Triangulované grafy

Definition (triangulovaný graf)

Graf je **triangulovaný**, pokud pokud pro něj existuje perfektní eliminační posloupnost.

Lemma (Alternativní definice triang. grafu)

Graf je triangulovaný, pokud každý jeho cyklus délky větší než tři má aspoň jednu tětivu.

Definition (Strom spojení)

Mějme množinu klik neorientovaného grafu G, kliky jsou organizovány do stromu T. T je **strom spojení**, pokud pro každé dva vrcholy $V, W \in T$ všechny uzly na cestě z V do W obsahují průnik $W \cap V$. Průnik dvou sousedních uzlů nazveme **separátor** těchto uzlů, separátorem V a W je $S_{V,W} = V \cap W$.

Theorem

Pokud kliky grafu G lze organizovat do stromu spojení, pak je G triangulovaný.

Theorem

Pokud je G triangulovaný, pak kliky grafu G lze organizovat do stromu spojení.

Důkaz indukcí, pro grafy s jedním vrcholem platí.

- Eliminuji simpliciální uzel X, jeho rodina F_X je klika (označíme jí C).
- Pro vzniklý graf $G' = G \setminus \{X\}$ najdu strom spojení T' dle indukčního předpokladu.
- Pokud je $C \setminus \{X\}$ klika G', k uzlu odpovídajícímu této klice v T' přidám 'popisku' X a mám strom spojení grafu G.
- Pokud $C \setminus \{X\}$ není klika G':
  - musí být částí kliky $C_?$ grafu G',
  - Ke stromu T' přidáme uzel C a připojíme separátorem $C \setminus \{X\}$ k uzlu kliky $C_?$. Vzniklý strom je strom spojení pro G.

Alternativní konstrukce

- Najdi kliky.
- Vytvoř graf, uzly=kliky, hrany váhy počtu veličin v průniku.
- Najdi kostru (spanning tree) nejvyšší váhy (Prim's or Kruskal's algorithm).

Tato kostra je strom spojení, protože

- Je-li proměnná X v j klikách, může být maximálně v $j-1$ separátorech stromu spojení.
- Číslo $j-1$ dosáhneme jen v případě, že všechny kliky obsahující X budou spojeny separátorem obsahujícím X.
- Proto má strom spojení nejvyšší možný součet velikostí separátorů přes všechny kostry grafu.

Vlastnost klouzavých průniků RIP

Definition (Vlastnost klouzavých průniků, Running Intersection Property RIP)

Řekneme, že posloupnost $C_1, \ldots, C_m$, $m \geq 1$, $C_i \subseteq V$ splňuje **vlastnost klouzavých průniků RIP**, pokud:

$$\forall (2 \leq i \leq m)\ \exists (1 \leq k < i)\ C_i \cap \left(\bigcup_{j<i} C_k \right) \subseteq C_k.$$

- Od RIP ke stromu spojení.
  - Každou kliku připoj k odpovídající C_k a hranu označ $C_i \cup C_k$.
- Přechod od stromu spojení k RIP.
  - Zvol uzel stromu spojení.
  - Pošli z něj 'zprávu' do všech uzlů atd. až do listů.
  - RIP posloupnost tvoř tak, že vždy jde dřív uzel, který dostal zprávu dřív. Separátor ve stromu spojení určuje jednu z možných C_k.

Strom spojení

Používám termín **strom spojení** ve třech významech:

- viz definice výše, strom klik splňující vlastnost průniků
- strom dle definice výše, kde jsou navíc hrany označeny separátory
- strom dle definice výše, kde je navíc v každé klice "schránka" na seznam pravděpodobnostních tabulek a v každém separátoru jsou dvě schránky na zprávy – tabulky – jdoucí jednotlivými směry. Tomuto se říká **junction tree**.

Strom spojení reprezentující bayesovskou síť

Definition (Strom spojení reprezentující bayesovskou síť)

- Mějme bayesovskou síť s množinou pravděpodobnostních tabulek Φ a evidenci e. Nechť množina tabulek Φ_e vznikne z Φ vložením evidence e do příslušných tabulek, tj. "vyříznutím"konkrétních 'řádků' v pravděpodobnostních tabulkách.
- **Strom spojení reprezentuje bayesovskou síť s evidencí** e, pokud každou tabulku $\phi \in \Phi_e$ přiřadíme do schránky některé z klik C_i takových, že $dom(\phi) \subseteq C_i$.
- Pokud některý uzel stromu spojení nemá žádnou tabulku, přiřadíme tabulku dávající identicky 1 na doméně dané kliky.

- Pozn: pokud strom spojení vznikl z moralizovaného a triangularizovaného grafu bayesovské sítě, tak takové klika vždy existuje.
- Pokud moralizovaný graf není triangulovaný, doplníme ho hranami na triangulovaný a z něj vytvoříme strom spojení.

Propagace ve stromu spojení

- Propagace (výpočet) ve stromu spojení spočívá v posílání zpráv, kterými se postupně plní schránky separátorů.
- Každý uzel (klika) posílá v každém směru právě jednu zprávu.
- Uzel (klika) může poslat zprávu v daném směru, pokud už ze všech ostatních směrů zprávy dostala.
- Protože se jedná o strom, vždycky někdo může poslat zprávu, nebo jsou již všechny schránky plné.

Poslání zprávy

Uvažujme kliku C se sousedními separátory $S_1, \ldots, S_k$, směr separátoru S_1 (bez újmy na obecnosti). **Poslat zprávu** z C do S_1 znamená zapsat do odchozí schránky S_1 tabulku, která vznikne součinem příchozích zpráv v separátorech $S_2, \ldots, S_k$ a tabulek obsažených v C. Tento součin marginalizujeme přes všechny veličiny $C \setminus S_1$ a výsledek zapíšeme do S_1.

Theorem

Nechť strom spojení reprezentuje bayesovskou síť a evidenci e, všechny schánky byly naplněny. Potom:

- *Nechť V je klika obsahující tabulky Φ_V a k ní směřující separátory $S_1, \ldots, S_k$ obsahují zprávy $\phi_1, \ldots, \phi_k$.*

$$P(V, e) = \Pi_{\phi \in \Phi_V} \cdot \Pi_{i=1}^{k} \phi_i$$

- *Nechť S je separátor se zprávami ϕ_1, ϕ_2.*

$$P(S, e) = \phi_1 \cdot \phi_2$$

Zprávy směřující do V odpovídají perfektní elim. posl., která má V na svém konci. Pro separátor, odchozí zpráva vznikla marginalizací z V, jen tam nebyla započtena zpráva přicházející z tohoto směru.

$$\begin{aligned} P(S_1, e) &= \sum_{V \setminus S_1} P(V, e) = \sum_{V \setminus S_1} (\Pi_{\phi \in \Phi_V} \phi \cdot \Pi_{i=1}^{k} \phi_i) \\ &= \sum_{V \setminus S_1} (\Pi_{\phi \in \Phi_V} \cdot \Pi_{i=2}^{k} \phi_i \cdot \phi_1) = (\sum_{V \setminus S_1} \Pi_{\phi \in \Phi_V} \cdot \Pi_{i=2}^{k} \phi_i) \cdot \phi_1 \end{aligned}$$

což je odchozí krát příchozí zpráva. Poslední řádek plyne z toho, že $dom(\phi_1) = S$.

Výpočet pomocí stromu spojení (shrnutí)

- BN moralizujeme
- doplníme hrany na triangulovaný graf
- vytvoříme strom spojení
- naplníme tabulkami
- vypočteme posíláním zpráv
- pravděpodobnost na veličině A zjistíme tak, že najdeme libovolnou kliku C obsahující A a marginalizujeme, tj. $P(A, e) = \sum_{C \setminus A} P(C, e)$
- pokud nás zajímá sdružená distibuce na množině, která není částí žádné kliky, musíme použít Eliminaci proměnných.
  Nebo předem zajistit výskyt v jedné klice:
  m3=compile(grain(plist),root=c('lung','bronc','tub'), propagate=TRUE).

Úkol:

- Propagujte: chestdag=propagate(chestdag) a ověřte, zda tabulky chestdag$equipot obsahují marginály na klikách stromu spojení.

Úkoly

- Klient má positivní xray, nekouří: určete pravděpodobnost 'lung' pro: bez info o 'asia', pro byl/nebyl v 'asia'. (14.2%, 9.3%, 14.3%)
- Pro kuřáky ... 64.6% .

Přibližný výpočet bayesovské sítě

- Základní myšlenkou je vygenerovat data dle zadaných podmíněných pravděpodobností a z nich spočítat pravděpodobnosti, které nás zajímají.
- Přesnost výpočtu samozřejmě závisí na počtu vygenerovaných vzorků.
- Metody generující náhodné vzorky se nazývají metody **Monte Carlo**.
- Základem je generátor náhodného výsledku podle zadané pravděpodobnosti, např. $\langle \frac{1}{4}, \frac{1}{2}, \frac{1}{4} \rangle$.

Přímé vzorkování bez evidence

- Uspořádáme vrcholy BN tak, aby každá hrana začínala v uzlu menšího čísla než končí.
- Vytvoříme N vzorků, každý následovně
  - Pro první uzel A_1 vygenerujeme náhodně výsledek a_1 podle $P(A_1)$.
  - Pro druhý uzel A_2 vygenerujeme náhodně výsledek a_2 podle $P(A_2|A_1 = a_1)$ (je–li hrana, jinak nepodmíněně)
  - Pro n–tý uzel vygenerujeme výsledek podle $P(A_n|pa(A_n))$, na rodičích už známe konkrétní hodnoty.
- Z N vzorků spočteme pravděpodobnost jevu, který nás zajímá. Pro N jdoucí k nekonečnu podíl výskytu jevu konverguje k správné pravděpodobnosti.

Přímé vzorkování s evidencí e (rejection sampling)

- $N(e)$ značí počet vzorků konzistentních s evidencí e, tj. nabývající na příslušných veličinách správné hodnoty.
- Vzorky tvoříme úplně stejně, jako dříve, jen ty, co nejsou konzistentní s e vyšktneme, tj. $\hat{P}(X|e) = \frac{N(X,e)}{N(e)}$
- Problém je v tom, že je–li $P(e)$ malé, tak většinu vzorků zahazujeme.

Vážení věrohodností (Likelihood weighting)

- Generuje jen vzorky konzistentní s e.
- Váhy vzorků jsou různé, podle $P(e|vzorek)$ (což je věrohodnost $L(vzorek|e)$, odtud likelihood weighting).

Algoritmus **vytvoření váženého vzorku pro** (bn, e)

$w = 1$
v pořadí topologického uspořádání bn, for $i = 1$ to n
    if A_i má evidenci a_i v e
        $w = w \cdot P(A_i = a_i|pa(A_i))$
    else
        a_i vyber podle rozložení $P(A_i = a_i|pa(A_i))$
return $(w, \langle a_1, \ldots, a_n \rangle)$

Učení parametrů

- Pokud známe strukturu a všechny veličiny jsou pozorované, odhad parametrů je (skoro) podíl odpovídajících četností.
- 'skoro' se vztahuje na nulové počty a dělení nulou. Proto máme možnost nastavit vyhlazování smooth=0.0001 - přičte ke všem četnostem, tj. nikde nebude nula.
- více o učení příště.

Úkol

- V kódu experimentujte se simulací, 'kontrolou' pravděpodobností/četností.
- Specifikujte jinou strukturu modelu, naučte parametry ze simulovaných dat a porovnejte (podmíněné) pravděpodobnosti v původním a novém modelu.

```
 coins1 <- loadHuginNet("two_coins_1.net")
sim.orig=simulate(coins1,n=1000)
novy.dag<-dag(~TwiceAHead,~Penny:TwiceAHead,~Dime:TwiceAHead)
md=grain(novy.dag,data=sim.orig,smooth=0)
```