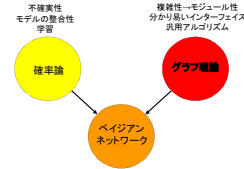


# 情報意味論 (9) ベイジアンネットワーク

慶應義塾大学工学部  
櫻井 彰人

## どこから生まれてきたか？

- 実問題の共通課題：
  - 不確実性 ← 確率的枠組み ← 確率変数を用いよう
  - 複雑性 ← すっきりと表現しよう



## 確率変数を用いて簡潔に

- 原理的にできそう：  
全変数の結合確率で表現  
しかし、naïve Bayes の時と同様に
- 「全変数の結合確率」ではパラメータが多すぎ  
かといっても、naïve Bayes は単純化しすぎ
- 「全変数の結合確率」と「naïve Bayes」の間  
はないか？
- Bayesian network は一つの答え

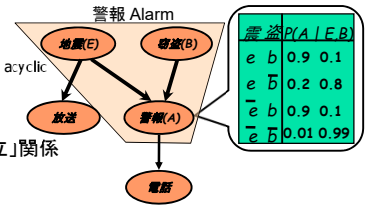
## BNとは何か？

条件付確率を用いた、結合確率のコンパクトな表現

### 定性的要素：

有向無閉路グラフ directed acyclic graph (DAG)

- ノード - 確率変数.
- エッジ - 非「条件付独立」関係



### あわせて：

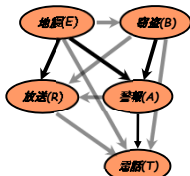
ある確率分布の因数分解(? 確率分布の積に分解)

定量的要素：  
条件付確率分布の集まり

Figure from N. Friedman

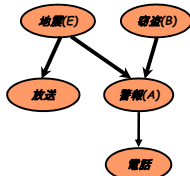
## コンパクトな表現

完全な結合の表現



$$p(e, b, r, a, t) = p(e)p(b|e)p(a|e, b)p(r|e, b, a)p(t|e, b, a, r)$$

コンパクトな表現



$$p(e, b, r, a, t) = p(e)p(b)p(a|e, b)p(r|e)p(t|a)$$

## 条件付き独立性

- 確率変数  $x$  と  $z$  が  $y$  を条件として条件付き独立であるとは、

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

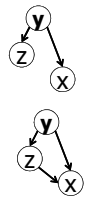
しばしば、 $x \perp\!\!\!\perp z|y$  と書く

- 条件付き確率の定義から

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

- 従って、 $x \perp\!\!\!\perp z|y$  if and only if

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



グラフィカルモデル (後編)

## なぜ役立つか？

- グラフ構造があるので
  - 知識をモジュール化して表現できる
  - 推論・学習に、局所的かつ分散的のアルゴリズムが使える
  - 直感的な (場合によっては因果的な) 解釈が可能
- 結合確率  $P(X_1, \dots, X_n)$  をそのまま表現するより、指数関数的に少ないパラメータで、表現可能 =>
  - 学習に必要なデータ数 (sample complexity) が少なくてすむ
  - 推論に必要な時間 (time complexity) が少なくてすむ

7

## 何に使うか？

- 事後確率推定
  - 証拠・現象 evidence から発生した事象 event の確率を推定
- 最も可能性が高い説明
  - 証拠・現象を説明するシナリオ
- 合理的な意思決定
  - 期待成果を最大化
  - 情報の価値

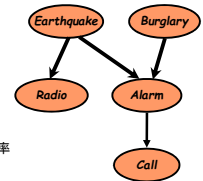
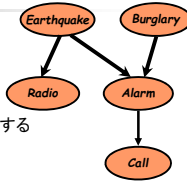


Figure from N. Friedman

## 他の類似モデルとの比較

- 多くの分類・回帰モデル
  - 未知サンプル  $(x_1, x_2, \dots, x_{n'})$ 
    - (欠けている) 目的変数の値を推定する
- 多くの確率モデル
  - 未知サンプル  $(x_1, x_2, \dots, x_{n'})$ 
    - (欠けている) 目的変数の分布を推定する
- BNでは
  - ある (未知) 状態  $(x_1, x_2, \dots, x_{n'}, y_1, \dots, y_{m'})$ 
    - 未観測の変数がある中で、(欠けている) 目的変数の分布を推定する



9

## 応用事例

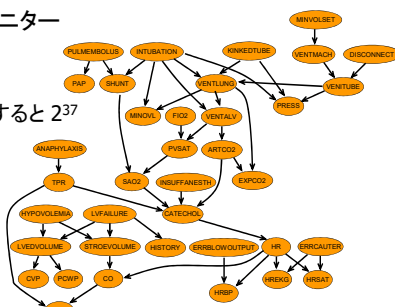
- "Microsoft's competitive advantage lies in its expertise in Bayesian networks"
  - Bill Gates, LA Times より, 1996
- MS Answer Wizards, (printer) troubleshooters
- 医療診断
- 遺伝子系統解析
- 音声認識 (HMMs)
- 遺伝子配列分析
- Turbocodes (通信路の符号化)

10

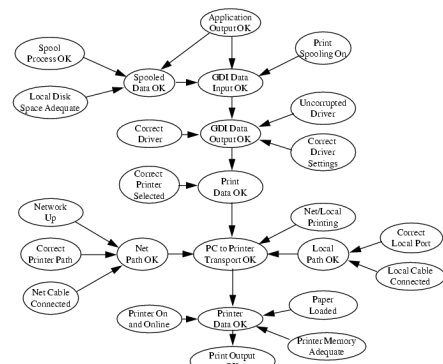
## 実例: Alarm (A Logical Alarm Reduction Mechanism)

分野: ICU でのモニター

- 37 変数
- 509 パラメータ
- ... 各変数2値とすると  $2^{37}$



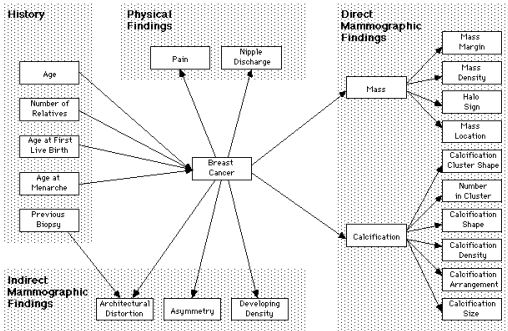
A Logical Alarm Reduction Mechanism Figure from N. Friedman



Microsoft Print Troubleshooter

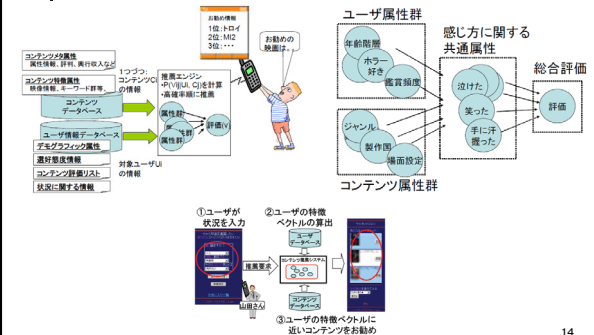
12

# MammoNet



<http://www.mcw.edu/midas/images/mammo.model.gif>  
<http://www.mcw.edu/midas/mammo.html>

# 状況に応じた映画推薦

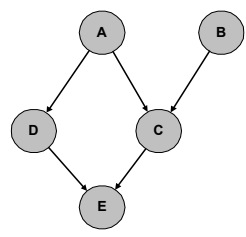


# 確率の変形規則...

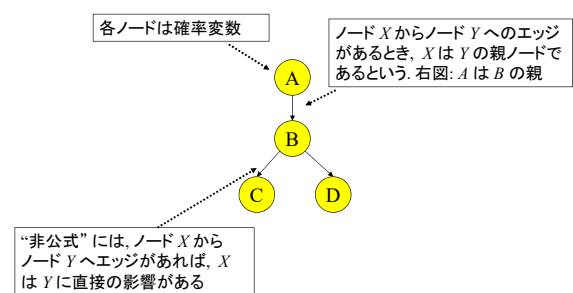
- ベイズ規則:  $\Pr(A, B) = \Pr(A|B) \cdot \Pr(B) = \Pr(B|A) \cdot \Pr(A)$
- 独立性  $A \perp B$  iff:  $\Pr(A|B) = \Pr(A)$   
 $\Pr(B|A) = \Pr(B)$   
 $\Rightarrow \Pr(A, B) = \Pr(A) \cdot \Pr(B)$
- チェーン規則:  $\Pr(A, B, C, D) = \Pr(A) \cdot \Pr(B, C, D|A)$   
 $= \Pr(A) \cdot \Pr(B|A) \cdot \Pr(C, D|A, B)$   
 $= \dots$   
 $= \Pr(A) \cdot \Pr(B|A) \cdot \Pr(C|A, B) \cdot \Pr(D|A, B, C)$
- 周辺化 marginalize:  $\Pr(A) = \sum_b \Pr(A, B = b)$

# BN の形式的定義

- DAG: 有向無閉路グラフ directed acyclic graph
  - ノード: 各ノードは確率変数。変数間にはある順序がある (離散値をとる)
  - エッジ: エッジの出る側は入る側の親と呼ぶ。子ノードには、親ノードを条件とする条件付確率表 (CPT) が定義されている。各エッジは変数間に直接的な関係がありうることを示している。正確には、「条件付独立関係がある」とは言えないことを示している。方向は、因果関係があれば、原因→結果、なければ、任意。ノード順序に矛盾しない
  - CPTs: 条件付確率表:  $\Pr(X|pa(X))$   
 右図:  $\Pr(C|A, B)$ ,  $\Pr(D|A)$ ,  $\Pr(E|C, D)$
  - 事前分布 a priori distribution: 親のないノードすべてに  
 右図:  $\Pr(A)$ ,  $\Pr(B)$
- 全変数の結合確率は、上記の条件付確率の積で表される  $P(X_1, \dots, X_n) = \prod P(X_i | pa(X_i))$



# DAG



# 条件付確率表 CPT

各ノード  $X_i$  には条件付確率表  $P(X_i | \text{Parents}(X_i))$  があり、親ノードの当該ノードへの影響を表現する

表中の条件付確率がパラメータとなる。

$P(D, C, B, A) = P(D|B)P(C|B)P(B|A)P(A)$

A	false	0.6
A	true	0.4

P(B A)		A	
		false	true
B	false	0.01	0.7
	true	0.99	0.3

P(C B)		B	
		false	true
C	false	0.40	0.90
	true	0.60	0.10

P(D B)		B	
		false	true
D	false	0.02	0.05
	true	0.98	0.95

## 条件付確率表 CPT

親ノードが B であるノード C, の条件付確率表

		B	
		false	true
C	false	0.4	0.9
	true	0.6	0.1

ここには直接の親ノードしか現れない

親ノード(左図ではB)の値のすべての組み合わせについて,  $P(C=true|B)$  と  $P(C=false|B)$  の和は 1 とならないといけない

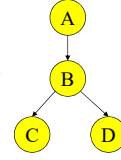
k 個の親が全てブール変数(2値変数)であれば, ブール値変数のCPTの要素数は  $2^k * 2 = 2^{k+1}$  となる

19

## BNの定義 (まとめると)

BN の構成要素:

1. 有向無閉路グラフ DAG directed acyclic graph



3. 全変数の結合確率は、各ノードに付随する条件付確率の積

$$P(D, C, B, A) = P(D|B)P(C|B)P(B|A)P(A)$$

もし構造がなければ  $\Pr(D|A, B, C) \cdot \Pr(C|A, B) \cdot \Pr(B|A) \cdot \Pr(A)$

2. 各ノードに付随する条件付確率表

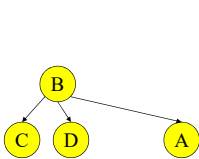
A	false	0.6
	true	0.4

P(B A)	A		
	false	true	
B	false	0.01	0.7
	true	0.99	0.3

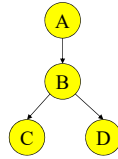
P(D B)	B		
	false	true	
D	false	0.02	0.05
	true	0.98	0.95

P(C B)	B		
	false	true	
C	false	0.40	0.90
	true	0.60	0.10

## 補足: naïve Bayes との比較



$$P(D, C, B, A) = P(D|B)P(C|B)P(A|B)P(B) = P(D|B)P(C|B)P(A, B)$$



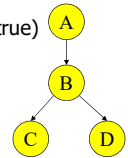
$$P(D, C, B, A) = P(D|B)P(C|B)P(B|A)P(A) = P(D|B)P(C|B)P(A, B)$$

21

## 計算例

先ほどの例で次の結合確率を計算する:

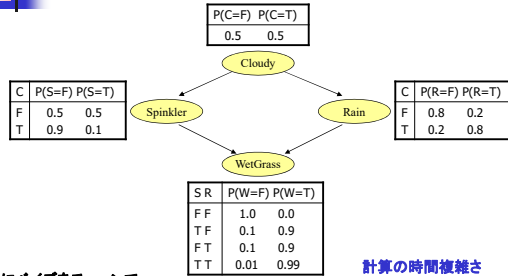
$$P(A = \text{true}, B = \text{true}, C = \text{true}, D = \text{true}) = P(A = \text{true}) * P(B = \text{true} | A = \text{true}) * P(C = \text{true} | B = \text{true}) * P(D = \text{true} | B = \text{true}) = (0.4) * (0.3) * (0.1) * (0.95)$$



例題が簡単すぎて、あまり簡単にならないが、...

22

## 他の例: Water-Sprinkler



単にベイズをチェーンで:

$$\Pr(C, R, S, W) = \Pr(C) \cdot \Pr(R|C) \cdot \Pr(S|R, C) \cdot \Pr(W|R, C, S)$$

計算の時間複雑さ

$$2 \times 4 \times 8 \times 16 = 1024$$

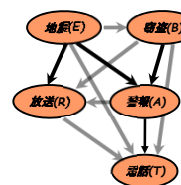
条件付独立性を使うと:

$$\Pr(C, R, S, W) = \Pr(C) \cdot \Pr(R|C) \cdot \Pr(S|C) \cdot \Pr(W|R, S)$$

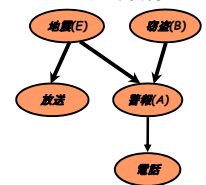
$$2 \times 4 \times 4 \times 8 = 256$$

## 条件付独立性とBN

完全な結合の表現



コンパクトな表現



$$p(e, b, a, r, t) = p(e)p(b)p(a|e, b)p(r|e, b, a)p(t|e, b, a, r)$$

$$p(e, b, a, r, t) = p(e)p(b)p(a|e, b)p(r|e)p(t|a)$$

$$\frac{p(b)}{b \perp\!\!\!\perp e}$$

$$\frac{p(r|e, b, a)}{b, a \perp\!\!\!\perp r|e}$$

24

## 条件付独立性の判定方法

全変数の結合確率表を作って計算すれば分かるのだが、それはしたくない

- D-separation: ある証拠が与えられたとき、それに対応する変数を条件として、他の変数が条件付独立であるための十分条件を与える。
  - 証拠: ある確率変数達について、実現した値
- DAG上で、2変数間を、証拠変数がささぎるか否かを判定し、それで、条件付独立か否かを表している。

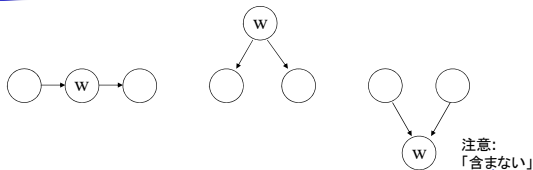
25

## D-separation

- D-separation は、DAG上の変数間の独立性を調べるグラフ的なテストである
- A, B: 変数集合. 独立性を調べる  
Z: 変数集合. 条件  
Aの全ての変数とBの全ての変数間の全てのpathを調べる
- AとBはZを条件として(i.e. Zが観測されるとき)独立である ( $A \perp\!\!\!\perp B \mid Z$ ) iff Aの全ての変数とBの全ての変数の間の全てのpathが通行止めである
- もしpathが一つでも通行可能であれば、独立も非独立もいえない
- D-separationが成立していないときに独立性を言おうと思えば、条件付確率表を調べるしかない(かな)
- ある pathが通行止めであるのは、このpath上のあるノード列が次のスライドに示す「通行止め」になっている場合である。

26

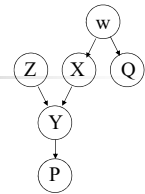
## 通行止め



	連続	分岐	合流
通行止め	$w \in Z$	$w \in Z$	$w \notin Z$ and 全子孫(w) $\notin Z$
通行可	$w \notin Z$	$w \notin Z$	$w \in Z$ or ある子孫(w) $\in Z$

27

## 例



正しい関係 D separation による説明

- ( $Q \perp\!\!\!\perp X, Y, Z, P \mid W$ ):  $Q \leftarrow W \rightarrow X$  は分岐. Wを条件として通行止め
- ( $Z \perp\!\!\!\perp X, W, Q \mid \emptyset$ ):  $Z \rightarrow Y \leftarrow X$  は合流. Y及びその子孫Pを条件としないので通行止め.
- ( $Z \perp\!\!\!\perp X, W, Q \mid P$ ):  $Z \rightarrow Y \leftarrow X$  は合流. Yの子孫Pを条件としているので通行可能.
- ( $Z, Y, P \perp\!\!\!\perp W, Q \mid X$ ):  $W \rightarrow X \rightarrow Y$  は連続. Xを条件として通行止め.
- ( $Z, Y, P \perp\!\!\!\perp W, Q \mid \emptyset$ ):  $W \rightarrow X \rightarrow Y$  は連続. Xを条件としないので通行可能.

28

## 推論

- ベイジアンネットワークで確率を計算することを推論という
- 一般に、推論では次の形のクエリーが扱われる:  
 $P(X \mid E)$

E = 証拠 evidence 変数

X = 問い合わせる変数  
周辺分布を求めるのが課題

29

## 再び 周辺分布・周辺化とは

語源(多分):

X \ Y	表	裏	p(Y)
表	4/25	6/25	10/25
裏	6/25	9/25	15/25
p(X)	10/25	15/25	

(表の)周辺 (margin)

1. 我々が考える確率分布は、より多くの確率変数を持つ分布の周辺分布だと考えた方がよい
2. 確率の書き方はプログラミング言語や数学の他の分野の書き方と大きく異なる。変数「名」が重要。「名前付き指数」+ $\alpha$

$\Pr(X) \neq \Pr(Y)$  as probability distributions,

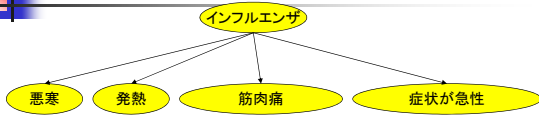
there is no such thing as  $\lambda x \Pr(x)$  or  $\Pr(\cdot)$ , and

$\Pr(X, Y) \equiv \Pr(Y, X)$  which is different from any programming language but similar to

$\Pr(X = x, Y = y) = \Pr(Y = y, X = x)$  in some programming language.

30

## 推論

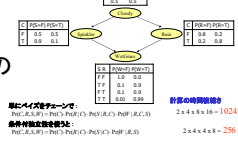


- クエリーは、例えば、:  
 $P(\text{インフルエンザ} = \text{true} \mid \text{発熱} = \text{true}, \text{急性症状} = \text{true})$
- 注: 悪寒と筋肉痛という変数がベイジアンネットワーク中に現れているが、クエリー中では値が与えられていない (i.e. 質問変数としても証拠変数としても現れていない)
- 未観測の確率変数として扱われる  
 周辺化されていると考える

31

## BNにおける推論

他の例: Water-Sprinkler



- WetGrass が真のとき、2つの説明が可能: Rain か Sprinkler

- どちらがよりありうるか?

$$\Pr(S=T \mid W=T) = \frac{\Pr(S=T, W=T)}{\Pr(W=T)} = \frac{\sum_{C,R} \Pr(C, R, S=T, W=T)}{\Pr(W=T)} = \frac{0.2781}{0.6471} = 0.430 \quad \text{Sprinkler}$$

$$\Pr(R=T \mid W=T) = \frac{\Pr(R=T, W=T)}{\Pr(W=T)} = \frac{\sum_{C,S} \Pr(C, S, R=T, W=T)}{\Pr(W=T)} = \frac{0.4581}{0.6471} = 0.708 \quad \text{Rain}$$

Rain が真であるのが理由である可能性がより高い

32

## BNにおける推論 (2)

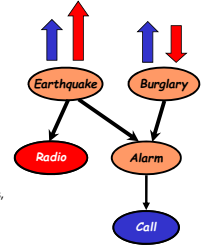
- Bottom-Up :**
  - 結果から原因へ → 診断 diagnostic
  - 例. エキスパートシステム, パターン認識, ...
  - 証拠・結果が与えられたとき、それを説明する最もありべき仮説を求める
- Top-Down :**
  - 原因から結果へ → 推論 causal
  - 例. 生成モデル, 計画, ...
  - ある仮説のもとどのような結果がどのような確率で起こるか?
- Explain Away :**
  - Sprinkler と Rain は、WetGrass が真であることの説明に際し、競合している → この二つは、共通の子供 (WetGrass) が観測されると条件付依存となる

33

## Explaining away effect

ある仮定(または仮定の集合)を支持する証拠が、その証拠とは相容れない(競合する)仮定の確からしさを減少させる効果、またはその現象

Explaining away effect



Call=true が観測されると、Earthquake=true の重要度(確からしさ)も Burglary=true の重要度も上昇する。しかし、Radio=true がさらに観測されると、Earthquake=true の重要度は上昇するが、Burglary=true の重要度は減少する。

to minimize the significance of by or as if by explanation <explains his faults, but does not try to explain them away— M. K. Spears>  
<http://www.merriam-webster.com/dictionary/explain%20away>  
 To dismiss or minimize the significance of (something) by means of an explanation or excuse: There is no way to explain away my carelessness.  
<http://www.thefreedictionary.com/explain+away>

## Explaining away effect 補足

the prior distribution

$$\Pr(X) = \sum_Y \Pr(X, Y)$$

$\Pr(X, Y = y_1) / \alpha \neq \Pr(X)$  as the probability distribution

the posterior distribution when  $Y = y_1$  is observed

$$\text{Note: } \alpha = \sum_x \Pr(X = x, Y = y_1)$$

- 因果関係(原因と結果)を表していると考えべきではない、または、
- 原因の推定方法が(暗黙裡に)書かれていると考え、結果が分からない時と分かっていた時とで推定される原因の分布が異なると考える

35

## 推論 - タイプ分け

- 因果推論  
Causal Inferences  $E \rightarrow O \rightarrow Q$
- 診断推論  
Diagnostic Inferences  $Q \rightarrow O \rightarrow E$
- 原因間推論  
Intercausal Inferences  $Q \rightarrow O$   
 $E \rightarrow O$
- 混合推論  
Mixed Inferences  $E \rightarrow Q \rightarrow E$

36

## 推論 – 結局のところ

- 条件付確率を求めること

$$P(Q|E) = \frac{P(Q, E)}{P(E)}$$

$Q$  と  $E$  は確率変数(または当該確率変数のある値)の集合で、重なりはない

- 注: 周辺分布を求める
- そのためには、周辺分布(のもとになる結合確率)が効率的速に計算できるとよい

37

## Naive な推論

BN で  $P(Q|E=e)$  を解く naive なアルゴリズム

- 条件付確率を全て乗じ、全変数に関する結合確率分布を求める

$$P(Q|E) = \frac{P(Q, E)}{P(E)} = \frac{P(Q, E)}{\sum_q P(Q=q, E)}$$

- BN 構造が使用されず、変数が多いときこのアルゴリズムは実効的ではない
- 一般にこの推論は NP-hard

全然、BN ではない。

38

手計算でやってみよう

## 因果推論 Causal Inferences

原因から結果への推論

例: 窃盗が入ったとして、

$P(J=\text{true}|B=\text{true})?$

$P(A=t|B=t)$

$$\begin{aligned} &= P(A=t, E=t|B=t) + P(A=t, E=f|B=t) \\ &= P(A=t|E=t, B=t)P(E=t|B=t) + P(A|E=f, B=t)P(E=f|B=t) \\ &= (0.95)(0.002) + (0.94)(0.998) \\ &= 0.94 \end{aligned}$$

$$\begin{aligned} P(J|B) &= P(J, A|B) + P(J, \neg A|B) \\ &= P(J|A, B)P(A|B) + P(J|\neg A, B)P(\neg A|B) \\ &= P(J|A)P(A|B) + P(J|\neg A)P(\neg A|B) \\ &= (0.9)(0.94) + (0.05)(0.06) \\ &= 0.85 \end{aligned}$$

略記:  $A$  とは  $A=t$ ,  $\neg A$  とは  $A=f$

同様に  $P(M|B)=0.67$  となる

39

手計算でやってみよう

## 診断推論 Diagnostic Inferences

結果から原因へ。

例: John が電話をした。では  $P(\text{burglary})?$

$P(B|J) = \frac{P(J|B)P(B)}{P(J)}$

$P(J)$  は? まず  $P(A)$  が必要:

$$\begin{aligned} P(A) &= P(A, B, E) + P(A, \neg B, E) + P(A, B, \neg E) + P(A, \neg B, \neg E) \\ &= \\ &= 0.002517 \end{aligned}$$

$$P(J) = P(J, A) + P(J, \neg A)$$

$$\begin{aligned} &= \\ &= 0.052 \end{aligned}$$

$$P(B|J) = \frac{(0.85)(0.001)}{(0.052)} = 0.016$$

false positives 多し

手計算でやってみよう

## 原因間推論 Intercausal Inferences

Explaining away effect が発生する。

Alarm が所与なら、 $P(B|A)=0.37$ 。

そこに Earthquake が真という証拠を加えれば、 $P(B|A, E)=0.003$ 。

すなわち、 $B$  と  $E$  は独立であるが、 $A$  を条件とした条件付独立ではないため、一方に証拠があれば、他方の確率分布は変化する可能性がある

$$\begin{aligned} P(B, \neg A) &= \\ &= 0.00094002 \\ P(B|A) &= P(B, A)/P(A) = 0.3735 \\ P(B, E, A) &= P(B)P(E)(0.95) = 0.00000019 \\ P(E, A) &= \\ &= 0.00058132 \\ P(B|A, E) &= P(B, A, E)/P(A, E) = 0.003268 \end{aligned}$$

手計算でやってみよう

## 混合推論 Mixed Inferences

原因間推論と診断推論を同時に

例: "John calls" かつ "Earthquake=false":

$P(A|J, \neg E) = 0.03$

$P(B|J, \neg E) = 0.017$

この計算はかなり込み入っている

$P(A, J, \neg E) =$

$=$

$= 0.001742$

$P(\neg A, J, \neg E) =$

$=$

$= 0.04980$

$P(A|J, \neg E) = P(A, J, \neg E)/(P(A, J, \neg E) + P(\neg A, J, \neg E))$

$= 0.03379$

42

## 行すべき推論

- 一部の 변수について、その値が観測される  $V_{e_1}, \dots, V_{e_n}$
- 仮に証拠変数と呼ぶ E
- 推論 – 証拠変数以外の 변수  $X_i$  すべてについて、条件付確率  $P(X_i | E)$  を求める(くどいですが、周辺分布)
- 一般には、計算量大 – (NP-hard)
- (ある条件のもと) 厳密値を与える計算方法がある
  - 確率伝播 **belief propagation**
- 近似計算も用いられる

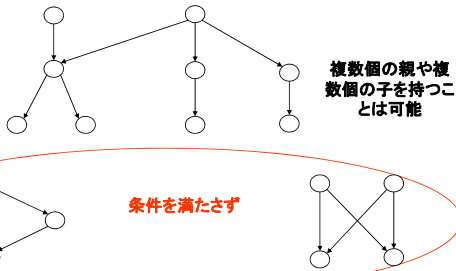
43

## 厳密な計算方法 – 信念伝播

- Judea Pearl, 1982 による
- 単結合グラフ singly-connected graph – どのノード間にもただか一つの無向路しか存在しない – についてのアルゴリズム.
- (下方に、上方に) (確率に基づく量を送る。これをメッセージと呼ぶ。(原理的には) 収束するまで繰り返す(単結合なら必ず収束する)
  - $\pi$ -message: ノード X の上方にある証拠や分布から計算される分布。下方に送られる
  - $\lambda$ -message: ノード X の下方にある証拠や分布から計算される分布。上方に送られる

44

## 単結合グラフ (または Polytrees)

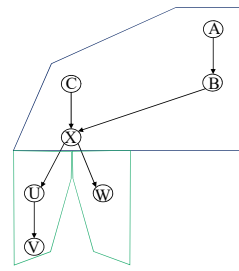


45

## 基本アイデアへの道

目的: 変数 X の分布を求める

ただし、観測された変数がある



- 左図のような簡単な polytree を考える
- ここでの簡単な polytree とは、有向エッジで表される「流れ」を考えると、
  - X の上流では、合流しかなく、
  - X の下流では、分岐しかない

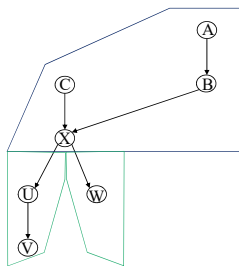
46

## 実際に見てみよう

変数は、X の上流と下流とに分かれ、結合確率は、上流分と下流分の積になる

$$\Pr(A, B, C, X, U, V, W)$$

$$= \Pr(A) \cdot \Pr(B | A) \cdot \Pr(C) \cdot \Pr(X | B, C) \cdot \Pr(U | X) \cdot \Pr(W | X) \cdot \Pr(V | U)$$

$$= \Pr(A, B, C, X) \cdot \Pr(U, V, W | X)$$


X においては合流と分岐があるが、X の上流では合流、X の下流では分岐しかないので、簡単に書ける

$$\Pr(A, B, C, X) = \Pr(A) \cdot \Pr(B | A) \cdot \Pr(C) \cdot \Pr(X | B, C)$$

$$\Pr(U, V, W | X) = \Pr(U, V | X) \cdot \Pr(W | X) = \Pr(V | U) \cdot \Pr(U | X) \cdot \Pr(W | X)$$

47

Skip

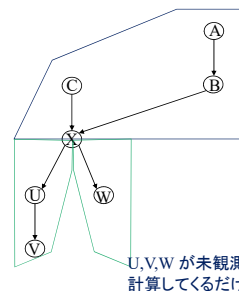
では周辺化してみよう

$$\sum_{A, B, C, U, V, W} \Pr(A, B, C, X, U, V, W)$$

$$= \sum_{A, B, C, U, V, W} \Pr(A) \cdot \Pr(B | A) \cdot \Pr(C) \cdot \Pr(X | B, C) \cdot \Pr(U | X) \cdot \Pr(W | X) \cdot \Pr(V | U)$$

$$= \left( \sum_{A, B, C} \Pr(A, B, C, X) \right) \cdot \left( \sum_{U, V, W} \Pr(U, V, W | X) \right)$$

U, V, W 中に証拠があれば1ではない  
例えば V が証拠  $v_1$  であれば、 $V=v_1$  とし、V に関する和はとらない



$$\sum_{A, B, C} \Pr(A, B, C, X)$$

$$= \sum_{B, C} \Pr(X | B, C) \cdot \Pr(C) \cdot \sum_A \Pr(A) \cdot \Pr(B | A)$$

$$\sum_{U, V, W} \Pr(U, V, W | X)$$

$$= \left( \sum_{U, V} \Pr(U, V | X) \right) \cdot \sum_W \Pr(W | X)$$

W が未観測であれば1  
U, V が未観測であれば1  
U, V が未観測であれば上流から計算してただでよい



Skip  $\sum_{A,B,C,U,W} \Pr(A, B, C, X, U, V = v_1, W)$

では周辺化してみよう  
ただし、観測値がある場合

$$= \sum_{A,B,C,U,W} \Pr(A) \cdot \Pr(B|A) \cdot \Pr(C) \cdot \Pr(X|B,C) \cdot \Pr(U|X) \cdot \Pr(W|X) \cdot \Pr(V = v_1 | U)$$

$$= (\sum_{A,B,C} \Pr(A, B, C, X)) \cdot (\sum_{U,W} \Pr(U, V = v_1, W | X))$$

いずれにせよ、上流部分と下流部分の積

この部分は下から上に伝えないといけない

$$\sum_{A,B,C} \Pr(A, B, C, X)$$

$$= \sum_{A,B,C} \Pr(X|B,C) \cdot \Pr(C) \cdot \Pr(A) \cdot \Pr(B|A)$$

$$= \sum_{B,C} \Pr(X|B,C) \cdot \Pr(C) \cdot (\sum_A \Pr(B|A) \cdot \Pr(A))$$

↑ 上流の方からの積

$$\sum_{U,W} \Pr(U, V = v_1, W | X)$$

$$= \sum_{U,W} \Pr(U, V = v_1 | X) \cdot \Pr(W | X)$$

$$= (\sum_U \Pr(U, V = v_1 | X)) \cdot \sum_W \Pr(W | X)$$

$$= (\sum_U \Pr(V = v_1 | U) \cdot \Pr(U | X)) \cdot \sum_W \Pr(W | X)$$

↓ 下流の方からの積

### 確率変数Xの分布の求め方

予想

- 注目点(ノードX)において、そのより「上流からの計算」と「下流からの計算」で求めた値の積
- その計算方法は、再帰的に適用できそう

50

### Pearlによる定義 メッセージと信念

- 証拠(観測値)を二つに分ける。  $e = e_x^+ \cup e_x^-$
- $\pi(X) = P(X | e_x^+)$  そして  $\lambda(X) = P(e_x^- | X)$  と定義する

$Bel(X) = P(X, E = e)$

$= \alpha P(E_x^- | X) P(X | E_x^+)$

$= \alpha \lambda(X) \pi(X)$

注:  $P(X | E) = \sum_{A_1, \dots, A_n, B_1, \dots, B_m} P(X, A_1, \dots, A_n, B_1, \dots, B_m | E)$

$P(E_x^- | X) = \sum_{B_1, \dots, B_m} P(E_x^-, B_1, \dots, B_m | X)$ ,  $P(X | E_x^+) = \sum_{A_1, \dots, A_n} P(X, A_1, \dots, A_n | E_x^+)$

- $\alpha$  は正規化定数
  - いちいち正規化する必要はない(最後に纏めて行えばよい)
  - ただ、正規化すれば、数値計算時、アンダーフローの可能性は減る

51

### 簡単な場合: 親子が一つ

$Bel(X) = p(x | e_x^+, e_x^-) = \alpha \pi(x) \lambda(x)$

ただし、

$\pi(x) = p(x | e_x^+)$ , 原因(確率)

$\lambda(x) = p(e_x^- | x)$ , 診断(尤度)

$\pi(x) = \sum_U p(x | U, e_x^+) p(U | e_x^+)$

$= \sum_U p(x | U) \pi(U)$

$\lambda(x) = \sum_Y p(e_x^- | x, Y) p(Y | x)$

$= \sum_Y p(e_y^- | y) p(y | x)$

$= \sum_Y \lambda(y) p(y | x)$

(p.158, Pearl 1988)

54

### 簡単な場合

例: 周辺分布  $p(x_5)$  の計算

$$p(x_5) = \sum_{x_3} \sum_{x_4} \sum_{x_2} \sum_{x_1} p(x_1, x_2, x_3, x_4, x_5)$$

$$= \sum_{x_3} \sum_{x_4} \sum_{x_2} p(x_1) p(x_2 | x_1) p(x_3 | x_2) p(x_4 | x_3) p(x_5 | x_3)$$

$$= \sum_{x_3} p(x_3 | x_5) \sum_{x_4} p(x_4 | x_3) \sum_{x_2} p(x_3 | x_2) \sum_{x_1} p(x_1) p(x_2 | x_1)$$

x3に対する下流からの影響

$\lambda_{44}$ ,  $\pi_3$ ,  $\pi_5$

55

### πメッセージ・λメッセージ

Skip  $\sum_{A,B,C,U,W} \Pr(A, B, C, X, U, V = v_1, W)$

$$= \sum_{A,B,C} \Pr(A, B, C, X) \cdot \Pr(U, V = v_1, W | X)$$

$$= (\sum_{A,B,C} \Pr(A, B, C, X)) \cdot (\sum_{U,W} \Pr(U, V = v_1, W | X))$$

$$= \sum_{U,W} \Pr(U, V = v_1, W | X) \cdot \Pr(W | X)$$

$\sum_{A,B,C,U,W} \Pr(A, B, C, X, U, V = v_1, W) = \pi_X \cdot \lambda_X$

$\pi_A = \Pr(A)$

$\pi_{A \rightarrow B} = \pi_A$

$\pi_B = \sum_A \Pr(B | A) \cdot \pi_{A \rightarrow B}$

$\pi_{B \rightarrow X} = \pi_B$

$\pi_X = \sum_{B,C} \Pr(X | B, C) \cdot \pi_{C \rightarrow X} \cdot \pi_{B \rightarrow X}$

$\lambda_X = \lambda_{X \rightarrow U} \cdot \lambda_{X \rightarrow W}$

$\lambda_{X \rightarrow U} = \sum_U \lambda_U \cdot \Pr(U | X)$

$\lambda_U = \lambda_{U \rightarrow V}$

$\lambda_{U \rightarrow V} = \Pr(V = v_1 | U)$

$\pi_{C \rightarrow X} = \pi_C$

$\lambda_{X \rightarrow W} = \sum_W \Pr(W | X)$

$= 1$

Skip

### πメッセージ・λメッセージ (少し一般化)

$Bel(X) = \alpha \cdot \pi_X \cdot \lambda_X$

$\pi_A = \Pr(A)$   
 $\pi_{A \rightarrow B} = \pi_A \cdot (\lambda_{A \rightarrow D} \cdot \lambda_{A \rightarrow D'})$   
 $\pi_B = \sum_A \Pr(B | A) \cdot \pi_{A \rightarrow B}$   
 $\pi_{B \rightarrow X} = \pi_B \cdot (\lambda_{B \rightarrow C} \cdot \lambda_{B \rightarrow C'})$   
 $\pi_X = \sum_{B, B'} \Pr(X | B, B') \cdot \pi_{B \rightarrow X} \cdot \pi_{B' \rightarrow X}$   
 $\lambda_X = \lambda_{X \rightarrow U} \cdot \lambda_{X \rightarrow U'}$   
 $\lambda_{X \rightarrow U} = \sum_U \lambda_U \cdot \sum_{W, W'} \Pr(U | X, W, W') (\pi_{W \rightarrow U} \pi_{W' \rightarrow U})$   
 $\lambda_U = \lambda_{U \rightarrow V} \cdot \lambda_{U \rightarrow V'}$   
 $\lambda_{U \rightarrow V} = \Pr(V = v_1 | U)$

一般の場合: 信念伝播

一般の場合: Polytree

添え字内の→はメッセージの伝播方向ではなく、エッジの方向を表している。

58

Skip

### 一般の場合: λ-メッセージ

以下, Pearl 1988 に比し, 下付添字を一文字多くしている

- 各ノード  $X$  は λ-メッセージを子ノードから受け取る
- λ-メッセージは, 次のように定義される  
 $Y_1, \dots, Y_c$  を  $X$  の子ノード,  
 $e_{\bar{X}} = e_{\bar{X}Y_1} \cup \dots \cup e_{\bar{X}Y_c}$  (排他的) であり  
 $e_{\bar{X}Y_i}$ : ノード  $Y_i$  を含むが  $X$  を含まない部分木上の全証拠 (polytree)  
以下  $Y_i$  を  $Y$  と略記し,  $Y$  の  $X$  以外の親を  $V_1, \dots, V_q$  とする  
 $\lambda_{XY}(x) = p(e_{\bar{X}Y} | x)$  と定める。この時  
 $\lambda_{XY}(x) = \sum_y \lambda_y(y) \sum_{v_1, \dots, v_q} p(y | x, v_1, \dots, v_q) \prod_{k=1}^q \pi_{V_k Y}(v_k)$

$x$  は, この式の中では, 確率変数を表す  
 $X$  はノード名である

59

Skip

### 一般の場合: π-メッセージ

以下, Pearl 1988 に比し, 下付添字を一文字多くしている

- 各ノード  $X$  は π-メッセージを親ノードから受け取る
- π-メッセージは次のように定義される:  
 $U_1, \dots, U_q$  を  $X$  の親ノード,  $e$  を全証拠,  
 $e_{UX}^+$  はノード  $U$  を含むが  $X$  を含まない部分木にある全証拠とし  
 $\pi_{UX}(u) = p(u | e_{UX}^+)$  と定める  
以下  $U_i$  を  $U$  と略記し,  $U$  の  $X$  以外の子  
 $V_1, \dots, V_c$  とする  
 $\pi_X(x) = \sum_{u_1, \dots, u_q} p(x | u_1, \dots, u_q) \prod_{k=1}^q \pi_{U_k X}(u_k)$

60

Skip 送信という観点

### 一般の場合: λ-メッセージ

以下, Pearl 1988 に比し, 下付添字を一文字多くしている

- 各ノード  $X$  は λ-メッセージを子ノードから受け取る
- λ-メッセージは, 次のように定義される  
 $Y_1, \dots, Y_c$  を  $X$  の子ノード,  
 $e_{\bar{X}} = e_{\bar{X}Y_1} \cup \dots \cup e_{\bar{X}Y_c}$  (排他的) であり  
 $e_{\bar{X}Y_i}$ : ノード  $Y_i$  を含むが  $X$  を含まない部分木上の全証拠 (polytree)  
 $X$  の親を  $U_1, \dots, U_q$  とする  
 $\lambda_{XY_j}(x) = p(e_{\bar{X}Y_j} | x)$  等とおく  
 $\lambda_{U_i X}(u_i) = \sum_x \lambda_X(x) \sum_{u_1, \dots, u_{i-1}, u_{i+1}, \dots, u_q} p(x | u_1, \dots, u_q) \prod_{k=1, k \neq i}^q \pi_{U_k X}(u_k)$   
 $\lambda_X(x) = \prod_{j=1}^c \lambda_{XY_j}(x)$  注  $\lambda_{U_i X}(u_i) = p(e_{\bar{U}_i X} | u_i)$

61

Skip 送信という観点

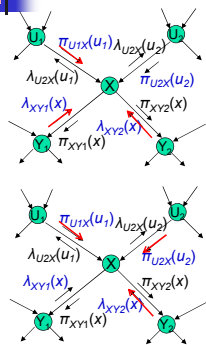
### 一般の場合: π-メッセージ

以下, Pearl 1988 に比し, 下付添字を一文字多くしている

- 各ノード  $X$  は π-メッセージを親ノードから受け取る
- π-メッセージは次のように定義される:  
 $Y_1, \dots, Y_c$  を  $X$  の子ノード,  
 $e_{X Y_i}^+$  はノード  $X$  を含むが  $Y_i$  を含まない部分木にある全証拠とし,  
 $\pi_{X Y_i}(x) = p(x | e_{X Y_i}^+)$  とおけば,  
 $\pi_{X Y_i}(x) = \alpha \prod_{j=1, j \neq i}^c \lambda_{X Y_j}(x) \pi_X(x) = \alpha \frac{Bel(x)}{\lambda_{X Y_i}(x)}$   
 $\pi_X(x) = \sum_{u_1, \dots, u_q} p(x | u_1, \dots, u_q) \prod_{k=1}^q \pi_{U_k X}(u_k)$

62

## λ-メッセージ と π-メッセージ



$$\lambda_{U_i, X}(u_i) = \sum_x \lambda_{U_i, X}(x) \sum_{u_1, \dots, u_{i-1}, u_{i+1}, \dots, u_q} p(x | u_1, \dots, u_q) \prod_{k=1, k \neq i}^q \pi_{U_i, X}(u_k)$$

親一人のときのλ-メッセージ: 親からのπ-メッセージ不要

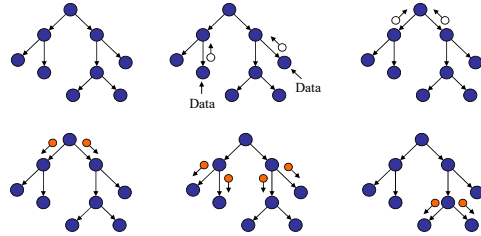
$$\pi_{X, Y_j}(x) = \alpha \prod_{j=1, j \neq i}^q \lambda_{X, Y_j}(x) \pi_X(x) = \alpha \frac{Bel(x)}{\lambda_{X, Y_j}(x)}$$

$$\pi_X(x) = \sum_{u_1, \dots, u_q} p(x | u_1, \dots, u_q) \prod_{k=1}^q \pi_{U_i, X}(u_k)$$

子一人のときのπ-メッセージ: 子からのλ-メッセージ不要

63

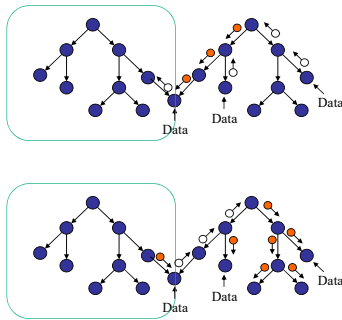
## 伝わり方: もう少し一般的な場合



Pearl, Judea (1982). "Reverend Bayes on inference engines: A distributed hierarchical approach". Proceedings of the Second National Conference on Artificial Intelligence. AAAI-82: Pittsburgh, PA. Menlo Park, California: AAAI Press. pp. 133-136

64

## 伝わり方: もう少し一般的な場合



66

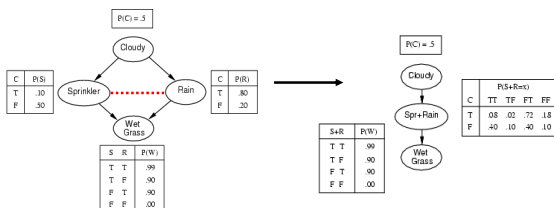
## 複雑度

- 単結合グラフ (polytree) 上では, BP アルゴリズムは収束する。収束速度はグラフの直径に比例する - 高々線形
- 各ノードごとの作業は CPT のサイズに比例する
- 従って BP の計算量はベイジアンネット中のパラメータ数に対し線形である
- 一般のベイジアンネットワークについては
  - 厳密な推論は NP-hard
  - 近似推論も(まともな近似は) NP-hard

69

## 単結合でない場合

- 信念伝播法が正しい値に収束するには、グラフが単結合でなければならない
- 一般的なグラフに対しては、それを junction tree に変換してから適用する方法が考えられている
- ただし、計算複雑度は、変換の結果発生するクラスター数の指数オーダーである → もし最適な junction tree を見出そうとすると、それは NP-hard



70

## 近似アルゴリズム

- なぜ?
  - ループを含むグラフに対して正確な計算を行おうとすると、指数関数時間かかるため
  - また、連続分布を考えた場合、非ガウスであると、message は閉じた形式では表現できないため
- どうやって?
  - 決定的な近似: loopy BP, 平均場近似 (変分ベイズ) 等
  - 統計的近似: MCMC (ギブスサンプリング), 等

- アルゴリズムにより、速度・精度のトレードオフがある(当然!)

71

## ランダムサンプリング Random Sampling

- For  $i = 1$  to  $n$ 
  1.  $X_i$  の親ノード ( $X_{p(i,1)}, \dots, X_{p(i,n)}$ ) を見つける
  2. 当該親ノードにランダムに(このアルゴリズムで)与えられた変数値を読み出す
  3. 次の値を表から読み出す  
 $P(X_i | X_{p(i,1)} = x_{p(i,1)}, \dots, X_{p(i,n)} = x_{p(i,n)})$
  4. この確率に従い  $x_i$  の値をランダムに設定する

72

## 確率的シミュレーション Stochastic Simulation

- 知りたいのは  $P(Q = q | E = e)$
- ランダムサンプリングを大量に行い次の個数を数える
  - $N_e$ :  $E = e$  となるサンプル数
  - $N_{e,q}$ :  $Q = q$  かつ  $E = e$  となるサンプル数
  - $N$ : ランダムサンプルの総数
- $N$  が充分大きければ
  - $N_e / N$  は  $P(E = e)$  の良い推定値
  - $N_{e,q} / N$  は  $P(Q = q, E = e)$  の良い推定値
  - $N_{e,q} / N_e$  は従って  $P(Q = q | E = e)$  の良い推定値

73

## 補足: 連続変数値

- 条件付確率表を考える場合は、離散変数を仮定している
- 連続値変数に対しては、例えば、ガウス分布を仮定する。その場合、平均値と分散を用いることになる
- しかし、基本的には、離散変数を用いる。実際問題として、連続値であっても離散化することが多いからである。とはいえ、離散化のよしあしが結果に大きく影響するので、簡単ではない。

74

## ベイジアンネットワークの学習

75

## 目次: BNの学習(構築)

- 入出力:
  - 入力: 訓練データと事前知識
  - 出力: ベイジアンネットワーク
    - グラフとパラメータ
- 事前知識:
  - 最善(期待できない): ネットワーク構造
  - 変数間の依存関係
  - 事前分布

76

## 場合分け

	構造は既知	構造が未知
完全データ	パラメータの統計的推測 (方程式)	構造を含めて離散最適化 (探索)
不完全データ	パラメータ最適化 (EM, 最急降下,...)	両方 (かなり大変,...)

77

## 構築

BN を構築する手続き:

- 適用領域を記述する変数集合を選ぶ
- 変数の順序を定める
- 空のネットワークから開始し、変数をネットワークに、指定した順序に従い、一個ずつ付加していく
- $i=1$  から順に下記を行う
  - 第  $i$  番目の変数  $X_i$  の付加:
    - すでにネットワーク中にある変数  $(X_1, \dots, X_{i-1})$  の中の変数から  $pa(X_i)$  を  $P(X_i | X_1, \dots, X_{i-1}) = P(X_i | pa(X_i))$  となるように定める
      - 領域知識を用いる
      - データから判断する
    - 有向弧を、 $pa(X_i)$  中の各変数から  $X_i$  に結ぶ

78

## 例: 領域知識を用いて

- 順序: B, E, A, J, M

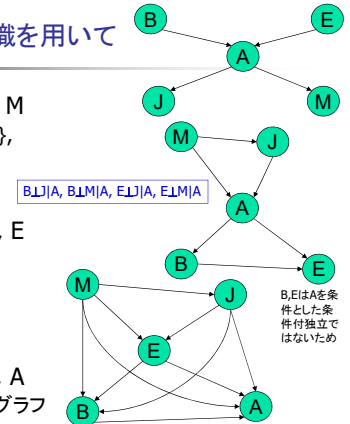
- $pa(B)=pa(E)=\{\}$ ,  
 $pa(A)=\{B,E\}$ ,  
 $pa(J)=\{A\}$ ,  
 $pa(M)=\{A\}$

- 順序: M, J, A, B, E

- $pa(M)=\{\}$ ,  
 $pa(J)=\{M\}$ ,  
 $pa(A)=\{M,J\}$ ,  
 $pa(B)=\{A\}$ ,  
 $pa(E)=\{A,B\}$

- 順序: M, J, E, B, A

- 完全に結合したグラフ



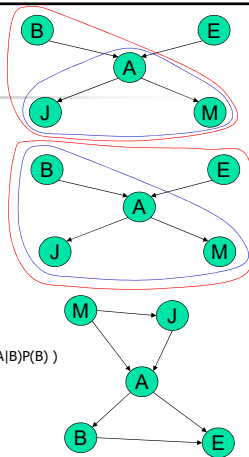
## 例: 説明

- 順序: M, J, A, B, E

$P(J|M)$ , 簡略化できず  
 $P(A|M, J)$ , 簡略化できず

$$\begin{aligned} P(B|M, J, A) &= \frac{P(M, J, A, B)}{P(M, J, A)} \\ &= \frac{P(J|A)P(M|A)P(A|B)P(B)}{(P(M|A)P(J|A)P(A))} \\ &= \frac{P(A, B)}{P(A)} \\ &= P(B|A) \end{aligned}$$

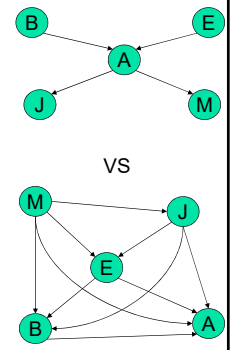
$$\begin{aligned} P(E|M, J, A, B) &= \frac{P(E, M, J, A, B)}{P(M, J, A, B)} \\ &= \frac{P(J|A)P(M|A)P(A|B, E)P(B)P(E)}{(P(J|A)P(M|A)P(A|B)P(B))} \\ &= \frac{P(A, B, E)}{P(A, B)} \\ &= P(E|A, B) \end{aligned}$$



## 変数順序が大切!

どの変数順序を用いるか?

- 視点: 確率を計算する自然な順序.  
 $M, J, E, B, A$  はよくない. なぜなら  $P(B | J, M, E)$  は自然でないから
- 視点: 弧の個数の最小化.  
 $M, J, E, B, A$  は宜しくない (弧が多すぎる), 初めの方がよい
- 視点: 因果関係反映, i.e. 原因が結果の前になる.  
 $M, J, E, B, A$  は宜しくない. というのも  $M$  と  $J$  は  $A$  の結果なのに  $A$  の前に来ている



## 領域知識がないとき

- データから判断する.
  - $P(X_i | X_1, \dots, X_{i-1}) = P(X_i | pa(X_i))$  となる最小の  $pa(X_i)$  を見つける
  - しかし、データの偏りのため、厳密に上記等号が成立することは期待できない
  - そこで、ある程度のエラーを許容することになる。
  - しかし、どれだけ許容したらよいか分からない。
- 様々な情報量規準を用いる
  - データだけ (多項分布を仮定する (後述) ので、実は頻度) を見ても、データ数の不足・統計的偏りのため、条件付独立性は結論できない。
  - 誤差を見込むことになる。どの程度の誤差なら、「条件付独立」と見なすかという間に対して、それによって、簡単になるなら「条件付独立」と見なそうと答える。
  - その時の、残余誤差と簡単さとの trade-off を考え、判断するために、情報量規準を用いる。
  - MDL やベイズアンネットにおけるその精密化である BD (Bayesian Dirichlet) score がよく用いられる
- 説明は「補足」に

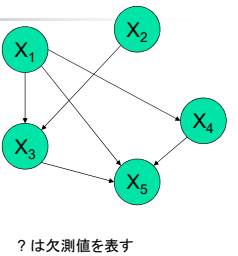
82

## パラメータ学習

例:

- ある BN の構造が所与
- データ集合

$X_1$	$X_2$	$X_3$	$X_4$	$X_5$
0	0	1	1	0
1	0	0	1	0
0	?	0	0	?
...	...	...	...	...



- 条件付確率  $P(X_i | pa(X_i))$  の推定

83

## パラメータの推定

- データには欠測値がないとする
- n 変数  $X_1, \dots, X_n$
- $X_i$  の状態数 or 変数値の数:  $r_i = |\Omega_{X_i}|$
- $X_i$  の親変数の状態総数:  $q_i = |\Omega_{pa(X_i)}|$
- 推定すべきパラメータ:  
 $\theta_{ijk} = P(X_i = j \mid pa(X_i) = k)$ ,  
 $i = 1, \dots, n; j = 1, \dots, r_i; k = 1, \dots, q_i$

84

## 簡単な例

例: BN を一つ. どの変数も2値 1, 2 をとるとする.

$$\theta_{ijk} = P(X_i = j \mid pa(X_i) = k)$$

親変数の状態組合せ

$\theta_{111} = P(X_1=1), \theta_{121} = P(X_1=2)$   
 $\theta_{211} = P(X_2=1), \theta_{221} = P(X_2=2)$   
 $pa(X_3) = 1 : \theta_{311} = P(X_3=1|X_1=1, X_2=1), \theta_{321} = P(X_3=2|X_1=1, X_2=1)$   
 $pa(X_3) = 2 : \theta_{312} = P(X_3=1|X_1=1, X_2=2), \theta_{322} = P(X_3=2|X_1=1, X_2=2)$   
 $pa(X_3) = 3 : \theta_{313} = P(X_3=1|X_1=2, X_2=1), \theta_{323} = P(X_3=2|X_1=2, X_2=1)$   
 $pa(X_3) = 4 : \theta_{314} = P(X_3=1|X_1=2, X_2=2), \theta_{324} = P(X_3=2|X_1=2, X_2=2)$

85

## 要は: 簡単な例

例: BN を一つ. どの変数も2値 1, 2 をとるとする.

$$\theta_{ijk} = P(X_i = j \mid pa(X_i) = k)$$

親変数の状態組合せ

P(X3 X1,X2)		X1, X2			
		1,1	1,2	2,1	2,2
X3	1	$\theta_{311}$	$\theta_{312}$	$\theta_{313}$	$\theta_{314}$
	2	$\theta_{321}$	$\theta_{322}$	$\theta_{323}$	$\theta_{324}$

P(X3 X1,X2)		X1, X2			
		1,1	1,2	2,1	2,2
X3	1	3	5	7	9
	2	7	15	23	31
		10	20	30	40

最尤推定

サンプル数

P(X3 X1,X2)		X1, X2			
		1,1	1,2	2,1	2,2
X3	1	3/10	5/20	7/30	9/40
	2	7/10	15/20	23/30	31/40

86

## BN におけるパラメータ推定

- 次が求まる:

$$\theta_{ijk}^* = \frac{m_{ijk}}{\sum_j m_{ijk}}$$

- 言葉でいえば,  
 $\theta_{ijk} = P(X_i = j \mid pa(X_i) = k)$  の最尤推定量は

$$\frac{X_i=j \text{ かつ } pa(X_i) = k \text{ となる事例数}}{pa(X_i) = k \text{ となる事例数}}$$

しかし、ご存じの通り、ちょっとした問題がある。

87

## BN におけるパラメータ推定

- 実は次の形がよく使われている (Laplace correction):

$$\theta_{ijk}^* = \frac{m_{ijk} + 1}{\sum_j m_{ijk} + r_i}$$

- 言葉でいえば,  
 $\theta_{ijk} = P(X_i = j \mid pa(X_i) = k)$  の最尤推定量は

$$\frac{X_i=j \text{ かつ } pa(X_i) = k \text{ となる事例数} + 1}{pa(X_i) = k \text{ となる事例数} + \lceil X_i \text{ の変数値の個数} \rceil}$$

なお、“+1” や “ $r_i$ ” にはもっと一般的な形がある。  
Dirichlet 分布を事前分布とすることに相当する。

88

補足

## ベイジアンネットワークの学習

領域知識がない場合

89

## BNの学習

BNをデータから構成する方法に2種類ある:

- ・ 制約を発見していく方法
  - 統計的検定を行って、条件付独立な変数組を発見していく
  - これを満たす DAG を見つける
- ・ スコア関数を用いる方法
  - DAG を比較するスコア関数を用いる。  
eg. Bayesian, BIC, MDL, MML
  - データに最もよくfitする DAG を選ぶ

注: 通常、Markov等価性(説明してありません)による制約を考える。というのも、Markov等価なDAGは統計的には区別できないからである。

90

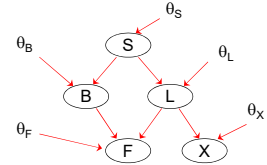
## Bayes的方法(1)

(Cooper and Herskovits, 1992)

データを用いて、条件付独立性に関する統計的推定を行う  
- 確率的関係をよりよく表現するモデルを探す

M - 構造を表す離散確率変数。値 m はありうる DAG 構造。  
Mの値は分布するとする。確率分布を P(m) で表す。

$\theta_m$  - モデル m に対応した連続ベクトル値の確率変数(パラメータ)。値  $\theta_m$  はそのパラメータ値。 $\theta_m$  の値も分布する。確率分布を  $P(\theta_m | m)$  で表す。



G.F. Cooper and E. Herskovits (1992)  
Machine Learning, 9, 309-47

## Bayes的方法(2)

訓練データ集合を D, DAG構造 m の事後確率は、D が与えられたとして:

$$P(m | D) = \frac{P(m)P(D | m)}{\sum_{m'} P(m')P(D | m')}$$

但し

$$P(D | m) = \int P(D | \theta_m, m) P(\theta_m | m) d\theta_m$$

は周辺尤度である。例によって事前分布 P(m) が一様分布であれば

$$P(m | D) \propto P(D | m)$$

従って、尤度最大化は事後確率最大化となる。

92

## Bayes的方法 (3)

Cooper and Herskovits (1992) によれば、周辺尤度は次の通り

$$P(D | m) = \prod_{i=1}^n \prod_{j=1}^{q_i} \frac{\Gamma(\alpha_{ij})}{\Gamma(\alpha_{ij} + N_{ij})} \prod_{k=1}^{r_i} \frac{\Gamma(\alpha_{ijk} + N_{ijk})}{\Gamma(\alpha_{ijk})}$$

n - 全ノード数

$q_i$  - ノード  $X_i$  の親ノード達の値全部の組合せ総数

$r_i$  - ノード(離散確率変数)  $X_i$  の値の総数

$\alpha$  - 事前分布である Dirichlet 分布のパラメータ ( $i$  はノード,  $1 \leq j \leq q_i$ )

N - データ数。ノード  $i$ , 親ノード値の組合せ  $j$ ,  $k$  番目の値

この P(D | m) は Bayesian scoring function として知られている。

G.F. Cooper and E. Herskovits (1992)  
Machine Learning, 9, 309-47

93

## 計算例

次の DAG  $m_1$  と訓練データ D を考える



$P(D | m_1)$  は

$$P(D | m_1) = \prod_{i=1}^n \prod_{j=1}^{q_i} \frac{\Gamma(\alpha_{ij})}{\Gamma(\alpha_{ij} + N_{ij})} \prod_{k=1}^{r_i} \frac{\Gamma(\alpha_{ijk} + N_{ijk})}{\Gamma(\alpha_{ijk})}$$

Y ( $i=2$ ) に対し  $q_i = 2$  (X は2値) かつ  $r_i = 2$  (Yは2値)。j = 1 に対応する項は

$$\frac{\Gamma(2)}{\Gamma(2+5)} \frac{\Gamma(1+4)}{\Gamma(1)} \frac{\Gamma(1+1)}{\Gamma(1)}$$

他の項も計算すれば  $P(D | m_1) = 7.22 \times 10^{-6}$

R.E. Neapolitan, Learning Bayesian Networks (2004)

データID	X	Y
1	1	1
2	1	2
3	1	1
4	2	2
5	1	1
6	2	1
7	1	1
8	2	2

## 計算例 (続)

$m_1$  は、変数 X と Y の間に(条件付)独立性がないことを示す DAG (の Markov同値クラス)の代表と考えることができる。

$m_2$  をエッジがない DAG とすると  $P(D | m_2) = 6.75 \times 10^{-6}$



さらに  $m_1$  と  $m_2$  の事前確率は等しい、すなわち  $P(m_1) = P(m_2) = 0.5$  とすると  $m_1$  の事後確率は  $m_2$  の事後確率より大きくなる。

Bayesの定理により

$$\begin{aligned} P(m_1 | D) &= \frac{P(D | m_1)P(m_1)}{P(D | m_1)P(m_1) + P(D | m_2)P(m_2)} \\ &= \frac{7.215 \times 0.5}{7.215 \times 0.5 + 6.7465 \times 0.5} \\ &= \frac{7.215}{13.4615} \approx 0.536 \end{aligned}$$

95

## 探索アルゴリズムの必要性

理想的には全DAGの空間を網羅的に探索し、前述の Bayesian scoring function を最大化するDAGを見つけたい。

しかし、ノード数を大きく(ほんの少し大きく)しただけで、DAGの数は莫大なものとなる:

ノード数	DAG総数
1	1
2	3
3	25
4	543
5	29,281
10	$4.2 \times 10^{18}$

様々な発見の方法が開発されている

96

## K2 Algorithm (1)

(Cooper and Herskovits, 1992)

$n$  変数  $\{X_1, X_2, \dots, X_n\}$  間に順序があると仮定する。すなわち、 $j > i$  ならば、 $X_j$  は  $X_i$  の親にはなれないとする。

### $X_2$ について

$X_2$  に親がないとして Bayesian score を求める

$X_2$  の親が  $X_1$  として Bayesian score を求める。これがより大きければ  $X_1$  から  $X_2$  へのエッジをつける。

### $X_1$ について

$X_1$  に親がないとして Bayesian score を求める

$X_1$  に親が一つだとして Bayesian score を求める。親がない場合より大きい score があればその最大値を与える  $X_j$  からのエッジをつける。

次に第二番目の親を選んで同様のことを試みる。これを score が大きくなるまで続ける。

## K2 Algorithm (2)

変数の順序を  $\{X, Y, Z\}$  とする

