

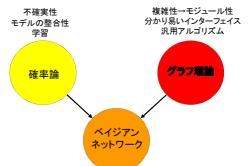
情報意味論 (9) ベイジアンネットワーク

慶應義塾大学理工学部
櫻井 彰人

1

どこから生まれてきたか？

- 実問題の共通課題：
 - 複雑性
 - 不確実性



2

何か？

条件付確率を用いた、結合確率のコンパクトな表現

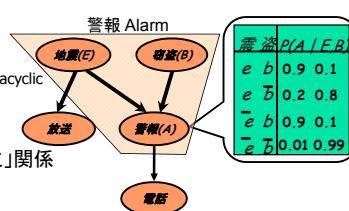
定性的要素：

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

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
 \Rightarrow
 - 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

4

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

これは、全確率変数の結合確率が分かっていればできること

これは、因果関係的な解釈ができる場合

最も可能性が高い説明

- 証拠・現象を説明するシナリオ

これは、因果関係的な解釈ができる場合

合理的な意思決定

- 期待成果を最大化
- 情報の価値

これは、全確率変数の結合確率が分かっていればできること

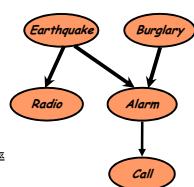
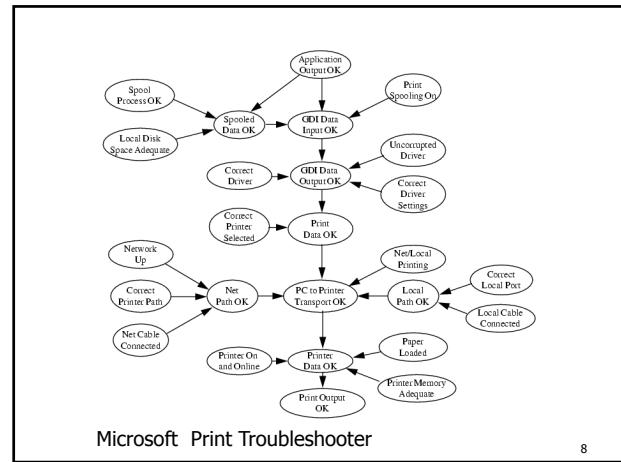
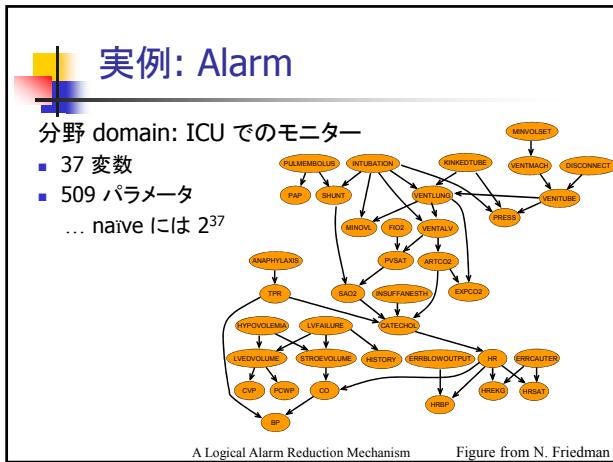


Figure from N. Friedman

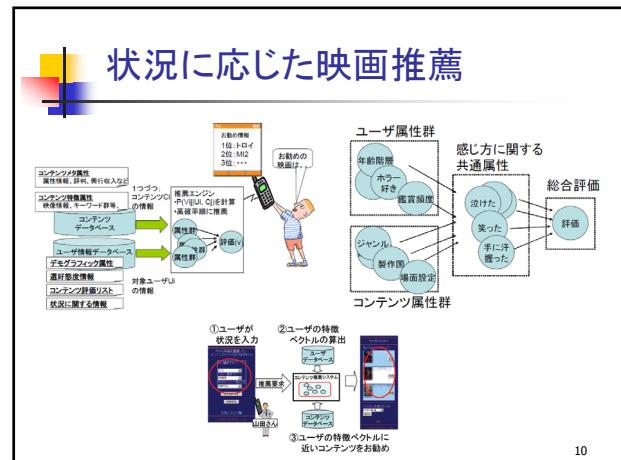
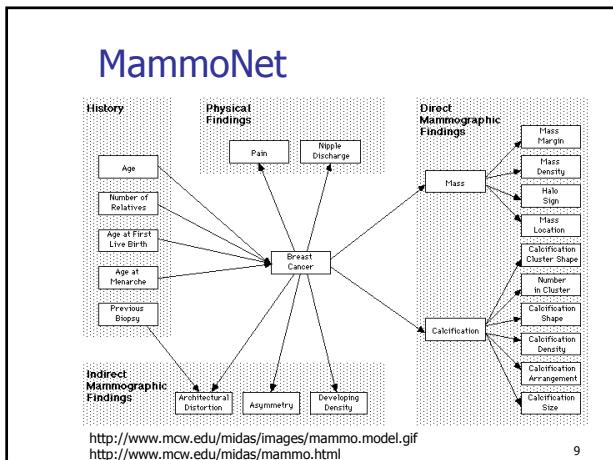
応用事例

- “Microsoft’s competitive advantage lies in its expertise in Bayesian networks”
 $\text{-- Bill Gates, LA Times より, 1996}$
- 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
- Turbo codes (通信路の符号化)

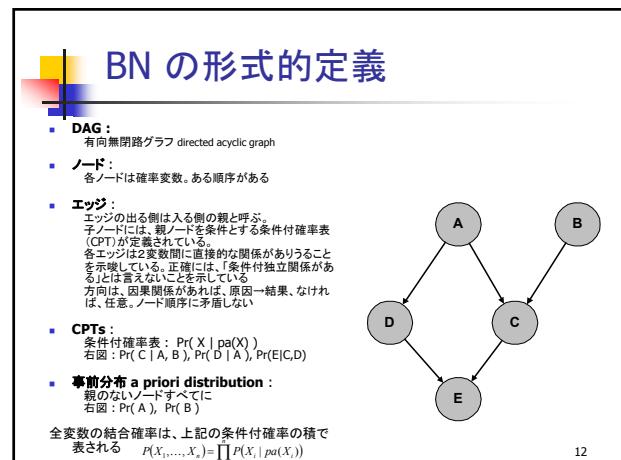
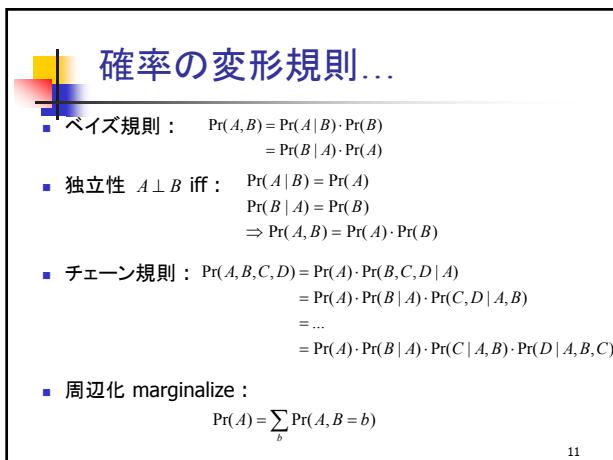
6

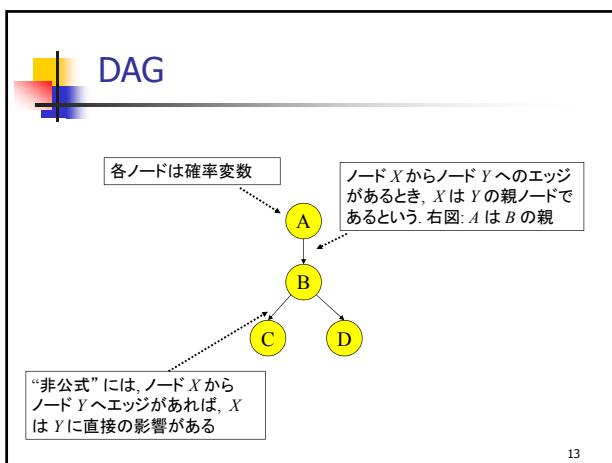


8



10





13

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

表中のパラメータが条件付確率である (CPTs)

A	false	0.6
	true	0.4

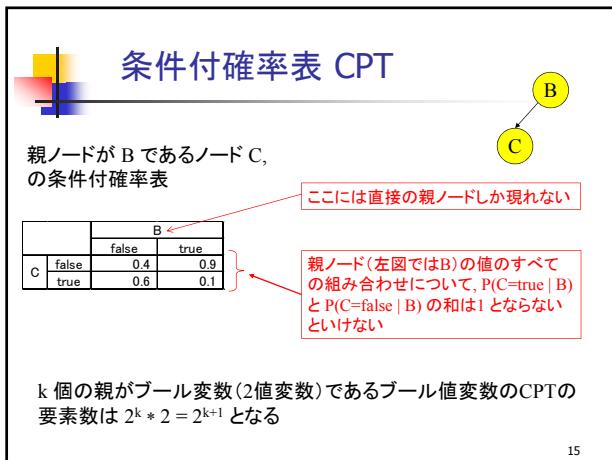
P(B A)		A	false	true
B	false	0.01	0.7	
	true	0.99	0.3	

P(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	

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

14



15

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
2. 各ノードに付随する条件付確率表
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
 $\Pr(D | A, B, C) \cdot \Pr(C | A, B) \cdot \Pr(B | A) \cdot \Pr(A)$

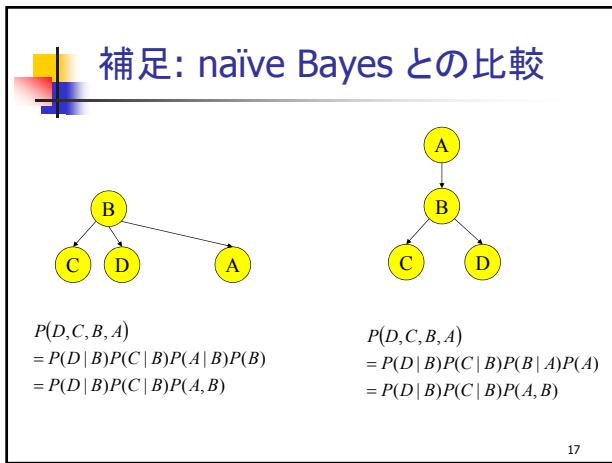
A	false	0.6
	true	0.4

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	

16



17

BN の特徴

主たる2つ:

1. 変数間の条件付独立の関係をグラフ構造で表現する
自分の親(エッジで直接繋がっている)と親以外の先祖(直接つながっていない)を分けて考え、自ノードは親ノードを条件として親以外の先祖ノードに対し、条件付独立
2. 変数間の結合確率をコンパクトに表現する
いろいろの変数間の結合確率・条件付確率が、コンパクトに表現できる

18

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、、、

19

別の計算例

$$\Pr(R = a | WG = b) = \frac{\Pr(R = a, WG = b)}{\Pr(WG = b)}$$

$$\Pr(WG = a, R = b) = \Pr(WG = a | R = b) * \Pr(R = b)$$

Rain	Pr(Rain)
F	0.5
T	0.5

Rain	Wet Grass	Pr(WetGrass, Rain)
F	F	0.50
F	T	0.00
T	F	0.05
T	T	0.45

Pr(Rain WetGrass)	Wet Grass
F	0.91
T	0.09

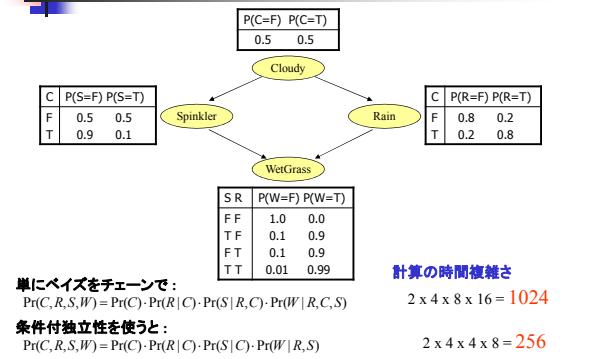
Pr(WetGrass Rain)	Rain
F	1.0
Wet Grass	0.0

Wet Grass	Pr(WetGrass)
F	0.55
T	0.45

$$\Pr(WG = a) = \sum_b \Pr(WG = a, R = b)$$

因果関係ではなく、相関関係または後事後分布の表現

他の例 : Water-Sprinkler



再び、DAGの意味

- DAG: 確率変数間にある半順序が定まっている。
 - つまり、確率変数間に矢印が定まっていて(全変数間である必要はない)、推移律に矛盾しない
- これと矛盾しない全順序がある
 - つまり、全確率変数間に矢印が定まっていて、推移律に矛盾しない。
 - 変数の名前を付け替えて、 $X_1 \rightarrow X_2 \rightarrow \dots \rightarrow X_n$
- そうすると、全変数の結合確率が次のように簡略化されると考える。
 - ただし、 $p(X_i)$ は X_i より上位の(i.e. 矢印が出ている)変数の集合
 - この式から逆にDAGを作ることができる。すなわち、両者は等価

$$\begin{aligned} p(X) &= p(X_1, \dots, X_n) \\ &= p(X_1)p(X_2 | X_1)p(X_3 | X_1, X_2)\dots \\ &= \prod_{i=1}^n p(X_i | X_1, \dots, X_{i-1}) \quad \longrightarrow \quad P(X_1, \dots, X_n) = \prod_{i=1}^n P(X_i | pa(X_i)) \end{aligned}$$

条件付独立

- 実際問題における条件付独立性:
 実際問題では、多くの場合、ある変数集合 $pa(X_i) = \{X_1, \dots, X_{i-1}\}$ を定めることができる。ただし $pa(X_i)$ が与えられたとき、 X_i は $\{X_1, \dots, X_{i-1}\} - pa(X_i)$ に含まれる変数に対して独立、i.e.
 $P(X_i | X_1, \dots, X_{i-1}) = P(X_i | pa(X_i))$
 とする。これを「 X_i と $\{X_1, \dots, X_{i-1}\} - pa(X_i)$ は、 $pa(X_i)$ を条件として、条件付き独立である」という
- このとき

$$P(X_1, \dots, X_n) = \prod_{i=1}^n P(X_i | pa(X_i))$$
- ペイジアンネットは、この式が成立するものと定義する
- なお、変数間の関係が不明なときは、次の変形しかできない

$$\begin{aligned} P(X_1, \dots, X_n) &= P(X_1)P(X_2 | X_1)P(X_3 | X_1, X_2)\dots \\ &= \prod_{i=1}^n P(X_i | X_1, \dots, X_{i-1}) \end{aligned}$$

23

条件付独立

- 条件付独立性の意味:
 例えば、上記の例では、 $P(\text{John Calls} | \text{順序が前の全変数}) = P(\text{John Calls} | \text{Alarm})$ であるが、これは、「John Calls の、順序が前にある全部の変数を条件とした確率を考えるとき、条件としては、Alarmだけを考えればよい」ということを意味している。
- つまり、(殆ど同語反復) Alarm を条件として、John Calls と順序が前の前である全部の変数が条件付独立ということは、Alarm の値(or 分布)が決まれば、これらの変数の値(分布)に関わりなく、John Calls の分布が決まるということである。

24

簡単な例

- 例(続): (順序 $B \rightarrow E \rightarrow A \rightarrow J \rightarrow M$)
 $P(B, E, A, J, M)$
 $= P(B) P(E|B) P(A|B, E) P(J|A, B, E) P(M|B, E, A, J)$
 $= P(B) P(E) P(A|B, E) P(J|A) P(M|A)$
 $= P(M|B, E, A, J) P(J|A, B, E) P(A|B, E) P(E|B) P(B)$
 $= P(M|A) P(J|A) P(A|B, E) P(E) P(B)$
- $pa(B) = \{\}$, $pa(E) = \{\}$, $pa(A) = \{B, E\}$,
 $pa(J) = \{A\}$, $pa(M) = \{A\}$
- 条件付確率表で定めるもの:
 $P(B)$, $P(E)$, $P(A | B, E)$, $P(M | A)$, $P(J | A)$

25

条件付独立性

- $P(A, B | C) = P(A | C) P(B | C)$ ならば、確率変数AとBは、確率変数Cを条件として独立であるという。このとき、
- $P(A | B, C) = P(A, B | C) / P(B | C) = P(A | C)$
- やっかいなのは、 $P(A, B) = P(A) P(B)$ 、すなわち、AとBが独立であっても、あるCに対して、 $P(A, B | C) \neq P(A | C) P(B | C)$ となる、すなわち、確率変数Cを条件とした条件付独立にならないことがある。

代表例は右図

independent causes are made dependent by conditioning on a common effect Pearl 1988

条件付独立性

- 確率変数AとBは、確率変数Cを条件として独立である
 $P(A, B | C) = P(A | C) P(B | C)$ or equivalently
 $P(A | B, C) = P(A | C)$ $P(A | B) = P(A)$: AとBは独立
- 右図では $P(A, B, C) = P(C | A, B) P(A | B)$

A	B	C	$P(A, B, C)$
T	T	T	0.392
T	F	T	0.147
F	T	T	0.147
F	F	T	0.018
T	T	F	0.098
T	F	F	0.063
F	T	F	0.063
F	F	F	0.072

	$P(C A, B)$	$P(A B)$	$P(A)$	$P(B)$
T	0.8	0.49	0.7	0.7
F	0.7	0.21	0.7	0.3
T	0.7	0.21	0.3	0.7
F	0.2	0.09	0.3	0.3
T	0.2	0.49	0.7	0.7
F	0.3	0.21	0.7	0.3
T	0.3	0.21	0.3	0.7
F	0.8	0.09	0.3	0.3

	$P(A B C)$	$P(A C)$	$P(B C)$
T	0.557	0.766	0.766
F	0.209	0.766	0.234
T	0.209	0.234	0.766
F	0.026	0.234	0.234
T	0.331	0.544	0.544
F	0.213	0.544	0.456
T	0.213	0.456	0.544
F	0.243	0.456	0.456

$P(A, B) = P(A) P(B)$ $P(A, B | C) \neq P(A | C) P(B | C)$

注意: $P(A, B) = \sum_c P(A, B, C=c)$, $P(C=T) = 0.704$, $P(C=F) = 0.296$ $P(X)$ は確率分布を表す関数である

条件付独立性

- $P(A, B | C) = P(A | C) P(B | C)$
- $\sum_c P(A, B, C=c) = \sum_c \sum_b P(A, B=b, C=c) \cdot \sum_a P(A=a, B, C=c)$
- $P(A, B | C) = P(A | C) P(B | C)$
- $\forall c \ P(A, B | C=c) = \sum_b P(A, B=b | C=c) \cdot \sum_a P(A=a, B | C=c)$

A	B	C	$P(A, B, C)$
T	T	T	0.392
T	F	T	0.147
F	T	T	0.147
F	F	T	0.018
T	T	F	0.098
T	F	F	0.063
F	T	F	0.063
F	F	F	0.072

	$P(C A, B)$	$P(A B)$	$P(A)$	$P(B)$
T	0.8	0.49	0.7	0.7
T	0.7	0.21	0.7	0.3
F	0.7	0.21	0.3	0.7
F	0.2	0.09	0.3	0.3
T	0.2	0.49	0.7	0.7
F	0.3	0.21	0.7	0.3
T	0.3	0.21	0.3	0.7
F	0.8	0.09	0.3	0.3

C=Tの場合のみ

A	B	C	$P(A, B, C)$
T	T	T	0.392
T	F	T	0.147
F	T	T	0.147
F	F	T	0.018

	$P(C A, B)$	$P(A C)$	$P(B C)$	$P(C)$	$P(A C)$	$P(B C)$
T	0.557	0.766	0.766	0.704	0.539	0.539
F	0.209	0.766	0.234	0.704	0.539	0.165
T	0.209	0.234	0.766	0.704	0.165	0.539
F	0.026	0.234	0.234	0.704	0.165	0.165

BN と条件付独立性

- 一般には:
 $P(E, D, C, A, B) = P(E | D, C) P(D | A) P(C | A, B) P(A | B) P(B)$
- 右図であれば:
 $P(E, D, C, A, B) = P(E | D, C) P(D | A) P(C | A, B) P(A) P(B)$
- 条件付独立らしきところ:
 $P(E | D, C, A, B) = P(E | D, C)$
 $P(D | C, A, B) = P(D | A, B)$
 ちょっと考えると
 $P(C | D, A, B) = P(C | A, B)$
- (直感的に):
 (D,C)を条件として、(A,B)とEは条件付独立
 (D,C)を何でもよいから定めると E の分布は定まり、当然、(A,B)の値によらない。
 [A,B]を条件としてCとDは条件付独立である。
 ただし、(A,B,E)を条件とした条件付独立である、とはいえない。
 (A,B)の値が定まると、CとDの分布はいずれも、Eの値に依存しているからである。

29

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、証拠変数がさえ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。

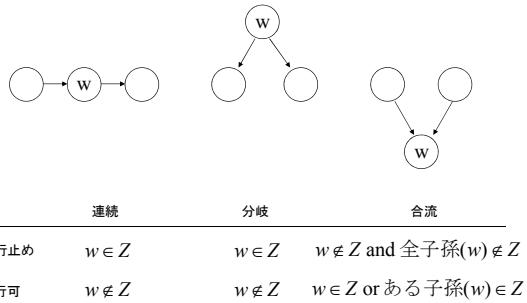
30

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
 $(A \perp\!\!\!\perp B | Z)$ 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
- 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。

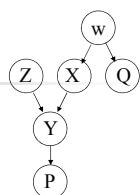
31

通行止め



32

例



正しい関係

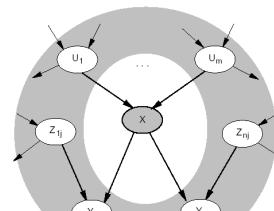
D separation による説明

- $(Q \perp\!\!\!\perp X, Y, Z, P | W)$: $Q \leftarrow W \rightarrow X$ は分岐. Wを条件として通行止め
- $(Z \perp\!\!\!\perp X, W, Q | \emptyset)$: $Z \rightarrow Y \leftarrow X$ は合流. Y及びその子孫Pを条件としないので通行止め。
- $(Z \not\perp\!\!\!\perp X, W, Q | P)$: $Z \rightarrow Y \leftarrow X$ は合流. Yの子孫Pを条件としているので通行可能。
- $(Z, Y, P \perp\!\!\!\perp W, Q | X)$: $W \rightarrow X \rightarrow Y$ は連続. Xを条件として通行止め。
- $(Z, Y, P \not\perp\!\!\!\perp W, Q | \emptyset)$: $W \rightarrow X \rightarrow Y$ は連続. Xを条件としていないので通行可能。

33

Markov Blanket

- Markov blanket: 親 + 子供 + 子供の親
- (中心にある)ノードは、Markov blanket 内の変数を条件として、ネットワーク内のどの変数からも、条件付独立である



34

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ふ
 - 一般に、推論では次の形のクエリーが扱われる:
 $P(X | E)$
- E = 証拠 evidence 変数
X = 問い合わせる変数

35

推論

- インフルエンザ
-
- クエリーは、例えば:
 $P(\text{インフルエンザ} = \text{true} | \text{発熱} = \text{true}, \text{急性症状} = \text{true})$
 - 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 (ie. 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

36

BNにおける推論

他の例 : Water-Sprinkler

BNにおける推論

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?

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

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

Rain が真であるのが理由である可能性がより高い

計算の時間複雑度
 $2^4 \times 3^8 \times 16 = 1024$

計算の時間複雑度
 $2^4 \times 3^8 \times 16 = 1024$

2.4 x 4 x 8 = 256

37

BNにおける推論 (2)

Bottom-Up :

- 結果から原因へ → 診断 diagnostic
- 例. エクスペリメントシステム, 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うべき仮説を求める

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

38

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

Call=true が観測されると、Earthquake=true への信頼度も Burglary=true への信頼度も上昇する。しかし、Radio=true がさらに観測されると、Earthquake=true への信頼度は上昇するが、Burglary=true への信頼度は減少する。

Explaining away effect

40

推論 – まとめると

- 因果推論 Causal Inferences
- 診断推論 Diagnostic Inferences
- 原因間推論 Intercausal Inferences
- 混合推論 Mixed Inferences

40

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

- そのためには、結合確率が高速に計算できるとい

41

Naïve な推論

BN で $P(Q|E = e)$ を解く naïve なアルゴリズム

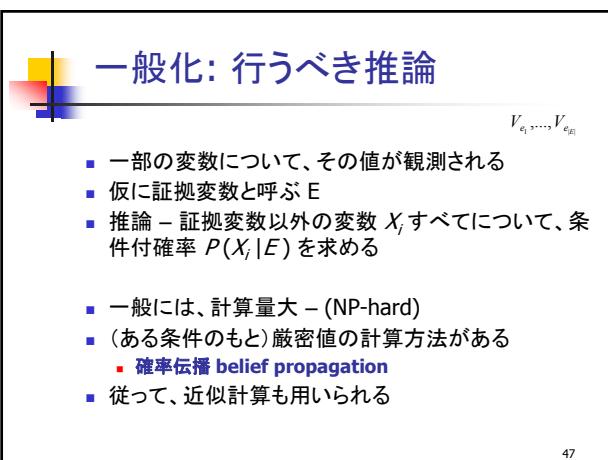
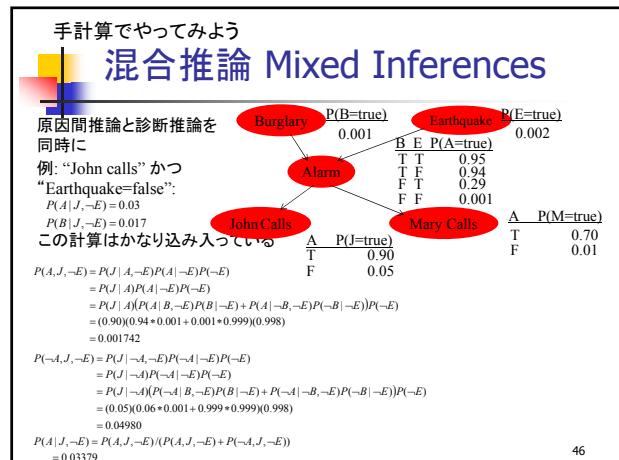
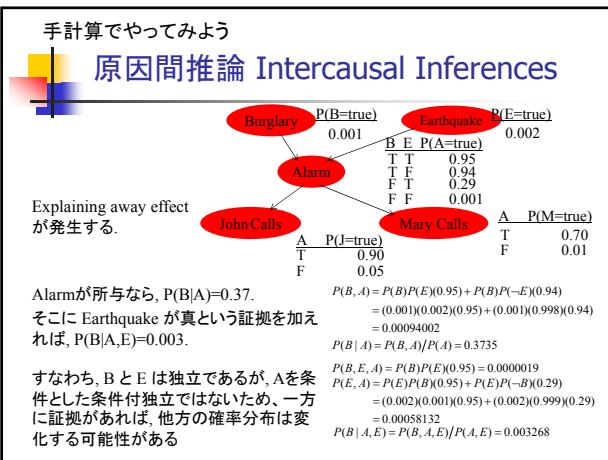
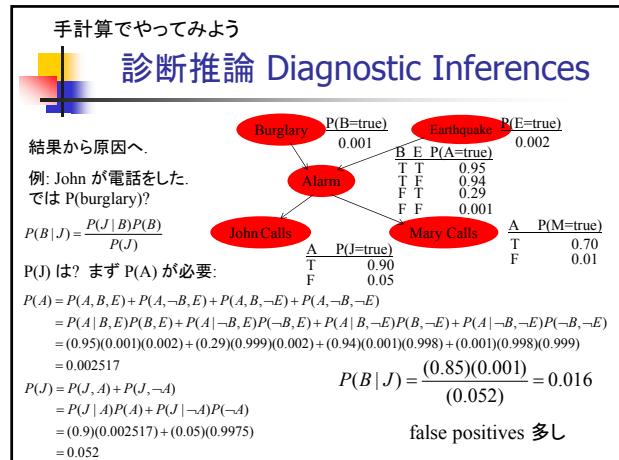
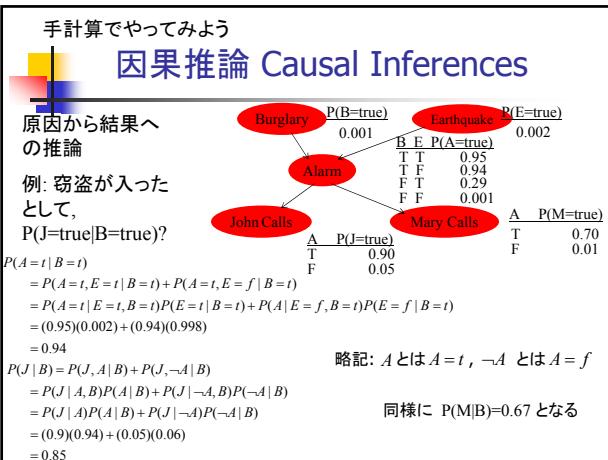
- 条件付確率を全て乗じ、全変数に関する結合確率分布を求める

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

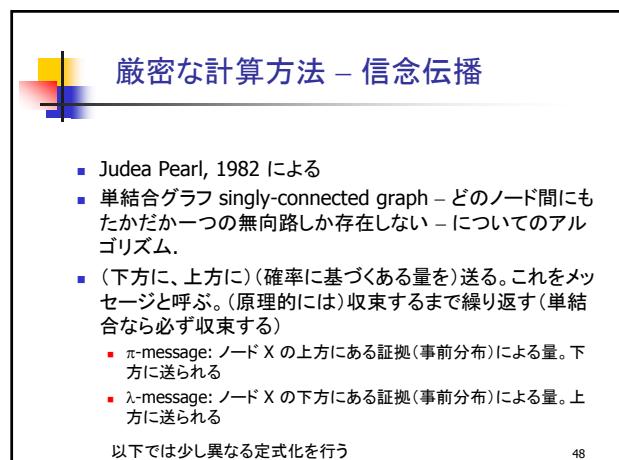
- BN 構造が使用されず、変数が多いときこのアルゴリズムは実効的ではない
- 一般にこの推論は NP-hard

全然、BN ではない。

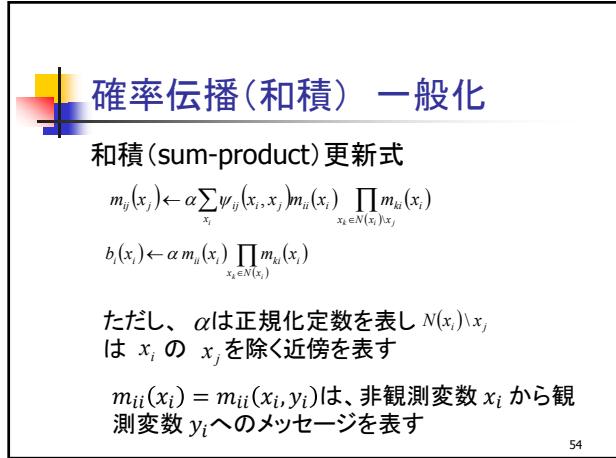
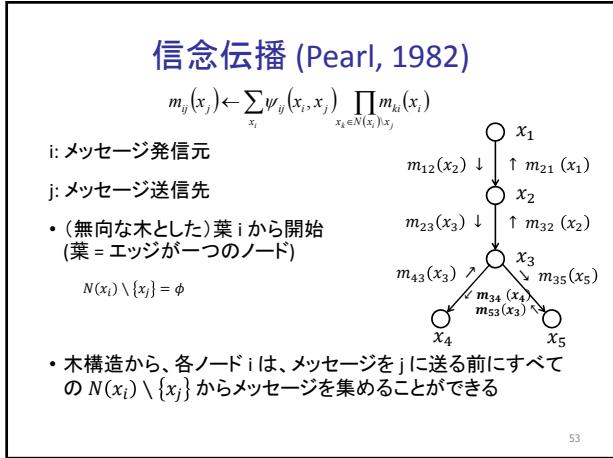
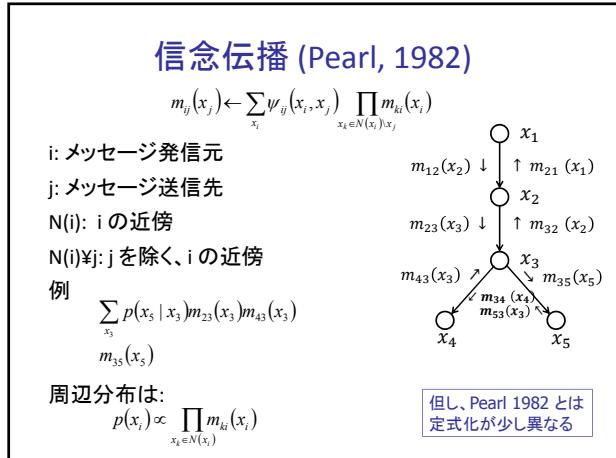
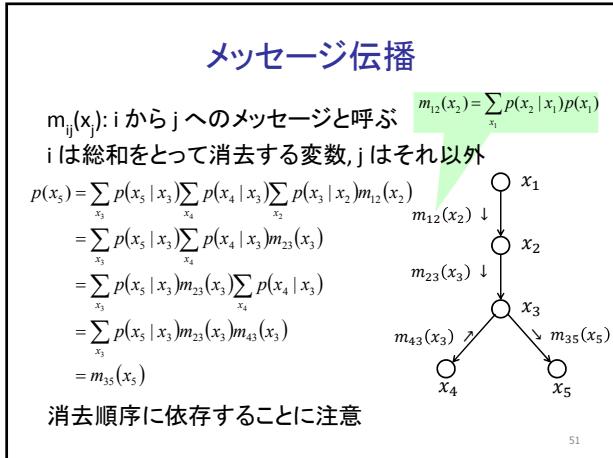
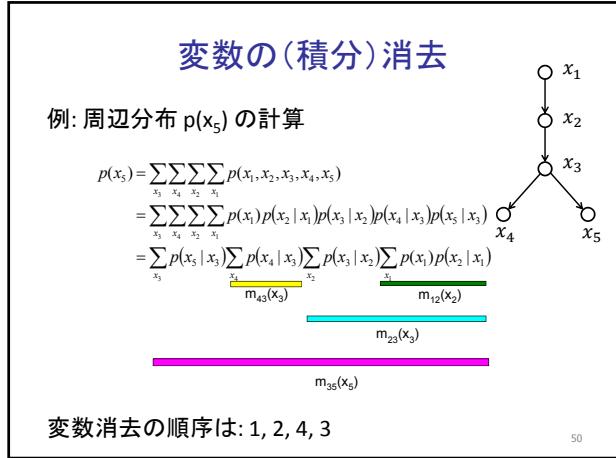
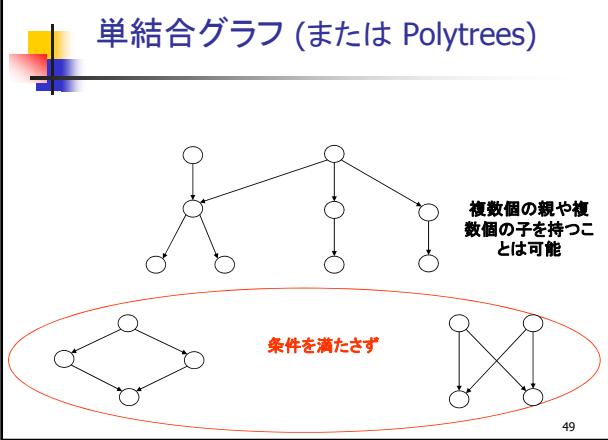
7



47



48



確率伝播(最大-積)

最大-積(max-product)更新式

$$m_{ij}(x_j) \leftarrow \alpha \max_{x_i} \Psi_{ij}(x_i, x_j) m_{ii}(x_i) \prod_{x_k \in N(x_i) \setminus x_j} m_{ki}(x_i)$$

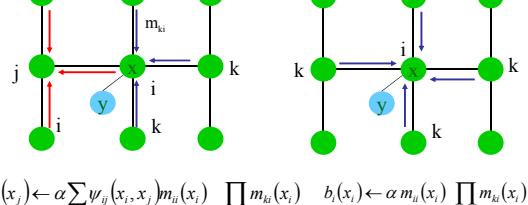
$$b_i(x_i) \leftarrow \alpha m_{ii}(x_i) \prod_{x_k \in N(x_i)} m_{ki}(x_i)$$

ただし、 α は正規化定数を表し $N(x_i) \setminus x_j$ は x_i の x_j を除く近傍を表す

$m_{ii}(x_i) = m_{ii}(x_i, y_i)$ は、非観測変数 x_i から観測変数 y_i へのメッセージを表す

55

確率伝播- 図示



$$m_{ij}(x_j) \leftarrow \alpha \sum_{x_i} \Psi_{ij}(x_i, x_j) m_{ii}(x_i) \prod_{x_k \in N(x_i) \setminus x_j} m_{ki}(x_i)$$

観測可能変数による証拠

$$b_i(x_i) \leftarrow \alpha m_{ii}(x_i) \prod_{x_k \in N(x_i)} m_{ki}(x_i)$$

56

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

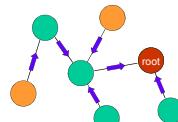
57

補足: 伝播の仕方

あるノードを選び、方向は無視して、それを根とする木を考える

- 2 パス : 収集し分配する
- Poly-tree に対してのみ有効

証拠の収集



証拠の分配

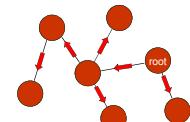
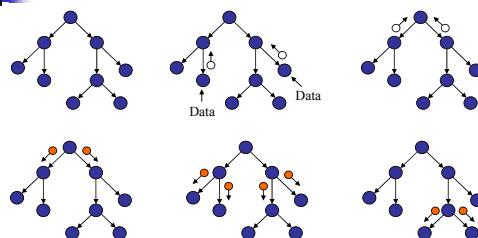


Figure from P. Green

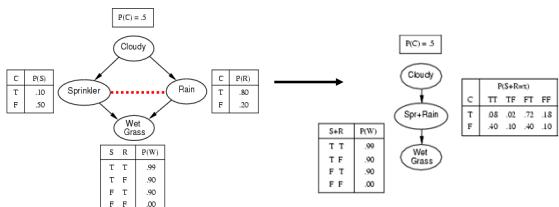
例



59

より一般的なグラフでは

- 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數オーダーである
→ もし最適な junction tree を見出そうとすると、それは NP-hard



60

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(ギップスサンプラー), 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！)

61

ランダムサンプリング Random Sampling

For $i = 1$ to n

- X_i の親ノード($X_{p(i, 1)}, \dots, X_{p(i, n)}$)を見つける
- 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
- 次の値を表から読み出す
 $P(X_i | X_{p(i, 1)}=x_{p(i, 1)}, \dots, X_{p(i, n)}=x_{p(i, n)})$
- この確率に従い x_i の値をランダムに設定する

62

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_c : $E = e$ となるサンプル数
 - N_s : $Q = q$ 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_c / N は $P(E = e)$ の良い推定値
 - N_s / N は $P(Q = q, E = e)$ の良い推定値
 - N_s / N_c は従って $P(Q = q | E = e)$ の良い推定値

63

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。とはいっても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。

64

BNの学習(構築) ad hocに

- 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)：ネットワーク構造
 - 変数間の依存関係
 - 事前分布

65

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,...)

66

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

67

例: 領域知識を用いて

順序: B, E, A, J, M

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

$B \perp\!\!\!\perp A, B \perp\!\!\!\perp M | A, E \perp\!\!\!\perp J | A, E \perp\!\!\!\perp M | A$

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

$P(J|M), \text{ 簡略化できず}$
 $P(A|M, J), \text{ 簡略化できず}$

$P(B|M, J, A)$
 $= P(M|A)P(M|J) / P(M, J, A)$
 $= P(J|A)P(M|A)P(A|B)P(B) / (P(M|A)P(J|A)P(A))$
 $= P(A, B) / P(A)$
 $= P(B|A)$

$P(E|M, J, A, B)$
 $= P(E|M, J, A, B) / P(M, J, A, B)$
 $= P(J|A)P(M|A)P(A|B)P(B)P(E) / (P(J|A)P(M|A)P(A|B)P(B))$
 $= P(A, B, E) / P(A, B)$
 $= P(E|A, B)$

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(弧が多くすぎる), 初めの方がよい
- 視点: 因果関係反映, i.e. 原因が結果の前にくる.
 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が分からない。
- 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やベイジ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
- 今回は説明省略

71

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	?
...

? は欠測値を表す

- 条件付確率 $P(X_i | pa(X_i))$ の推定

72

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: $q_i = |\Omega_{\text{pa}(X_i)}|$
- 推定すべきパラメータ:
 $\theta_{ijk} = P(X_i = j | \text{pa}(X_i) = k)$,
 $i = 1, \dots, n; j = 1, \dots, r_i; k = 1, \dots, q_i$

73

簡単な例

例: BN を一つ。どの変数も2値 1, 2 をとるとする。

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

親変数の状態組合せ

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

74

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。

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

親変数の状態組合せ

		X1, X2			
		1,1	1,2	2,1	2,2
X3	1	θ311	θ312	θ313	θ314
	2	θ321	θ322	θ323	θ324

		X1, X2			
		1,1	1,2	2,1	2,2
X3	1	3	5	7	9
	2	7	15	23	31

75

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,

$\theta_{ijk} = P(X_i = j | \text{pa}(X_i) = k)$ の最尤推定量は

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

しかし、ご存じの通り、ちょっとした問題がある。

76

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,

$\theta_{ijk} = P(X_i = j | \text{pa}(X_i) = k)$ の最尤推定量は

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

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。

77

ベイジアンネットワークの学習

少し数学的に

78

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(説明していません)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。

79

Bayes的方法(1)

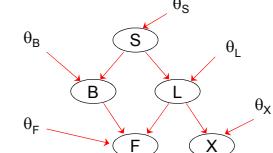
(Cooper and Herskovits, 1992)

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.

θ_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 θ_m はそのパラメータ値. θ_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.

81

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 X_i の離散確率変数 X_i の値の総数

α - 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; 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

82

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

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

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

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

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

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

$$\begin{aligned} P(m_1 | D) &= \frac{P(D | m_1)P(D | m_1)}{P(D | m_1)P(D | m_1) + P(D | m_2)P(D | m_2)} \\ &= \frac{7.215 \times 0.5}{7.215 \times 0.5 + 6.7465 \times 0.5} \\ &= 0.517 \end{aligned}$$

84

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を見つけてたい。

しかし、ノード数を大きく(ほんの少しだけ)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	4.2×10^{18}

様々な発見的方法が開発されている

85

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_i 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が大きくならなければ止める。

K2 Algorithm (2)

変数の順序を $\{X, Y, Z\}$ とする

