Bayes 的に最適な学習のための Chow-Liu アルゴリズム

鈴木 譲

大阪大学

Kruscal のアルゴリズム

V: 有限集合
$w_{i,j} \in \mathbb{R}$: $w_{i,j} = w_{j,i}$ $i \neq j$

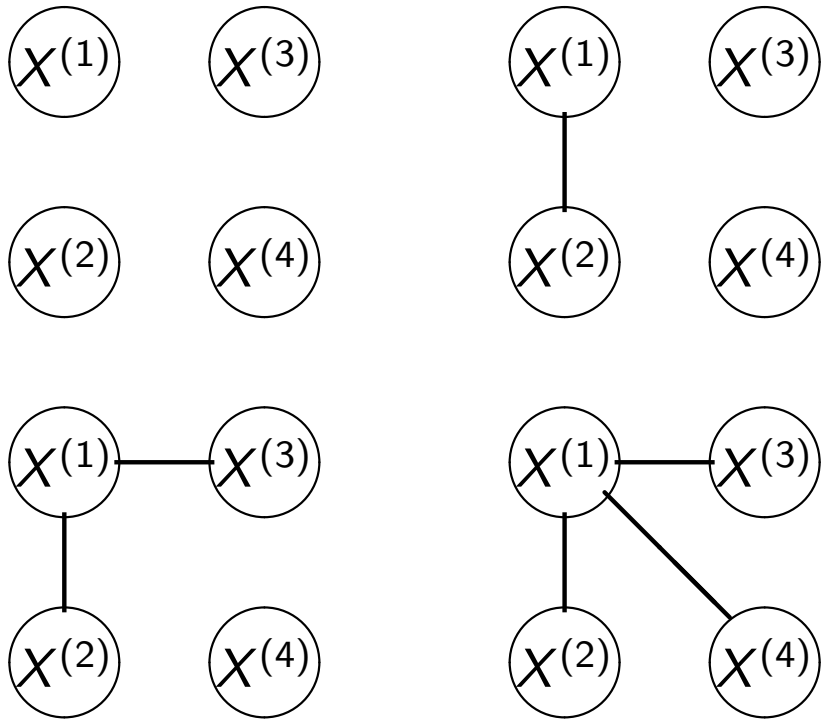
Kruscal のアルゴリズム ($w_{i,j} \geq 0$ のとき)

1. $\mathcal{E} \leftarrow \{\{i,j\} | i,j \in V, i \neq i\}$
2. $E \leftarrow \{\}$
3. while($\mathcal{E} \neq \phi$) for $w_{i,j}$ が最大となる $\{i,j\} \in \mathcal{E}$
   1. $\mathcal{E} \leftarrow \mathcal{E} \backslash \{i,j\}$
   2. $G = (V, E \cup \{i,j\})$ が木 $\Longrightarrow E \leftarrow E \cup \{i,j\}$

最終的な G が $\sum_{\{i,j\} \in E} w_{i,j}$ を最大にする木

$w_{i,j} = I(i,j)$(相互情報量) のとき

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



Dendroid 分布

$G := (V, E)$ 木
$V := \{1, \cdots, N\}$

$$Q_{1,\cdots,N}(x^{(1)}, \cdots, x^{(N)}) = \frac{\prod_{\{i,j\}\in E} P_{i,j}(x^{(i)}, x^{(j)})}{\prod_{i\in V} P_i(x^{(i)})^{d_i - 1}}$$

$d_i := |\{j \in V | \{i, j\} \in E\}|$

Chow-Liu アルゴリズム: 近似

真の分布 $P_{1,\cdots,N}$ を Dendroid 分布 $Q_{1,\cdots,N}$ で近似

Chow-Liu, 1968

重みとして $w_{i,j} := I(i,j)$ を用いて、Kruscal のアルゴリズムを適用すると、$D(P_{1,\cdots,N}||Q_{1,\cdots,N})$ が最小になる

$$
\begin{aligned}
& - \sum_{x^{(1)},\cdots,x^{(N)}} P_{1,\cdots,N}(x^{(1)},\cdots,x^{(N)}) \log Q_{1,\cdots,N}(x^{(1)},\cdots,x^{(N)}) \\
= & \sum_{i \in V} H(i) - \sum_{\{i,j\} \in E} I(i,j)
\end{aligned}
$$

Chow-Liu アルゴリズム: 最尤推定

- $\{(X_i^{(1)}, \cdots, X_i^{(N)})\}_{i=1}^{\infty}$: i.i.d
- $P_{1,\cdots,N}$ は未知
- n 組のサンプル $\{(x_i^{(1)}, \cdots, x_i^{(N)})\}_{i=1}^{n}$ が利用可能

$I_n(i,j)$: $X^{(i)}, X^{(j)}$ の経験的相互情報量
$\hat{P}_{1,\cdots,N}$: $P_{1,\cdots,N}$ の最尤推定

1. $I_n(i,j)$ の大きい $i,j \in V$ $(i \neq j)$ から辺を結ぶ
2. ループを作る場合、結ばない

Chow-Liu アルゴリズム: 最尤推定

$D(\hat{P}_{1,\cdots,N} || \hat{Q}_{1,\cdots,N})$ を最小にする $\hat{Q}_{1,\cdots,N}$ を表現する木が得られる

$\{x_k^{(i)}\}_{k=1}^n$, $\{(x_k^{(i)}, x_k^{(j)})\}_{k=1}^n$ の予測分布 $R^n(i)$, $R^n(i,j)$

$A^{(i)}$: $X^{(i)}$ の取りうる値の集合
$\alpha^{(i)} := |A^{(i)}|$
$c_n[x]$: $\{x_k^{(i)}\}_{k=1}^n$ における $x \in A^{(i)}$ の頻度
$c_n[x,y]$: $\{(x_k^{(i)}, x_k^{(j)})\}_{k=1}^n$ における $(x,y) \in A^{(i)} \times A^{(j)}$ の頻度

パラメータの分布として Direchlet($a = 1/2$) を仮定

$$R^n(i) := \frac{\Gamma(n + \alpha^{(i)} a)\Gamma(a)^{\alpha^{(i)}}}{\Gamma(\alpha^{(i)} a)\prod_{x \in A^{(i)}} \Gamma(c_n[x] + a)}$$

$$R^n(i,j) := \frac{\Gamma(n + \alpha^{(i)}\alpha^{(j)} a)\Gamma(a)^{\alpha^{(i)}\alpha^{(j)}}}{\Gamma(\alpha^{(i)}\alpha^{(j)} a)\prod_{x \in A^{(i)}} \Gamma(c_n[x,y] + a)}$$

$\{(x_k^{(1)}, \cdots, x_k^{(N)})\}_{k=1}^n$ の記述長 L

$$-\log R^n(i) \approx nH_n(i) + \frac{\alpha^{(i)} - 1}{2} \log n$$

$$-\log R^n(i,j) \approx nH_n(i,j) + \frac{\alpha^{(i)}\alpha^{(j)} - 1}{2} \log n$$

$$\begin{aligned}
& -\log R^n(i,j) + logR^n(j) \\
\approx\ & n\{H_n(i,j) - H_n(j) + \frac{\alpha^{(j)}(\alpha^{(i)} - 1)}{2n} \log n\} \\
=\ & n\{H_n(i) + \frac{\alpha^{(i)} - 1}{2n} \log n - I_n(i,j) + \frac{(\alpha^{(j)} - 1)(\alpha^{(i)} - 1)}{2n} \log n\}
\end{aligned}$$

$$\begin{aligned}
L \approx\ & n \sum_{i \in V} \{H_n(i) + \frac{\alpha^{(i)} - 1}{2n} \log n\} \\
& -n \sum_{\{i,j\} \in E} \{I_n(i,j) - \frac{1}{2n}(\alpha^{(i)} - 1)(\alpha^{(j)} - 1) \log n\}
\end{aligned}$$

Chow-Liu アルゴリズム: 記述長最小

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

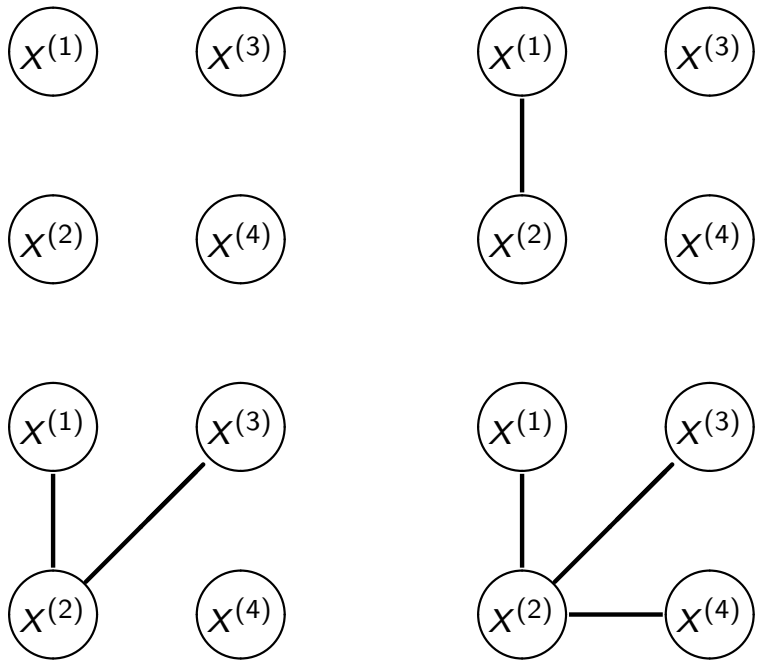
- $J_n(i,j)$ は負になりうるので、拡張版 Kruscal のアルゴリズムを適用
- (V,E) は、木ではなく、一般的な森

1. $J_n(i,j)$ の大きい $i,j \in V (i \neq j)$ から辺を結ぶ
2. ループを作る場合、結ばない
3. $J_n(i,j) < 0$ の場合、結ばない

Chow-Liu アルゴリズム: 記述長最小 (Suzuki, 1993)

記述長を最小にする $\hat{Q}_{1,\cdots,N}$ を表現する森が得られる

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



Radon-Nikodym の定理

$(\Omega, \mathcal{F})$: 可測空間
μ, ν: F 上の測度
$\mathcal{B}$: $\mathbb{R}$ の Borel 集合

μ が ν に対して絶対連続 $(\mu << \nu)$ 各 $A \in \mathcal{F}$ について、
$$\nu(A) = 0 \Longrightarrow \mu(A) = 0$$

ν が σ-有限 $\Omega = \cup_i A_i$ かつ $\nu(A_i) < \infty$ なる $A_i \in \mathcal{F}$, $i = 1, 2, \cdots$ が存在

$f : \Omega \to \mathbb{R}$ が $\mathcal{F}$ 可測 任意の $D \in \mathcal{B}$ について、
$$\{\omega \in \Omega | X(\omega) \in D\} \in \mathcal{F}$$

Radon-Nikodym の定理

μ, ν が σ 有限で、 $\mu \ll \nu$ のとき、各 $A \in \mathcal{F}$ で $\mu(A) = \int_A f d\nu$ となるような、$\mathcal{F}$ 可測な $\dfrac{d\mu}{d\nu} := f \geq 0$ が存在

Kullback-Leibler 情報量

$(\Omega, \mathcal{F}, \mu)$: 確率空間

$X : \Omega \to \mathbb{R}$ が確率変数

X が $\mathcal{F}$ 可測

離散、連続、どちらでもないものが存在する

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

Kullback-Leibler 情報量

$\mu \ll \nu$ のとき、

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

一般の Chow-Liu アルゴリズム: 近似

$D^{(1)}, \cdots, D^{(N)} \in \mathcal{B}$

$$\mu_i(D^{(i)}) := \mu(X^{(i)} \in D^{(i)})$$

$$\mu_{ij}(D^{(i)}, D^{(j)}) := \mu(X^{(i)} \in D^{(i)}, X^{(j)} \in D^{(j)})$$

$$\mu_{1,\cdots,N}(D^{(1)}, \cdots, D^{(N)}) := \mu(X^{(i)} \in D^{(i)}, i = 1, \cdots, N)$$

$$\mu_1 \otimes \cdots \otimes \mu_N(D^{(1)}, \cdots, D^{(N)}) := \prod_{i=1}^{N} \mu_i(D^{(i)})$$

$\mu_{ij} \ll \mu_i \otimes \mu_j$ のとき、相互情報量 $I(i,j) := D(\mu_{ij}||\mu_i \otimes \mu_j)$ が定義

$$d\nu_{1,\cdots,N}(x^{(1)},\cdots,x^{(N)}) = \frac{\prod_{\{i,j\}\in E} d\mu_{i,j}(x^{(i)},x^{(j)})}{\prod_{i\in V} d\mu_i(x^{(i)})^{d_i-1}}$$

定理 1

$$D(\mu_{1,\cdots,N}||\nu_{1,\cdots,N}) = D(\mu_{1,\cdots,N}||\mu_1 \otimes \cdots \otimes \mu_N) - \sum_{\{i,j\}\in E} I(i,j)$$

ユニバーサル測度 (有限の確率変数の場合)

X: 有限集合 A に値をとる確率変数
$c_n[x]$: 長さ n の系列 $x^n := \{x_k\}_{k=1}^n \in A^n$ における $x \in A$ の頻度

$$R^n(x^n) := \frac{\Gamma(n + |A|/2)\Gamma(1/2)^{|A|}}{\Gamma(|A|/2)\prod_{x\in A}\Gamma(c_n[x] + 1/2)}$$

H: X を i.i.d で発生させたときのエントロピー
X の分布 P によらず、$-\frac{1}{n}\log R^n(x^n) \to H\ (n \to \infty)$
$P^n(x^n)$: $X^n = x^n$ の真の確率
Shannon-McMillan-Breiman 定理
X の分布 P によらず、$-\frac{1}{n}\log P^n(x^n) \to H\ (n \to \infty)$

ユニバーサル測度 R

X の分布 P によらず、$\frac{1}{n}\log\frac{P^n(x^n)}{R^n(x^n)} \to 0\ (n \to \infty)$

ユニバーサル測度 (確率密度関数が存在する場合)

$X \in [0,1)$ のとき、$A_0 := \{[0,1)\}$ として

$$
\begin{aligned}
A_1 &= \{[0,1/2), [1/2,1)\} \\
A_2 &= \{[0,1/4), [1/4,1/2), [1/2,3/4), [3/4,1)\} \\
&\cdots \\
A_k &= \{[0,2^{-(k-1)}), [2^{-(k-1)}, 2\cdot 2^{-(k-1)}), \\
&\quad \cdots, [(2^{k-1}-1)2^{-(k-1)}, 1)\} \\
&\quad \cdots
\end{aligned}
$$

各レベルの予測分布を幅で割ったものの重みづけ和

ユニバーサル確率密度 g (Ryabko, 2009)

X の確率密度 f によらず、$\frac{1}{n}\log\frac{f^n(x^n)}{g^n(x^n)} \to 0 \ (n \to \infty)$

ユニバーサル測度 (確率密度関数が存在しない場合)

η: σ 有限, $\mu \ll \eta$

Suzuki, 2011

確率密度関数 $f := \dfrac{d\mu}{d\lambda}$ ではなく、一般の $\dfrac{d\mu}{d\eta}$ を推定

(アルゴリズムとして、λ を η に変えるだけ)

$$D(\mu_k||\eta) \to D(\mu||\eta)\ (k \to \infty) \Longrightarrow \frac{1}{n} \log \frac{d\mu^n}{d\nu^n}(x^n) \to 0\ (n \to \infty)$$

(μ を、η に関するユニバーサル測度とよぶ)

Bayes 的な相互情報量の推定

(x^n, y^n): 確率変数 (X, Y) の n 個のサンプル

$\mu_X \ll \eta_X$, $\mu_Y \ll \eta_Y$

$\frac{d\mu_X^n}{d\eta_X^n}(x^n)$, $\frac{d\mu_Y^n}{d\eta_Y^n}(y^n)$, $\frac{d\mu_{XY}^n}{d\eta_X^n d\eta_Y^n}(x^n, y^n)$ を、

ユニバーサル測度 $\frac{d\nu_X^n}{d\eta_X^n}(x^n)$, $\frac{d\nu_Y^n}{d\eta_Y^n}(y^n)$, $\frac{d\nu_{XY}^n}{d\eta_X^n d\eta_Y^n}(x^n, y^n)$ で推定

定理 2: 提案する相互情報量の一般的な推定量

$$\frac{1}{n} \log \frac{\frac{d\nu_{XY}^n}{d\eta_X^n d\eta_Y^n}(x^n, y^n)}{\frac{d\nu_X^n}{d\eta_X^n}(x^n) \frac{d\nu_Y^n}{d\eta_Y^n}(y^n)} \to I(X, Y) \tag{1}$$

一般的な意味での記述長最小基準

1. 可測空間 $(\Omega, \mathcal{F})$ で、$\mu(\Omega) = 1$ となる異なる測度の列 $\{\mu[m]\}_{m\in\mathbb{N}}$ に

$$\frac{1}{n} \log \frac{\mu^n[m]}{\nu^n[m]}(x^n) \to 0$$

なる $\{\nu[m]\}_{m\in\mathbb{N}}$ が用意されている。

2. ある $m^* \in \mathbb{N}$ について、$\mu[m^*]$ にしたがって独立に発生した n 個のサンプル x^n から、記述長

$$-\log \frac{d\nu^n[m]}{d\eta^n}(x^n)$$

を最小にする m を見出し、m^* の推定値とする。

$$-\log \frac{d\nu[m^*]^n}{d\eta^n}(x^n) \leq -\log \frac{d\nu[m]^n}{d\eta^n}(x^n) \Longleftrightarrow -\log \frac{d\nu[m^*]^n}{d\nu[m]^n}(x^n) \leq 0$$

一般の Chow-Liu アルゴリズム: 記述長最小

$K_n(i,j)$: $X^{(i)}, X^{(j)}$ の Bayes 的な相互情報量の推定値
$\{(x_k^{(1)}, \cdots, x_k^{(N)})\}_{k=1}^n$ の記述長は

$$\sum_{i \in V} -\log \frac{d\nu_i^n}{d\eta_i^n}(\{x_k^{(i)}\}_{k=1}^n) - n \sum_{\{i,j\} \in E} K_n(i,j)$$

1. $K_n(i,j)$ の大きい $i,j \in V$ $(i \neq j)$ から辺を結ぶ
2. ループを作る場合、結ばない
3. $K_n(i,j) < 0$ の場合、結ばない

一般の Chow-Liu アルゴリズム: 記述長最小

生成された森の測度 ν^* が、他の森の測度 ν^n に対して

$$-\log \frac{d\nu_*^n}{d\nu^n}(x^n) \leq 0$$

まとめ

離散や連続を仮定しない一般の確率変数について、

- Bayes 的な相互情報量の提案
- 記述長最小基準の提案
- Bayes 的な Chow-Liu アルゴリズムの提案

今後の課題:

- 数値実験