{P}(CRP \mid K)$ je identická, jelikož CRP je taktéž binární proměnná. Dohromady je tedy k charakterizaci faktorizované pravděpodobnostní funkce potřeba $2 + 3 + 3 = 8$ parametrů.

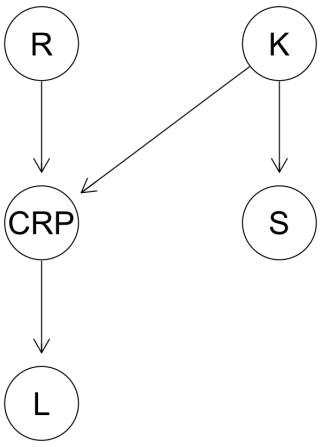
Další významnou výhodou rozkladu na součin je *modularita* (obdobně by se dala tato vlastnost popsat jako *skladebnost*). Po přidání proměnné CRP by v případě sdružené pravděpodobnostní funkce musely být její hodnoty znovu specifikovány. Kdežto u rozkladu pomocí součinu můžeme využít znalosti rozdělení $\mathsf{P}(K)$ a $\mathsf{P}(S \mid K)$ a pouze připojit další *modul* $\mathsf{P}(CRP \mid K)$.

Ukažme si jednoduchý příklad výpočtu v tomto modelu. Předpokládejme, že rozdělení pravděpodobností $\mathsf{P}(K)$ a $\mathsf{P}(S \mid K)$ zůstávají stejná (viz tabulky 2.2 a 2.3). Hodnoty pravděpodobností pro $\mathsf{P}(CRP \mid K)$ jsou uvedeny v následující tabulce 2.4.

	crp^0	crp^1
k^0	0.97	0.03
k^1	0.24	0.76
k^2	0.45	0.55

Tabulka 2.4: CPD pro $(CRP \mid K)$

Ptáme se na pravděpodobnost, že pacient má vlhký kašel, nesípe a jeho hladina CRP se nenachází v šedé zóně:
$p(k^1, s^0, crp^0) = p(k^1)\, p(s^0 \mid k^1)\, p(crp^0 \mid k^1) = 0.3 \cdot \frac{1}{3} \cdot 0.24 = 0.024.$



Obrázek 2.1: DAG pro příklad CRP

Na obrázku 2.1 vidíme DAG pro kompletní příklad, tedy i se zbývajícími proměnnými R a L. Graf je vytvořen jednoduchým způsobem, kdy jednotlivé uzly představují proměnné modelu a orientované hrany podmíněnost, kterou známe ze závislostní struktury.

Dosud jsme pracovali s proměnnými K, S a CRP a jejich sdruženou pravděpodobnostní funkci jsme rozdělili na součin dle vztahu (2.4). Pokud se zaměříme na obrázku 2.1 pouze na tyto 3 proměnné, vidíme, že hrany z uzlu K reprezentují podmíněná rozdělení $\mathsf{P}(S \mid K)$ a $\mathsf{P}(CRP \mid K)$.

Nyní bychom chtěli opět faktorizovat sdruženou pravděpodobnostní funkci, tentokrát ovšem i s proměnnými R a L. Dojdeme k tomu obdobně jako v předchozím, potřebujeme znát podmíněná rozdělení jednotlivých proměnných. Konkrétně $\mathsf{P}(K), \mathsf{P}(R), \mathsf{P}(S \mid K), \mathsf{P}(CRP \mid K, R)$ a $\mathsf{P}(L \mid CRP)$.

Obecně je každá proměnná v modelu spojena s CPD, kdy podmíněnost rozdělení určují všechny kombinace realizací rodičů dané proměnné. Pro uzel bez rodičů se CPD mění na marginální rozdělení, v našem příkladu se jedná o rozdělení $\mathsf{P}(K)$ a $\mathsf{P}(R)$.

Faktorizovaná sdružená pravděpodobnostní funkce celého modelu tedy vypadá následovně:

$$\mathsf{P}(K, R, S, CRP, L) = \mathsf{P}(K)\,\mathsf{P}(R)\,\mathsf{P}(S \mid K)\,\mathsf{P}(CRP \mid K, R)\,\mathsf{P}(L \mid CRP). \tag{2.5}$$

2.1.2 Podmíněné nezávislosti

V této části popíšeme a rozebereme jaké proměnné, a proč v našem příkladu splňují vztah podmíněné nezávislosti. Podmínky nám budou tvořit konkrétní realizace rodičů pozorovaného uzlu.

Proměnná L - nasazení správných léků - je podmíněná proměnnou CRP.

Tato informace je přeložena do grafu skutečností, že do uzlu L vede jediná hrana, a to z uzlu CRP. V řeči teorie pravděpodobnosti řekneme, že L je podmíněně nezávislá s proměnnými K, R a S vzhledem k CRP. Pokud tedy znám hodnotu, které nabyla proměnná CRP, žádná další informace o zbývajících třech proměnných již neovlivní pravděpodobnosti pro nasazení správného či nesprávného léku. Formálně zapsáno:

$$(L \perp K, R, S \mid CRP). \tag{2.6}$$

Obdobná situace je u proměnné S - sípání. Zde předpokládáme, že sípání je závislé pouze na kašli. Opět řekneme, že S je podmíněně nezávislá s proměnnými R, CRP a L vzhledem ke K:

$$(S \perp R, CRP, L \mid K). \tag{2.7}$$

Pokud bychom stejným způsobem pokračovali u proměnné CRP, tak bychom mohli tvrdit, že CRP je podmíněně nezávislá s S a L vzhledem ke svým rodičům, v tomto případě proměnným K a R. Tento úsudek však není pravdivý.

Předpokládejme, že pacient trpí vlhkým kašlem i rýmou, tedy podmiňujeme CRP realizacemi k^1 a r^1. Jsou v tomto případě CRP a L nezávislé? Nejsou, a to z následujícího důvodu. Pokud lékař nasadí správné léky (realizace l^1), potom by se pravděpodobnost, že hodnota CRP vyšla v šedé nerozhodné zóně, měla zmenšit: $p(crp^1 \mid k^1, r^1, l^1) < p(crp^1 \mid k^1, r^1)$.

Tudíž vidíme, že uzel/proměnná může i při daných realizacích svých rodičů záviset na svých potomcích, ale ne už na ostatních uzlech *nepotomcích*.

Pokud víme, že pacient má suchý kašel, skutečnost o stavu jeho sípání nijak neovlivní pravděpodobnosti pro proměnnou CRP. Čili platí:

$$(CRP \perp S \mid K, R). \tag{2.8}$$

Uzly K a R nemají rodiče. Již víme, že nejsou nezávislé se svými potomky, ale jsou nezávislé se svými *nepotomky*. Pro uzel K to je uzel R a pro uzel R jsou to uzly K a S. Dostáváme tedy vztah:

$$(K, S \perp R), \tag{2.9}$$

Z předešlých úvah lze odvodit, že jestliže známe realizace rodičů daného uzlu, žádná další informace spojená přímo či nepřímo s jeho rodiči nebo dalšími předchůdci neovlivní jeho hodnoty pravděpodobností. Ale můžou být ovlivněny realizacemi potomků. Vztahy (2.6) - (2.9) tvoří pro náš příklad onu množinu lokálních nezávislostí $I(G)$ z definice ze začátku kapitoly.

2.2 Graf a rozdělení pravděpodobnosti

Podstata propojení mezi závislostní strukturou proměnných či dat a jejich grafickou podobou tkví v již představené podmíněné nezávislosti a grafickou separací. V této části je čerpáno ze zdrojů ([1], kapitola 3.2.3) a [2].

2.2.1 I-zobrazení

Obdobně jako množinu lokálních nezávislostí $I(G)$, která je definována grafem G, zavedeme další množinu lokálních nezávislostí - $I(\mathsf{P})$, která však bude definována pomocí rozdělení pravděpodobnosti P. Tudíž tvrzení, že

v rámci rozdělení pravděpodobnosti P platí vztahy lokálních nezávislostí spojené s grafem G, lze přepsat jako $I(G) \subseteq I(\mathsf{P})$. V tomto případě hovoříme o grafu G jako o *I-zobrazení* (z angl. *I-map* - independency map = zobrazení nezávislosti) rozdělení P. Aby graf G mohl být *I-zobrazením*, musí každý vztah nezávislosti ním daný platit také v rozdělení P. Naopak, P může obsahovat vztahy, které v G reprezentovány nejsou. Na následujícím příkladu si představíme koncept I-zobrazení.

Mějme dány dvě binární proměnné X a Y. K těmto proměnným existují 3 DAGy, které můžeme sestrojit:

- $G_\emptyset$, což jsou nespojené uzly X a Y;
- $G_{X \to Y}$, který má hranu $X \to Y$;
- $G_{Y \to X}$, který obsahuje $Y \to X$.

Graf $G_\emptyset$ znázorňuje nezávislost mezi oběma proměnnými - $(X \perp Y)$. Zbývající grafy nereprezentují žádný vztah nezávislosti. Uvažujme následující dvě rozdělení pravděpodobnosti:

X	Y	$\mathsf{P}_1(X,Y)$
x^0	y^0	0.12
x^0	y^1	0.48
x^1	y^0	0.08
x^1	y^1	0.32

Tabulka 2.5: Rozdělení P_1

X	Y	$\mathsf{P}_2(X,Y)$
x^0	y^0	0.35
x^0	y^1	0.25
x^1	y^0	0.25
x^1	y^1	0.15

Tabulka 2.6: Rozdělení P_2

Nejprve zjistěme, zda platí nezávislost v rozdělení určené tabulkou 2.5:
$\mathsf{P}_1(x^1) = 0.08 + 0.32 = 0.4$,
$\mathsf{P}_1(y^1) = 0.48 + 0.32 = 0.8$,
$\mathsf{P}_1(x^1, y^1) = 0.32 = 0.4 \cdot 0.8 = \mathsf{P}_1(x^1) \cdot \mathsf{P}_1(y^1)$.

Obdobně bychom si počínali i u dalších kombinací realizací X a Y. Došli jsme tedy k závěru, že $(X \perp Y) \in I(\mathsf{P}_1)$, tudíž $G_\emptyset$ je I-zobrazení P_1. Ve skutečnosti jsou všechny 3 grafy I-zobrazení P_1, jelikož $I(G_{X\to Y})$ i $I(G_{Y\to X})$ neobsahují žádný vztah nezávislosti, čili splňují předpoklad z definice I-zobrazení: $I(G) \subseteq I(\mathsf{P})$.

V tabulce 2.6 bychom stejnou úvahou došli k tomu, že $(X \perp Y) \notin I(\mathsf{P}_2)$. Množina $I(\mathsf{P}_2)$ je tedy prázdná, takže $G_\emptyset$ není I-zobrazení rozdělení P_2. Jak už bylo řečeno, zbylé grafy $G_{X\to Y}$ a $G_{X\to Y}$ nereprezentují žádný vztah nezávislosti, takže jsou I-zobrazením rozdělení P_2.

Faktorizace pomocí I-zobrazení

Víme tedy, že struktura bayesovské sítě, DAG G, reprezentuje množinu vztahů podmíněné nezávislosti a v každém rozdělení pravděpodobnosti, pro které je G I-zobrazením, musí tyto vztahy platit.

Podle věty o násobení pravděpodobnosti (viz [6], str. 35, Věta 1.3) můžeme rozložit sdruženou pravděpodobnostní funkci z našeho příkladu:

$$\begin{aligned}\mathsf{P}(K,R,S,CRP,L) = \mathsf{P}(K)\,\mathsf{P}(R \mid K)\,\mathsf{P}(S \mid K,R)\,\mathsf{P}(CRP \mid K,R,S)\\ \mathsf{P}(L \mid K,R,S,CRP).\end{aligned} \tag{2.10}$$

Uvažujme libovolné rozdělení pravděpodobnosti P, pro které je DAG z obrázku 2.1 I-zobrazením. Nyní využijeme vztahů (2.6) - (2.9) a jednotlivá podmíněná rozdělení upravíme.

Ze vztahu (2.9) a z definice I-zobrazení víme, že $(K \perp R) \in I(\mathsf{P})$, tudíž můžeme druhý člen pravé strany (2.10) upravit na $\mathsf{P}(R \mid K) = \mathsf{P}(R)$. Ze vztahu (2.7) dostaneme, že $\mathsf{P}(S \perp R, CRP, L \mid K) \in I(\mathsf{P})$. Člen $\mathsf{P}(S \mid K,R)$ změníme na $\mathsf{P}(S \mid K)$. Člen $\mathsf{P}(CRP \mid K,R,S) = \mathsf{P}(CRP \mid K,R)$ dle vztahu (2.8) a $\mathsf{P}(L \mid K,R,S,CRP) = \mathsf{P}(L \mid CRP)$ ze vztahu (2.6). Těmito úpravami

jsme faktorizovali sdruženou pravděpodobnostní funkci do náledující podoby, která je identická se vztahem 2.5:

$$\mathsf{P}(K, R, S, CRP, L) = \mathsf{P}(K)\,\mathsf{P}(R)\,\mathsf{P}(S \mid K)\,\mathsf{P}(CRP \mid K, R)\,\mathsf{P}(L \mid CRP). \tag{2.11}$$

Každý z faktorů představuje podmíněnou pravděpodobnost pro danou proměnnou vzhledem ke svým rodičům. Tato faktorizace platí pro každé rozdělení pravděpodobnosti, pro které je daná struktura bayesovské sítě I-zobrazením.

Schopnost faktorizace sdružené pravděpodobnostní funkce nazýváme *markovská vlastnost*. Obecně ji lze zapsat jako:

$$\mathsf{P}(X_1, \ldots, X_n) = \prod_{i=1}^{n} \mathsf{P}(X_i \mid Pa_{X_i}).$$

2.2.2 d-separace

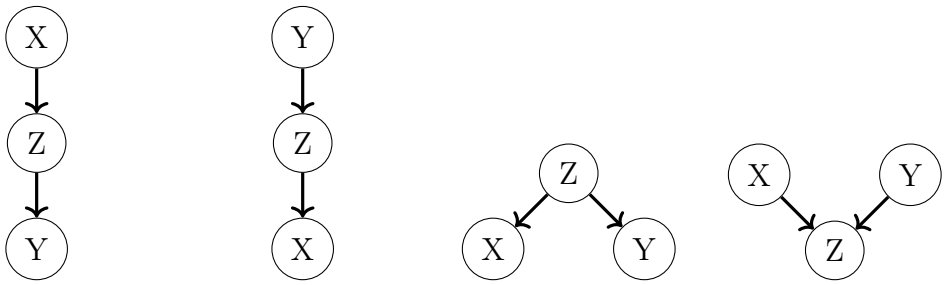
Náš cíl v této části je porozumět, kdy můžeme zaručit platnost podmíněné nezávislosti v rozdělení pravděpodobnosti P spojené s grafem G. K lepšímu pochopení nám může pomoct zvážit opačnou situaci. Zaměříme se tedy na analýzu situací, ve kterých X může ovlivnit Y vzhledem k Z. Čerpáno z ([1], kapitola 3.3.1), [4] a [5].

Přímé propojení

Začneme jednoduchým případem, kdy jsou X a Y přímo propojeny hranou, řekněme $X \longrightarrow Y$. Pro libovolnou síťovou strukturu G obsahující hranu $X \longrightarrow Y$ je možné zkonstruovat rozdělení pravděpodobnosti, kde jsou X a Y spolu vzájemně ve vztahu bez ohledu na to, zda pozorujeme realizace jiných proměnných v síti. Jinými slovy, pokud jsou X a Y přímo propojeny, vždy

se ovlivňují navzájem, bez ohledu na Z.

Nepřímé propojení



Obrázek 2.2: 4 možnosti nepřímého propojení uzlů X a Y přes uzel Z: (a) nepřímý kauzální efekt, (b) nepřímý důkazní efekt, (c) společná příčina, (d) společný efekt

Nyní uvažujme složitější případ, kdy X a Y nejsou přímo propojeny. Budeme uvažovat síť o 3 proměnných, X a Y nejsou přímo propojeny, ale existuje mezi nimi trasa přes Z. Na těchto jednoduchých příkladech si ukážeme koncept nepřímé interakce v bayesovských sítích.

Existují čtyři možnosti, jak mohou být X a Y propojeny přes Z (viz obrázek 2.2). První dva odpovídají kauzálním řetězcům, třetí společné příčině a čtvrtý společnému efektu.

Nepřímý kauzální efekt (obrázek 2.2a): Vraťme se k našemu příkladu CRP, kde máme kauzální cestu $K \longrightarrow CRP \longrightarrow L$. Začněme případem, kdy proměnná CRP není pozorována. Pokud pozorujeme, že pacient netrpí kašlem, jsme více nakloněni k tomu věřit, že hladina CRP nebude vysoká, tudíž ani v šedé zóně, a tedy i rozhodnutí lékaře ohledně nasazení léků bude pravděpodobněji správné. Jinými slovy, pokud pozorujeme konkrétní realizaci proměnné K, tak pravděpodobnosti realizací dalších proměnných v řetězci se mění.

Nyní předpokládejme, že Z je pozorováno. Jak jsme již viděli v části 2.1.2, pokud známe realizaci CRP, pak stav pacientova kašle již neovlivňuje nasazení léků. Tedy docházíme k závěru, že X nemůže ovlivnit Y, pokud je Z pozorováno.

Nepřímý důkazní efekt (obrázek 2.2b): Jak může X ovlivnit Y v tomto případě? Nyní si pomůžeme cestou $R \longrightarrow CRP \longrightarrow L$. Jestliže víme, že lékař nasadil špatné léky, pravděpodobnost, že se CRP nacházelo v šedé zóně stoupne, a tím i pravděpodobnost, že pacient měl rýmu. Pokud však je nám známa realizace CRP, fakt o nasazení léku už nijak neovlivní hodnoty pro proměnnou R.

Výsledek je tedy stejný, X může ovlivnit hodnoty pravděpodobnosti proměnné Y, pouze pokud Z není pozorováno. Závěr si můžeme potvrdit i skutečností, že ne/závislost je symetrická vlastnost - pokud neplatí $(X \perp Y \mid Z)$, tak neplatí ani $(Y \perp X \mid Z)$.

Společná příčina (obrázek 2.2c): Vraťme se do části 2.1.1, kde jsme rozebírali vztah mezi kašlem, sípáním a hladinou CRP. Tento příklad můžeme reprezentovat grafem 2.2c: uzel Z je kašel, uzel Y sípání a uzel X hladina CRP. Sípání a hladina CRP mají mezi sebou úzký vztah, jelikož informace o sípání ovlivní pravděpodobnosti jednotlivých realizací kašle, což následně má vliv na rozdělení pravděpodobnosti proměnné CRP. Jakmile ovšem pozorujeme danou realizaci kašle, tak informace o sípání již nijak neovlivní pravděpodobnosti CRP. Závěr je tedy stejný jako v předchozích dvou případech nepřímého propojení: proměnné X a Y se mohou ovlivnit pouze a jen tehdy, jestliže nepozorujeme konkrétní realizaci Z.

Společný efekt (obrázek 2.2d): Ve všech předchozích případech jsme došli k tomu, že X ovlivňuje Y skrz Z jen tehdy, pokud Z není pozorováno. Mohli bychom očekávat, že tento vzorec bude platit i v tomto posledním případě,

ale opak je pravdou.

Uvažujme tu část grafu 2.1, která odpovídá struktuře 2.2d. Pracujeme tedy s uzly R, K a jejich společným přímým potomkem CRP. Když nepozorujeme žádnou realizaci CRP, tak víme ze vztahu 2.9, že K a R jsou nezávislé. Pravděpodobnostní vliv tak nemůže *proudit* mezi proměnnými ve struktuře $X \longrightarrow Z \longleftarrow Y$, pokud Z není pozorováno.

Co se stane, když pozorujeme konkrétní realizaci Z? Mějme situaci, ve které uzel X bude představovat zemětřesení, uzel Y vloupání do domu a uzel Z spuštění alarmu. Jestliže nás uprostřed noci vzbudí alarm a zároveň cítíme otřesy, tak naše přesvědčení o tom, že za spuštěním alarmu stojí lupiči, výrazně klesne. X tedy ovlivňuje Y při daném pozorování Z.

Upravme příklad s alarmem přidáním hrany z uzlu Z do nového uzlu W - příjezd policie (spuštění alarmu podmiňuje příjezd policie). Představme si, že jsme se vzbudili díky příjezdu policejní hlídky. Tím se zvýší i pravděpodobnost, že se spustil alarm. Pokud navíc uslyšíme podezřelé zvuky z jiného pokoje v domě, tak pravděpodobnost, že došlo k zemětřesení, se sníží. Proměnné se tedy ovlivňují i při daném pozorování potomků Z.

Pokud se v daném grafu nachází struktura totožná s obrázkem 2.2d, nazýváme ji kvůli jejímu tvaru *v-struktura*. Uzel Z a jeho potomky můžeme považovat za jakési spínače, které umožňují či znemožňují ve v-strukturách tok pravděpodobnostního vlivu mezi proměnnými.

Jestliže pravděpodobnostní vliv může proudit z uzlu X do Y přes Z, řekneme, že trasa $X \rightleftharpoons Z \rightleftharpoons Y$ je *aktivní*. Shrňme závěry analýzy všech čtyř případů:

- **Kauzální trasa** $X \to Z \to Y$: aktivní právě tehdy, když uzel Z není pozorován.
- **Důkazní trasa** $X \leftarrow Z \leftarrow Y$: aktivní právě tehdy, když Z není pozo-

rován.

- **Společná příčina** $X \leftarrow Z \rightarrow Y$: aktivní právě tehdy, když Z není pozorován.

- **Společný efekt** $X \rightarrow Z \leftarrow Y$: aktivní právě tehdy, když je pozorován buď Z, nebo některý z jeho potomků.

Ač to může být pro člověka neintuitivní, tak vidíme, že v některých případech proudí pravděpodobností vliv proti směru hrany. Hrany v grafu tedy reprezentují pouze závislostní strukturu sítě, ovšem ne kompletní schéma pravděpodobnostního vlivu mezi proměnnými.

Pokud nastane situace, kdy máme více možných tras mezi dvěma uzly, tak vliv může proudit, jestliže je alespoň jedna z tras aktivní. Předchozí úvahy shrneme do nového pojmu - *d-separace*, který nese význam oddělení množiny uzlů jinou množinou uzlů v rámci orientovaných (angl. *directed*, proto d-separace) grafů.

Definice (d-separace) Nechť **X**, **Y**, **Z** jsou tři množiny uzlů v grafu G. Řekneme, že **X** a **Y** jsou *d-separovány* množinou uzlů **Z**, značíme

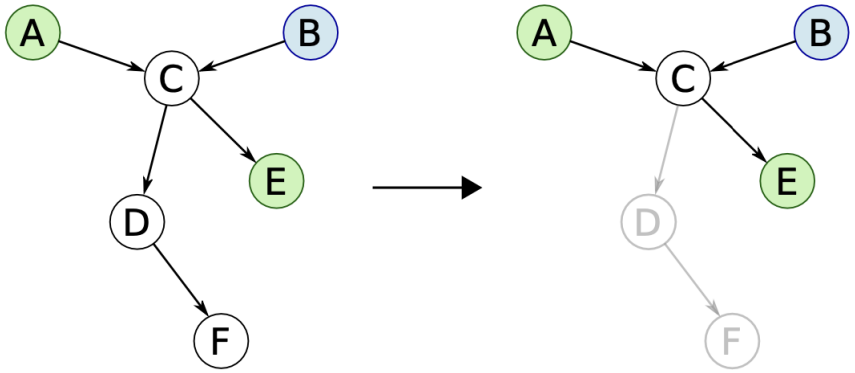
$$(\mathbf{X} \perp_G \mathbf{Y}) \mid \mathbf{Z},$$

pokud mezi libovolným uzlem $X \in \mathbf{X}$ a libovolným uzlem $Y \in \mathbf{Y}$ není žádná aktivní trasa, která vede přes **Z**.

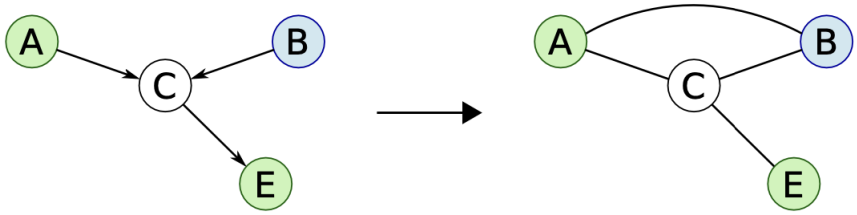
Algoritmus d-separace

Chceme zjistit, zda uzel A a E (zeleně na obrázcích 2.3 a 2.4 jsou d-separovány uzlem B (modře). Postupujeme následovně:

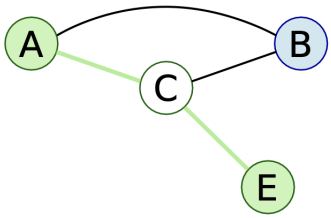
- Krok 1 - uzly, které nejsou předchůdci uzlů A, E i B, smažeme a vytvoříme tím podgraf (obrázek 2.3)



Obrázek 2.3: Algoritmus d-separace: krok 1, převzato z [4]



Obrázek 2.4: Algoritmus d-separace: kroky 2 a 3, převzato z [4]



Obrázek 2.5: Algoritmus d-separace: krok 4, převzato z [4]

- Krok 2 - spojíme všechny uzly, které mají společné dítě a není mezi nimi hrana, neorientovanou hranou (obrázek 2.4)
- Krok 3 - všechny původně orientované hrany změníme na neorientované (obrázek 2.4)
- Krok 4 - hledáme cestu mezi uzly A a E, která neprochází uzlem B. Pokud nalezneme alespoň jednu takovou cestu (A, C, E na obrázku 2.5), vlastnost d-seperace neplatí.

Kroky 2 a 3 nazýváme *moralizací* grafu. Název vznikl tak, že nově přidaná hrana spojující rodiče uzlu je brána jako symbol sňatku mezi rodiči.

Markovův obal

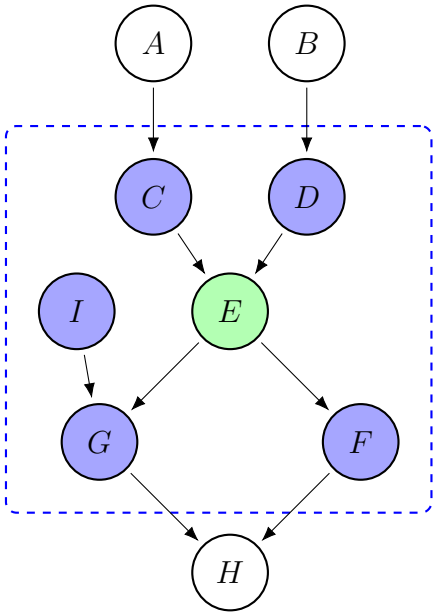
Množinu uzlů $\mathbf{Y}$, která kompletně d-separuje vybraný uzel X (v našem případě uzel E na obrázku 2.6), od ostatních uzlů, označme $\mathbf{Z}$, nazýváme *Markovův obal* (angl. *Markov blanket*) uzlu E. Markovův obal zahrnuje rodiče daného uzlu (uzly C, D na obrázku 2.6), jeho přímé potomky (uzly G a F) a také ostatní rodiče jeho potomků (uzel I).

2.2.3 Třídy ekvivalence

Z části 2.2.1 jsme již seznámeni s tím, že DAG jednoznačně určuje podobu faktorizace sdružené pravděpodobnostní funkce. Naopak to však nemusí být nutně pravda. Čerpáno ze zdrojů [2] a [4].

Mějme následující DAG reprezentující strukturu bayesovské sítě pro množinu proměnných $\mathbf{X} = \{A, B, C, D, E, F\}$:

Faktorizujeme sdruženou pravděpodobnostní funkci a pomocí definice podmíněné pravděpodobnosti a dalších úprav (např. přidání členu $\frac{\mathsf{P}(F)}{\mathsf{P}(F)}$) upravíme:

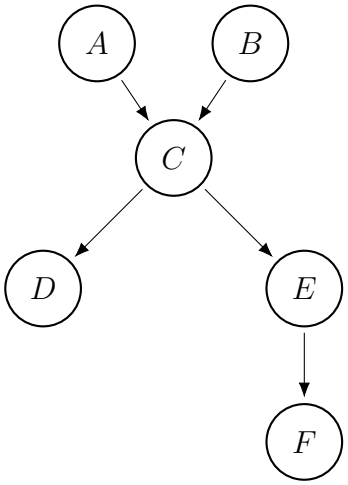


Obrázek 2.6: Markovův obal pro uzel E

$$\begin{aligned}
\mathsf{P}(\mathbf{X}) &= \mathsf{P}(A)\,\mathsf{P}(B)\,\mathsf{P}(C \mid A, B)\,\mathsf{P}(D \mid C)\,\mathsf{P}(E \mid C)\,\mathsf{P}(F \mid E) \\
&= \mathsf{P}(A)\,\mathsf{P}(B)\,\frac{\mathsf{P}(A,B,C)}{\mathsf{P}(A)\mathsf{P}(B)}\,\frac{\mathsf{P}(C,D)}{\mathsf{P}(C)}\,\frac{\mathsf{P}(C,E)}{\mathsf{P}(C)}\,\frac{\mathsf{P}(E,F)}{\mathsf{P}(E)} \\
&= \frac{\mathsf{P}(A,B,C)}{\mathsf{P}(B,C)}\,\frac{\mathsf{P}(B,C)}{\mathsf{P}(C)}\,\frac{\mathsf{P}(C,D)}{\mathsf{P}(C)}\,\frac{\mathsf{P}(C,E)}{\mathsf{P}(E)}\,\frac{\mathsf{P}(E,F)}{\mathsf{P}(F)}\,\mathsf{P}(F) \\
&= \mathsf{P}(F)\,\mathsf{P}(E \mid F)\,\mathsf{P}(E \mid C)\,\mathsf{P}(D \mid C)\,\mathsf{P}(B \mid C)\,\mathsf{P}(A \mid B, C).
\end{aligned} \tag{2.12}$$

Na obrázku 2.8 vidíme DAG reprezentující jinou, ovšem pravděpodobnostně ekvivalentní, pravděpodobnostní funkci.

Jelikož struktury typu $X \to Y \to Z$ a $X \leftarrow Z \to Y$ jsou pravděpodobnostně ekvivalentní (první, resp. poslední, řádek pravé strany rovnice (2.13)) můžeme obrátit směry jejich hran dle libosti, pokud tím nevytvoříme žádnou novou v-strukturu $X \to Y \leftarrow Z$. Tu totiž nikdy nevytvoříme pomocí úprav

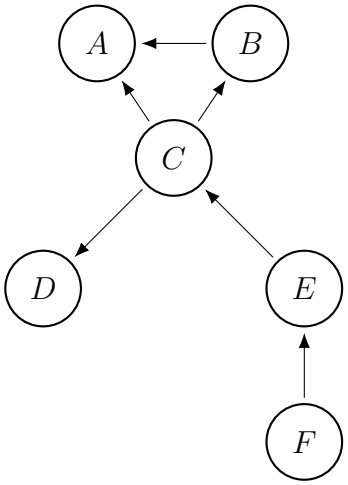


Obrázek 2.7: Původní DAG

použitých v rovnici (2.13), jestliže nepřidáme novou hranu. Tuto situaci vidíme jasně na obrázcích 2.7 a 2.8. Ve v-struktuře $A \to B \leftarrow C$ jsme obrátili obě hrany, ale aby mohla platit pravděpodobnostní ekvivalence, musela být do modelu přidána hrana $A \leftarrow B$.

$$\begin{aligned} \mathsf{P}(X,Y,Z) &= \mathsf{P}(X)\,\mathsf{P}(Y \mid X)\,\mathsf{P}(Z \mid Y) = \\ &= \mathsf{P}(X)\,\frac{\mathsf{P}(X,Y)}{\mathsf{P}(X)}\,\frac{\mathsf{P}(Y,Z)}{\mathsf{P}(Y)} = \\ &= \mathsf{P}(X \mid Y)\,\mathsf{P}(Y)\,\frac{1}{\mathsf{P}(Y)}\,\mathsf{P}(Z \mid Y)\,\mathsf{P}(Y) = \\ &= \mathsf{P}(Y)\,\mathsf{P}(X \mid Y)\,\mathsf{P}(Z \mid Y). \end{aligned} \tag{2.13}$$

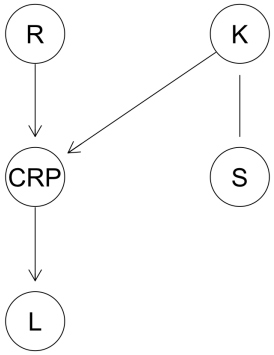
Třída ekvivalence sdružuje pravděpodobnostně ekvivalentní DAGy. Je reprezentována pomocí *úplného částečně orientovaného acyklického grafu* (ve zkratce CPDAG, z angl. *completed partially directed acyclic graph*), ve kterém jsou orientovány pouze hrany ve v-strukturách a ty, které by po změně orientace zavedly nové v-struktury nebo cykly. Takové hrany se nazývají *vynucené* (angl. *compelled*, protože jejich směr je pevně dán, i když nejsou



Obrázek 2.8: DAG pro upravenou pravděpodobnostní funkci

součástí žádné v-struktury. DAGy na obrázcích 2.7 a 2.8 nepatří do stejné třídy ekvivalence, jelikož upravený DAG neobsahuje původní v-strukturu. Změna směru jakékoli hrany, která není vynucená, vede k vytvoření jiného DAGu ve stejné třídě ekvivalence.

Jak je vidět z obrázku 2.9, jediná nevynucená hrana pro příklad CRP je mezi uzly K a S. Hrany $R \rightarrow CRP$ a $K \rightarrow CRP$ již jsou součástí v-struktury a hrana $CRP \rightarrow L$ by při změně orientace vytvořila dvě nové v-struktury $R \rightarrow CRP \leftarrow L$ a $K \rightarrow CRP \leftarrow L$.



Obrázek 2.9: CPDAG pro příklad CRP

Kapitola 3

Bayesovské sítě v R

V této kapitole si ukážeme základy práce s bayesovskými sítěmi v softwaru R. Budeme využívat balíčky *bnlearn*, *gRain*, *gRbase* a *Rgraphviz*. Celý kód je k dispozici jako příloha k této práci. Pro tuto kapitolu byly využity zdroje [3] a [4].

3.1 Vytvoření sítě

První způsob, jak zadefinovat graf pro bayesovskou síť, je pomocí množiny proměnných, respektive uzlů, a matice hran.

```
> nazvy.promennych = c('R','K','CRP','S','L')

# vytvoření grafu pouze s uzly
> crp1 = empty.graph(nodes = nazvy.promennych)

# definování hran pomocí jmen uzlů
> hrany = matrix(c('R','CRP',
                   'K', 'CRP',
                   'K','S',
                   'CRP','L'),
                    byrow = TRUE, ncol = 2,
```

```
                    dimnames = list(NULL, c("z", "do")))
> hrany
     z     do
[1,] "R"   "CRP"
[2,] "K"   "CRP"
[3,] "K"   "S"
[4,] "CRP" "L"
> arcs(crp1) = hrany    # vložení hran do grafu

# nebo pomocí matice sousednosti
amat(crp1) = matrix(c(0, 0, 1, 0, 0,
                      0, 0, 1, 1, 0,
                      0 ,0, 0, 0, 1,
                      0, 0, 0, 0, 0,
                      0, 0, 0, 0, 0),
         byrow = TRUE, nrow = 5, ncol = 5,
         dimnames = list(nodes(crp1), nodes(crp1)))
```

Nebo druhým způsobem pomocí funkce *model2network* a zadáním jednotlivých podmíněností mezi proměnnými.

```
> crp2 = model2network("[K][R][S|K][CRP|K:R][L|CRP]")
```

Výstup charakteristky modelu:

- schéma modelu (shodné s argumentem funkce *model2network*)
- počet uzlů v modelu
- počet hran celkem
- počty neorientovaných a orientovaných hran
- průměrný počet uzlů v Markovově obalu
- průměrný počet sousedů (rodiče + přímí potomci)
- průměrný počet hran vycházející z uzlu

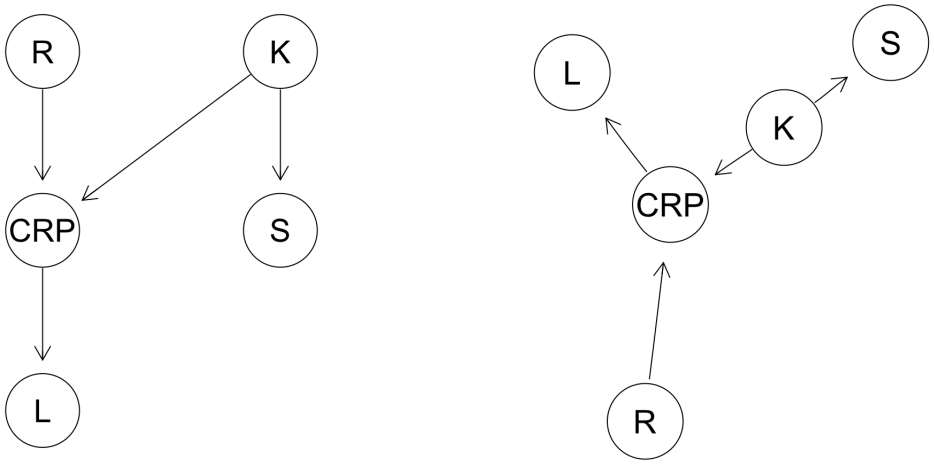
Vidíme, že charakteristiky jsou pro oba způsoby zadefinování modelu identické.

```
> crp1

  Random/Generated Bayesian network

  model:
   [R][K][CRP|R:K][S|K][L|CRP]
  nodes:                                5
  arcs:                                 4
    undirected arcs:                    0
    directed arcs:                      4
  average markov blanket size:          2.00
  average neighbourhood size:           1.60
  average branching factor:             0.80

  generation algorithm:                 Empty

> crp2

  Random/Generated Bayesian network

  model:
   [K][R][CRP|K:R][S|K][L|CRP]
  nodes:                                5
  arcs:                                 4
    undirected arcs:                    0
    directed arcs:                      4
  average markov blanket size:          2.00
  average neighbourhood size:           1.60
  average branching factor:             0.80

  generation algorithm:                 Empty
```

K vykreslení grafu používáme funkci *graphviz.plot*. Jedním z jejich argumentů je *layout*, neboli způsob rozložení grafu. Mezi možnosti patří *dot*, *circo*, *neato*, *twopi* a *fdp*.

```
> graphviz.plot(crp1, layout = "dot")
```

Na obrázcích 3.1a a 3.1b vidíme srovnání mezi rozloženími typu *dot* a *neato*.



(a) Graf s rozložením *dot* (b) Graf s rozložením *neato*

Obrázek 3.1: Srovnání mezi rozloženími grafu

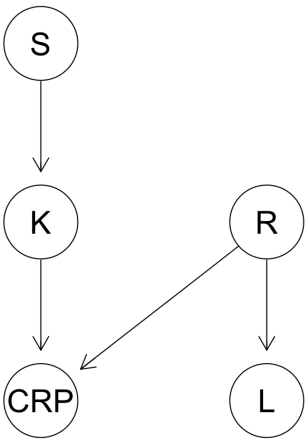
Můžeme provádět úpravy jednotlivých hran. Lze je přidat, odstranit i obracet jejich směr (viz obrázek 3.2).

```
> crp3 = set.arc(crp1, from = "R", to = "L")
> crp3 = drop.arc(crp3, from = "CRP", to = "L")
> crp3 = reverse.arc(crp3, from = "K", to = "S")
> graphviz.plot(crp3, layout = "dot")
```

Jestliže chceme vykreslit CPDAG, uděláme to následovně pomocí funkce *cpdag*. Výsledkem bude graf na obrázku 2.9.

```
> graphviz.plot(cpdag(crp1))
```

Díky funkci *mb* obdržíme seznam uzlů, které tvoří Markovův obal pro konkrétní uzel, *parents* seznam rodičů, *children* seznam přímých potomků



Obrázek 3.2: Graf CRP modelu po úpravách hran

a *vstructs* vrací uzly, které tvoří v síti v-struktury (výstup značí schéma $X \rightarrow Z \leftarrow Y$).

```
> mb(crp1, "CRP")
[1] "R" "K" "L"
> parents(crp1, "R")
character(0)
> children(crp1, "K")
[1] "CRP" "S"
> vstructs(crp1)
     X   Z     Y
[1,] "R" "CRP" "K"
```

Nyní definujeme jednotlivá podmíněná rozdělení pravděpodobnosti. Nejprve zavedeme názvy kategorií realizací proměnných. Poté vytvoříme pole s hodnotami podmíněných pravděpodobností. Nakonec seskupíme všechna tato pole do seznamu a pomocí funkce *custom.fit* vytvoříme bayesovskou síť pro příklad CRP.

```
> boolean = c("ne", "ano")
> kasel.kategorie = c("žádný", "vlhký", "suchý")
> crp.kategorie = c("není v šedé zóně", "je v šedé zóně")
```

```
> lek.kategorie = c("špatný", "správný")

> R.pst = array(c(0.25, # pravděpodobnost, že pacient nemá rýmu
+                 0.75), # p-st, že rýmu má
+               dim = 2,
+               dimnames = list(R = boolean))

> K.pst = array(c(0.2, # p-st na žádný kašel
+                 0.3, # p-st na vlhký kašel
+                 0.5), # p-st na suchý kašel
+               dim = 3,
+               dimnames = list(K = kasel.kategorie))

> CRP.pst = array(c(0.99, 0.01, # p-st, že CRP není, resp.
                                # je v šedé zóně, když
+                                # nemá rýmu ani kašel
+                   0.25, 0.75, # p-st, že CRP není/je
                                # v šedé zóně, když
+                                # nemá rýmu, ale má vlhký kašel
+                   0.2, 0.8,   #...nemá rýmu, ale má
                                #   suchý kašel
+                   0.3, 0.7,   # ...má rýmu, ale nemá kašel
+                   0.15, 0.85, # ...má rýmu a vlhký kašel
+                   0.1, 0.9),  # ...má rýmu a suchý kašel
+                 dim = c(2, 3, 2),
+                 dimnames = list(CRP = crp.kategorie,
                                  K = kasel.kategorie,
                                  R = boolean))

> S.pst = array(c(0.95, 0.05,
+                 0.4, 0.6,
+                 0.15, 0.85),
+               dim = c(2, 3),
+               dimnames = list(S = boolean,
                                K = kasel.kategorie))

> L.pst = array(c(0.05, 0.95,
+                 0.8, 0.2),
+               dim = c(2, 2),
+               dimnames = list(L = lek.kategorie,
```

```
                                    CRP = crp.kategorie))

> rozdeleni.psti = list(R = R.pst, K = K.pst, S = S.pst,
                        CRP = CRP.pst, L = L.pst)

> bayes.sit = custom.fit(crp1, rozdeleni.psti)
```

Takto vypadá podmíněné rozdělení pravděpodobnosti pro proměnnou CRP (příkaz bayes.sit$CRP).

```
  Parameters of node CRP (multinomial distribution)

Conditional probability table:

, , K = žádný

                   R
CRP                   ne  ano
  není v šedé zóně 0.99 0.30
  je v šedé zóně   0.01 0.70

, , K = vlhký

                   R
CRP                   ne  ano
  není v šedé zóně 0.25 0.15
  je v šedé zóně   0.75 0.85

, , K = suchý

                   R
CRP                   ne  ano
  není v šedé zóně 0.20 0.10
  je v šedé zóně   0.80 0.90
```

3.2 Pravděpodobnost při daném pozorování

Nyní si ukážeme, jak určit pravděpodobnosti v bayesovské síti vzhledem k pozorování konkrétních realizací daných proměnných. Tedy budeme provádět bayesovskou inferenci. Využijeme zde k tomu funkce *querygrain* z balíčku *gRain*. Počítá totiž přesné hodnoty podmíněných pravděpodobností, narozdíl od funkce *cpquery*, která je pouze odhaduje a využívá se tak zejména u gaussovských bayesovských sítí, kterým se tato práce nevěnuje.

Nejdřív musíme síť mít jako objekt typu *grain*. Poté můžeme provést inference.

Vraťme se k našemu příkladu s CRP. Jako první se podíváme na situaci, kdy pacient nebude mít žádný kašel. K tomu slouží funkce *setEvidence*. Funkce *querygrain* nám vrátí hodnoty marginálních rozdělení (*type = "marginal"* nastaven jako default, zde vypsáno pro názornost).

```
> crp.inference = compile(as.grain(bayes.sit))
> crp.evidence = setEvidence(crp.inference, nodes = "K",
                             states = "žádný")
> querygrain(crp.evidence, type = "marginal")

R
  ne  ano
0.25 0.75

CRP
není v šedé zóně   je v šedé zóně
          0.4725           0.5275

S
  ne  ano
0.95 0.05

L
  špatný  správný
0.445625 0.554375
```

Vidíme, že hodnoty u CRP a L se změnily, ale u S a R tomu tak není. U proměnné S jsme obdrželi původní hodnoty podmíněné žádným kašlem, které jsme do modelu zadávali. Pravděpodobnosti R se nezměnily, jelikož je součástí v-struktury $R \longrightarrow CRP \longleftarrow K$, která v tomto případě není aktivní, jelikož nepozorujeme konkrétní realizaci CRP. Přidejme tedy k žádnému kašli i fakt, že hladina CRP se nenachází v šedé zóně. Očekáváme, že se nám změní přesvědčení o rýmě, ale i o nasazení správného léku.

```
> crp.evidence2 = setFinding(crp.inference,
+                            nodes = c("K","CRP"),
+                            states = c("žádný",
+                                       "není v šedé zóně"))
> querygrain(crp.evidence2)

R
       ne       ano
0.5238095 0.4761905

S
  ne  ano
0.95 0.05

L
 špatný správný
   0.05    0.95
```

Dle očekávání se změnily hodnoty pravděpodobností u proměnné R i L, ale u S zůstaly stejné. Zde nám totiž struktura $CRP \longleftarrow K \longrightarrow S$ tvoří již dříve rozebíranou společnou příčinu, která není aktivní z důvodu pozorování prostředního uzlu K, a tedy není možnost již nijak ovlivnit pravděpodobnosti proměnné S.

Pokud nás zajímají pouze vybrané uzly, použijeme parametr *nodes*. Můžeme také změnit typ rozdělení na sdružené. Zde je ukázka sdruženého rozdělení pravděpodobnosti pro proměnné R a L za stejných podmínek jako

v předchozím.

```
> querygrain(crp.evidence2,
+            nodes = c("R","L"),
+            type = "joint")
     L
R         špatný  správný
  ne  0.02619048 0.497619
  ano 0.02380952 0.452381
```

Jako poslední si ukážeme vliv pozorování uzlu L na celý model.

```
> crp.evidence3 = setEvidence(crp.inference,
+                             nodes = c("L"),
+                             states = c("špatný"))
> querygrain(crp.evidence3)
R
       ne       ano
0.2023138 0.7976862

CRP
není v šedé zóně   je v šedé zóně
      0.01629399       0.98370601

K
    žádný     vlhký     suchý
0.1386350 0.3120747 0.5492903

S
       ne       ano
0.3389267 0.6610733
```

Závěr

V této práci jsme prozkoumali základní teorii diskrétních bayesovských sítí a vysvětlili ji na vlastních reálných příkladech. Skrze tuto analýzu jsme demonstrovali úspěšné fungování této teorie, její praktičnost a efektivitu. Software R jsme využili k implementaci a vizualizaci diskrétních bayesovských sítích.

Když jsem si toto téma vybral, tak jsem netušil, čeho se grafické modely týkají, ale vůbec tohoto výběru nelituji, ba naopak. Ačkoli některé části teorie byly ze začátku pro mě velmi neintuitivní a dlouho trvalo než jsem jim porozuměl, tak mi toto téma poskytlo cenné poznatky a vědomosti v oblasti grafických modelů, které určitě využiji při jejich dalším zkoumání.

Doufám, že se mi podařilo téma diskrétních bayesovských sítí přiblížit a vysvětlit co nejasněji, a že tato práce poskytne čtenáři pevný základ pro další studium bayesovských sítí či jiných grafických modelů.

Literatura

[1] Koller, D., Friedman, N.: *Probabilistic Graphical Models: Principles and Techniques.* MIT Press, Cambridge, Massachusetts, 2012.

[2] Nagarajan, R., Scutari, M., Lèbre, S.: *Bayesian networks in R with applications in Systems Biology.* Springer, New York, 2013.

[3] Højsgaard, S., Edwards, D., Lauritzen, S.L.: *Graphical models with R.* Springer, Boston, 2012.

[4] Scutari M.: *Understanding Bayesian Networks with Examples in R.* [online]. Dostupné z: `https://www.bnlearn.com/about/teaching.html`

[5] Scutari M., Strimmer K.: *Introduction to Graphical Modelling.* [online.] Dostupné z: `https://arxiv.org/pdf/1005.1036.pdf`.

[6] Hron, K., Kunderová, P. a Vencálek, O.: *Základy počtu pravděpodobnosti a metod matematické statistiky (4. doplněné vydání)*. Univerzita Palackého v Olomouci, Přírodovědecká fakulta, Olomouc, 2021.