17. Probabilistički grafički modeli

Strojno učenje 1, UNIZG FER, ak. god. 2021./2022.

Jan Šnajder, predavanja, v2.0

Prošli put bavili smo se **Bayesovim klasifikatorom**. Bayesov klasifikator je probabilistički i generativni model. Danas nastavljamo s probabilističkim modelima. Bavit ćemo se jednom širom porodicom probabilističkih modela koju zovemo **probabilistički grafički modeli** (engl. *probabilistic graphic models*), ili kraće **PGM-ovi**. PGM-ovi uključuju Bayesov klasifikator, ali i druge složenije generativne modele, a također uključuju i neke diskriminativne modele. Možda ste čuli za skrivene Markovljeve modele (HMM), ili Latentnu Dirichletovu alokaciju (LDA), ili pak za uvjetna slučajna polja (CRF). Sve su to PGM-ovi koji se vrlo često koriste u praksi i na mnogim klasifikacijskim zadatcima daju vrlo dobre rezultate. S konceptualne strane, PGM-ovi su zanimljivi jer predstavljaju jedan dobar (možda najbolji?) radni okvir za modeliranje **kauzalnosti**, a kauzalnost je ono što bi nas u znanosti trebalo najviše zanimati. 1

PGM-ovi su ogromno područje unutar strojnog učenja: mnogo se ljudi bavi PGM-ovima i postoje mnoge korisne primjene. Zapravo, PGM-ovi su bili dominantna paradigma u strojnom učenju do negdje 2010. godine, kada se fokus prebacio na duboko učenje. Budući da se radi o velikom području, a da je naše vrijeme ograničeno, radit ćemo na dvije razine: temeljne ideje ćemo temeljito obraditi, dok ćemo neke napredne koncepte samo spomenuti.

1 Uvod

Kao uvod u PGM-ove poslužit će nam **naivan Bayesov klasifikator**. Prisjetimo se najprije tog modela. Npr., model naivnog Bayesovog klasifikatora za tri značajke jest:

$$h(\mathbf{x}) = P(x_1|y)P(x_2|y)P(x_3|y)P(y)$$

Također, prisjetimo se **polunaivnog Bayesovog klasifikatora**, kod kojega ne pretpostavljamo da su sve značajke uvjetno nezavisne za zadanu klasu, nego dopuštamo da su neke značajke možda zavisne, pa ih modeliramo zajedničkim faktorom. Npr., ako ne želimo pretpostaviti uvjetnu nezavisnost značajki x_2 i x_3, onda ćemo model definirati ovako:

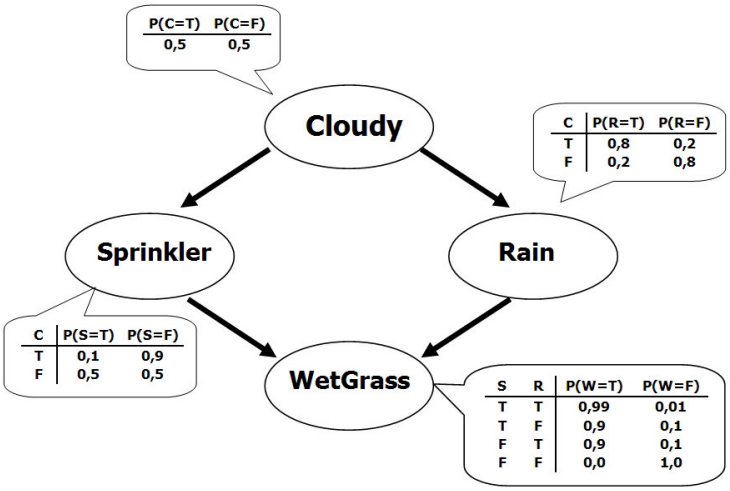
$$h(\mathbf{x}) = P(x_1|y)P(x_2, x_3|y)P(y)$$

U oba ova slučaja radi se o tome da smo na neki način (potpuno ili manje potpuno) faktorizirali zajedničku distribuciju $P(x_1, x_2, x_3)$. Ispostavlja se da su oba ova modela samo posebni slučajevi probabilističkog grafičkog modela (PGM).

PGM je **sažet način zapisa zajedničke distribucije pomoću grafa**. Ta zajednička distribucija može biti vrlo složena i može biti definirana nad visokodimenzijskim prostorima. Pritom graf služi kao kostur za faktorizaciju zajedničke distribucije: zasniva se na ideji da svaka varijabla interagira samo s manjim brojem drugih varijabli, tj. zavisnosti između varijabli su **lokalne**. Čvorovi u grafu su varijable, dok bridovi u grafu kodiraju zavisnosti između varijabli. Budući da se graf može prikazati grafički, to se ovi modeli zovu **grafički modeli**. To je malo nesretno ime, ali sad je gotovo. 2

▸ **PRIMJER**

Evo jednog jednostavog motivirajućeg primjera, koji se pojavljuje u svoj literaturi o PGM-ovima:



Ova mreža prikazuje zajedničku vjerojatnost na temelju zavisnosti između varijabli. Možemo reći da te zavisnosti zapravo modeliraju **kauzalnost** (uzročnost). Svaki čvor odgovara jednoj slučajnoj varijabli i pohranjuje uvjetne vjerojatnosti za tu varijablu – to su **uvjetne vjerojatnosne distribucije** (engl. *conditional probability distributions, CPDs*). Parametri uvjetnih vjerojatnosti za svaki čvor zapravo su parametri modela, koje možemo naučiti iz podataka, npr. procjeniteljem MLE. Ako su varijable diskretne, kao što je to ovdje slučaj, onda su parametri njihovi distribucija zapravo parametri μ_k kategoričke razdiobe, i njih možemo prikazati **tablicom uvjetne vjerojatnosti** (engl. *conditional probability table, CPT*), koju smo već bili spomenuli prošli put. U ovom primjeru, graf je izgrađen ručno, na temelju našeg znanja o kauzalnosti između događaja, no postoje postupci da se i sam graf uči na temelju podataka.

Svrha PGM-a jest omogućiti **probabilističko zaključivanje** o vrijednostima jedne ili više varijabli, moguće uz fiksirane vrijednosti nekih drugih varijabli. Na primjer, na temelju mreže iz gornjeg primjera možemo zaključiti da, ako vidimo da je trava mokra, a ne znamo je li bilo oblačno, koliko je vjerojatno da je radila prskalica, a koliko da je padala kiša (konkretno, ispada da je vjerojatnost da je radila prskalica 43%, a vjerojatnost da je padala kiša 70%).

Postoje tri aspekta PGM-a o kojima trebamo pričati: **prikazivanje (reprezentacija)**, **zaključivanje** (engl. *inference*) i **učenje**.

Što se reprezentacije tiče, tu postoje dvije osnovne porodice modela: **Bayesove mreže** (engl. *Bayes(ian) networks*), koje koriste usmjerene grafove, i **Markovljeve mreže** (engl. *Markov networks*), koje koriste neusmjerene grafove. Ova dva pristupa razlikuju se po načinu kako prikazuju zavisnosti između varijabli. Ta razlika nije samo estetske prirode: svaki od ovih modela može prikazati zavisnosti koje onaj drugi ne može. Ovaj gornji primjer je Bayesova mreža.

Zaključivanje znači da, na temelju kompaktnog zapisa zajedničke distribucije, odredimo vrijednosti nepoznatih varijabli na temelju poznatih varijabli – vrijednosti značajki. To se također naziva **upiti nad distribucijom** (engl. *querying the distribution*): upit su varijable koje su nam poznate, a odgovor su vrijednosti varijabli koje su nam prethodno bile nepoznate. (Primijetite da ovo nije isto kao "zaključivanje" u statistici, koje se zapravo svodi na procjenu parametara distribucije na temelju uzorka iz populacije.)

Konačno, htjeli bismo moći učiti ovakve modele na temelju podataka. Učenje modela može biti **generativno** ili **diskriminativno**. Bayesove mreže (usmjereni grafovi) tipično učimo generativno, a Marovljeve mreže (neusmjereni grafovi) tipično učimo diskriminativno. Osim učenja parametara, možemo učiti i samu strukturu modela (graf).

Danas ćemo se fokusirati na Bayesove mreže, poput ove u gornjem primjeru. Bayesove mreže također se nazivaju **usmjereni grafički modeli**. Također se nazivaju **mreže vjerovanja** (engl. *belief networks*) i **kauzalne mreže**, jer lijepo mogu opisati naše vjerovanje o kauzalnim vezama između događaja. (Teorija vjerojatnosti je dobar radni okvir za opisivanje kauzalnosti.)

Kao što smo rekli, postoje tri aspekta o kojima možemo pričati kod PGM-ova, pa tako i kod Bayesovih mreža: reprezentacija, zaključivanje i učenje. Danas ćemo pričati samo o reprezentaciji, dok ćemo o zaključivanju i učenju pričati idući put.

2 Bayesove mreže: reprezentacija

2.1 Usmjereni graf i uvjetne (ne)zavisnosti

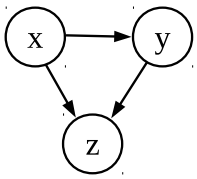
Kao primjer, razmotrit ćemo zajedničku distribuciju triju varijabli:

$$p(x, y, z)$$

Općenitosti radi, ovdje pišemo gustoće vjerojatnosti p, ali sve što ćemo dalje raditi vrijedi i za vjerojatnosti P, tj. za slučaj kada su varijable diskretne. Primjenom **pravila umnoška** ovu distribuciju možemo napisati kao:

$$p(x, y, z) = p(x)p(y|x)p(z|x, y)$$

Iz ovoga sada crtamo usmjereni graf: čvorovi grafa odgovaraju varijablama (ili skupovima varijabli), a bridovi odgovaraju zavisnostima, i to tako da crtamo brid od varijable koja uvjetuje prema varijabli koja je uvjetovana:



Npr., za faktor $p(z|x, y)$ imamo bridove iz x i y u z, dok za faktor $p(x)$ nemamo ulaznih bridova.

Primijetite da smo faktorizaciju mogli napraviti i na neki drugi način. Mi smo odabrali redoslijed varijabli x, y, z. Međutim, mogli smo odabrati neki drugi redoslijed varijabli (od ukupno $3! = 6$ mogućih), npr.:

$$p(x, y, z) = p(y)p(z|y)p(x|y, z)$$

Ovo naravno funkcionira i kad imamo više od tri varijable. Naime, prisjetimo se da svaku zajedničku distribuciju uvijek možemo faktorizirati primjenom **pravila lanca** (koji je zapravo samo višestruka primjena pravila umnoška):

$$\begin{aligned} p(x_1, \ldots, x_n) &= p(x_1)p(x_2|x_1)p(x_3|x_1, x_2) \cdots p(x_n|x_1, \ldots, x_{n-1}) \\ &= \prod_{k=1}^{n} p(x_k|x_1, \ldots, x_{k-1}) \end{aligned}$$

Kako izgleda Bayesova mreža koja odgovara takvoj faktorizaciji? Svaki čvor ima ulazne bridove iz svih čvorova koji mu u odabranom poretku prethode (svih čvorova s manjim rednim brojem indeksa). Drugim riječima, svaki čvor x_i povezan je sa svim čvorovima $x_{i-1}, \ldots, x_1$. To zapravo znači da je Bayesova mreža **potpuno povezan graf** (svaki par čvorova je povezan).

Međutim, iz naših prošlotjednih razmatranja Bayesovog klasifikatora, već znamo što je problem s takvim modelima: presloženi su. Broj kombinacija vrijednosti varijabli (a tome odgovara i broj parametara modela koje moramo procijeniti iz podataka) raste ekponencijalno s

brojem varijabli. Nadalje, takvi modeli uopće ne mogu generalizirati: za kombinacije vrijednosti značajki koje se nisu pojavile u skupu za učenje vjerojatnost je jednaka nuli. Da bi model mogao generalizirati, moramo uvesti neke **induktivne pristranosti**, odnosno neke pretpostavke. Kod probabilističkih grafičkih modela, jednako kao i kod Bayesovog klasifikatora, te pretpostavke dolaze u obliku **uvjetnih nezavisnosti**. Prisjetimo se, slučajne varijable X i Y uvjetno su nezavisne uz varijablu Z, ako:

$$\begin{aligned} X \perp Y|Z &\Leftrightarrow P(X,Y|Z) = P(X|Z)P(Y|Z) \\ &\Leftrightarrow P(X|Y,Z) = P(X|Z) \\ &\Leftrightarrow P(Y|X,Z) = P(Y|Z) \end{aligned}$$

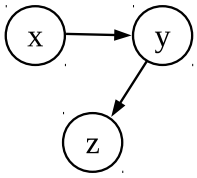
Uvođenjem uvjetnih nezavisnosti između parova varijabli, zajednička distribucija može se jednostavnije faktorizirati, tj. faktorizirati u faktore koji imaju manji broj varijabli (a time i parametara) nego što bi to imali faktori bez pretpostavki o uvjetnoj nezavisnosti. Npr., ako u ranijem primjeru pretpostavimo da vrijedi:

$$x \perp z|y$$

onda $p(z|x,y) = p(z|y)$, pa dobivamo:

$$p(x,y,z) = p(x)p(y|x)p(z|\cancel{x},y) = p(x)p(y|x)p(z|y)$$

što nam daje i jednostavniju Bayesovu mrežu:



Nadalje, ako bismo još pretpostavili da $x \perp y$ (marginalna nezavisnost), onda $p(y|x) = p(y)$, pa se faktorizacija i pripadna Bayesova mreža još više pojednostavljuju:

$$p(x,y,z) = p(x)p(y|\cancel{x})p(z|y) = p(x)p(y)p(z|y)$$

Vidimo da uvođenje dodatnih pretpostavki o nezavisnosti zapravo pojednostavljuje Bayesovu mreža – i to uklanjanjem bridova. Gdjegod između para varijabli nedostaje neki brid (u odnosu na potpuno povezan graf), tu su varijable međusobno uvjetno nezavisne. Konstatiramo, dakle, da Bayesove mreže – a to vrijedi i za PGM-ovove općenito – sažeto prikazuju zajedničku distribuciju na temelju pretpostavki o **uvjetnoj nezavisnosti**. Te pretpostavke **pojednostavljuju** model i u konačnici omogućavaju **generalizaciju**.

Vidimo da iz faktorizacije zajedničke vjerojatnosti uvijek možemo konstruirati pripadnu Bayesovu mrežu. Također je očito da možemo napraviti i obrnuto: iz zadane Bayesove mreže možemo iščitati faktorizaciju zajedničke distribucije.

2.2 Uređajno Markovljevo svojstvo

Definirajmo sada malo formalnije taj odnos između Bayesove mreže (usmjerenog grafa) i njoj odgovarajuće zajedničke distribucije. Za Bayesovu mrežu sa n čvorova (varijabli), zajednička distribucija dana je sa:

$$p(\mathbf{x}) = \prod_{k=1}^{n} p(x_k|\mathrm{pa}(x_k))$$

gdje $\mathrm{pa}(x_k)$ označava čvorove roditelje čvora x_k. Da bi ova faktorizacija bila moguća, bitno je da se čvorovi mogu poredati u **potpuni (linearni) uređaj** tako da roditelji dolaze prije djece

(po indeksima). Inače bi se jedna te ista varijabla trebala u faktorizaciji koristiti više puta, što ne bi dalo faktorizaciju. Ovaj uvjet u Bayesovoj mreži svodi se na to da u grafu ne smije biti usmjerenih ciklusa (diciklusa), tj. Bayesova mreža je **usmjereni aciklički graf** (engl. *directed acyclic graph, DAG*). Uređaj čvorova koji možemo iščitati iz DAG-a zapravo je tzv. **topološki uređaj**: potpuni (linearni) uređaj kod kojeg čvorovi roditelji dolaze prije čvorova djece. Svaki DAG ima barem jedan topološki uređaj, ali ih može imati i više.

Za zadani topološki uređaj, faktorizacija koja odgovara Bayesovoj mreži zapravo pretpostavlja da svaki čvor x_k ovisi samo o svojim **direktnim roditeljima**, a ne i o svim svojim prethodnicima. To možemo formalno zapisati ovako:

$$x_k \perp \text{pred}(x_k) \backslash \text{pa}(x_k) \mid \text{pa}(x_k)$$

gdje je $\text{pred}(x_k)$ skup prethodnika čvora x_k po topološkom uređaju. Dakle, ako su poznate vrijednosti čvorova roditelja, vrijednost čvora x_k nezavisna je od vrijednosti čvorova koji prethode roditeljskim čvorovima. Navedeno svojstvo naziva se **uređajno Markovljevo svojstvo** (engl. *ordered Markov property*).

Razmotrima sada jedan primjer koji će nam pomoći da bolje shvatimo povezanost između grafa Bayesove mreže i uvjetnih nezavisnosti koje su kodirane tom mrežom.

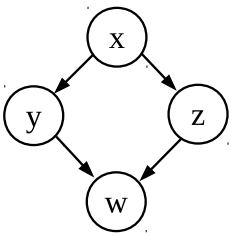
▸ **PRIMJER**

Pretpostavimo da se zajednička vjerojatnost od četiri varijable može faktorizirati na sljedeći način:

$$p(x, y, z, w) = p(x)p(y|x)p(z|x)p(w|y, z)$$

Vidimo da, kada bismo pokušali pomnožiti ove faktore primjenom pravila umnoška, ne bismo uspjeli dobiti vjerojatnost na lijevoj strani jednakosti. Drugim riječima, ova faktorizacija nije dobivena izravnom primjenom pravila lanca, a to onda znači da su u ovu faktorizaciju ugrađene neke uvjetne nezavisnosti.

Bayesova mreža koja odgovara ovoj faktorizaciji je sljedeća:



Koji bridovi ovdje nedostaju (u odnosu na potpuno povezan graf)? Nedostaju dva brida: od x do w i od y do z (ili obrnuto, ovisno za koji se topološki uređaj odlučimo). Topološki uređaji čvorova su x, y, z, w i x, z, y, w. Odaberimo ovaj prvi.

Možemo li iz Bayesove mreže rekonstruirati zajedničku distribuciju $p(x, y, z, w)$? To znači da želimo napisati zajedničku distribuciju kao umnožak faktora, poštujući pritom pretpostavke o uvjetnoj nezavisnosti koje su implicitno (kroz odsustvo bridova) kodirani u Bayesovoj mreži.

Prvi način na koji to možemo napraviti jest da krenemo od faktorizacije Bayesove mreže i primjenjujemo **pravilo umnoška**, proširujući faktore gdjegod je to potrebno. Gdjegod proširujemo faktore, zapravo otkrivamo uvjetnu nezavisnost koja je bila kodirana u tom faktoru dok on nije bio proširen. Konkretno:

$$\begin{aligned}
p(x)p(y|x)p(z|x)p(w|y,z) &= p(x,y)\underline{p(z|x)}p(w|y,z) \\
y \perp z|x \Rightarrow\ & p(x,y)\underline{p(z|x,y)}p(w|y,z) \\
&= p(x,y,z)\underline{p(w|y,z)} \\
x \perp w|y,z \Rightarrow\ & p(x,y,z)\underline{p(w|x,y,z)} \\
&= p(x,y,z,w)
\end{aligned}$$

Potcrtani su faktori koje proširujemo i koje smo dobili proširivanjem. To su faktori kod kojih smo dodajemo varijable u uvjetni dio uvjetne vjerojatnosti, kako bismo mogli primijeniti pravilo umnoška. Npr., u prvom koraku faktor $p(z|x)$ proširili smo u $p(z|x,y)$, tj. dodali smo varijablu y u uvjetni dio, jer tek tako prošireni faktor možemo pomnožiti s faktorom $p(x,y)$ kako bismo, na temelju pravila umnoška, dobili faktor $p(x,y,z)$. Budući da smo faktor $p(z|x)$ napisali kao faktor $p(z|x,y)$, to znači da mi zapravo pretpostavljamo da vrijedi $p(z|x) = p(z|x,y)$, a to, prema definiciji uvjetne nezavisnosti, znači da prepostavljamo da vrijedi uvjetna nezavisnost $y \perp z|x$. Slično smo postupili kod faktora $p(w|y,z)$, koji smo proširili varijablom x u uvjetnom dijelu. Na koncu smo doista uspijeli doći do zajedničke vjerojatosti primjenom pravila umnoška. Pritom smo uveli dvije prepostavke o uvjetnoj nezavisnosti: $y \perp z|x$ i $x \perp w|y,z$. To su pretpostavke koje su implicitno ugrađene u graf Bayesove mreže. Primijetite da te pretpostavke ne možemo samo tako iščitati iz Bayesove mreže: naime, u Bayesovoj mreži nedostaju bridovi između varijabli x i w i varijabli y i z, pa bismo mogli zaključiti da su ti parovi varijabli nezavisni, međutim nije jasno čime je ta nezavisnost uvjetovana. (Vidjet ćemo nešto kasnije da postoje pravila pomoću kojih bismo to ipak mogli zaključiti.)

Alternativno, umjesto da pokušamo rekonstruirati zajedničku distribuciju i pritom implicitno otkrivamo uvjetne nezavisnosti, uvjetne nezavisnosti možemo izvesti izravno iz Markovljevog svojstva:

$$x_k \perp \text{pred}(x_k) \backslash \text{pa}(x_k) \mid \text{pa}(x_k)$$

To ćemo napraviti tako da izrazimo Markovljevo svojstvo za svako od četiri varijable:

$$\begin{aligned} y \perp \{x\} \backslash \{x\} \mid \{x\} & \\ z \perp \{x,y\} \backslash \{x\} \mid \{x\} & \Rightarrow y \perp z|x \\ w \perp \{x,y,z\} \backslash \{y,z\} \mid \{y,z\} & \Rightarrow x \perp w|y,z \end{aligned}$$

Ovdje smo koristili topološki uređaj x, y, z, w, pa je tako, na primjer, $\text{pred}(z) = \{x, y\}$. Primijetite da vrijedi komutativnost (uvjetne) nezavisnosti, pa je $z \perp y|x$ isto što i $y \perp z|x$. Također primijetite da u prvom izrazu nismo uspjeli izvesti nezavisnost za y, ali u drugom jesmo. To ovisi o tome koji smo topološki uređaj odabrali. Da smo se odlučili za drugi topološki uređaj (x, z, y, w), dobili bismo u konačnici iste dvije uvjetne nezavisnosti.

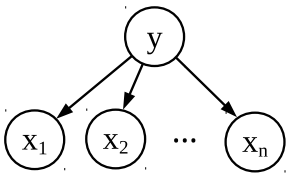
Pogledajmo sada nekoliko primjera Bayesovih mreža.

3 Primjeri Bayesovih mreža

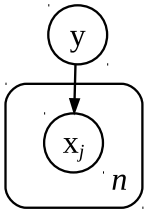
Prošli tjedan pričali smo o **naivnom Bayesovom klasifikatoru**. Kod naivnog Bayesovog klasifikatora pretpostavljamo da su značajke uvjetno nezavisne za danu klasu. Ta nam je pretpostavka omogućila ovakvu faktorizaciju zajedničke vjerojatnosti:

$$p(\mathbf{x}, y) = p(y) \prod_{j=1}^{n} p(x_j|y)$$

Ovoj faktorizaciji odgovara sljedeća Bayesova mreža:



Čvorova x_j ima onoliko koliko ima značajki, odnosno možemo reći da je čvor x_j multipliciran n puta. Često se događa da se neki čvor tako multiplicira: za svaku značajku ili za svaki primjer ili nešto treće. Kako bi se u takvim slučajevima pojednostavio grafički izgled Bayesove mreže, uvodimo malo sintaktičkog šećera: **pladnjeve** (engl. *plate notation*). Varijablu koja se ponavlja crtamo u okviru, varijabli dodajemo indeks, i u kutu pladnja pišemo broj ponavljanja:



Prošli put rekli smo da, u slučajevima kada između nekih varijabli postoji zavisnost (točnije: kada ne postoji uvjetna nezavisnost), bolje je ne prepostaviti takvu nezavisnost, odnosno ne faktorizirati u potpunosti. To nas je dovelo do **polunaivnog Bayesovog klasifikatora**. Npr., ako znamo (ili ako utvrdimo, npr., mjerom uzajamne informacije) da varijable x_2 i x_3 nisu uvjetno nezavisne za klasu y, tj $.x_2 \not\perp x_3|y$, onda nam je bolje da umjesto

$$p(x_1, x_2, x_3, y) = p(x_1|y)p(x_2|y)p(x_3|y)p(y)$$

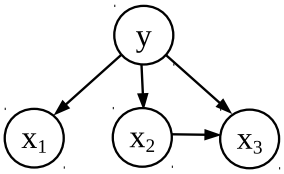
zajedničku vjerojatnost faktoriziramo kao:

$$p(x_1, x_2, x_3, y) = p(x_1|y)\underline{p(x_2, x_3|y)}p(y)$$

što je jednako kao (primjenom **pravila umnoška**):

$$p(x_1, x_2, x_3, y) = p(x_1|y)\underline{p(x_2|y)p(x_3|x_2, y)}p(y)$$

Ovoj drugoj faktorizaciji odgovara sljedeća Bayesova mreža:

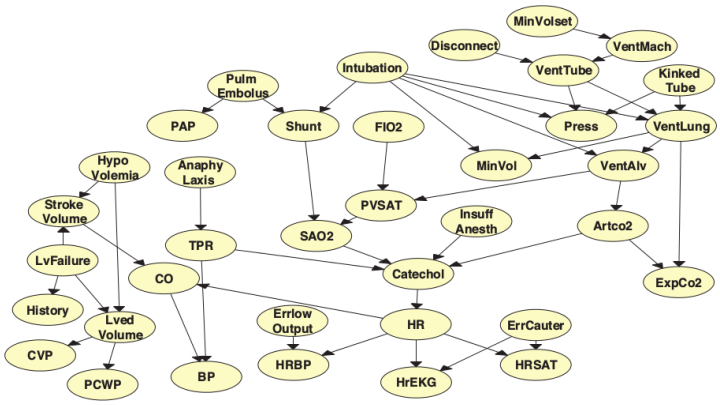


(Prvoj faktorizaciji odgovara Bayesova mreža koja u jednom čvoru ima kombinaciju dvije varijable, x_2, x_3, tj. ima čvor za zajedničku vjerojatnost $p(x_2, x_3|y)$. Ovo razlika je samo grafičke prirode; modeli su zapravo istovjetni te imaju isti broj parametara).

A evo sada i nekoliko primjera složenijih Bayesovih mreža iz literature.

▸ **PRIMJER**

"Alarm network" dijagnostički je model za nadzor pacijenata, temeljen na Bayesovoj mreži. Mreža ima 37 varijabli, od čega 8 varijabli predstavlja moguće dijagnoze, 16 varijabli predstavljaju simptome, a ostalih 13 varijabli su dodatne varijable: 5

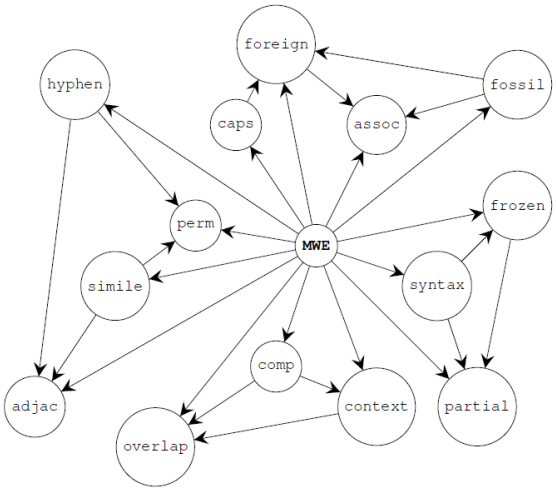


Varijable simptoma očitavaju fiziološke podatke pacijenta, kao što su krvni tlak, frekvenciju srca, itd. Sve varijable su kategoričke (kontinuirane varijable su diskretizirane). Mreža ukupno ima 504 parametara.

▸ **PRIMJER**

Drugi primjer je model iz područja obrade prirodnog jezika. Model je razvijen u svrhu prepoznavanja višerječnih izraza (engl. *multiword expressions, MWE*) hrvatskoga jezika, npr. izraza poput "godišnji odmor", "povišica plaće", "uputstva za uporabu" ili "Zavod za elektroniku, mikroelektroniku, računalne i inteligentne sustave". Višerječni izrazi uključuju i tzv. semantički neprozirne izraze, kod kojih je značenje izraza nadilazi jednostavnu kombinaciju značenja riječi od kojih su ti izrazi sastavljeni, npr., izrazi poput "ležeći policajac" ili "medeni mjesec". Prepoznavanje višerječnih izraza važno je u obradi prirodnog jezika, kako bi se takvi izrazi mogli obrađivati kao cjeline, umjesto da ih se rastavlja na pojedinačne riječi. Prepoznavanje se tipično provodi analizom velikih količina tekstova (tzv. korpusa), gdje se analiziraju statistička stvojstva nizova riječi, kako bi se utvrdilo čini li neki niz riječi višerječnu jedinicu. 6

Bayesova mreža funkcionira kao binarni klasifikator: za zadani niz od n riječi mreža treba odlučiti je li taj niz višerječni izraz hrvatskoga jezika, ili je naprosto slučajna kombinacija od dvije ili više riječi. U tu svrhu za nizove riječi iz korpusa izračunali smo niz značajki, poput frekvencije zajedničkog pojavljivanja riječi, pojavljuju li se te riječi u permutiranom redoslijedu, sadrži li neka od riječi spojnicu, je li neka od riječi strana riječ, itd. To su varijable u našoj Bayesovoj mreži. Ciljna varijabla je varijabla MWE, koja određuje je li zadani niz riječi doista višerječni izraz hrvatskoga jezika. Mreža izgleda ovako:



Varijabla MWE (u sredini mreže) je varijabla klase koja nas zanima, a ona uvjetuje sve druge varijable, bilo direktno ili indirektno. Ova je mreža izrađena ručno, na temelju lingvističke intuicije (postupci za učenje strukture mreže nisu dali dobre rezultate). Parametri mreže naučeni su na ručno označenom skupu podataka (skupu koji za nizove riječi ima označeno jesu li to doista višerječni izrazi ili nisu).

▸ **PRIMJER**

Microsoftov Office je od 1997. godine imao ugrađenog virtualnog asistenta Clippy, koji je pružao podršku korisnicima pri izvođenju određenih akcija, ovisno o ciljevima korisnika. Clippy je bio temeljen na Bayesovoj mreži. Varijable te mreže bile su akcije korisnika (npr., gdje je trenutačno pozicioniran pokazivač miša, je li korisnik u prošlosti odbijao ponudu za pomoć i sl.). 7

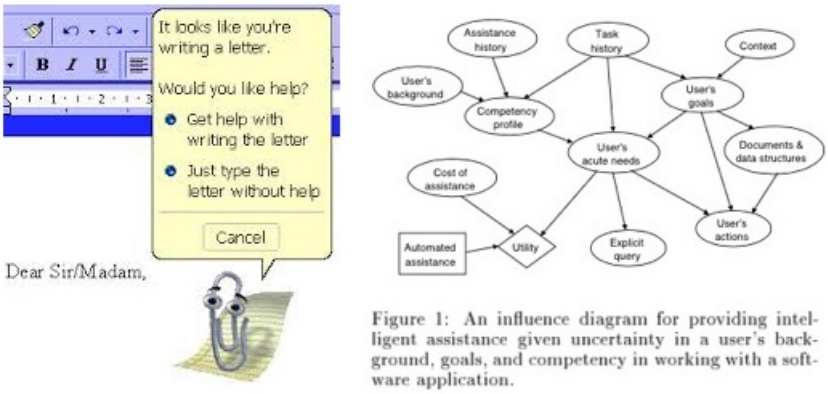
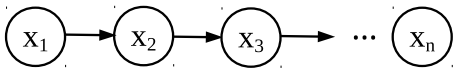


Figure 1: An influence diagram for providing intelligent assistance given uncertainty in a user's background, goals, and competency in working with a software application.

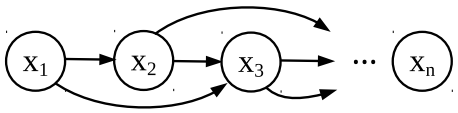
Još jedna važna vrsta PGM-ova su modeli za **slijedne podatke**. To mogu biti vremenski nizovi ili, npr., riječi u rečenici (kod obrade jezika) ili fonemi u riječi (kod prepoznavanja govora). Tipično se za slijedne podatke koriste Bayesove mreže koje nazivamo **Markovljev model**. Specifično, **Markovljev model prvog reda** zajedničku distribuciju faktorizira u lanac, na ovakav način:

$$p(\mathbf{x}) = p(x_1) \prod_{k=2}^{n} p(x_k|x_{k-1})$$

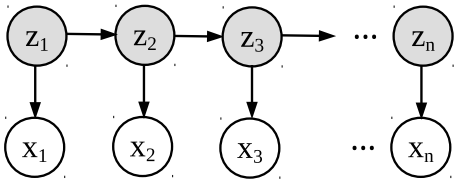


Lanac može odgovarati točkama u vremenu ili općenito bilo kakvom slijedu (npr., niz riječ u rečenici). Npr., ako su u pitanju riječi, ovim modelom možemo modelirati da vjerojatnost da se u rečenici pojavi neka riječ ovisi o riječi koja joj neposredno prethodi. Ako postoje zavisnosti koje idu dalje od samo jedne prethodne varijable, npr., ako želimo modelirati da riječ u rečenici ovisi o dvije prethodne riječi, možemo koristiti **Markovljev model drugog reda**:

$$p(\mathbf{x}) = p(x_1)p(x_2|x_1) \prod_{k=3}^{n} p(x_k|x_{k-1}x_{k-2})$$



No, što ako postoje još dalje zavisnosti? To onda postaje problematično: ne možemo u nedogled povećavati red modela, jer to, kao što znamo, brzo dovodi do eksplozije broja parametara. Rješenje je **skriveni Markovljev model** (engl. *Hidden Markov model, HMM*). Kod tog modela stanja procesa razdvojena su od podataka koje opažamo. Pretpostavljamo da je proces koji generira podatke doista Markovljev model prvog reda, no mi ne opažamo direktno stanja tog procesa, nego opažamo podatke koje taj proces generira i na koje utječe šum. Bayesova mreža HMM-a je ovakva:



Sam proces koji generira opažene podatke x_j nije opažen, tj. skriven je, i taj proces prolazi kroz stanja $z_1, z_2, \ldots, z_n$ (varijable koje nisu odgovaraju u gornjem prikazu sivo obojanim čvorovima). Ono što mi opažamo jesu podatci koje taj proces generira, tj. opažamo slijed $x_1, x_2, \ldots, x_n$. Ovoj mreži odgovara sljedeća faktorizacija zajedničke distribucije:

$$p(\mathbf{x}, \mathbf{z}) = p(z_1)p(x_1|z_1) \prod_{k=2}^{n} p(z_k|z_{k-1})p(x_k|z_k)$$

HMM se, dakle, sastoji od **modela prijelaza** $p(z_k|z_{k-1})$ i **modela opažanja** $p(x_k|z_k)$. Model prijelaza opisuje prijelaz skrivenog procesa iz jednog stanja u drugi. Budući da varijable z_k ne opažamo u podatcima, već opažamo samo varijable x_k, to malo komplicira postupak zaključivanja i učenja. O tome više idući put.

4 D-odvajanje

Uvjetne nezavisnosti glavni su sastojak Bayesovih mreža. Iz zadane Bayesove mreže možemo izračunati zajedničku distribuciju. No, što ako nas zanima jesu li neke dvije varijable uvjetno nezavisne? Npr., u Bayesovoj mreži s travom i prskalicom, jesu li varijable "cloudy" i "wet grass" nezavisne? A jesu li nezavisne ako opažamo varijable "sprinkler" i "rain"? Sada se možda pitate zašto bi nam bilo bitno znati da su neke varijable uvjetno nezavisne. To je bitno s praktične strane, jer ako to znamo, možemo pri zaključivanju ukloniti one varijable koje su nezavisne od upita i time smanjiti broj varijabli koje moramo analizirati. No, važnije, znati koje su varijable nezavisne bitno je s konceptualne strane, jer na taj način možemo zaključivati o tome koje statistička svojstva (zavisnost/nezavisnost varijabli) proizlaze iz određene kauzalne strukture (definirane Bayesovom mrežom). Na primjer, ako varijable "cloudy" i "wet grass" 8
nisu uvjetno nezavisne, onda znamo da između njih postoji neka kauzalna veza. S druge strane, ako su varijable "cloudy" i "wet grass" nezavisne uz opažene varijable "sprinkler" i "rain", onda znamo da su prskalica ili kiša već same po sebi dovoljne da uzrokuju mokru travu, tj. da se nalaze negdje u kauzalnom lancu između oblaka i mokre trave. Za ovakvu vrstu zaključivanja o kauzalnim odnosima nužno je da možemo ispitati nezavisnosti i uvjetne nezavisnosti varijabli
i skupova varijabli u Bayesovoj mreži. 9

Očito, kako smo pokazali u ranijem primjeru, primjenom **uređajnog Markovljevog svojstva** (ili višestrukom primjenom pravila umnoška i pravila zbroja) mogli bismo iz zadane Bayesove mreže izvesti sve uvjetne nezavisnosti koje su u njoj kodirane. Međutim, to je vrlo nepraktično. Osim toga, na taj način ne možemo odrediti zavisnost ili nezavisnost dviju proizvoljnih varijabli iz Bayesove mreže, koje u toj mreži mogu biti vrlo udaljene jedna od druge. Srećom, Bayesove mreže nude jednu elegantnu mogućnost da se uvjetne nezavisnosti izvedu direktno iz strukture usmjerenog grafa. Taj se postupak temelji na principu tzv. **d-odvajanja** (engl. *directed separation*).

Ideja d-odvajanja jest da analiziramo **staze** u grafu između čvorova x i y, koji odgovaraju varijablama čiju nezavisnost želimo ispitati, i da onda na temelju toga odredimo jesu li varijable uvjetno nezavisne. Staze između čvorova mogu biti **povezane** ili **odvojene**, već ovisno o tome koje su varijable u mreži opažene. Ako su sve staze između dva čvora odvojene, onda su čvorovi uvjetno nezavisni. Naravno, da bi ovo funkcioniralo, moramo definirati kada su staze povezane a kada su odvojene. I to ćemo napraviti na temelju uređajnog Markovljevog svojstva. Prema tome, na d-odvajanje možemo gledati kao na proširenje uređajnog Markovljevog svojstva, koje opisuje lokalnu uvjetnu nezavisnost varijabli, na nezavisnost između varijabli koje su u mreži proizvoljno udaljene, a koje su povezane nekom stazom.

4.1 Pravila d-odvajanja

Konkretno, definirat ćemo tri jednostavna pravila. Ona opisuju tri osnovna slučaja koja mogu nastupiti između tri čvora u mreži: **račvanje**, **lanac** i **sraz**. Pravila ćemo izvesti razmatrajući sljedeće tri faktorizacije:

$$\begin{array}{lll} (1) \quad p(x,y,z) = p(x|z)p(y|z)p(z) & x \leftarrow z \rightarrow y & \text{račvanje} \\ (2) \quad p(x,y,z) = p(x)p(z|x)p(y|z) & x \rightarrow z \rightarrow y & \text{lanac} \\ (3) \quad p(x,y,z) = p(x)p(y)p(z|x,y) & x \rightarrow z \leftarrow y & \text{sraz} \end{array}$$

Ideja je promatramo kada je par varijabli x i y zavisan ili nezavisan, ovisno o tome je li varijabla u sredini, z, poznata (opažena) ili nije. Pogledajmo najprije **račvanje**. Iz uređajnog Markovljevog svojstva za varijablu y slijedi:

$$y \perp x|z \qquad \Leftrightarrow x \perp y|z$$

Iz ovoga sad možemo izvesti pravilo za račvanje na čvorove x i y od čvora z: ako je varijabla z **opažena**, onda su varijable x i y uvjetno nezavisne uz varijablu z, tj. čvorovi su **odvojeni**. Inače, ako varijabla z nije opažena, onda varijable x i y nisu nezavisne i njihovu su čvorovi povezani.

Pogledajmo sada **lanac**. Prema uređajnom Markovljevom svojstvu za varijablu y ponovo dobivamo:

$$y \perp x|z \qquad \Leftrightarrow x \perp y|z$$

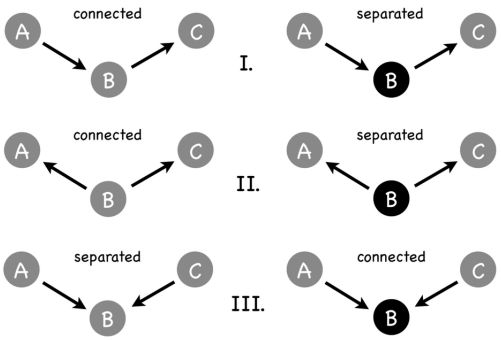
Iz toga izvodimo pravilo za lanac: ako čvorovi x i y čine lanac preko čvora z, i ako je varijabla z opažena, onda su varijable x i y uvjetno nezavisne uz varijablu z, tj. čvorovi su **odvojena**. Inače, ako varijabla z nije opažena, varijable x i y nisu nezavisne i njihovi su čvorovi povezani.

Konačno, pogledajmo **sraz**. Uređajno Markovljevo svojstvo za varijablu y daje nam:

$$y \perp x|\varnothing$$

tj. varijable x i y su marginalno nezavisne (primijetimo da ti čvorovi nemaju roditelja, a da je čvor x topološki prethodnik čvora y). Dakle, za razliku od prethodna dva slučaja (račvanje i lanac), u ovom trećem slučaju (sraz) nismo dobili da su varijable x i y uvjetno nezavisne uz z. Umjesto toga, dobili smo da su varijable marginalno nezavisne. Dakle, slučaj sraza je upravo suprotan slučaju račvanja i slučaju lanca: opažanje srednje varijable u lancu čini varijable na krajevima lanca zavisnima, dok su inače nezavisne. To nas onda dovodi do našeg trećeg pravila: ako su čvorovi x i y u srazu preko čvora z, i ako je varijabla z **neopažena**, onda su varijable x i y nezavisne i njihovi su čvorovi **odvojeni**. Inače, ako je varijabla z opažena, varijable x i y su zavisne i njihovi su čvorovi povezani. 10

Sažeto, pravila su dakle ova:



Kako god, moramo priznati da je pravilo za sraz malo čudno. Pogledajmo to malo detaljnije.

4.2 Efekt objašnjavanja

O srazu je korisno razmišljati kao o situaciji u kojoj se dvije varijable **natječu** za objašnjavanje (odnosno uzrokovanje) treće varijable. Budući da i varijabla x i varijabla y svaka samostalno može objasniti/uzrokovati varijably z, ono što se događa jest to da, ako znamo da se ostvarila z, onda se naše vjerovanje o tome je li se ostvario x mijenja ovisno o tome je li se ostvario y ili nije. Formalno:

$$x \not\perp y|z \quad \Leftrightarrow \quad p(x|z) \neq p(x|y,z)$$

Ako opazimo da su se ostvarili i y i z, to će smanjiti vjerojatnost da se ostvario x u odnosu na situaciju kada se ostvario samo z. (Također, vrijedi i obrnuto: ako opazimo da su se ostvarili i x i z, to će smanjiti vjerojatnost da se ostvario y u odnosu na situaciju kada se ostvario samo z). Ovaj se fenomen naziva **efekt objašnjavanja** (engl. *explaining away*). Efekt je zapravo vrlo intuitivan, kao što će pokazati sljedeći primjer.

▸ **PRIMJER**

Bacamo dva novčića. Neka glava odgovara vrijednosti 1, a pismo vrijednosti 0. Opažamo zbroj tih novčića. Zbroj može biti 0, 1 ili 2. Neka su novčići x i y, a njihov zbroj je z. Bayesova mreža odgovara strukturi sraza:

$$x \rightarrow z \leftarrow y$$

Očito, oba novčića utječu na zbroj. Apriorno, novčići su nezavisni. No, čim vidimo zbroj, oni postaju zavisni: npr., ako je zbroj 1, a prvi novčić je 0, onda drugi mora biti 1. Dakle, vrijedi $x \not\perp y|z$, i to je efekt objašnjavanja (opažanje zbroj novčića i jednog od dva novčića objašnjava vrijednost drugog novčića).

Efekt objašnjavanja samo je jedan primjer tzv. **međukauzalnog zaključivanja**, koje ljudi zapravo vrlo često (nesvjesno) koriste. Ipak, u nekim situacijama efekt se manifestira na neintuitivan način, i u takvim je situacijama poznat pod nazivom **Berksonov paradoks**. Sljedeći primjer ilustrira takvu situaciju. [11] [12]

▸ **PRIMJER**

Ekstreman primjer, koji se često navodi u literaturi, je sljedeći. Zamislimo da fakultet prima studente koji su visokointeligentni ili sportaši (ili oboje!). Neka P označava da je netko primljen na fakultet, što će biti istina ako je netko visokointeligentan (V) ili sportaš (S). Pretpostavimo da su u općenitoj populaciji V i S nezavisni. Odnose možemo modelirati strukturom sraza:

$$V \rightarrow P \leftarrow S$$

U općenitoj populaciji $V \perp S$. Međutim, ako pogledamo samo studente koji su primljeni na fakultet (dakle one za koje opažamo $P = 1$), naći ćemo da su visokointeligentni studenti manje vjerojatno sportaši i obrnuto:

$$P(S = 1|P = 1, V = 1) \leqslant P(S = 1|P = 1)$$

To je zato što je svako svojstvo od ova dva samo po sebi dovoljno da objasni/uzrokuje P. Zbog toga ćemo u studentskoj populaciji imati negativnu koreliranost visokointeligentnih i sportaša.

4.3 D-odvajanje čvorova

Sada kada smo se upoznali sa sva tri slučaja koja se mogu dogoditi na stazi od tri čvora, možemo to poopćiti na slučaj dvaju čvorova koja se nalaze bilo gdje u grafu. To nas dovodi do ideje **d-odvajanja čvorova**.

▸ **D-odvajanje čvorova**

Raspolažemo skupom varijabli E koje su opažene. Za **stazu** P od čvora x do čvora y kažemo da je **d-odvojena** (engl. *d-separated*) ako i samo ako vrijedi barem jedno od sljedećeg:

(1) P sadrži **lanac** $x \to z \to y$ ili $x \leftarrow z \leftarrow y$ i $z \in E$

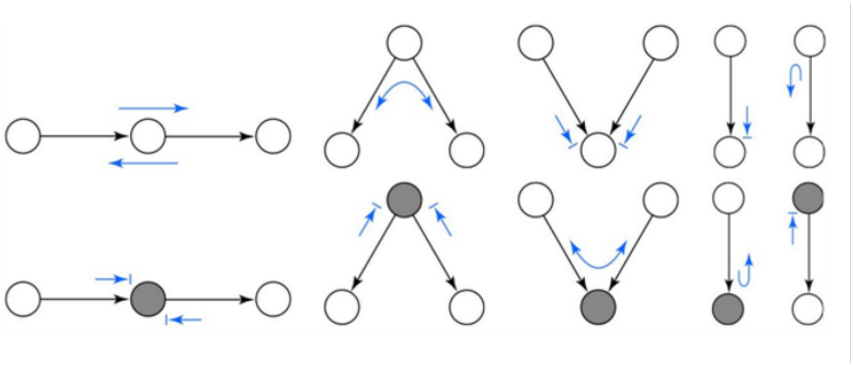
(2) P sadrži **račvanje** $x \leftarrow z \to y$ i $z \in E$

(3) P sadrži **sraz** $x \to z \leftarrow y$ i varijabla z nije u E i niti jedan sljedbenik od z nije u E

Za **par čvorova** x i y kažemo da su čvorovi x i y **d-odvojeni** za dani E ako su sve staze između ta dva čvora d-odvojene za dani E. Čvorovi x i y su uvjetno nezavisni za dani E ako i samo ako su d-odvojeni za dani E.

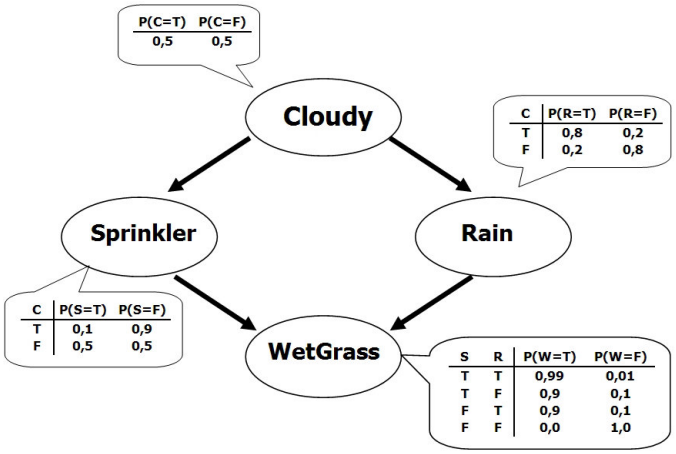
Dakle, da bismo utvrdili jesu li, za neko zadano opažanje varijabli, dva čvora uvjetno nezavisna, trebamo analizirati jesu li sve staze između njih d-odvojene. Ako jesu, onda su čvorovi uvjetno nezavisni, inače to nisu.

Ispitivanje d-odvojenosti lako se može pretočiti u algoritam. Naivan algoritam bio bi da ispitamo sve moguće staze u mreži između dva čvora. Međutim, za velike mreže to bi moglo biti skupo, pa su razvijeni učinkovitji algoritmi čija složenost linearno ovisi o broju čvorova. Jedan takav algoritam je **Bayesova kuglica** (engl. *Bayes-Ball*). Algoritam funkcionira tako da 13
"aktiviramo" sve čvorove čije varijable opažamo stavimo kuglice u jedan skup varijabli, i onda ih pustimo da se kreću po mreži (neovisno o usmjerenju bridova) prema pravilima koja odgovaraju d-odvajanju. Ukupno postoji 10 pravila:



Pravila definiraju kroz koje čvorove kuglica može proći (plave strelice), a kroz koje ne može (plave strelice s vertikalnom crtom), u ovisnosi o tome koje su varijable opažene (sivi čvorovi odgovaraju opaženim varijablama). Ako ijedna kuglica dosegne drugi skup čvorova, to onda znači da odgovarajuće varijable nisu d-separirane, tj. da nisu uvjetno nezavisne. Nećemo ići u detalje tog algoritma. Umjesto toga, pogledajmo opet primjer sa travom i prskalicom, i pokažimo na toj mreži kako funkcionira d-odvajanje.

▸ **PRIMJER**



Analizirajmo kada će varijabla "cloudy" biti uvjetno nezavisna od varijable "wet grass". Između te dvije varijable, odnosno između njihovih čvorova, imamo dva lanca. Prema pravilu d-odvajanja, oba će ova lanca biti odvojena ako opažamo i varijablu "sprinkler" i varijably "rain". Dakle, varijable "cloudy" i "wet grass" su uvjetno nezavisne ako opažamo varijable "sprinkler" i "rain" (obje!). Intuitivno, to znači da možemo zaključiti da je trava mokra a da pritom ne znamo je li oblačno samo ako znamo radi li prskalica za travu i je li pada kiša. Drugačije rečeno, prskalica i kiša su mogući uzroci (objašnjenja) za mokru travu, i ako opažamo i jedno i drugo, onda imamo dovoljnu informaciju da zaključimo i njihovoj kauzalnoj posljedici (mokroj travi).

Pogledajmo sada jesu li varijable "sprinkler" i "rain" nezavisne? Ove su varijable povezane dvjema stazama: jednim račvanjem (preko varijable "cloudy") i jednim srazom (preko varijable "wet grass"). Sukladno pravilima d-odvajanja, ako opažamo "cloudy", onda je ta staza odvojena. Suprotno, ako *ne* opažamo varijablu "wet grass", onda je i ta staza odvojena. Dakle, variable "sprinkler" i "rain" uvjetno su nezavisne ako opažamo varijably "cloudy", a ne opažamo varijablu "wet grass". S druge strane, ako opažamo varijablu "wet grass", varijable "sprinkler" i "rain" postaju uvjetno nezavisne. To je efekt objašnjavanja: premda su varijable "sprinkler" i "rain" nezavisne (za opažanu varijablu "cloudy"), ako opažamo njihov zajednički čvor-dijete, "wet grass", onda te varijable postaju uvjetno zavisne. Npr. , ako je trava mokra a znamo da pada kiša, onda se smjanjuje vjerojatnost da je trava mokra zbog prskalice.

Sažetak

- **Probabilistički grafički modeli (PGM-ovi)** na sažet način prikazuju **zajedničku distribuciju** kao graf u kojem su čvorovi varijable a prijelazi **zavisnosti** između njih
- Tri aspekta PGM-a: **reprezentacija**, **zaključivanje** i **učenje**
- **Bayesove mreže** su PGM-ovi s usmjerenim acikličkim grafovima, a **Markovljeve mreže** s neusmjerenim grafovima
- Bayesove mreže kodiraju uvjetne nezavisnosti između varijabli preko **uređajnog Makovljevog svojstva**
- Pomoću pravila **d-odvajanja** možemo odrediti je li par varijabli u mreži (uvjetno) nezavisan

Bilješke

[1] Glavna i doista sveobuhvatna referenca za probabilističke grafičke modele je [Koller and Friedman, 2009], koju toplo preporučam da razmotrite kao jedan cjeloljetni projekt. Izvrsna referenca je i [Darwiche, 2009], koja je mnogo sažetija. Kraće preglede Bayesovih mreža i PGM-ova, po kojima

je uglavnom i strukturirano ovo predavanje, možete naći u [Alpaydin, 2020] (poglavlje 14), [Bishop, 2006] (poglavlja 8.1–2, 11.2–3 i 13.1) te [Murphy, 2012] (poglavlje 10).

2 Za PGM-ove se često kaže da predstavljaju (sretan) brak teorije vjerojatnosti i teorije grafova. Osnovna ideja je **modularizacija**: složeni sustav gradi se kombinacijom jednostavnijih. Teorija vjerojatnosti služi kao "ljepilo" koje povezuje dijelove sustava i povezuje model s podatcima. S druge strane, teorija grafova nudi sučelje prema čovjeku, koje omogućava modeliranje na visokoj razini apstrakcije.

3 Ovdje jedna terminološka napomena o nazivu **Bayesove mreže** (na engleskom "Bayes networks", ali češće "Bayesian networks"), i alternativnim nazivima **mreže vjerovanja** (engl. *belief networks*) i **kauzalne mreže** (engl. *causal networks*). Uglavnom je prihvaćen stav da ovi nazivi nisu baš najsretniji. Kevina Murphy u Murphy [2012] vrlo jasno objašnjava zašto: *"A directed graphical model or DGM is a GM whose graph is a DAG. These are more commonly known as Bayesian networks. However, there is nothing inherently "Bayesian" about Bayesian networks: they are just a way of defining probability distributions. These models are also called belief networks. The term "belief" here refers to subjective probability. Once again, there is nothing inherently subjective about the kinds of probability distributions represented by DGMs. Finally, these models are sometimes called causal networks, because the directed arrows are sometimes interpreted as representing causal relations. However, there is nothing inherently causal about DGMs."* Murphy stoga predlaže korisiti naziv **directed graphical model (DGM)**. Mi ćemo ipak koristiti naziv **Bayesove mreže**, primjećujući, usput, da time radimo manju pogrešku nego da koristimo naziv **Bayesovske mreže**, budući da Bayesove mreže ne impliciraju da koristimo bayesovsku statistiku (koju i nećemo koristiti).

4 Bayesove mreže mogu se koristiti za modeliranje **kauzalnosti**, no one same po sebi, bez dodatnih pretpostavki i mehanizama, nisu nužno kauzalni model. Ako vas zanima veza između Bayesovih mreža i kauzalnosti, preporučam pročitati radove izraelsko-američkoga znanstvenika i filozofa Judea Pearla, koji je najpriznatiji stručnjak u području kauzalnog zaključivanja. O vezi između Bayesovih mreža i kauzalnih mreža možete pročitati u [Pearl, 1995], a o kauzalnim modelima općenito u [Pearl, 2000]. Pristupačan uvod u Pearlov rad je [Pearl and Mackenzie, 2018]. Više možete naći na njegovoj web-stranici: `http://bayes.cs.ucla.edu/jp_home.html`. Kauzalnost je važan koncept u strojnom učenju, pogotovo u kontekstu izgradnje pravednih (nepristranih) i tumačivih (interpretabilnih) modela. O kauzalnosti u kontekstu strojnog učenja možete pročitati u [Schölkopf, 2019].

5 Model "alarm network" opisan je u [Beinlich et al., 1989].

6 Model za prepoznavanje višerječnih jedinica u tekstovima na hrvatskome jeziku opisan je u [Buljan and Šnajder, 2017].

7 Clippy je bio razvijen u okviru Microsoftovog projekta Lumière, čiji je fokus bio primjena Bayesovog zaključivanja za automatiziranu pomoć korisnicima. Voditelj tima bio je američki znanstvenik Eric Horvitz, danas vrlo poznat u AI krugovima, posebno po svojim radovima u području zaključivanja uz ograničene resurse. Unatoč početnom entuzijazmu, deset godina nakon uvođenja u Office 97, Clippy je u tišini eliminiran, jer je uglavnom živcirao korisnike. Razlozi za to su razni, uključivo i loša predikcija cilja korisnika. Za to krivnju ne treba svaljivati na Bayesove mreže, već na prejjednostavan model. Naime, neke od očitih nedostataka modela bile su što model nije u obzir uzimao profil korisnika (npr. stupanj stručnosti, kao i činjenicu da korisnik kroz vrijeme poboljšava svoju vještinu korištenja Officea) te je akcije modelirao kao atomičke cjeline, a ne kao složenije nizove atomičkih akcija. Službeno Microsoftovo objašnjenje bilo je, vrlo mudro, da je "Office XP toliko jednostavan za korištenje da Clippy više nije niti potreban niti koristan" (`https://tinyurl.com/yxqbnqw2`). A ovdje je detaljna analiza zašto korisnici nisu voljeli Clippyja: `http://xenon.stanford.edu/~lswartz/paperclip/`.

8 Modeliranje kauzalnih struktura kompleksnih sustava (bioloških, društvenih, psiholoških, tehničkih) osnovna je motivacija za razvoj Bayesovih mreža, odnosno PGM-ova općenito. Kod takvih je modela ključno imati na raspolaganju metode pomoću kojih možemo izvesti "statističke posljedice" zadane kauzalne strukture. Znanstvenici su se ovim metodama bavili otpočekta razvoja umjetne inteligencije; važi su u tom kontekstu radovi Herberta Simona i Hyberta Blalocka s kraja šezdesetih godina. Međutim, problem je adekvatno riješen tek uvođenjem koncepta **d-odvajanja** sredinom osamdesetih

godina, u radovima Judea Pearla, Dana Geigera i Thomasa Verma sa Sveučilišta u Kaliforniji. Ovo je njihov izvorni rad: [Geiger et al., 1990].

9 Već smo rekli da Bayesove mreže općenito ne moraju nužno modelirati stvarnu kauzalnost. Je li neki odnos između dvaju fenomenta, kojima odgovaraju dvije slučajne varijable, doista kauzalan je svojstvo koja ne proizlazi iz same Bayesove mreže već dolazi izvana, već je pretpostavka koju unosimo mi sami, kod analize mreže, a na temelju našeg postojećeg znanja ili pretpostavki o kauzalnim odnosima u stvarnome svijetu. Da je Bayesovoj mreži doista svejedno kako modeliramo odnose između varijabli postaje jasno kada shvatimo da je su sljedeće dvije mreže ekvivalentne:

$$x \to y \to z$$
$$x \leftarrow y \leftarrow z$$

Naime, obje mreže modeliraju zajedničku vjerojatnost $p(x, y, z)$ i obje mreže imaju isti broj parametara. Vidimo da je smjer bridova nevažan, međutim kod modeliranja kauzalnosti smjer je, naravno, bitan. To znači da, ako pokažemo da su dvije varijable duž neke staze u Bayesovoj mreži zavisne, jesu li te varijable kauzalno povezane i koji je smjer kauzalnosti možemo reći samo ako se oslonimo na dodatno znanje ili pretpostavke.

10 Tri **pravila d-odvajanja** izveli smo primjenom uređajnog Markovljeg svojstva. Ista smo pravila, međutim, mogli izvesti i primjenom pravila umnožaka i pravila zbroja. Pokažimo to. Krenimo od pravila za **račvanje**. Podijelimo lijevu i desnu stranu izraza za zajedničku vjerojatnost za $p(z)$:

$$\frac{p(x, y, z)}{p(z)} = \frac{p(x|z)p(y|z)p(z)}{p(z)}$$
$$p(x, y|z) = p(x|z)p(y|z)$$

Iz ovoga, po definiciji uvjetne nezavisnosti, slije $x \perp y|z$. Primijetite međutim da ne vrijedi marginalna nezavisnost, tj. ne vrijedi $x \perp y$. U to se možemo uvjeriti marginalizacijom lijeve i desne strane po varijabli z (tj. primjenom pravila zbroja):

$$\sum_z p(x, y, z) = \sum_z p(x|z)p(y|z)p(z)$$
$$p(x, y) = \sum_z p(x|z)p(y|z)p(z)$$

Desna strana se općenito ne faktorizira u umnožak $p(x)p(y)$, pa zaključujemo $x \not\perp y$. Slično možemo izvesti i za druga dva slučaja. Za slučaj lanca, prema uređajnom Markovljevom svojstvu dobivamo:

$$y \perp x|z \qquad \Leftrightarrow x \perp y|z$$

I opet smo to mogli izvesti tako da lijevu i desnu stranu podijelimo sa $p(z)$:

$$p(x, y, z) = p(x)p(z|x)p(y|z)$$
$$\frac{p(x, y, z)}{p(z)} = \frac{\overbrace{p(x)p(z|x)}^{p(x,z)}\, p(y|z)}{p(z)}$$
$$p(x, y|z) = p(x|z)p(y|z)$$

Slično kao i u prvom slučaju, maginalizacijom po z nećemo moći izvesti faktorizaciju $p(x)p(y)$, iz čega zaključujemo da varijable x i y nisu marginalno nezavisne, $x \not\perp y$. Konačno u trećem slučaju, uređajno Markovljevo svojstvo daje nam:

$$y \perp x|\varnothing$$

tj. x i y su marginalno nezavisni (primijetimo da oni nemaju roditelja, a da je x prethodnik od y). Isto smo mogli dobiti marginalizacijom po varijabli z:

$$\sum_z p(x, y, z) = \sum_z p(x)p(y)p(z|x, y)$$
$$p(x, y) = p(x)p(y)\sum_z p(z|x, y) = p(x)p(y)$$

S druge strane, ako uvjetujemo na varijablu z:

$$\frac{p(x,y,z)}{p(z)} = \frac{p(x)p(y)p(z|x,y)}{p(z)}$$

Ovo se dalje ne faktorizira na $p(x|z)p(y|z)$, pa zaključujemo da varijable x i y nisu uvjetno nezavisne za zadani z, tj. $x \not\perp y|z$.

11 Razmatranje **efekta objašnjavanja** s aspekta **međukauzalnog zaključivanja** možete naći u [Wellman and Henrion, 1993]. No, evo možda uvjerljivijeg primjera koji ilustrira da u svakodnevnom zaključivanju često koristimo efekt objašnjavanja. Ako imate visoku temperaturu, i bojite se da možda imate COVID-19, bit će vam jako drago kada vam liječnik kaže da imate običnu upalu grla. Očito, to što imate upalu grla ne sprječava da imate i COVID-19. Međutim, to što imate upalu grla dobro objašnjava vaše simptome, i time značajno smanjuje vjerojatnost da imate COVID-19. Svejedno, često perite ruke.

12 Prema američkom fizičaru i statističaru Josephu Barksonu, koji je taj fenomen prvi primijetio. Berkson je također uveo naziv **logit** (logaritam omjera šansi), koji smo bili spomenuli prošli put.

13 Algoritam **Bayesove kuglice** osmislio je 1990. godine Ross D. Shachter sa Sveučilišta Stanford [Shachter, 1998]. Za alternativne algoritme, također temeljene na d-odvajanju, pogledajte [Koller and Friedman, 2009] (poglavlje 3.3.3) i [Darwiche, 2009] (poglavlje 4.5).

Literatura

E. Alpaydin. *Introduction to machine learning.* MIT press, 2020.

I. A. Beinlich, H. J. Suermondt, R. M. Chavez, and G. F. Cooper. The alarm monitoring system: A case study with two probabilistic inference techniques for belief networks. In *AIME 89*, pages 247–256. Springer, 1989.

C. M. Bishop. *Pattern recognition and machine learning.* Springer, 2006.

M. Buljan and J. Šnajder. Combining linguistic features for the detection of croatian multiword expressions. In *Proceedings of the 13th Workshop on Multiword Expressions (MWE 2017)*, pages 194–199, 2017.

A. Darwiche. *Modeling and reasoning with Bayesian networks.* Cambridge university press, 2009.

D. Geiger, T. Verma, and J. Pearl. d-separation: From theorems to algorithms. In *Machine Intelligence and Pattern Recognition*, volume 10, pages 139–148. Elsevier, 1990.

D. Koller and N. Friedman. *Probabilistic graphical models: principles and techniques.* MIT press, 2009.

K. P. Murphy. *Machine learning: a probabilistic perspective.* MIT press, 2012.

J. Pearl. From bayesian networks to causal networks. In *Mathematical models for handling partial knowledge in artificial intelligence*, pages 157–182. Springer, 1995.

J. Pearl. *Causality: Models, reasoning and inference.* 2000.

J. Pearl and D. Mackenzie. *The book of why: the new science of cause and effect.* Basic Books, 2018.

B. Schölkopf. Causality for machine learning. *arXiv preprint arXiv:1911.10500*, 2019.

R. Shachter. Bayes-ball: The rational pastime for determining irrelevance and requisite information in belief networks and influence diagrams. In *Uncertainty in Artificial Intelligence*, 1998.

M. P. Wellman and M. Henrion. Explaining 'explaining away'. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 15(3):287–292, 1993.