SID – Wykład XI
Sieci Bayesowskie

Dominik Ślęzak

Wydział Matematyki, Informatyki i Mechaniki UW
slezak@mimuw.edu.pl

Niepewność

Niech akcja A_t = wyjedź na lotnisko t minut przed odlotem.
Czy A_t pozwoli mi zdążyć na czas?

Problemy:

1. informacja częściowa (stan ulic, plany innych kierowców, etc.)
2. niedokładne informacje (raport o korkach)
3. niepewność działania akcji (złapanie gumy, etc.)
4. ogromna złożoność modelowania i przewidywania ruchu

Stąd czysto logiczne podejście albo
1) ryzykuje fałszywość: "A_{25} pozwoli mi zdążyć na czas"
albo 2) prowadzi do wniosków zbyt słabych do podjęcia decyzji:
"A_{25} pozwoli mi zdążyć na czas jeśli nie będzie wypadku na moście i nie będzie padać i nie złapię gumy itd."

Metody wnioskowania w niepewności

Reguły z czynnikiem ryzyka:

$A_{25} \longmapsto_{0.3}$ zdąży na czas
$\mathit{Zraszacz} \longmapsto_{0.99} \mathit{MokryTrawnik}$
$\mathit{MokryTrawnik} \longmapsto_{0.7} \mathit{Deszcz}$

Pytania: Problemy z kombinowaniem, np. czy *Zraszacz* powoduje *Deszcz*??

Prawdopodobieństwo

Dla dostępnych przesłanek

A_{25} zdąży na czas z prawdopodobieństwem 0.04

Mahaviracarya (IX w.), Cardamo (1565) teoria ryzyka

(Logika rozmyta zarządza stopniem prawdziwości NIE niepewnością, np. *MokryTrawnik* jest prawdą w stopniu 0.2)

Prawdopodobieństwo

Stwierdzenia prawdopodobne warto iterować w przypadku
- ograniczeń: niemożność wyliczenia wyjątków, warunków, etc.
- braku wiedzy: brak istotnych faktów, warunków początkowych, etc.

Prawdopodobieństwa mogą się odnosić do czyjegoś stanu wiedzy,

np. $P(A_{25}|\text{brak zgłoszonych wypadków}) = 0.06$

Takie prawdopodobieństwa można określać na podstawie doświadczenia, analizy podobnych sytuacji, danych statystycznych itd.

Prawdopodobieństwo zdarzenia zmienia się wraz z nową przesłanką:

np. $P(A_{25}|\text{brak zgłoszonych wypadków}, \text{5-ta rano}) = 0.15$

Podejmowanie decyzji w niepewności

Załóżmy, że wierzę w następujące zdania:

$$P(A_{25} \text{ pozwoli zdążyć na czas} | \ldots) = 0.04$$
$$P(A_{90} \text{ pozwoli zdążyć na czas} | \ldots) = 0.70$$
$$P(A_{120} \text{ pozwoli zdążyć na czas} | \ldots) = 0.95$$
$$P(A_{1440} \text{ pozwoli zdążyć na czas} | \ldots) = 0.9999$$

Którą akcję wybrać?

Zależy od moich preferencji co do spóźnienia, kuchni na lotnisku, itd.

Teoria użyteczności jest używana do reprezentacji i wnioskowania o tego rodzaju preferencjach

Teoria decyzji = teoria użyteczności + teoria prawdopodobieństwa

Podstawy prawdopodobieństwa

Zmienna losowa jest funkcją z przestrzeni próbek w pewien zbiór wartości,
np. rzeczywistych lub boolowskich
np. $Odd(1) = true$.

P indukuje rozkład prawdopodobieństwa dla dowolnej zmiennej losowej X:
Przypadek dyskretny:

$$P(X = x_i) = \sum_{\{\omega : X(\omega) = x_i\}} P(\omega)$$

np. $P(Odd = true) = 1/6 + 1/6 + 1/6 = 1/2$

Podstawy prawdopodobieństwa

Zdania reprezentują pewne zdarzenia (podzbiory przestrzeni próbek), w których są prawdziwe

Dla danych zmiennych boolowskich A i B:

zdarzenie a = zbiór punktów próbkowych gdzie $A(\omega) = \mathit{true}$
zdarzenie $\neg a$ = zbiór punktów próbkowych gdzie $A(\omega) = \mathit{false}$
zdarzenie $a \wedge b$ = zbiór punktów gdzie $A(\omega) = \mathit{true}$ i $B(\omega) = \mathit{true}$

Dla zmiennych boolowskich, punkt próbkowy = model rachunku zdań
np. $A = \mathit{true}$, $B = \mathit{false}$, lub $a \wedge \neg b$.

Zdanie = alternatywa zdarzeń atomowych, w których to zdanie jest prawdą
np. $(a \vee b) \equiv (\neg a \wedge b) \vee (a \wedge \neg b) \vee (a \wedge b)$
$\implies P(a \vee b) = P(\neg a \wedge b) + P(a \wedge \neg b) + P(a \wedge b)$

Składnia zdań

Boolowskie zmienne losowe

np. *Cavity* (czy jestem osłabiony?)

Dyskretne zmienne losowe (skończone lub nieskończone)

np. *Weather* ma jedną wartość z $\langle sunny, rain, cloudy, snow\rangle$

$Weather = rain$ jest zdaniem

Wartości muszą być kompletne i wzajemnie się wykluczać

Ciągłe zmienne losowe (ograniczone lub nieograniczone)

np. $Temp = 21.6$; można także $Temp < 22.0$.

Można stworzyć dowolne kombinacje boolowskie prostych zdań, np. ścieżki lub grupy ścieżek w drzewach decyzyjnych.

Prawdopodobieństwo bezwarunkowe

Bezwarunkowe prawdopodobieństwo zdań

np. $P(Cavity = true) = 0.1$ i $P(Weather = sunny) = 0.72$

odpowiada przekonaniom przed dostarczeniem jakiejkolwiek (nowej) przesłanki

Rozkład prawdopodobieństwa daje wartości dla wszystkich przypisań:

$\mathbf{P}(Weather) = \langle 0.72, 0.1, 0.08, 0.1 \rangle$ (znormalizowana: sumuje się do 1)

Łączny rozkład prawdopodobieństwa dla zbioru zm. los. daje prawdopodobieństwa każdego zdarzenia atomowego na tych zm. los. (tzn. każdy punkt próbkowy)

$\mathbf{P}(Weather, Cavity)$ = macierz wartości 4×2:

Weather =	sunny	rain	cloudy	snow
Cavity = true	0.144	0.02	0.016	0.02
Cavity = false	0.576	0.08	0.064	0.08

Prawdopodobieństwo warunkowe

Prawdopodobieństwo warunkowe (a posteriori)

np. $P(Cavity|Toothache) = 0.8$

tzn. zakładając, że *Toothache* to to, o czym wiem

Notacja rozkładów warunkowych:

$\mathbf{P}(Cavity|Toothache)$ = 2-elementowy wektor 2-elementowych wektorów (bo zarówno *Cavity* jak i *Toothache* mają dwie wartości)

Jeśli wiemy więcej, np. *Cavity* też jest dane, wtedy mamy

$P(cavity|toothache, cavity) = 1$

Nowe przesłanki mogą być nieistotne, umożliwiając upraszczanie, np.

$P(cavity|toothache, 49ersWin) = P(cavity|toothache) = 0.8$

Prawdopodobieństwo warunkowe

Definicja podstawowa:

$$P(a|b) = \frac{P(a \wedge b)}{P(b)} \text{ if } P(b) \neq 0$$

Sformułowanie alternatywne:

$$P(a \wedge b) = P(a|b)P(b) = p(b|a)P(a)$$

Prawdopodobieństwa warunkowe pojawiają się przy ocenie reguł i drzew decyzyjnych. Nie można ich jednak mylić z implikacjami.

Prawdopodobieństwo warunkowe

Ogólna wersja dla całych rozkładów, np.

$$\mathbf{P}(Weather, Cavity) = \mathbf{P}(Weather|Cavity)\mathbf{P}(Cavity)$$

(zbiór 4×2 równań, 4 wartości *Weather*, 2 wartości *Cavity*)

Reguła łańcuchowa otrzymywana przez kolejne zastosowania reguły produkcji:

$$\begin{aligned}
\mathbf{P}(X_1, \ldots, X_n) &= \mathbf{P}(X_1, \ldots, X_{n-1})\mathbf{P}(X_n|X_1, \ldots, X_{n-1}) \\
&= \mathbf{P}(X_1, \ldots, X_{n-2})\mathbf{P}(X_{n-1}|X_1, \ldots, X_{n-2})\mathbf{P}(X_n|X_1, \ldots, X_{n-1}) \\
&= \ldots \\
&= \prod_{i=1}^{n} \mathbf{P}(X_i|X_1, \ldots, X_{i-1})
\end{aligned}$$

Wnioskowanie przez wyliczanie

Zacznij od rozkładu łącznego

	tootache		$\neg$*tootache*	
	catch	$\neg$*catch*	*catch*	$\neg$*catch*
cavity	.108	.012	.072	.008
$\neg$*cavity*	.016	.064	.144	.576

Dla dowolnego zdania ϕ, sumuj zdarzenia atomowe, w których to zdanie jest prawdziwe:

$$P(\phi) = \sum_{\omega:\omega\models\phi} P(\omega)$$

$P(toothache) = 0.108 + 0.012 + 0.016 + 0.064 = 0.2$
$P(cavity \lor toothache) = 0.108 + 0.012 + 0.072 + 0.008 + 0.016 + 0.064 = 0.28$

Wnioskowanie przez wyliczanie

Zacznij od rozkładu łącznego

	tootache		$\neg$*tootache*	
	catch	$\neg$*catch*	*catch*	$\neg$*catch*
cavity	.108	.012	.072	.008
$\neg$*cavity*	.016	.064	.144	.576

Można również policzyć prawdopodobieństwa warunkowe:

$$P(\neg cavity | toothache) = \frac{P(\neg cavity \wedge toothache)}{P(toothache)}$$
$$= \frac{0.016 + 0.064}{0.108 + 0.012 + 0.016 + 0.064} = 0.4$$

Wnioskowanie przez wyliczanie

Zacznij od rozkładu łącznego

	tootache		$\neg$*tootache*	
	catch	$\neg$*catch*	*catch*	$\neg$*catch*
cavity	.108	.012	.072	.008
$\neg$*cavity*	.016	.064	.144	.576

Mianownik można również traktować jako stałą normalizacji α:

$$\begin{aligned}
\mathbf{P}(Cavity|toothache) &= \alpha\,\mathbf{P}(Cavity, toothache) \\
&= \alpha\,[\mathbf{P}(Cavity, toothache, catch) + \mathbf{P}(Cavity, toothache, \neg catch)] \\
&= \alpha\,[\langle 0.108, 0.016\rangle + \langle 0.012, 0.064\rangle] \\
&= \alpha\,\langle 0.12, 0.08\rangle = \langle 0.6, 0.4\rangle
\end{aligned}$$

Wnioskowanie przez wyliczanie

Zazwyczaj interesuje nas

rozkład warunkowy zadanych zmiennych **Y**
przy danych specyficznych wartościach **e** dla zmiennych-przesłanek **E**

Zmienne ukryte $\mathbf{H} = \mathbf{X} - \mathbf{Y} - \mathbf{E}$
Ogólny pomysł: ustalamy zmienne-przesłanki i sumujemy prawdopodobieństwa po wartościach zmiennych ukrytych:

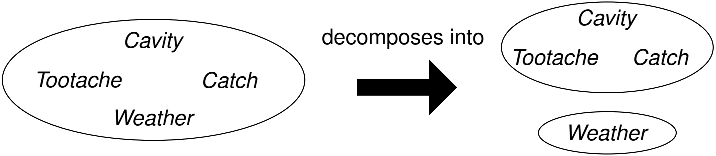
$$\mathbf{P}(\mathbf{Y}|\mathbf{E}=\mathbf{e}) = \alpha\mathbf{P}(\mathbf{Y},\mathbf{E}=\mathbf{e}) = \alpha\sum_{\mathbf{h}}\mathbf{P}(\mathbf{Y},\mathbf{E}=\mathbf{e},\mathbf{H}=\mathbf{h})$$

Wyrażenia w sumowania są wartościami łącznego rozkładu ponieważ **Y**, **E** i **H** razem wyczerpują cały zbiór zmiennych losowych.
Problemy:

1. Złożoność czasowa $O(d^n)$ gdzie d jest maks. liczbą wartości zmiennej
2. Złożoność pamięciowa $O(d^n)$, żeby pamiętać łączny rozkład
3. Jak zbudować słownik wartości prawdopodobieństw dla $O(d^n)$ punktów próbkowych???

Niezależność

A i B są niezależne $\iff$
$\mathbf{P}(A|B) = \mathbf{P}(A)$ lub $\mathbf{P}(B|A) = \mathbf{P}(B)$ lub $\mathbf{P}(A, B) = \mathbf{P}(A)\mathbf{P}(B)$



$\mathbf{P}(Toothache, Catch, Cavity, Weather) = \mathbf{P}(Toothache, Catch, Cavity)\mathbf{P}(Weather)$

32 wartości prawdopodobieństw zredukowane do 12; dla n niezależnych rzutów monetą $2^n \to n$

Pełna niezależność zmiennych jest bardzo efektywna, ale bardzo rzadka.

Niezależność warunkowa

$\mathbf{P}(Toothache, Cavity, Catch)$ wymaga $2^3 - 1 = 7$ niezależnych wartości

Jeśli mam osłabienie, prawdopodobieństwo, że złapię wtedy przeziębienie jest niezależne od tego, czy mam ból zęba:

$$(1)\ P(catch|toothache, cavity) = P(catch|cavity)$$

Ta sama niezależność pozostaje, jeśli nie mam osłabienia:

$$(2)\ P(catch|toothache, \neg cavity) = P(catch|\neg cavity)$$

Catch jest warunkowo niezależne od *Toothache* przy danym *Cavity*:

$$\mathbf{P}(Catch|Toothache, Cavity) = \mathbf{P}(Catch|Cavity)$$

Równoważne zdania:

$$\mathbf{P}(Toothache|Catch, Cavity) = \mathbf{P}(Toothache|Cavity)$$
$$\mathbf{P}(Toothache, Catch|Cavity) = \mathbf{P}(Toothache|Cavity)\mathbf{P}(Catch|Cavity)$$

Niezależność warunkowa

Używając pełnego łącznego rozkładu i reguły łańcuchowej:

$$\begin{aligned}
&\mathbf{P}(\mathit{Toothache}, \mathit{Catch}, \mathit{Cavity}) \\
&= \mathbf{P}(\mathit{Toothache}|\mathit{Catch}, \mathit{Cavity})\mathbf{P}(\mathit{Catch}, \mathit{Cavity}) \\
&= \mathbf{P}(\mathit{Toothache}|\mathit{Catch}, \mathit{Cavity})\mathbf{P}(\mathit{Catch}|\mathit{Cavity})\mathbf{P}(\mathit{Cavity}) \\
&= \mathbf{P}(\mathit{Toothache}|\mathit{Cavity})\mathbf{P}(\mathit{Catch}|\mathit{Cavity})\mathbf{P}(\mathit{Cavity})
\end{aligned}$$

Tzn. 2 + 2 + 1 = 5 niezależnych wartości (równania 1 i 2 usuwają 2)

W większości przypadków użycie prawdopodobieństwa warunkowego redukuje rozmiar reprezentacji łącznego rozkładu z wykładniczego od n do linowego od n.

Niezależność warunkowa jest najbardziej podstawową i efektywną formą wiedzy o niepewnym środowisku.

Reguła Bayesa

Reguła produkcyjna $P(a \wedge b) = P(a|b)P(b) = P(b|a)P(a)$

$$\Longrightarrow \text{ reguła Bayesa } P(a|b) = \frac{P(b|a)P(a)}{P(b)}$$

lub dla rozkładów

$$\mathbf{P}(Y|X) = \frac{\mathbf{P}(X|Y)\mathbf{P}(Y)}{\mathbf{P}(X)} = \alpha \mathbf{P}(X|Y)\mathbf{P}(Y)$$

Użyteczne przy szacowaniu prawdopodobieństwa <u>diagnostycznego</u> na podstawie prawdopodobieństwa <u>przyczynowego</u>:

$$P(Cause|Effect) = \frac{P(Effect|Cause)P(Cause)}{P(Effect)}$$

Np. M dolegliwość meningitis, S sztywnienie szyi:

$$P(m|s) = \frac{P(s|m)P(m)}{P(s)} = \frac{0.8 \times 0.0001}{0.1} = 0.0008$$

Reguła Bayesa i niezależność warunkowa

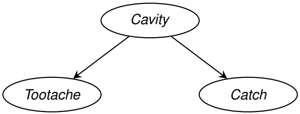
$\mathbf{P}(Cavity|toothache \wedge catch)$

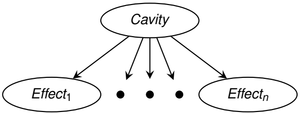
$= \alpha\,\mathbf{P}(toothache \wedge catch|Cavity)\mathbf{P}(Cavity)$

$= \alpha\,\mathbf{P}(toothache|Cavity)\mathbf{P}(catch|Cavity)\mathbf{P}(Cavity)$

Model wnioskowania naiwny Bayesowski (zakłada niezależność obserwacji):

$$\mathbf{P}(Cause, Effect_1, \ldots, Effect_n) = \mathbf{P}(Cause)\prod_i \mathbf{P}(Effect_i|Cause)$$





Całkowita liczba parametrów liniowa od n.

Sieci bayesowskie

Prosta, grafowa notacja do reprezentacji stwierdzeń o niezależności warunkowej i do zwartej specyfikacji pełnych rozkładów wielu zmiennych losowych.

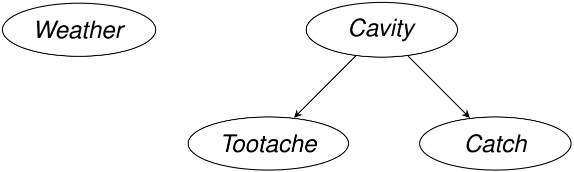
Składnia:

- zbiór węzłów, jeden dla każdej zmiennej losowej
- skierowany graf acykliczny (strzałka $\approx$ "bezpośrednio wpływa na")
- dla każdego węzła rozkład warunkowy na podstawie rodziców: $\mathbf{P}(X_i | Parents(X_i))$

W najprostszym przypadku rozkład warunkowy reprezentowany jest jako tablica prawdopodobieństwa warunkowego (TPW) dająca rozkład X_i dla każdej kombinacji wartości rodziców.

Sieci bayesowskie: przykład

Topologia sieci koduje stwierdzenie o warunkowej niezależności:



Weather jest niezależna od innych zmiennych.

Toothache i *Catch* są warunkowo niezależne przy danym *Cavity*.

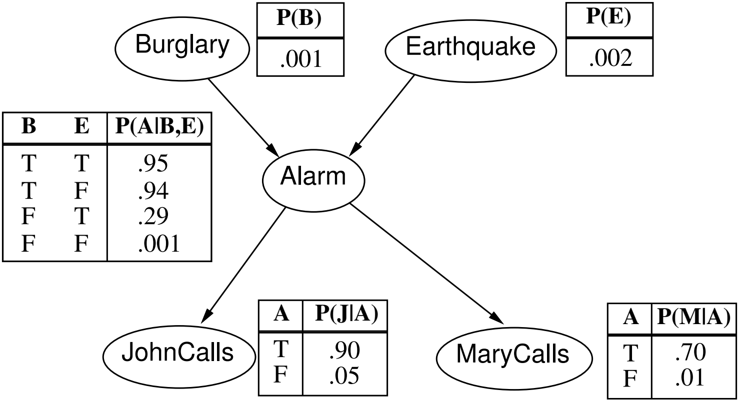
Sieci bayesowskie: przykład

Jestem w pracy, sąsiad John dzwoni do mnie, mówiąc mi, że mój alarm domowy się włączył, ale sąsiadka Mary nie dzwoni. Czasami alarm włącza się przy drobnych trzęsieniach ziemi. Czy to jest włamanie?

Zmienne: *Burglar*, *Earthquake*, *Alarm*, *JohnCalls*, *MaryCalls*
Topologia sieci odzwierciedla wiedzę "przyczynowo-skutkową":

- Włamywacz może uruchomić alarm
- Trzęsienie ziemi może uruchomić alarm
- Uruchomiony alarm może spowodować, że Mary zadzwoni
- Uruchomiony alarm może spowodować, że John zadzwoni

Sieci bayesowskie: przykład



P(B)
.001

P(E)
.002

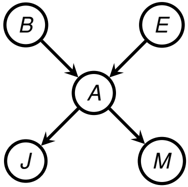
B	E	P(A|B,E)
T	T	.95
T	F	.94
F	T	.29
F	F	.001

A	P(J|A)
T	.90
F	.05

A	P(M|A)
T	.70
F	.01

Zwartość reprezentacji sieci

TPW dla boolowskiej zmiennej
X_i z k boolowskimi zmiennymi-rodzicami ma 2^k wierszy
będących kombinacjami wartości zmiennych-rodziców.



Każdy wiersz
TPW wymaga jednej wartości prawd. p dla $X_i = true$
(prawdopodobieństwo dla $X_i = false$ jest $1 - p$).

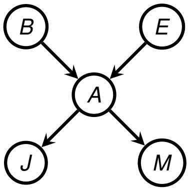
Jeśli każda zmienna ma co najwyżej k rodziców, to pełna
sieć wymaga $O(n \cdot 2^k)$ wartości prawdopodobieństw.

Tzn. rośnie liniowo z n, vs. $O(2^n)$ dla pełnego rozkładu łącznego.

Dla sieci z włamaniem, $1 + 1 + 4 + 2 + 2 = 10$ wartości prawdopodobieństw
(vs. $2^5 - 1 = 31$).

Globalna semantyka

Globalna semantyka definiuje pełny
rozkład łączny jako produkt lokalnych rozkładów warunkowych:

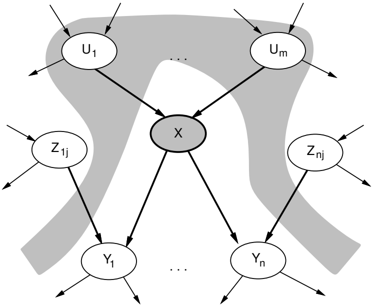


$$\mathbf{P}(X_1, \ldots, X_n) = \prod_{i=1}^{n} \mathbf{P}(X_i | Parents(X_i))$$

np. $P(j \wedge m \wedge a \wedge \neg b \wedge \neg e)$
$= P(j|a)P(m|a)P(a|\neg b, \neg e)P(\neg b)P(\neg e)$

Lokalna semantyka

Lokalna semantyka: każdy węzeł jest warunkowo niezależny przy danych rodzicach od pozostałych węzłów nie będących jego potomkami



Twierdzenie: Lokalna semantyka $\iff$ globalna semantyka

Konstruowanie sieci bayesowskiej

Wymaga metody takiej, że ciąg lokalnie testowalnych zależności warunkowych nadaje znaczenie globalne.

1. Wybierz uporządkowanie zmiennych los. $X_1, \ldots, X_n$
2. Dla każdego $i = 1$ do n
   - dodaj X_i do sieci
   - wybierz rodziców $X_1, \ldots, X_{i-1}$ takich, że $\mathbf{P}(X_i | Parents(X_i)) = \mathbf{P}(X_i | X_1, \ldots, X_{i-1})$

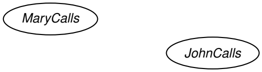
Wybór rodziców gwarantuje znaczenie globalne:

$$\mathbf{P}(X_1, \ldots, X_n) = \prod_{i=1}^{n} \mathbf{P}(X_i | X_1, \ldots, X_{i-1}) \qquad \text{(reguła łańcuchowa)}$$

$$= \prod_{i=1}^{n} \mathbf{P}(X_i | Parents(X_i)) \qquad \text{(przez konstrukcję)}$$

Konstruowanie sieci bayesowskiej: przykład

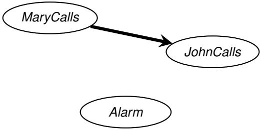
Załóżmy, że wybieramy M, J, A, B, E



$P(J|M) = P(J)$?

Konstruowanie sieci bayesowskiej: przykład

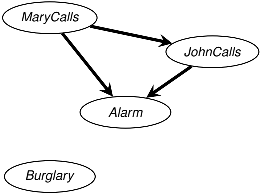
Załóżmy, że wybieramy M, J, A, B, E



$P(J|M) = P(J)$? Nie
$P(A|J, M) = P(A|J)$? $P(A|J, M) = P(A)$?

Konstruowanie sieci bayesowskiej: przykład

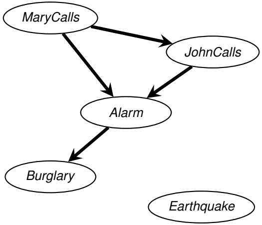
Załóżmy, że wybieramy M, J, A, B, E



$P(J|M) = P(J)$? Nie
$P(A|J,M) = P(A|J)$? $P(A|J,M) = P(A)$? Nie
$P(B|A,J,M) = P(B|A)$?
$P(B|A,J,M) = P(B)$?

Konstruowanie sieci bayesowskiej: przykład

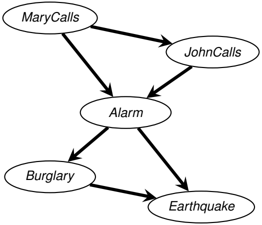
Załóżmy, że wybieramy M, J, A, B, E



$P(J|M) = P(J)$? Nie
$P(A|J,M) = P(A|J)$? $P(A|J,M) = P(A)$? Nie
$P(B|A,J,M) = P(B|A)$? Tak
$P(B|A,J,M) = P(B)$? Nie
$P(E|B,A,J,M) = P(E|A)$?
$P(E|B,A,J,M) = P(E|A,B)$?

Konstruowanie sieci bayesowskiej: przykład

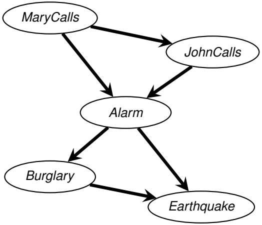
Załóżmy, że wybieramy M, J, A, B, E



$P(J|M) = P(J)$? Nie
$P(A|J,M) = P(A|J)$? $P(A|J,M) = P(A)$? Nie
$P(B|A,J,M) = P(B|A)$? Tak
$P(B|A,J,M) = P(B)$? Nie
$P(E|B,A,J,M) = P(E|A)$? Nie
$P(E|B,A,J,M) = P(E|A,B)$? Tak

Konstruowanie sieci bayesowskiej: przykład

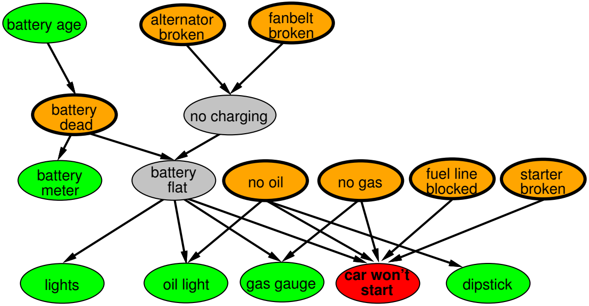
Załóżmy, że wybieramy *M*, *J*, *A*, *B*, *E*



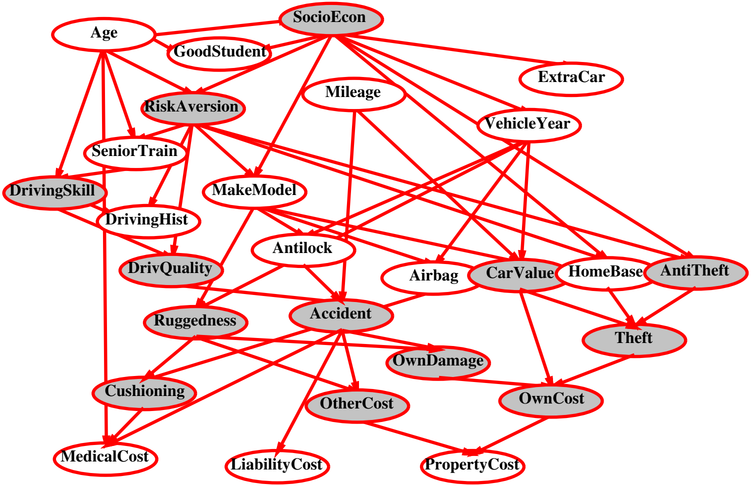
Rozpoznawanie warunkowych niezależności i oszacowanie prawdopodobieństw warunkowych jest trudne dla ludzi w nie przyczynowo-skutkowych kierunkach.
Sieć jest mniej zwarta: $1 + 2 + 4 + 2 + 4 = 13$ wartości prawdopodobieństw

Sieć bayesowska: diagnoza samochodu

Początkowa przesłanka: samochód nie zapala
Zmienne testowalne (zielone), zmienne “zepsute, napraw to” (pomarańczowe), zmienne ukryte (szare) rozrzedzają strukturę, redukują parametry



Siec bayesowska: ubezpieczenie samochodu



Zwarty rozkład warunkowy

TPW rośnie wykładniczo wraz z liczbą zmiennych-rodziców
TPW staje się nieskończona dla rodzica lub syna z wartościami ciągłymi

Rozwiązanie: kanoniczne rozkłady, które są zdefiniowane w zwarty sposób

Deterministyczne węzły są najprostszym przypadkiem:

$X = f(Parents(X))$ dla pewnej funkcji f

Np. funkcje boolowskie

$$NorthAmerican \iff Canadian \vee US \vee Mexican$$

Np. numeryczne powiązania pomiędzy zmiennymi ciągłymi

$$\frac{\partial Level}{\partial t} = \text{inflow + precipitation - outflow - evaporation}$$

Zwarty rozkład warunkowy

Rozkłady noisy-OR modelują wiele niezależnych przyczyn

- Rodzice $U_1 \ldots U_k$ obejmują wszystkie możliwe przyczyny
- Niezależne prawdopodobieństwo porażki q_i dla każdej przyczyny

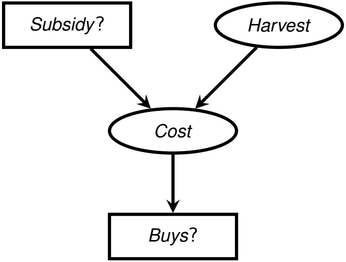
$$\Longrightarrow \; P(X|U_1 \ldots U_j, \neg U_{j+1} \ldots \neg U_k) = 1 - \prod_{i=1}^{j} q_i$$

Cold	*Flu*	*Malaria*	$P(Fever)$	$P(\neg Fever)$
F	F	F	**0.0**	1.0
F	F	T	0.9	**0.1**
F	T	F	0.8	**0.2**
F	T	T	0.98	$0.02 = 0.2 \times 0.1$
T	F	F	0.4	**0.6**
T	F	T	0.94	$0.06 = 0.6 \times 0.1$
T	T	F	0.88	$0.12 = 0.6 \times 0.2$
T	T	T	0.988	$0.012 = 0.6 \times 0.2 \times 0.1$

Liczba parametrów liniowa od liczby rodziców.

Sieci hybrydowe (zmienne dyskretne+ciągłe)

Dyskretne (*Subsidy*? i *Buys*?); ciągłe (*Harvest* i *Cost*)



Opcja 1: dyskretyzacja zm. ciągłych — możliwe duże błędy, duże TPW
Opcja 2: skończenie parametryzowalne rodziny funkcji kanonicznych

1) Zmienne ciągłe, zmienne-rodzice dyskretne+ciągłe (np. *Cost*)
2) Zmienne dyskretne, zmienne-rodzice ciągłe (np. *Buys*?)

Zmienne-dzieci ciągłe

Wymaga jednej funkcji warunkowej gęstości dla zmiennej będącej dzieckiem przy ciągłych zmiennych-rodzicach, dla każdego możliwego przypisania na zmiennych-rodzicach dyskretnych

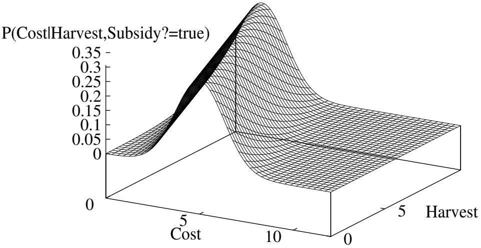
Najbardziej powszechny jest model gaussowski liniowy (LG), np.:

$$P(Cost = c|Harvest = h, Subsidy? = true)$$
$$= N(a_t h + b_t, \sigma_t)(c)$$
$$= \frac{1}{\sigma_t \sqrt{2\pi}} exp\left(-\frac{1}{2}\left(\frac{c-(a_t h + b_t)}{\sigma_t}\right)^2\right)$$

Średnia zmiennej *Cost* zmienia się liniowo w zależności od wartości *Harvest*, wariancja jest stała.

Liniowa zmienność jest nieodpowiednia dla pełnego zakresu wartości *Harvest* ale działa dobrze, jeśli prawdopodobny zakres tych wartości jest wąski.

Zmienne-dzieci ciagle



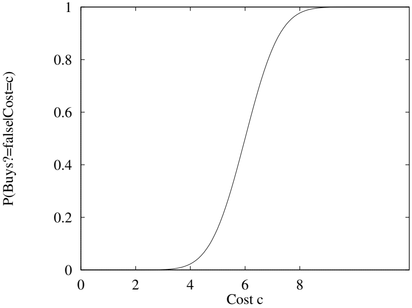
Sieć tylko ze zmiennymi ciągłymi z rozkładami LG

$\Longrightarrow$ pełny rozkład gaussowski wielu zmiennych

Sieć LG zmiennych dyskretnych+ciągłych jest siecią gaussowską warunkową tzn. gaussowski rozkład wszystkich zmiennych ciągłych dla każdej kombinacji wartości zmiennych dyskretnych.

Zmienne dyskretne z ciągłymi zmiennymi-rodzicami

Prawdopodobieństwo *Buys*? dla danego *Cost* powinno być “miękkim” progiem:



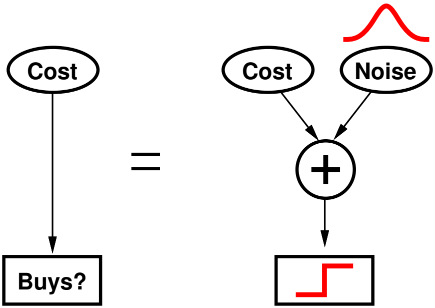
Rozkład probitowy używa całkowania funkcji gaussowskiej:

$\Phi(x) = \int_{-\infty}^{x} N(0,1)(x)dx$

$P(Buys? = true|Cost = c) = \Phi((-c+\mu)/\sigma)$

Dlaczego rozkład probitowy?

1. Ma właściwy kształt
2. Może być traktowany jako sztywny próg, którego położenie jest zakłócone

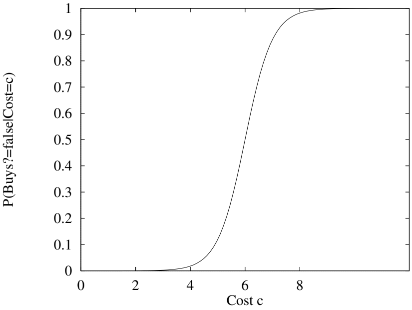


Zmienne dyskretne z ciągłymi zmiennymi-rodzicami

Rozkład sigmoidalny (lub logitowy) używany również w sieciach neuronowych:

$$P(Buys? = true | Cost = c) = \frac{1}{1 + exp(-2\frac{-c+\mu}{\sigma})}$$

Rozkład sigmoidalny ma kształt podobny do probitowego, ale dłuższe ogony:



Rodzaje zadań dla sieci bayesowskiej

Zapytania proste: oblicz brzegową wartość warunkową $\mathbf{P}(X_i|\mathbf{E}=\mathbf{e})$
np. $P(NoGas|Gauge = empty, Lights = on, Starts = false)$

Zapytania koniunkcyjne: $\mathbf{P}(X_i, X_j|\mathbf{E}=\mathbf{e}) = \mathbf{P}(X_i|\mathbf{E}=\mathbf{e})\mathbf{P}(X_j|X_i, \mathbf{E}=\mathbf{e})$

Decyzje optymalizacyjne: sieć zawiera informację o użyteczności;
np. wnioskowanie probabilistyczne dla $P(outcome|action, evidence)$

Wartość informacji: którą przesłankę sprawdzić jako następną?

Analiza wrażliwości: które wartości prawdopodobieństwa są najbardziej krytyczne?

Wyjaśnienie: Dlaczego potrzebuję nowego rozrusznika?

Wnioskowanie w sieci bayesowskiej

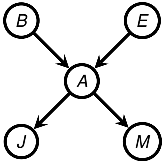
- Wnioskowanie dokładne
  - Przez wyliczanie wartości
  - Przez eliminację zmiennych
- Wnioskowanie aproksymacyjne
  - Przez symulację stochastyczną
  - Metodą Monte Carlo z łańcucha Markowa

Wnioskowanie przez wyliczanie wartosci

Sumowanie iloczynów z prawdopodobieństw brzegowych bez faktycznego konstruowania ich jawnej reprezentacji, przy użyciu prawdopodobieństw warunkowych z sieci bayesowskiej.

Proste zapytanie w sieci z alarmem domowym:

$$\begin{aligned}&\mathbf{P}(B|j,m)\\&=\mathbf{P}(B,j,m)/P(j,m)\\&=\alpha\mathbf{P}(B,j,m)\\&=\alpha\sum_e\sum_a\mathbf{P}(B,e,a,j,m)\end{aligned}$$



Przechodząc po zmiennych w kolejności zgodnej z siecią (np. B, E, A, J, M) wyciągamy sumowanie po kolejnych zmiennych na zewnątrz wyrażenia i używamy wartości prawdopodobieństw z tablic TPW:

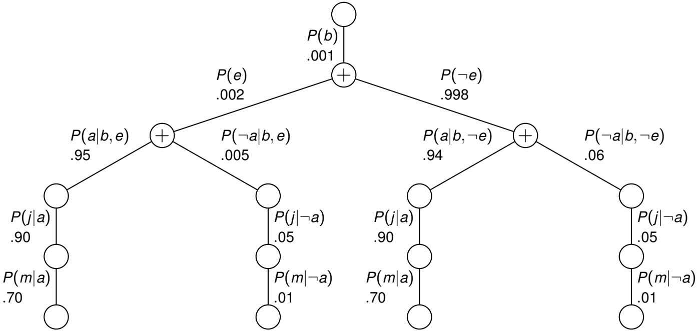
$$\begin{aligned}\mathbf{P}(B|j,m)&=\alpha\sum_e\sum_a\mathbf{P}(B)P(e)\mathbf{P}(a|B,e)P(j|a)P(m|a)\\&=\alpha\mathbf{P}(B)\sum_e P(e)\sum_a\mathbf{P}(a|B,e)P(j|a)P(m|a)\end{aligned}$$

Wyliczanie wartości: algorytm

```
function ENUMERATION-ASK(X, e, bn) returns a distribution over X
   inputs: X - the query variable
           e - observed values for variables E
           bn - a Bayesian network with variables {X} ∪ E ∪ Y
   Q(X) ← a distribution over X, initially empty
   for each value x_i of X do
      extend e with value x_i for X
      Q(x_i) ← ENUMERATE-ALL(VARS[bn], e)
   end for
   return NORMALIZE(Q(X))
end function
```

```
function ENUMERATE-ALL(vars, e) returns a real number
   if EMPTY?(vars) then
      return 1.0
   end if
   Y ← FIRST(vars)
   if Y has value y in e then
      return P(y|Parent(Y)) × ENUMERATE-ALL(REST(vars), e)
   else
      return Σ_y P(y|Parent(Y)) × ENUMERATE-ALL(REST(vars), e_y)
         where e_y is e extended with Y = y
   end if
end function
```

Wyliczanie wartości: działanie



Rekurencyjne wyliczanie zmiennych w głąb sieci: $O(n)$ pamięci, $O(d^n)$ czasu.

Wyliczanie jest nieefektywne: powtarza obliczenia,
np. liczy $P(j|a)P(m|a)$ dla każdej wartości e.

Wnioskowanie przez eliminację zmiennych

Eliminacja zmiennych: wykonuje sumowanie z prawej do lewej, pamięta wyniki pośrednie (czynniki) w celu uniknięcia powtórzeń

$$
\begin{aligned}
\mathbf{P}(B|j,m) &= \alpha \underbrace{\mathbf{P}(B)}_{B} \sum_e \underbrace{P(e)}_{E} \sum_a \underbrace{\mathbf{P}(a|B,e)}_{A} \underbrace{P(j|a)}_{J} \underbrace{P(m|a)}_{M} \\
&= \alpha \mathbf{P}(B) \sum_e P(e) \sum_a \mathbf{P}(a|B,e) P(j|a) f_M(a) \\
&= \alpha \mathbf{P}(B) \sum_e P(e) \sum_a \mathbf{P}(a|B,e) f_{JM}(a) \\
&= \alpha \mathbf{P}(B) \sum_e P(e) f_{\bar{A}JM}(b,e) \\
&= \alpha \mathbf{P}(B) f_{\bar{E}\bar{A}JM}(b) \\
&= \alpha f_B(b) \times f_{\bar{E}\bar{A}JM}(b)
\end{aligned}
$$

Wnioskowanie przez eliminację zmiennych

$$f_M(A) = \begin{pmatrix} P(m|a) \\ P(m|\neg a) \end{pmatrix}, \qquad f_{JM}(A) = f_J(A) \times f_M(A) = \begin{pmatrix} P(j|a)P(m|a) \\ P(j|\neg a)P(m|\neg a) \end{pmatrix}$$

$f_A(A, B, E)$ jest macierzą $2 \times 2 \times 2$ dla wszystkich wartości A, B, E

$$f_{\bar{A}JM}(B, E) = f_A(a, B, E) \times f_{JM}(a) + f_A(\neg a, B, E) \times f_{JM}(\neg a)$$

$$f_{\bar{E}\bar{A}JM}(B, E) = f_E(e) \times f_{\bar{A}JM}(B, e) + f_E(\neg e) \times f_{\bar{A}JM}(B, \neg e)$$

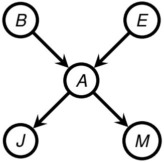
Eliminacja zmiennych: algorytm

```
function ELIMINATION-ASK(X, e, bn) returns a distribution over X
    inputs: X - the query variable
                e - evidence specified as an event
                bn - a belief network specifying join distribution P(X_1, ..., X_n)
    factors ← []
    vars ← REVERSE(VARS[bn])
    for each var in vars do
        factors ← [MAKE-FACTOR(var, e)|factors]
        if var is a hidden variable then
            factors ← SUM-OUT(var, factors)
        end if
    end for
    return NORMALIZE(POINTWISE-PRODUCT(factors))
end function
```

Eliminacja zmiennych: zmienne nieistotne

Rozważmy zapytanie $P(JohnCalls|Burglary = true)$

$$P(J|b) = \alpha P(b) \sum_e P(e) \sum_a P(a|b,e) P(J|a) \sum_m P(m|a)$$



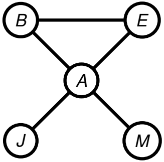
Suma po m jest równa 1; M jest <u>nieistotne</u> dla zapytania
$\Longrightarrow$ Można pominąć sumowanie po zmiennych nieistotnych

Tw 1: Y jest nieistotne jeśli $Y \notin Ancestors(\{X\} \cup \mathbf{E})$

Tutaj $X = JohnCalls$, $\mathbf{E} = \{Burglary\}$
i $Ancestors(\{X\} \cup \mathbf{E}) = \{Alarm, Earthquake\}$
więc M jest nieistotne.

Eliminacja zmiennych: zmienne nieistotne

Def: moralny graf sieci bayesowskiej (nieskierowany): zawiera krawędzie z oryginalnej sieci bez kierunku oraz krawędzie pomiędzy każdą parą rodziców mającą wspólne dziecko.



Def: **A** jest m-odseparowane od **B** przez **C** $\iff$ gdy jest odseparowane przez **C** w grafie moralnym.

Tw 2: Y jest nieistotne jeśli jest m-odseparowane od X przez **E**.

Dla $P(JohnCalls|Alarm = true)$,
obie *Burglary* i *Earthquake* są nieistotne.

Złożoność dokładnego wnioskowania

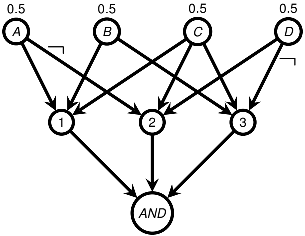
Sieci pojedynczych połączeń (polidrzewa):

- każde dwa wierzchołki połączone są co najwyżej jedną ścieżką
- złożoność czasowa i pamięciowa algorytmu eliminacji zmiennych $O(d^k n)$

Sieci wielokrotnych połączeń:

- można zredukować 3SAT do dokładnego wnioskowania $\Longrightarrow$ NP-trudne
- równoważne zliczaniu modeli 3SAT $\Longrightarrow$ #P-zupełne

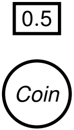
1. $\mathbf{A} \vee \mathbf{B} \vee \mathbf{C}$
2. $\mathbf{C} \vee \mathbf{D} \vee \neg\mathbf{A}$
3. $\mathbf{B} \vee \mathbf{C} \vee \neg\mathbf{D}$



Wnioskowanie przez symulację stochastyczną

Podstawowy pomysł:

1. Losuj N próbek z rozkładem próbkowym S
2. Oblicz aproksymacyjne prawdopodobieństwo wynikowe $\hat{P}$
3. Udowodnij zbieżność do prawdopodobieństwa faktycznego P



Wnioskowanie stochastyczne bezwarunkowe (bez przesłanek):

- Próbkowanie bezpośrednie

Wnioskowanie stochastyczne warunkowe (z przesłankami):

- Próbkowanie z odrzucaniem: odrzuca próbki niezgodne z przesłankami
- Ważenie prawdopodobieństwa próbek:
  używa przesłanek do ważenia prawdopodobieństwa próbek
- Monte Carlo z łańcucha Markowa (MCMC):
  próbkuje z procesu stochastycznego, w którym prawdopodobieństwo stacjonarne jest rzeczywistym prawdopodobieństwem warunkowym

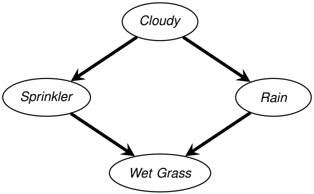
Próbkowanie bezpośrednie

```
function DIRECT-SAMPLING(X, bn, N) returns an estimate of P(X)
    local variables: N - a vector of counts over X, initially zero
    for j = 1 to N do
        x ← PRIOR-SAMPLE(bn)
        N[x] ← N[x] + 1 where x is the values of X in x
    end for
    return NORMALIZE(N[X])
end function
```

```
function PRIOR-SAMPLE(bn) returns an event sampled from bn
    inputs: bn - a belief network specifying joint distribution P(X1, ..., X_n)
    x ← an event with n elements
    for i = 1 to n do
        x_i ← a random sample from P(X_i | Parents(X_i))
    end for
    return x
end function
```

Próbkowanie bezpośrednie: przykład

$P(C)$
.50



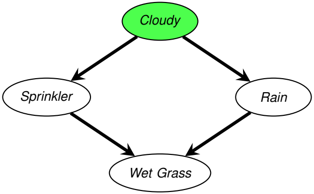
C	$P(S\|C)$
T	.10
F	.50

C	$P(R\|C)$
T	.80
F	.20

S	R	$P(W\|S,R)$
T	T	.99
T	F	.90
F	T	.90
F	F	.01

Próbkowanie bezpośrednie: przykład

$P(C)$
.50



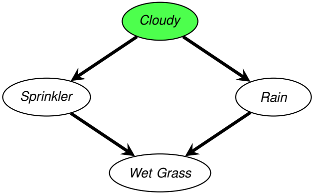
C	$P(S\|C)$
T	.10
F	.50

C	$P(R\|C)$
T	.80
F	.20

S	R	$P(W\|S,R)$
T	T	.99
T	F	.90
F	T	.90
F	F	.01

Próbkowanie bezpośrednie: przykład

$P(C)$
.50



C	$P(S\|C)$
T	.10
F	.50

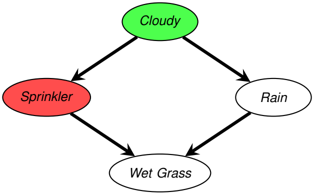
C	$P(R\|C)$
T	.80
F	.20

S	R	$P(W\|S,R)$
T	T	.99
T	F	.90
F	T	.90
F	F	.01

Próbkowanie bezpośrednie: przykład

$P(C)$
.50

C	$P(S\|C)$
T	.10
F	.50

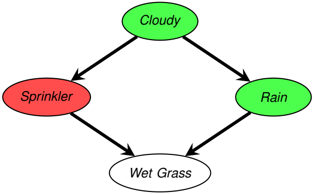


C	$P(R\|C)$
T	.80
F	.20

S	R	$P(W\|S,R)$
T	T	.99
T	F	.90
F	T	.90
F	F	.01

Próbkowanie bezpośrednie: przykład

$P(C)$
.50



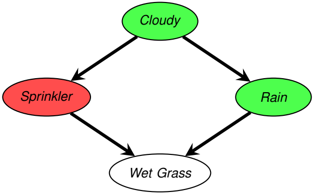
C	$P(S \mid C)$
T	.10
F	.50

C	$P(R \mid C)$
T	.80
F	.20

S	R	$P(W \mid S, R)$
T	T	.99
T	F	.90
F	T	.90
F	F	.01

Próbkowanie bezpośrednie: przykład

$P(C)$
.50



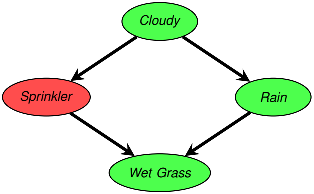
C	$P(S\|C)$
T	.10
F	.50

C	$P(R\|C)$
T	.80
F	.20

S	R	$P(W\|S,R)$
T	T	.99
T	F	.90
F	T	.90
F	F	.01

Próbkowanie bezpośrednie: przykład

$P(C)$
.50



C	$P(S \mid C)$
T	.10
F	.50

C	$P(R \mid C)$
T	.80
F	.20

S	R	$P(W \mid S, R)$
T	T	.99
T	F	.90
F	T	.90
F	F	.01

Próbkowanie bezpośrednie: własności

Prawdopodobieństwo, że PRIOR-SAMPLE generuje dane zdarzenie

$$S_{PS}(x_1 \ldots x_n) = \prod_{i=1}^{n} P(x_i | Parents(X_i)) = P(x_1 \ldots x_n)$$

to odpowiada prawdopodobieństwu faktycznemu tego zdarzenia

Np. $S_{PS}(t, f, t, t) = 0.5 \times 0.9 \times 0.8 \times 0.9 = 0.324 = P(t, f, t, t)$

$N_{PS}(x_1 \ldots x_n)$ — liczba próbek wygenerowanych dla zdarzenia $x_1, \ldots, x_n$

Wtedy

$$\begin{aligned}
\lim_{N\to\infty} \hat{P}(x_1, \ldots, x_n) &= \lim_{N\to\infty} N_{PS}(x_1, \ldots, x_n)/N \\
&= S_{PS}(x_1, \ldots, x_n) \\
&= P(x_1 \ldots x_n)
\end{aligned}$$

Powyższą własność algorytmu DIRECT-SAMPLING nazywamy <u>spójnością</u>

Notacja: $\hat{P}(x_1, \ldots, x_n) \approx P(x_1 \ldots x_n)$

Próbkowanie z odrzucaniem

$\hat{\mathbf{P}}(X|\mathbf{e})$ szacowane z próbek zgodnych z przesłankami **e**

```
function REJECTION-SAMPLING(X, e, bn, N) returns an estimate of P(X|e)
    local variables: N - a vector of counts over X, initially zero
    for j = 1 to N do
        x ← PRIOR-SAMPLE(bn)
        if x is consistent with e then
            N ← N + 1 where x is the value of X in x
        end if
    end for
    return NORMALIZE(N[X])
end function
```

Np. oszacowanie $\mathbf{P}(Rain|Sprinkler = true)$ przy użyciu 100 próbek

27 próbek ma $Sprinkler = true$

Z tego, 8 ma $Rain = true$ i 19 ma $Rain = false$.

$\hat{\mathbf{P}}(Rain|Sprinkler = true) = \text{NORMALIZE}(\langle 8, 19\rangle) = \langle 0.296, 0.704\rangle$

Próbkowanie z odrzucaniem: własności

$$\begin{aligned}\hat{\mathbf{P}}(X|\mathbf{e}) &= \alpha \mathbf{N}_{PS}(X,\mathbf{e}) \qquad \text{(wynik algorytmu REJECTION-SAMPLING)}\\ &= \mathbf{N}_{PS}(X,\mathbf{e})/N_{PS}(\mathbf{e}) \qquad \text{(normalizowane przez } N_{PS}(\mathbf{e})\text{)}\\ &\approx \mathbf{P}(X,\mathbf{e})/P(\mathbf{e}) \qquad \text{(własność PRIOR-SAMPLE)}\\ &= \mathbf{P}(X|\mathbf{e}) \qquad \text{(prawdopodobieństwo faktyczne)}\end{aligned}$$

Zatem próbkowanie z odrzucaniem ma własność spójności tzn. oszacowanie zbiega do faktycznego prawdopodobieństwa warunkowego

Problem: bardzo kosztowne jeśli $P(\mathbf{e})$ jest małe

$P(\mathbf{e})$ rozpada się wykładniczo wraz z liczbą zmiennych!

Ważenie prawdopodobieństwa próbek

Pomysł: ustala zmienne z przesłanek, próbkuje tylko zmienna spoza przesłanek, i waży prawdopodobieństwo każdej próbki stosownie do przesłanek.

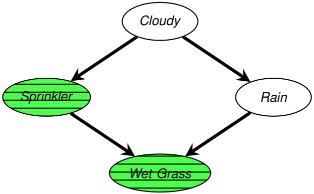
function LIKELIHOOD-WEIGHTING(X, $\mathbf{e}$, bn, N) **returns** an estimate of $P(X|\mathbf{e})$
  local variables: W - a vector of weighted counts over X, initially zero
  for $j = 1$ to N **do**
    $\mathbf{x}, w \leftarrow$ WEIGHTED-SAMPLE(bn)
    $\mathbf{W}[x] \leftarrow \mathbf{W}[x] + w$ where x is the value of X in $\mathbf{x}$
  end for
  return NORMALIZE($\mathbf{W}[X]$)
end function

---

function WEIGHTED-SAMPLE(bn) **returns** an event and a weight
  $\mathbf{x} \leftarrow$ an event with n elements
  $w \leftarrow 1$
  for $i = 1$ to n **do**
    if X_i has a value x_i in **e then**
      $w \leftarrow w \times P(X_i = x_i | Parents(X_i))$
    else
      $x_i \leftarrow$ a random sample from $\mathbf{P}(X_i | Parents(X_i))$
    end if
  end for
  return x, w
end function

Próbkowanie bezpośrednie: przykład

$P(C)$
.50



C	$P(S\|C)$
T	.10
F	.50

C	$P(R\|C)$
T	.80
F	.20

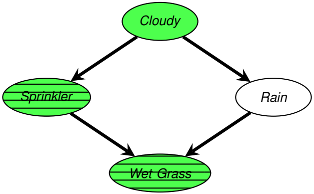
S	R	$P(W\|S,R)$
T	T	.99
T	F	.90
F	T	.90
F	F	.01

$w = 1.0$

Próbkowanie bezpośrednie: przykład

$P(C)$
.50

C	$P(S\|C)$
T	.10
F	.50



C	$P(R\|C)$
T	.80
F	.20

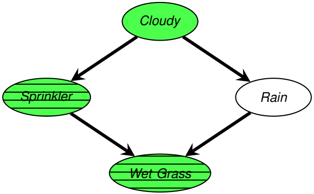
S	R	$P(W\|S,R)$
T	T	.99
T	F	.90
F	T	.90
F	F	.01

$w = 1.0$

Próbkowanie bezpośrednie: przykład

$P(C)$
.50

C	$P(S\|C)$
T	.10
F	.50



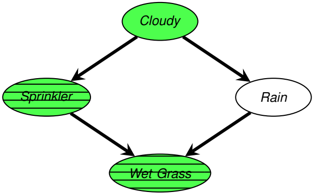
C	$P(R\|C)$
T	.80
F	.20

S	R	$P(W\|S,R)$
T	T	.99
T	F	.90
F	T	.90
F	F	.01

$w = 1.0$

Próbkowanie bezpośrednie: przykład

$P(C)$
.50



C	$P(S\|C)$
T	.10
F	.50

C	$P(R\|C)$
T	.80
F	.20

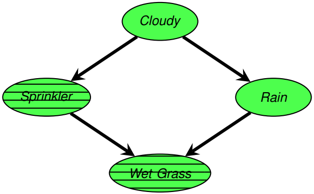
S	R	$P(W\|S,R)$
T	T	.99
T	F	.90
F	T	.90
F	F	.01

$$w = 1.0 \times 0.1$$

Próbkowanie bezpośrednie: przykład

$P(C)$
.50

C	$P(S\|C)$
T	.10
F	.50



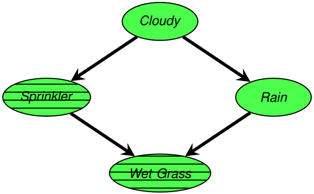
C	$P(R\|C)$
T	.80
F	.20

S	R	$P(W\|S,R)$
T	T	.99
T	F	.90
F	T	.90
F	F	.01

$$w = 1.0 \times 0.1$$

Próbkowanie bezpośrednie: przykład

$P(C)$
.50



C	$P(S\|C)$
T	.10
F	.50

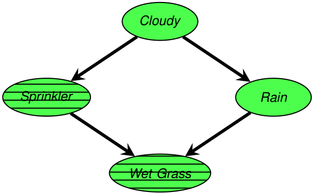
C	$P(R\|C)$
T	.80
F	.20

S	R	$P(W\|S,R)$
T	T	.99
T	F	.90
F	T	.90
F	F	.01

$$w = 1.0 \times 0.1$$

Próbkowanie bezpośrednie: przykład

$P(C)$
.50



C	$P(S \mid C)$
T	.10
F	.50

C	$P(R \mid C)$
T	.80
F	.20

S	R	$P(W \mid S, R)$
T	T	.99
T	F	.90
F	T	.90
F	F	.01

$$w = 1.0 \times 0.1 \times 0.99 = 0.099$$

Ważenie prawdopodobieństwa próbek: własności

Prawdopodobieństwo próbki ważonej WEIGHTED-SAMPLE wynosi

$$S_{WS}(\mathbf{z},\mathbf{e}) = \prod_{i=1}^{l} P(z_i|Parents(Z_i))$$

Uwaga: S_{WS} uwzględnia tylko przesłanki z <u>przodków</u> z_i

$\Longrightarrow$ daje prawdopodobieństwo pośrednie pomiędzy prawdopodobieństwem a priori i a posteriori

Waga dla danej próbki $\mathbf{z},\mathbf{e}$ wynosi

$$w(\mathbf{z},\mathbf{e}) = \prod_{i=1}^{m} P(e_i|Parents(E_i))$$

Ważone prawdopodobieństwo próbkowe:

$$\begin{aligned} & S_{WS}(\mathbf{z},\mathbf{e})w(\mathbf{z},\mathbf{e}) \\ & \quad = \prod_{i=1}^{l} P(z_i|Parents(Z_i)) \quad \prod_{i=1}^{m} P(e_i|Parents(E_i)) \\ & \quad = P(\mathbf{z},\mathbf{e}) \text{ (ze standardowej, globalnej semantyki sieci)} \end{aligned}$$

Stąd ważenie prawdopodobieństwa też ma własność spójności ale efektywność nadal maleje przy dużej liczbie przesłanek ponieważ bardzo mało próbek ma dużą wagę.

Monte Carlo dla łańcucha Markowa

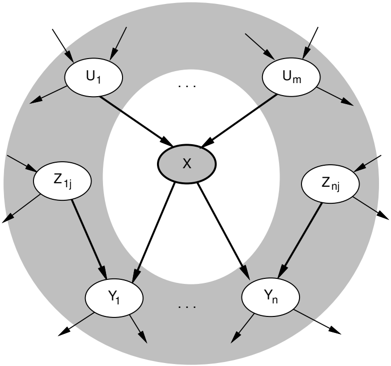
"Stan" sieci: bieżące przypisanie wszystkich zmiennych

Łańcuch Markowa: ciąg stanów sieci, następny stan jest generowany poprzez próbkowanie jednej zmiennej nie będącej przesłanką na podstawie jej koca Markowa

function MCMC-ASK(X, $\mathbf{e}$, bn, N) **returns** an estimate of $P(X|\mathbf{e})$
  local variables: $\mathbf{N}[X]$ - a vector of counts over X, initially zero
    $\mathbf{Z}$ - the nonevidence variables in bn
    $\mathbf{x}$ - the current state of the network, initially copied from $\mathbf{e}$
  initialize $\mathbf{x}$ with random values for the variables in $\mathbf{Y}$
  for $j = 1$ to N **do**
    $\mathbf{N}[x] \leftarrow \mathbf{N}[x] + 1$ where x is the value of X in $\mathbf{x}$
    for each Z_i in $\mathbf{Z}$ **do**
      sample the value of Z_i in $\mathbf{x}$ from $\mathbf{P}(Z_i|MB(Z_i))$ given the values of $MB(Z_i)$ in $\mathbf{x}$
    end for
  end for
  return NORMALIZE($\mathbf{N}[X]$)
end function

Koc Markowa

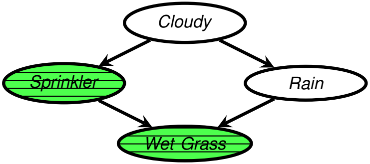
Każdy węzeł jest warunkowo niezależny od wszystkich pozostałych przy danym jego kocu Markowa: rodzice + dzieci + inni rodzice dzieci



Koc Markowa: przykład

Koc Markowa dla *Cloudy*:
Sprinkler i *Rain*

Koc Markowa dla *Rain*:
Cloudy, *Sprinkler* i *WetGrass*

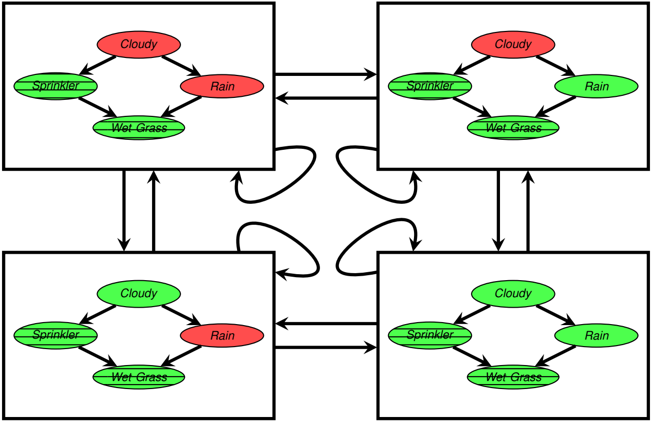


Prawdopodobieństwo warunkowe przy danym kocu Markowa:

$$P(x_i'|MB(X_i)) = P(x_i'|Parents(X_i)) \prod_{Z_j \in Children(X_i)} P(z_j|Parents(Z_j))$$

Łańcuch Markowa

Przy przesłankach *Sprinkler* $=$ *true*, *WetGrass* $=$ *true*
łańcuch Markowa zawiera 4 stany:



Monte Carlo dla łańcucha Markowa: przykład

Szacowanie $\mathbf{P}(Rain|Sprinkler = true, WetGrass = true)$

Algorytm powtarza próbkowanie zmiennych *Cloudy* i *Rain* na podstawie ich koca Markowa. Zlicza, ile razy *Rain* było true i false w kolejnych stanach sieci.

Np. odwiedza 100 stanów

31 ma $Rain = true$, 69 ma $Rain = false$

$$\hat{\mathbf{P}}(Rain|Sprinkler = true, WetGrass = true)$$
$$= \text{NORMALIZE}(\langle 31, 69 \rangle) = \langle 0.31, 0.69 \rangle$$

Monte Carlo dla łańcucha Markowa: własności

Twierdzenie: łańcuch zbiega do rozkładu stacjonarnego (≈ spójność):

proporcja czasu spędzonego w danym stanie w czasie długiego działania sieci jest dokładnie proporcjonalna do faktycznego prawdopodobieństwa warunkowego.

◇ Zalety

- Metoda nie jest wrażliwa na topologię sieci
- Można stosować do zmiennych dyskretnych i ciągłych

◇ Wady

- Zbieżność może być wolna
- Trudno określić moment, w którym algorytm daje już bliskie rozwiązanie
- Może być czasowo rozrzutny, jeśli występują duże koce Markowa: $P(X_i|MB(X_i))$ nie zmienia się dużo (Prawo Wielkich Liczb) a jest liczone za każdym razem

Dziękuję za uwagę!