# 情報意味論 (第12回)
## 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

## イントロ

- 対象とする問題領域は、変数リスト $X_1, \ldots, X_n$ で表現される
- 問題領域に関する知識は結合確率 $P(X_1, \ldots, X_n)$ で表現される

## イントロ

例: Alarm

- 物語: LA では窃盗(burglary)と地震(earthquake) は珍しいことではない. どちらも警報をならす可能性がある. 警報が鳴ると, 隣人の二人 John と Mary が電話をする
- 問題: 誰かが電話したまたは電話しなかったという条件のもとで窃盗が起こった確率を推定せよ
- 変数: Burglary (B), Earthquake (E), Alarm (A), JohnCalls (J), MaryCalls (M)
- 問題をとくに必要な知識:
  P(B, E, A, J, M)

$P(B, E, A, J, M)$

| B | E | A | J | M | Prob | B | E | A | J | M | Prob |
|---|---|---|---|---|---|---|---|---|---|---|---|
| y | y | y | y | y | .00001 | n | y | y | y | y | .0002 |
| y | y | y | y | n | .000025 | n | y | y | y | n | .0004 |
| y | y | y | n | y | .000025 | n | y | y | n | y | .0004 |
| y | y | y | n | n | .00000 | n | y | y | n | n | .0002 |
| y | y | n | y | y | .00001 | n | y | n | y | y | .0002 |
| y | y | n | y | n | .000015 | n | y | n | y | n | .0002 |
| y | y | n | n | y | .000015 | n | y | n | n | y | .0002 |
| y | y | n | n | n | .0000 | n | y | n | n | n | .0002 |
| y | n | y | y | y | .00001 | n | n | y | y | y | .0001 |
| y | n | y | y | n | .000025 | n | n | y | y | n | .0002 |
| y | n | y | n | y | .000025 | n | n | y | n | y | .0002 |
| y | n | y | n | n | .0000 | n | n | y | n | n | .0001 |
| y | n | n | y | y | .00001 | n | n | n | y | y | .0001 |
| y | n | n | y | n | .00001 | n | n | n | y | n | .0001 |
| y | n | n | n | y | .00001 | n | n | n | n | y | .0001 |
| y | n | n | n | n | .00000 | n | n | n | n | n | .996 |

## イントロ

- Maryが電話をしたという条件のもとでの窃盗があった確率 P(B = y | M = y) は?
- 周辺確率を計算し
  P(B , M) = $\sum_{E, A, J}$ P(B, E, A, J, M)
- 条件付確率の定義を用いよ
- 答え:

| B | M | Prob |
|---|---|---|
| y | y | .000115 |
| y | n | .000075 |
| n | y | .00015 |
| n | n | .99971 |

$$P(B = y \mid M = y) = \frac{P(B = y, M = y)}{P(M = y)} = 0.61$$

## イントロ

- 難しさ: モデル構築と推論の複雑さ
- Alarm の例:
  - 31 個の数値が必要
  - P(B = y | M = y) の計算に 29 回の加算が必要
- 一般に
  - $P(X_1, \ldots X_n)$ なる結合確率を指定するには最小 $2^n - 1$ 個の数値が必要
  - (変数の個数に対し) 指数関数的な記憶域と推論時間が必要

## 条件付独立

- 指数関数的サイズの問題を条件付独立性を用いて克服する
- 確率の chain rule:

$$p(x) = p(x_1,...,x_n)$$
$$= p(x_1)p(x_2 \mid x_1)p(x_3 \mid x_1,x_2)\cdots$$
$$= \prod_{i=1}^{n} p(x_i \mid x_1,...,x_{i-1})$$

## 条件付独立

- 問題領域における条件付独立性:問題領域では一般にある変数集合 $pa(X_i) \subseteq \{X_1, \ldots, X_{i-1}\}$ 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 $\{X_1, \ldots, X_{i-1}\} - pa\{X_i\}$ に含まれる変数に対して独立, i.e.
  $P(X_i \mid X_1, \ldots, X_{i-1}) = P(X_i \mid pa(X_i))$
  とする. このとき

$$P(X_1,\ldots,X_n) = \prod_{i=1}^{n} P(X_i \mid pa(X_i))$$

## 条件付独立

- その結果, 結合確率 $P(X_1, \ldots, X_n)$ は条件付確率 $P(X_i \mid pa(X_i))$ で表現することができる
- 例 (続) :
  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)
- pa(B) = {}, pa(E) = {}, pa(A) = {B, E}, pa{J} = {A}, pa{M} = {A}
- 条件付確率表が定めるもの:
  P(B), P(E), P(A | B, E), P(M | A), P(J | A)

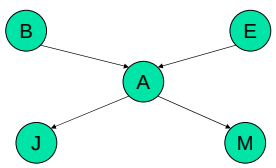
## 条件付独立

この結果:

- モデルサイズがより小さくなる
- モデル構築がより容易になる
- 推論がより容易になる

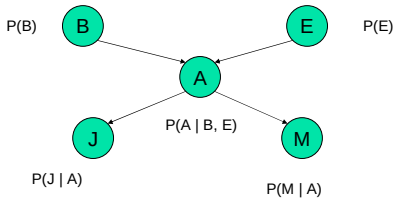
## グラフ表現

- 条件付独立性の関係をグラフで表現するには, 有向グラフを用い, $X_j$ から $X_i$ への有向弧を, $X_j \in pa(X_i)$ かつそのときに限り, 描けばよい
- pa(B) = {}, pa(E) = {}, pa(A) = {B, E}, pa{J} = {A}, pa{M} = {A}




## グラフ表現

- ノード $X_i$ には条件付確率表 $P(X_i \mid pa(X_i))$ もおく
- 結果: ベイジアンネットワーク Bayesian network



## 形式的定義

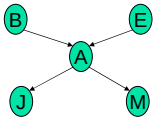
一つの Bayesian network は:

- サイクルのない有向グラフ (DAG) であって
- 各ノードは確率変数を表し
- 各ノードには、当該ノードの親ノード (の表す確率変数) の条件付確率表が付随したもの

## 直感的には

一つの BN は各弧が直接の確率的依存性を表現する DAG であると考えることができる

- 弧でつながっていなければ確率的に独立: 変数は, 親が与えられたとき, 子孫以外から条件付独立である
- グラフから: $B \perp E,\ J \perp B \mid A,\ J \perp E \mid A$
- 正しくない: $J \perp B,\ J \perp E$




## 構築

BN を構築する手続き:

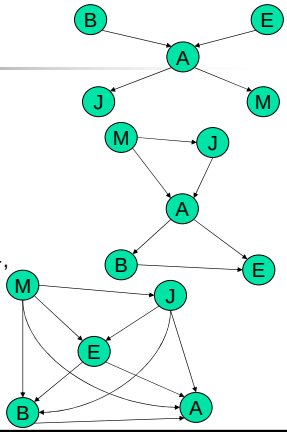
- 応用領域を記述する変数集合を選ぶ
- 変数の順序を定める
- 空のネットワークから開始し、変数をネットワークに、指定した順序に従い、一個ずつ付加していく

## 構築

- 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($X_1, \ldots, X_{i-1}$) の中の変数から $pa(X_i)$ を $P(X_i \mid X_1, \ldots, X_{i-1}) = P(X_i \mid pa(X_i))$ となるように定める(領域知識が必要)
  - 有向弧を、$pa(X_i)$ 中の各変数から $X_i$ に結ぶ

## 例

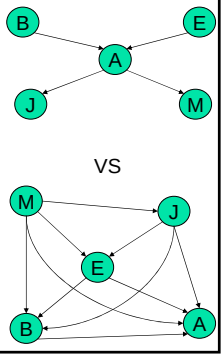
- 順序: B, E, A, J, M
  - pa(B)=pa(E)={},
    pa(A)={B,E},
    pa(J)={A},
    pa{M}={A}
- 順序: M, J, A, B, E
  - pa{M}={}, pa{J}={M},
    pa{A}={M,J},
    pa{B}={A},
    pa{E}={A,B}
- 順序: M, J, E, B, A
  - 完全に結合したグラフ




## 構築

どの変数順序を用いるか?

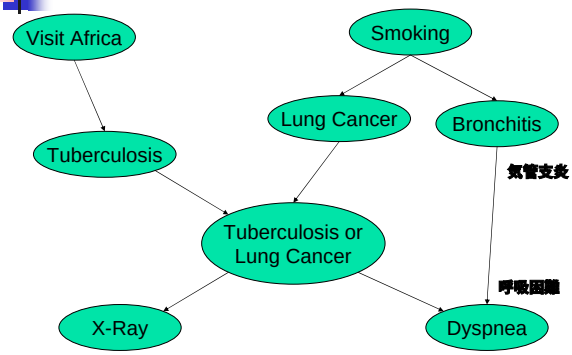
- 確率を計算する自然な順序
  M, J, E, B, A はよくない. なぜなら P(B | J, M, E) は自然でないから
- 弧の個数を最小化する
  M, J, E, B, A は宜しくない (弧が多すぎる), 初めの方がよい
- 因果関係を用いる: 原因が結果の前にくる
  M, J, E, B, A は宜しくない. というのも M と J は A の結果なのに A の前に来ている



## 因果的 Bayesian Networks

- 因果 Bayesian network, または単に因果ネットワークは, Bayesian network であってその弧が因果関係を表すと解釈できるもの
  - 一般に BN は因果ネットワークではない
- 因果ネットワークの構築:
  - 応用領域を記述する変数を選ぶ
  - 直接原因 (を表す変数) から当該変数に有向弧を結ぶ (領域知識が必要)

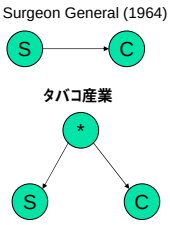
## 例




## 因果 BN

- 因果関係は, 良く理解された概念というわけではない.
  - 広く受容れられる定義はない.
  - 世界の性質なのか我々の心の中の性質なのか, 共通認識はない
- 時には, 因果関係は明確:
  - 警報が人間をビルから退避させる.
  - 肺がんは, 胸部X線の瘤の原因となる.
- 多くの場合, 因果関係は明確ではない.

医者は喫煙は肺がんを引起すと信じている. しかしタバコ産業は異なったストーリを持っている:




## 推論

- BN への事後確率問合せ
  - ある変数の値を観測したとしよう
  - そのとき, 他の変数の事後確率分布はどうなるであろうか ?
- 例: John も Mary も警報を聞いたという
  - 窃盗が入った確率 P(B|J=y,M=y) は?

## 推論

- 問合せの一般形 P(Q | E = e) = ?
- Q は問合せる変数のリスト
- E は証拠となる (観測された) 変数のリスト
- e は観測された変数値

## さまざまな推論の形

- 診断推論: P(B | M = y)
- 予測/因果推論: P(M | B = y)
- 原因間推論 (共通の結果の原因間)
  P(B | A = y, E = y)
- 混合推論 (上記のものの混合)
  P(A | J = y, E = y) (診断と因果)
- このすべての型が同一の推論方法で取り扱うことができる

## Naïve な推論

BN で P(Q|E = e) を解く naïve なアルゴリズム

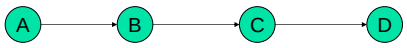
- 条件付確率を全て乗じ, 全変数に関する結合確率分布を求める

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

- 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

## 簡単な例

- 条件付確率: P(A),P(B|A),P(C|B),P(D|C)
- 問合せ: P(D) = ?
- P(D) = $\sum_{A, B, C}$ P(A, B, C, D)
  = $\sum_{A, B, C}$ P(A)P(B|A)P(C|B)P(D|C) (1)
  = $\sum_C$P(D|C) $\sum_B$P(C|B) $\sum_A$P(A)P(B|A) (2)
- 複雑性:
  - (1) を使用: $2^3 + 2^2 + 2$
  - (2) を使用: 2 + 2 + 2




## 推論

- 一般に正確な推論は NP-hard であるが, ある場合には tractable になる. 例えば, もし BN の構造が (poly)-tree であれば効率的なアルゴリズムが存在する
  (poly tree とは, サイクルのない有向グラフで, どの2つのノードもそれをつなぐ道がたかだか一つしかないもの)
- 他の実用的な手法: 確率的シミュレーション

## 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($X_{p(i,1)}$, …, $X_{p(i,n)}$ ) 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($X_i$ | $X_{p(i,1)}$= $x_{p(i,1)}$, …, $X_{p(i,n)}$= $x_{p(i,n)}$ )
  4. この確率に従い $x_i$ の値をランダムに設定する

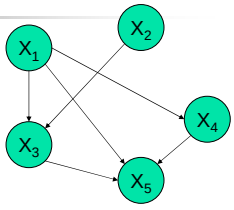
## 確率的シミュレーション Stochastic Simulation

- 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

## 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 | ? |
| … | … | … | … | … |




? は欠損値を表す

- 条件付確率 P($X_i$ | pa($X_i$)) の推定

## パラメータ学習

- 欠損値がない場合を考える
- 最尤推定 (ML) アルゴリズムとベイズ推定を用いる
- 学習のモード:
  - オンラインモード
  - バッチモード
- ベイズ推定はどちらのモードにも適している
- ML はバッチモードに適している

## BN における ML

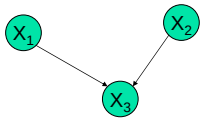
- データには欠損がないとする
- n 変数 $X_1, \ldots, X_n$
- $X_i$ の状態数 : $r_i = |\Omega_{X_i}|$
- $X_i$ の親変数の状態総数: $q_i = |\Omega_{pa(X_i)}|$
- 推定すべきパラメータ数:
  $\theta_{ijk} = P(X_i = j \mid pa(X_i) = k)$,
  $i = 1, \ldots, n; j = 1, \ldots, r_i; k = 1, \ldots, q_i$

## BN における ML

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



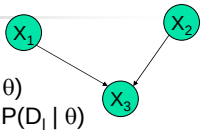

$$
\begin{aligned}
\theta_{111} &= P(X_1{=}1), \theta_{121} = P(X_1{=}2) \\
\theta_{211} &= P(X_2{=}1), \theta_{221} = P(X_2{=}2) \\
pa(X_3) = 1 : \theta_{311} &= P(X_3{=}1 \mid X_1 = 1, X_2 = 1), \theta_{321} = P(X_3{=}2 \mid X_1 = 1, X_2 = 1) \\
pa(X_3) = 2 : \theta_{312} &= P(X_3{=}1 \mid X_1 = 1,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

## BN における ML

- 欠測値のない場合: $D_l$ を値のベクトル, すなわち各要素値組合せのベクトル, とする (事例).
  例: $D_l = (X_1 = 1, X_2 = 2, X_3 = 2)$
- 所与:
  事例集合: $D = \{D_1, \ldots, D_m\}$
- 求む: パラメータ $\theta$ の最尤推定量

## BN における ML

- 対数尤度 loglikelihood:
  $l(\theta \mid D) = \log L(\theta \mid D) = \log P(D \mid \theta)$
  $= \log \prod_l P(D_l \mid \theta) = \sum_l \log P(D_l \mid \theta)$




- 項 $\log P(D_l \mid \theta)$:
  - $D_4 = (1, 2, 2)$

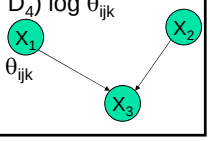
  $\log P(D_4 \mid \theta) = \log P(X_1 = 1, X_2 = 2, X_3 = 2 \mid \theta)$
  $= \log P(X_1{=}1 \mid \theta)\, P(X_2{=}2 \mid \theta)\, P(X_3{=}2 \mid X_1{=}1, X_2{=}2, \theta)$
  $= \log \theta_{111} + \log \theta_{221} + \log \theta_{322}$
- パラメータ:
  $\theta = \{\theta_{111}, \theta_{121}, \theta_{211}, \theta_{221}, \theta_{311}, \theta_{312}, \theta_{313}, \theta_{314}, \theta_{321}, \theta_{322}, \theta_{323}, \theta_{324}\}$

## BN における ML

- $D_l$ の特徴関数を定める :

$$\chi(i, j, k : D_l) = \begin{cases} 1 & \text{if } X_i = j,\ pa(X_i) = k \text{ in } D_l \\ 0 & \text{otherwise} \end{cases}$$

- $l = 4$ , $D_4 = \{1, 2, 2\}$ としよう
  $\chi(1,1,1{:}D_4) = \chi(2,2,1{:}D_4) = \chi(3,2,2{:}D_4) = 1$,
  $\chi(i, j, k{:}\, D_4) = 0$ (他の全ての i, j, k)
- 従って $\log P(D_4 \mid \theta) = \sum_{ijk} \chi(i, j, k{:}\, D_4) \log \theta_{ijk}$
- 一般に,
  $\log P(D_l \mid \theta) = \sum_{ijk} \chi(i, j, k{:}\, D_l) \log \theta_{ijk}$



## BN における ML

- **定義**: $m_{ijk} = \sum_l \chi(i, j, k: D_l)$
  $X_i = j$ **かつ** $pa(X_i) = k$ **となる事例の総数**
- **そうすると** $l(\theta \mid D) = \sum_l \log P(D_l \mid \theta)$
  $= \sum_l \sum_{i,j,k} \chi(i, j, k : D_l) \log \theta_{ijk}$
  $= \sum_{i,j,k} \sum_l \chi(i, j, k : D_l) \log \theta_{ijk}$
  $= \sum_{i,j,k} m_{ijk} \log \theta_{ijk} = \sum_{i,k} \sum_j m_{ijk} \log \theta_{ijk}$

## BN における ML

- **求めたいものは**:
  $\operatorname{argmax}_\theta l(\theta \mid D) = \operatorname{argmax}_{\theta_{ijk}} \sum_{i,k} \sum_j m_{ijk} \log \theta_{ijk}$
- **仮定**: $\theta_{ijk} = P(X_i = j \mid pa(X_i) = k)$ **は** $\theta_{i'j'k'}$ **と無関係.**
  **但し** $i \neq i'$ **または** $k \neq k'$ **という条件のもとで考える**
- **そうすると総和** $\sum_{i,k}[\ldots]$ **の中の各項を別個に最大化すればよい**
  $\operatorname{argmax}_{\theta_{ijk}} \sum_j m_{ijk} \log \theta_{ijk}$

## BN における ML

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

## BN に関するその他の話題

- **欠測値があるときのパラメータ学習**
- **訓練事例からの** BN **の構造の学習**
- **その他多数**…

## 参考文献


- Pearl, Judea, *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*, Morgan Kaufmann, San Mateo, CA, 1988.
- Heckerman, David, "A Tutorial on Learning with Bayesian Networks," Technical Report MSR-TR-95-06, Microsoft Research, 1995.
  ftp://ftp.research.microsoft.com/pub/tr/tr-95-06.pdf