一般的な確率変数に対する Chow-Liu アルゴリズム

鈴木譲

2009 年 12 月 3 日 (木)

研究のねらい

まず、Chow-Liu アルゴリズムを一般の確率変数で

情報理論というと、離散分布と連続分布ばかり

さらに、サンプルから学習する場合の一般化

Chow-Liu アルゴリズムを MDL 基準に適用
(鈴木, 大嶽, 平澤, 情報処理学会論文誌 1992 年 11 月号)
有限の確率変数を仮定

無向グラフ

V: 有限集合
$E \subseteq \mathcal{E} := \{\{X, Y\} \| X, Y \in V, X \neq Y\}$
$G = (V, E)$: 無向グラフ

- V G の頂点集合 (要素: G の頂点)
- E G の辺集合 (要素: G の辺)

G は森

無向グラフ G に巡回経路が存在しない

G は木

森であって連結されている

Chow-Liu アルゴリズム

$X := (X^{(1)}, \cdots, X^{(N)})$, $X^{(j)}$: 有限の値をとる

Kullback-Leibler 情報量

$$D(P||Q) := \sum_{x^{(1)},\cdots,x^{(N)}} P(x^{(1)}, \cdots, x^{(N)}) \log \frac{P(x^{(1)}, \cdots, x^{(N)})}{Q(x^{(1)}, \cdots, x^{(N)})}$$

$$Q(x^{(1)}, \cdots, x^{(N)}) := \prod_{\pi(j)=0} P_j(x^{(j)}) \prod_{\pi(i)\neq 0} P_{i|\pi(i)}(x^{(i)}|x^{(\pi(i))})$$

Dendroid 分布 Q への近似 (木への近似)

$D(P||Q)$ 最小の $\pi : \{1, \cdots, N\} \to \{0, 1, \cdots, N\}$

$$\pi^k(j) \neq j \ , \ k \geq 1 \ , \ j \in \{1, \cdots, N\}$$

相互情報量

X, Y 間の相互情報量

$$I(X,Y) := \sum_{x,y} \mu_{XY}(x,y) \log \frac{P_{XY}(x,y)}{P_X(x)P_Y(y)}$$

$$D(P||Q) = -\sum_{\pi(i)\neq 0} I(X^{(i)}, X^{(\pi(i))}) + (\pi \text{ によらない定数})$$

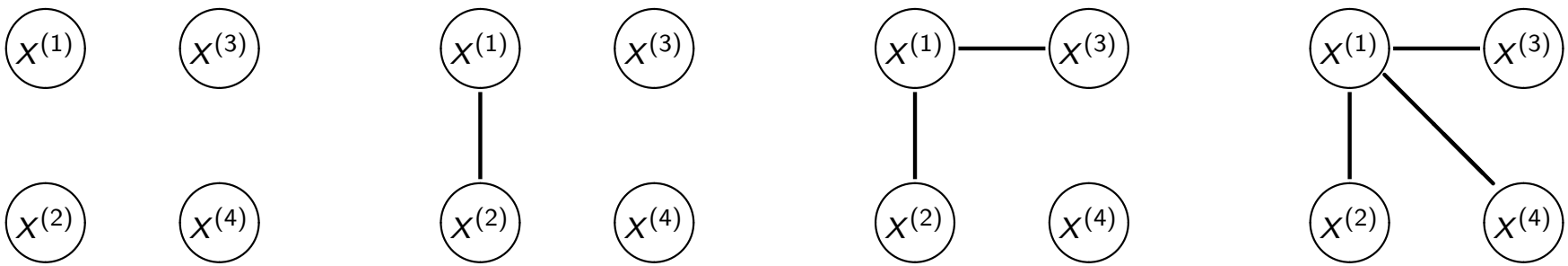
Chow-Liu アルゴリズム (近似)

$\sum_{\pi(i)\neq 0} I(X^{(i)}, X^{(\pi(i))})$ 最大の木を生成する問題に帰着

Chow-Liu アルゴリズム: 例

1. $I(1,2)$ が最大なので、$X^{(1)}, X^{(2)}$ を結ぶ。
2. それ以外では $I(1,3)$ が最大なので、$X^{(1)}, X^{(3)}$ を結ぶ。
3. それ以外では $I(2,3)$ が最大だが、結ぶと巡回経路。
4. それ以外では $I(1,4)$ が最大なので、$X^{(1)}, X^{(4)}$ を結ぶ。
5. これ以上辺を結ぶと巡回経路。

i	1	1	2	1	2	3
j	2	3	3	4	4	4
$I(i,j)$	12	10	8	6	4	2



Chow-Liu アルゴリズム: 手順

$V = \{1, \cdots, N\}$
$I(i,j) := I(X^{(i)}, X^{(j)})$ $(i \neq j)$ 相互情報量

1. $E := \{\}$;
2. $\mathcal{E} := \{\{i,j\} | i \neq j\}$;
3. $I_{i,j}$ を最大にする $\{i,j\} \in \mathcal{E}$ について、$\mathcal{E} := \mathcal{E} \backslash \{\{i,j\}\}$;
4. $(V, E \cup \{\{i,j\}\})$ が巡回経路を含まないなら、
   $E := E \cup \{\{i,j\}\}$;
5. $\mathcal{E} \neq \{\}$ なら 3. へ, そうでなければ終わる;

Chow-Liu アルゴリズム (木の近似)

$G = (V, E)$ から得られる Q は、$D(P||Q)$ を最小

サンプルからはじめたとき、尤度を最大にする

$$P(x^{(1)}, \cdots, x^{(N)} | \theta, \pi) \ , \ \theta = (\theta_1, \cdots, \theta_k)$$

$$x^n := \{(x_i^{(1)}, \cdots, x_i^{(N)})\}_{i=1}^n$$

$\hat{\theta}(x^n)$: 最尤推定量 ($\log P(x^n|\theta, \pi)$ 最大の θ)

$$\begin{aligned} I_n(i,j) &:= \sum_{x,y} P_{i,j}(x,y|\hat{\theta},\pi) \log \frac{P_{i,j}(x,y|\hat{\theta},\pi)}{P_i(x|\hat{\theta},\pi)P_j(y|\hat{\theta},\pi)} \\ &= \frac{1}{n} \sum_{x,y} c_{i,j}(x,y) \log \frac{c_{i,j}(x,y)}{c_i(x)c_j(y)} \end{aligned}$$

$$H(\pi, x^n) := -\log P(x^n|\hat{\theta}(x^n), \pi) = -\sum_{\pi(i)\neq 0} I_n(i, \pi(i)) + (\text{const})$$

Chow-Liu アルゴリズム: 尤度最大 (経験的エントロピー最小)

$I(i, \pi(i))$ を $I_n(i, \pi(i))$ におきかえるだけ

サンプルからはじめたとき、記述長を最大にする

$$Q(x) := \int P(x|\theta,\pi)w(\theta)d\theta \ , \int w(\theta)d\theta = 1$$

$\alpha^{(i)}$: $X^{(i)}$ のとりうる値の数 $(i = 1, \cdots, N)$, $(\alpha^{(0)} := 1)$

$$k = \sum_{\pi(i)\neq 0} (\alpha^{(i)} - 1)(\alpha^{(\pi(i))} - 1)$$

$$J_n(i,j) := I_n(i,j) - \frac{1}{2}(\alpha^{(i)} - 1)(\alpha^{(j)} - 1)\log n$$

$$\begin{aligned} L(\pi, x^n) &:= -\log Q(x^n) \\ &= -\log P(x^n|\hat{\theta}(x^n),\pi) + \frac{k(\pi)}{2}\log n + (\text{const}) \\ &\quad - \sum_{\pi(i)\neq 0} \{I_n(i,\pi(i)) - \frac{1}{2}(\alpha^{(i)} - 1)(\alpha^{(\pi(i))} - 1)\} \\ &= -\sum_{\pi(i)\neq 0} J_n(i,\pi(i)) \end{aligned}$$

サンプルからはじめたとき、記述長を最大にする (続)

1. $E = \{\}$;
2. $\mathcal{E} := \{\{i,j\} | i \neq j\}$;
3. $J_n(i,j)$ を最大にする $\{i,j\} \in \mathcal{E}$ について $\mathcal{E} := \mathcal{E} \backslash \{\{i,j\}\}$;
4. $J_n(i,j) \geq 0$ and $(V, E \cup \{\{i,j\}\})$ がループを含まなければ, $E := E \cup \{\{i,j\}\}$;
5. $\mathcal{E} \neq \{\}$ であれば 3. へ, そうでなければ終わる

- 木でなく森をつくる ($J_n < 0$ で停止)
- 各辺を結ぶか否かで、適合度だけでなく、辺の複雑さを考慮

(鈴木 大嶽 平澤, 1992)

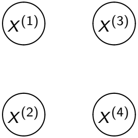
辺を選択する順序

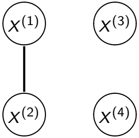
$\{I_n(i,j)\}_{i \neq j}, \{J_n(i,j)\}_{i \neq j}$ で辺を選択する順序が異なる

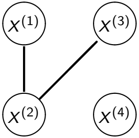
サンプルからはじめたとき、記述長を最大にする (続)

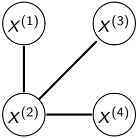
1. $J_n(1,2)=8$ が最大なので、$X^{(1)}, X^{(2)}$ を結ぶ。
2. $J_n(2,3)=6$ が最大なので、$X^{(2)}, X^{(3)}$ を結ぶ。
3. $X^{(1)}, X^{(3)}$ を結ぶと巡回経路が出来る。
4. $I_n(2,4)=1$ が最大なので、$X^{(2)}, X^{(4)}$ を結ぶ。
5. $J_n<0$ となる辺、巡回経路ができる辺は結ばない。

i	j	$I_n(i,j)$	$\alpha^{(i)}$	$\alpha^{(j)}$	$J_n(i,j)$
1	2	12	5	2	8
1	3	10	5	3	2
2	3	8	2	3	6
1	4	6	5	4	-6
2	4	4	2	4	1
3	4	2	3	4	-4









離散や連続は、むしろ特殊ケース

確率 $1/2$ で $X = -1$
確率 $1/2$ で $X = x \geq 0$ $(\int_0^\infty f(x)dx = 1)$

$$F_X(x) = \begin{cases} 0 & x < -1 \\ \frac{1}{2} & 1 \leq x < 0 \\ \int_0^x \frac{1}{2} f(t)dt & 0 \leq x \end{cases}$$

F_X には確率密度関数は存在しないし、離散分布でもない

一般的な確率変数

$(\Omega, \mathcal{F}, \mu)$: 確率空間
$\mathcal{B}$: $\mathbb{R}$ の Borel 集合族

$X : \Omega \to \mathbb{R}$ が $(\Omega, \mathcal{F}, \mu)$ における確率変数

$$D \in \mathcal{B} \Longrightarrow \{\omega \in \Omega | X(\omega) \in D\} \in \mathcal{F}$$

X の確率測度 $\mu_X : \mathcal{B} \to \mathbb{R}$

$$\mu_X(D) := \mu(\{\omega \in \Omega | X(\omega) \in D\})$$

一般的な確率変数に対する Chow-Liu アルゴリズム

Kullback-Leibler 情報量

$\mu << \nu$ のとき

$$D(\mu||\nu) := \int_\Omega d\mu \log \frac{d\mu}{d\nu}$$

$\frac{d\mu}{d\nu} := f$ s.t. $\mu = \int f d\nu$ (Radon-Nykodim)

X, Y の相互情報量

$$I(X,Y) := \int_\Omega d\mu_{XY} \log \frac{d^2\mu_{XY}}{d\mu_X d\mu_Y}$$

$\frac{d\mu_{XY}}{d\mu_X d\mu_Y} := g$ s.t. $\mu_{XY} = \int g d\mu_X d\mu_Y$ (Radon-Nykodim)

一般的な Dendoroid 分布近似

一般の Dendoroid 分布

任意の $D_1, \cdots, D_N \in \mathcal{B}$ について

$$\nu(D_1, \cdots, D_N) = \prod_{\pi(i)\neq 0} \frac{\mu_{i,\pi(i)}(D_i, D_{\pi(i)})}{\mu_i(D_i)\mu_{\pi(i)}(D_{\pi(i)})} \cdot \prod_{i=1}^{N} \mu_i(D_i)$$

$\eta(D_1, \cdots, D_N) := \prod_{i=1}^{N} \mu_i(D_i)$ とおくと、$\mu << \eta$ より

$$\frac{d\nu}{d\eta} = \prod_{\pi(i)\neq 0} \frac{d^2\mu_{i,\pi(i)}}{d\mu_i d\mu_{\pi(i)}}$$

と同値になる (証明略)

一般的な確率変数にも Chow-Liu アルゴリズムは適用できる

スケッチ:

$$\frac{d\mu}{d\nu} = \frac{d\mu}{d\eta} \Big/ \frac{d\nu}{d\eta} = [\prod_{\pi(i)\neq 0} \frac{d^2\mu_{i,\pi(i)}}{d\mu_i d\mu_{\pi(i)}}]^{-1} \frac{d\mu}{d\eta}$$

$$\begin{aligned} E \log \frac{d\mu}{d\nu}(X^{(1)}, \cdots, X^{(N)}) &= -\sum_{\pi(i)\neq 0} E \log \frac{d^2\mu_{i,\pi(i)}}{d\mu_i d\mu_{\pi(i)}}(X^{(i)}, X^{(\pi(i))})] \\ &\quad + E \log \frac{d\mu}{d\eta}(X^{(1)}, \cdots, X^{(N)}) \end{aligned}$$

$$D(\mu||\nu) = -\sum_{\pi(i)\neq 0} I(X^{(i)}, X^{(\pi(i))}) + (\text{const})$$

例: すべての確率変数が Gauss 分布の場合

$$X^{(i)} \sim N(0, \sigma^2)$$

$$(X^{(i)}, X^{(j)}) \sim N(0, \Sigma), \Sigma = \begin{bmatrix} \sigma_{ii} & \sigma_{ij} \\ \sigma_{ji} & \sigma_{jj} \end{bmatrix}$$

$$\rho_{ij} := \frac{\sigma_{ij}}{\sqrt{\sigma_{ii}\sigma_{jj}}} \text{ (相関関数)}$$

$$\begin{aligned} I(i,j) &= \int\int f_{i,j}(x^{(i)}, x^{(j)}) \log \frac{f_{i,j}(x^{(i)}, x^{(j)})}{f_i(x^{(i)}) f_j(x^{(j)})} dx^{(i)} dx^{(j)} \\ &= \log \frac{\sqrt{\sigma_{ii}\sigma_{jj}}}{|\Sigma|^{\frac{1}{2}}} = -\frac{1}{2} \log (1 - {\rho_{ij}}^2) \end{aligned}$$

Chow-Liu アルゴリズムへの適用

ρ_{ij} から $I(i,j)$ を求めればよい

記述長最小: 2 確率変数が Gauss 分布の場合

$$\bar{x}^{(i)} = \frac{1}{n}\sum_{h=1}^{n} x_h^{(i)},\ \hat{\sigma}_{ij} := \frac{1}{n}\sum_{h=1}^{n}(\bar{x} - x_h^{(i)})(\bar{x} - x_h^{(j)}),\ \hat{\rho}_{ij} := \frac{\hat{\sigma}_{ij}}{\sqrt{\hat{\sigma}_{ii}\hat{\sigma}_{jj}}}$$

$$I_n(i,j) := -\frac{1}{2}\log\left(1 - \hat{\rho_{ij}}^2\right)$$

$$H(\pi, x^n) = -\sum_{\pi(i)\neq 0} I_n(i, \pi(i)) + (\mathrm{const})$$

記述長最小: 2 確率変数が Gauss 分布の場合 (続)

$$J_n(i,j) = I_n(i,j) - \frac{1}{2}\log n$$

$$\begin{aligned} L(\pi, x^n) &= -\sum_{\pi(i)\neq 0} I_n(i,\pi(i)) + \frac{k}{2}\log n + (\text{const}) \\ &= -\sum_{\pi(i)\neq 0} \{I_n(i,\pi(i)) - \frac{1}{2}\log n\} + (\text{const}) \\ &= -\sum_{\pi(i)\neq 0} J_n(i,\pi(i)) + (\text{const}) \end{aligned}$$

2 確率変数が Gauss 分布: 確率パラメータは 1 個増える

$$J_n(i,j) := I_n(i,j) - \frac{1}{2}\log n$$

$\{I_n(i,j)\}_{i\neq j}, \{J_n(i,j)\}_{i\neq j}$ で辺を選択する順序は同じ

記述長最小: 2 確率変数が Gauss 分布と有限分布の場合

- $X^{(i)}$: Gauss
- $X^{(j)}$: 有限 ($\alpha^{(j)}$ 通り)

$$I(i,j) = \sum_{y \in X^{(j)}} \mu_j(y) \int_{x \in X^{(i)}} f_{i,j}(x|y) \log \frac{f_{i,j}(x|y)}{\sum_{z \in X^{(j)}} \mu_j(z) f_{i,j}(x|z)} dx$$

$g : X^{(j)}(\Omega) \to \mathbb{R}$

- $X^{(i)} = \epsilon_i \sim \mathcal{N}(g(X^{(j)}), \phi_i)$
- $Eg(X^{(j)}) = 0$ (パラメータ数 1 個少ない)

記述長最小: 2 確率変数が Gauss 分布と有限分布の場合 (続)

$\alpha^{(j)}$ 個のパラメータ $g(y)$, $y \in X^{(j)}(\Omega)$ を推定

$$\sum_{y \in X^{(j)}(\Omega)} \mu_j(y) g(y) = 0 \text{ の制約}$$

確率パラメータは $\alpha^{(j)} - 1$ 個だけ増える

$$J_n(i,j) := I_n(i,j) - \frac{(\alpha^{(j)} - 1)}{2} \log n$$

$\{I_n(i,j)\}_{i \neq j}, \{J_n(i,j)\}_{i \neq j}$ で辺を選択する順序は異なる

記述長最小: まとめ

$d_n := \log n$ は任意の非負実数列 $\{d_n\}$ でよい

$X^{(i)}, X^{(j)}$: 有限 $J_n(i,j) = I_n(i,j) - \frac{1}{2}(\alpha^{(i)} - 1)(\alpha^{(j)} - 1)d_n$

$X^{(i)}, X^{(j)}$: Gauss $J_n(i,j) = I_n(i,j) - \frac{1}{2}d_n$

$X^{(i)}$: Gauss, $X^{(j)}$: 有限 $J_n(i,j) = I_n(i,j) - \frac{1}{2}(\alpha^{(j)} - 1)d_n$

辺を選択する順序

有限分布にしたがう確率変数が存在すれば、
$\{I_n(i,j)\}_{i \neq j}, \{J_n(i,j)\}_{i \neq j}$ で辺を選択する順序が異なる

本講演のまとめ

一般的な Chow-Liu アルゴリズム

相互情報量 $I = (I_{ij})_{i \neq j}$

$$I_{ij} = \int d\mu_{ij} \log \frac{d^2\mu_{ij}}{d\mu_i d\mu_j}$$

Dendroid 測度が得られる

一般的な Chow-Liu アルゴリズム (記述長最小バージョン)

- 有限分布のみ
- Gauss 分布のみ
- 有限と Gauss が混在

将来的に、おもしろいバリエーションが得られそう