ガウス型ベイジアンネットワークの構造学習の一致性について

鈴木譲

大阪大学

2009 年 8 月 26 日
和歌山大学紀南サテライト

あらまし

確率空間 $(\Omega, \mathcal{F}, \mu)$

Ω: 全体集合

$\mathcal{F}$ が Ω 上の σ 集合体

1. $\Omega, \phi \in \mathcal{F}$
2. $A, B \in \mathcal{F} \Longrightarrow A \cup B, A \cap B, A \backslash B \in \mathcal{F}$

$\mathcal{F}$ の要素を事象という

μ が $\mathcal{F}$ 上の測度

1. $\mu(\phi) = 0$
2. $A \in \mathcal{F} \Longrightarrow \mu(A) \geq 0$
3. $A, B \in \mathcal{F}, A \cap B = \phi \Longrightarrow \mu(A \cup B) = \mu(A) + \mu(B)$

$\mu(\Omega) = 1$ を仮定 (確率測度)

確率的学習

X: 確率変数
μ_X: X の確率測度
$x_1, \cdots, x_n \in X(\Omega)$

帰納と演繹

- $x_1, \cdots, x_n \longmapsto \mu_X$ (確率的学習、設計段階)
- $\mu_X \longmapsto x_1, \cdots, x_n$ (乱数生成、運用段階)

モデル選択をともなう問題: 条件付確率の学習、ARMA の学習

条件付確率	ARMA
有限型 BN	ガウス型 BN

本研究の目標

ガウス型 BN の構造学習の誤り率の公式を証明
(有限型 BN の構造学習の公式から予想できる)

条件付確率の学習

X, Y: 確率変数
$\mu_{Y|X}$: Y の X のもとでの条件付測度

X のもとでの Y の条件付学習

$$x \sim x' \Longleftrightarrow \mu_{Y|X}(y|x) = \mu_{Y|X}(y|x') \ , \ y \in Y(\Omega)$$

n 個の例から $X(\Omega)$ の同値関係 $\sim$ を見出す

仮定

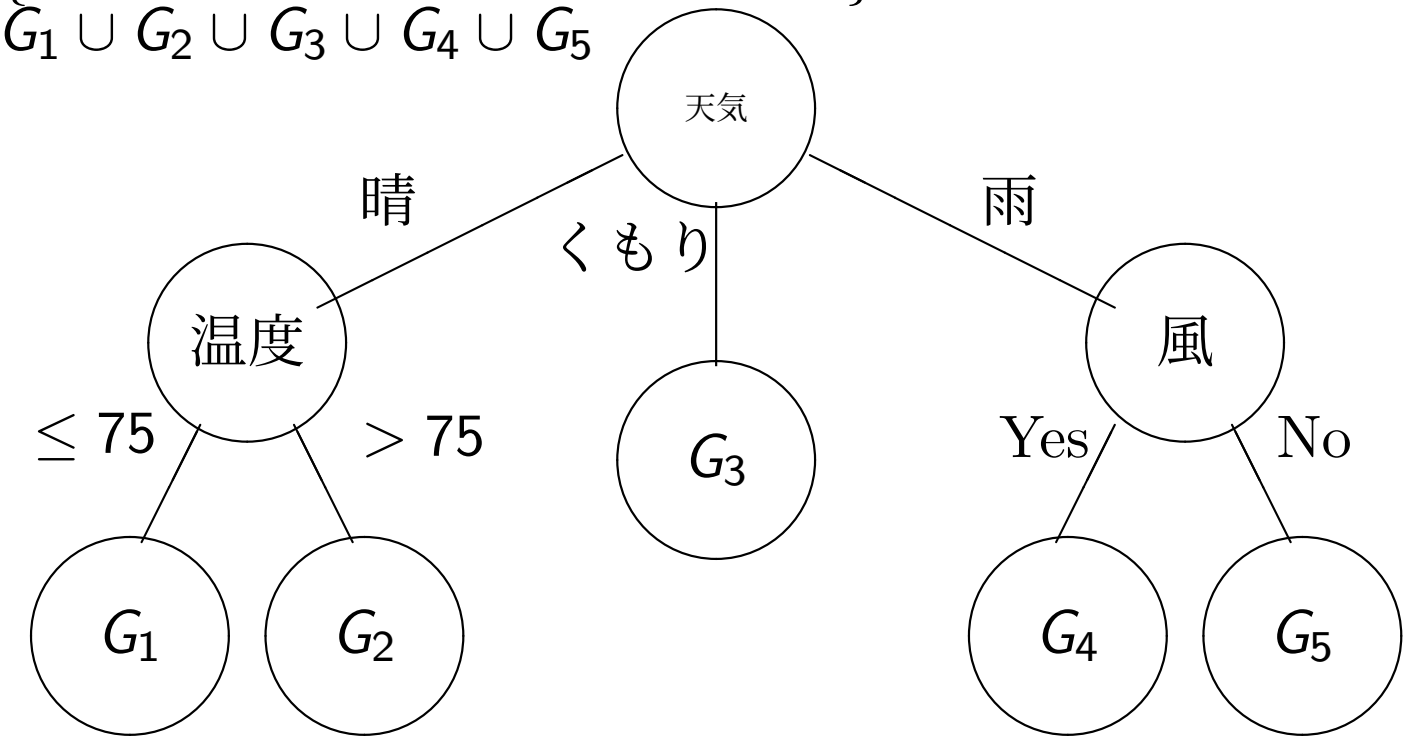
$|Y(\Omega)| < \infty$

条件付確率の学習: 応用

確率的決定木 $Y|X$

例から $X(\Omega)$ の分割を見出す

$Y(\Omega) = \{$ ゴルフできる, ゴルフできない $\}$
$X(\Omega) = G_1 \cup G_2 \cup G_3 \cup G_4 \cup G_5$

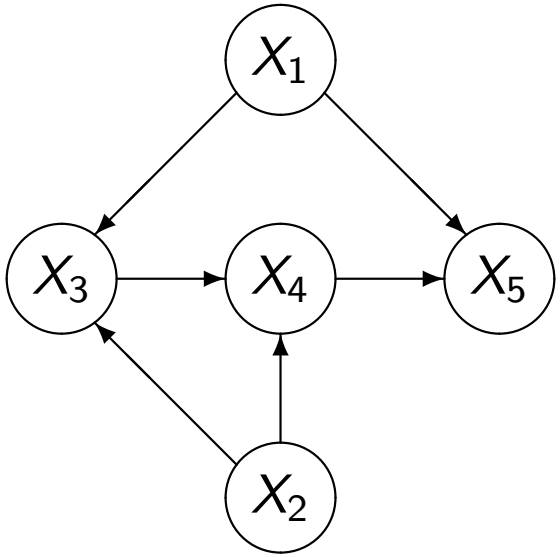


条件付確率の学習: 応用

有限型 BN $X_i|X_j, j \in \pi(i)$

例から $\pi(i) \subseteq \{1, \cdots, i-1\}$ を見出す

1. X_2 は X_1 とは独立
2. X_3 は X_1, X_2 に依存
3. X_4 は X_2, X_3 に依存
4. X_5 は X_1, X_4 に依存



条件付確率の学習: 定式化

n 個の例 $z^n := (z_1, \cdots, z_n) \in Z^n(\Omega)$ から、$X(\Omega)$ の分割を見出す

$$z_i := (x_i, y_i) \in Z(\Omega) := X(\Omega) \times Y(\Omega)$$

仮定

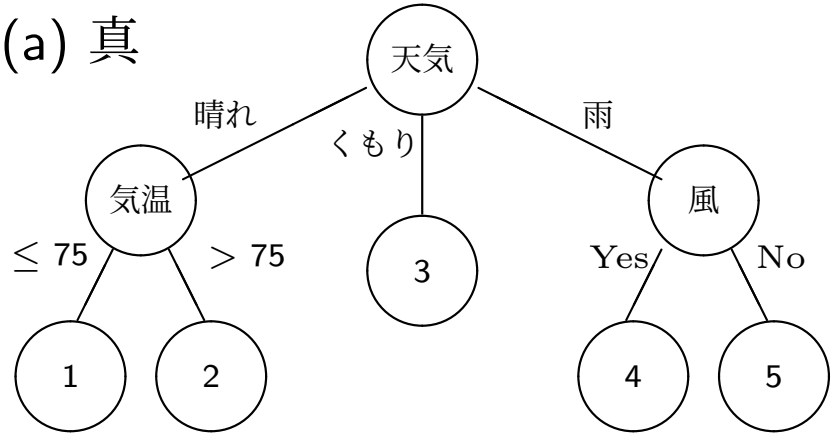
有限個に分割される

2 種類の誤り

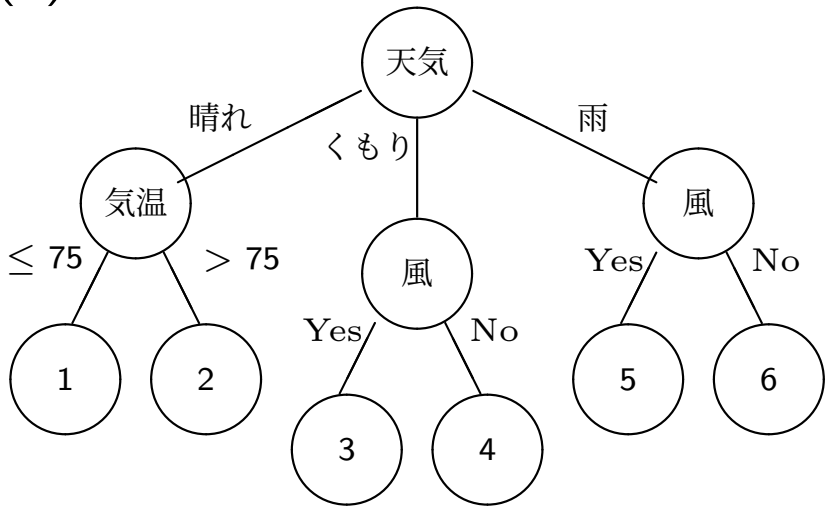
- 真のものより細かく分割される (過学習)
- 荒く分割されてる箇所がある (未学習)

例: Quinlan の Q4.5

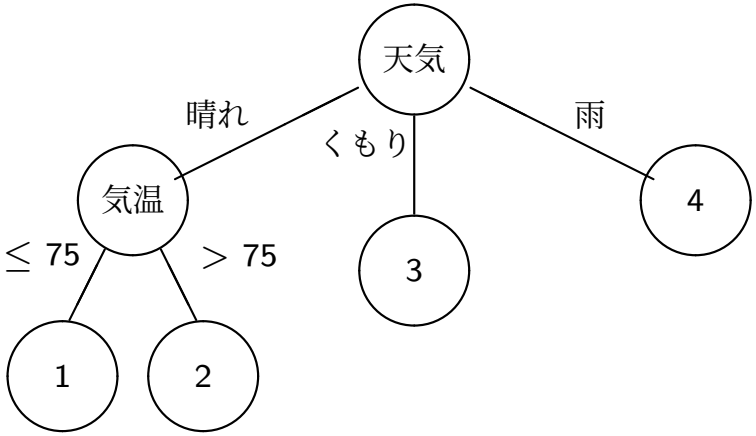
(a) 真



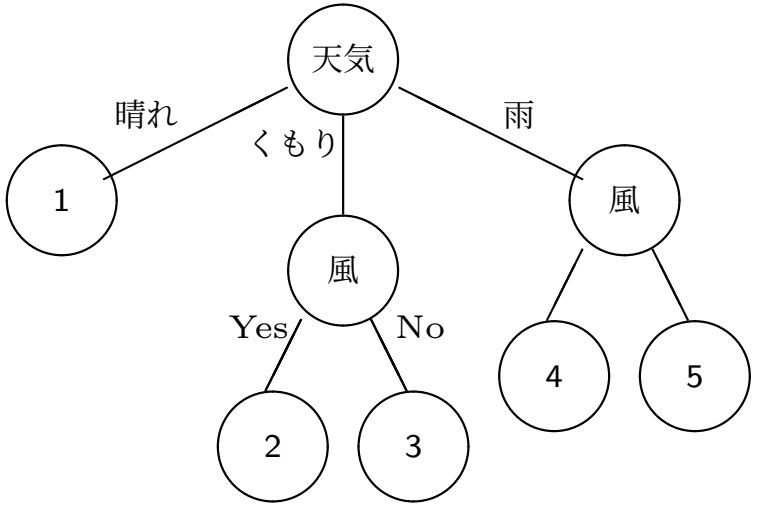
(b) 過学習



(c) 未学習



(d) 未学習



情報量基準の適用

例 $z^n \in Z^n(\Omega)$ から、

$$I(\mathcal{G}, z^n) := H(\mathcal{G}, z^n) + \frac{k(\mathcal{G})}{2} d_n$$

を最小にする分割 $\mathcal{G}$ を見出す
$H(\mathcal{G}, z^n)$: 経験的エントロピー (例 z^n の分割 $\mathcal{G}$ への適合性)
$k(\mathcal{G})$: パラメータ数 ($\mathcal{G}$ の簡潔さ)
$d_n \geq 0$: $\dfrac{d_n}{n} \to 0$

$d_n = \log n$ BIC/MDL
$d_n = 2$ AIC

一致性

一致性 ($n \to \infty$ で、推定結果が真のそれに一致)

- 弱一致性 確率収束 ($O(1) < d_n < o(n)$)
- 強一致性 概収束 (MDL/BIC etc.)

AIC ($d_n = 2$) は、$\{d_n\}$ が小さすぎて、一致性を満足しない

問題

強一致性を満足する最小の $\{d_n\}$ は何か？

答え (Suzuki, 2006)

$d_n = 2 \log \log n$

(重複対数の法則)

誤り確率

$\mathcal{G}^*$: 真の分割

$$\mu\{\omega \in \Omega | I(\mathcal{G}, Z^n(\omega)) < I(\mathcal{G}^*, Z^n(\omega))\}$$

分割 $\mathcal{G}$ が $\mathcal{G}^*$ の過学習

$$\int_{(K(\mathcal{G})-K(\mathcal{G}^*))d_n}^{\infty} f_{K(\mathcal{G})-K(\mathcal{G}^*)}(x)dx$$

f_l: 自由度 l の χ^2 分布の確率密度関数

分割 $\mathcal{G}$ が $\mathcal{G}^*$ の未学習

n とともに指数的に 0 に低減

ARMA の学習

$k \geq 0$
$\{\lambda_j\}_{j=1}^{k}$: $\lambda_i \in \mathbb{R}$
$\sigma^2 \in \mathbb{R}_{>0}$

ARMA (Autoregressive Moving Average, 自己回帰移動平均)

$\{X_n\}_{n=-\infty}^{\infty}$: $X_n + \sum_{j=1}^{k} \lambda_j X_{n-j} = \epsilon_i \sim \mathcal{N}(0, \sigma^2)$

ARMA の学習

n 個の例から

次数 k が既知　係数 $\{\lambda_j\}_{j=1}^{k}$ を見出す

次数 k が未知　次数 k と係数 $\{\lambda_j\}_{j=1}^{k}$ を見出す

Yule-Walker 方程式

次数 k が既知のとき、以下を $\{\hat{\lambda}_{j,k}\}_{j=1}^{k}$ および $\hat{\sigma}_k^2$ について解く。

$$\bar{x} := \frac{1}{n}\sum_{i=1}^{n} x_i$$

$$c_j := \frac{1}{n}\sum_{i=1}^{n-j}(x_i - \bar{x})(x_{i+j} - \bar{x}) \ , \ j = 0, \cdots, k$$

$$\begin{bmatrix} -1 & c_1 & c_2 & \cdots & c_k \\ 0 & c_0 & c_1 & \cdots & c_{k-1} \\ 0 & c_1 & c_0 & \cdots & c_{k-2} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & c_{k-1} & c_{k-2} & \cdots & c_0 \end{bmatrix} \begin{bmatrix} \hat{\sigma}_k^2 \\ \hat{\lambda}_{1,k} \\ \hat{\lambda}_{2,k} \\ \vdots \\ \hat{\lambda}_{k,k} \end{bmatrix} = \begin{bmatrix} -c_0 \\ -c_1 \\ -c_2 \\ \vdots \\ -c_k \end{bmatrix}$$

情報量基準の適用

次数 k が未知のとき、例 $x^n \in X^n(\Omega)$ から、$I(k, x^n)$ 最小の k を見出す

$$I(k, x^n) := \frac{1}{2} \log \hat{\sigma}_k^2 + \frac{k}{2} d_n$$

$\hat{\sigma}_k^2$: Yule-Walker 方程式から

$d_n \geq 0$: $\dfrac{d_n}{n} \to 0$

$d_n = \log n$ BIC/MDL

$d_n = 2$ AIC

$d_n = 2 \log \log n$ Hannan-Quinn (1979)

$\Longrightarrow$ 強一致性を満足する最小の $\{d_n\}$

- Suzuki (2006)
- Hannan-Quinn (1979)

誤り確率 (未学習, ARMA)

k_*: 真の次数

$$\mu\{\omega \in \Omega | I(k, X^n(\omega)) < I(k_*, X^n(\omega))\}$$

$k_* > k$ (未学習, ARMA)

n とともに指数的に 0 近づく (Hannan-Quinn, 1979)

研究のねらい

	条件付確率 (領域分割 $\mathcal{G}$)	ARMA (次数 k)
強一致性 のための 最小の d_n	$2 \log \log n$ (Suzuki, 2006)	$2 \log \log n$ (Hannan-Quinn, 1979)
誤り確率 (未学習)	指数的に 0 (Suzuki, 2006)	指数的に 0 (Hannan-Quinn, 1979)
誤り確率 (過学習)	$\int_{(K(\mathcal{G})-K(\mathcal{G}^*))d_n}^{\infty} f_{K(\mathcal{G})-K(\mathcal{G}^*)}(x)dx$ (Suzuki, 2006)	$\int_{(k-k_*)d_n}^{\infty} f_{k-k_*}(x)dx$?

$k^* < k$ (過学習, ARMA)

$$\int_{(k-k_*)d_n}^{\infty} f_{k-k_*}(x)dx$$

を証明する

証明のスケッチ

$k = k_* + 1, k_* + 2, \cdots$ に対して、確率 1 で、

$$2\{I(k, x^n) - I(k-1, x^n)\} = -n\hat{\lambda}_{k,k}^2 + d_n$$

が成立する (Hannan-Quinn, 1979) ので、

$$\mu_k := \sqrt{n}\hat{\lambda}_{k,k} \sim \mathcal{N}(0, 1)$$

でしかも独立であることをいえば、

$$\sum_{j=k_*+1}^{k} 2\{I(k, x^n) - I(k-1, x^n)\} = \sum_{j=k_*+1}^{k} \mu_j^2 \sim \chi^2_{k-k_*}$$

が成立する。

定常エルゴードな確率過程の中心極限定理

$\{X_i\}_{i=-\infty}^{\infty}$: 定常エルゴード
$S_n := \sum_{j=1}^{n} X_j$

Hyde, 1974

1. $E[X_0] = 0,\ E[X_0^2] < \infty$

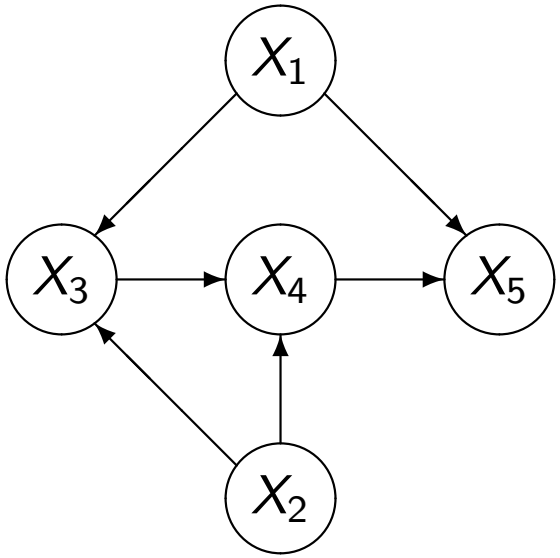
X_0 が $\mathcal{G}$ 上可測 $(\mathcal{G} \subseteq \mathcal{F})$ であるとして、

2. $\sum_{j=1}^{\infty} E[X_j E[X_N|\mathcal{G}]]$ が、各 $N \geq 1$ で収束
3. $\sum_{j=J}^{\infty} E[X_j E[X_N|\mathcal{G}]]$ が、J について一様に、$N \to \infty$ で 0 に収束

$\Longrightarrow S_n/(\sigma\sqrt{n}) \sim \mathcal{N}(0,1)$

ガウス型 BN

1. $X_2 = \epsilon^{(2)} \sim \mathcal{N}(0, \sigma_2^2)$
2. $\lambda_1^{(3)} X_1 + \lambda_2^{(3)} X_2 + X_3 = \epsilon^{(3)} \sim \mathcal{N}(0, \sigma_3^2)$
3. $\lambda_2^{(4)} X_2 + \lambda_3^{(4)} X_3 + X_4 = \epsilon^{(4)} \sim \mathcal{N}(0, \sigma_4^2)$
4. $\lambda_1^{(5)} X_1 + \lambda_4^{(5)} X_4 + X_5 = \epsilon^{(5)} \sim \mathcal{N}(0, \sigma_5^2)$



ガウス型 BN の学習

$i = 1, \cdots, N$

$$\sum_{j \in \pi(i)} \lambda_j^{(i)} X_j + X_i = \epsilon^{(i)}$$

n 個の例 $x^n = (x_1, \cdots, x_n)$
$x_m = (x_{m,1}, \cdots, x_{m,N}) \in X_1(\Omega) \times \cdots \times X_N(\Omega),\ m = 1, \cdots, n$

Yule-Walker 方程式

$c_{j,h} := \frac{1}{n} \sum_{m=1}^{n} x_{m,j} x_{m,h}\ ,\ j, h \in \pi(i) \cup \{i\}$

$$\sum_{j \in \pi(i)} \lambda_j^{(i)} c_{j,h} + c_{i,h} = \sigma_i^2 \delta_{i,h}\ ,\ h \in \pi(i) \cup \{i\}$$

($|\pi(i)| + 1$ 個の変数、$|\pi(i)| + 1$ 式の連立方程式)

ガウス型 BN の構造学習の誤り率

正しい $\pi_*(i) = \pi(i)$

過学習 $\pi_*(i) \subset \pi(i)$

未学習 $\pi_*(i) \not\subseteq \pi(i)$

強一致性のための最小の d_n

$d_n = 2 \log \log n$

誤り確率 (過学習)

$$\int_{(|\pi(i)-\pi_*(i)|)d_n}^{\infty} f_{|\pi(i)-\pi_*(i)|}(x)dx$$

誤り確率 (未学習)

n とともに指数的に 0

まとめ

条件付確率	ARMA
有限型 BN	ガウス型 BN

証明したこと: ARMA

- 誤り確率 (過学習)

証明したこと: ガウス型 BN

- 強一致性のための最小の d_n
- 誤り確率 (過学習)
- 誤り確率 (未学習)

課題

なぜ似てくるのか、指数分布族で共通の特徴があるのか