情報論的学習理論テクニカルレポート 2009
Technical Report on Information-Based Induction Sciences 2009 (IBIS2009)

Chow-Liu アルゴリズムの一般化と、木の複雑さを考慮した修正版について

A generalized version of Chow-Liu algorithm for data-mining.

石田　悠*　　鈴木　譲†
Yuu Ishida　　Joe Suzuki

1　まえがき

確率空間 $(\Omega, \mathcal{F}, \mu)$ を決めれば、$\mathcal{F}$ 上で可測な写像 $X : \Omega \to \mathbb{R}$ は確率変数とよばれる。しかしながら、工学的な応用では、X の値域が離散の値をとるとか、その分布関数 F_X が絶対連続 (確率密度関数 f_X が存在) といった特殊な場合を扱うことが多いようである。たとえば、確率変数 X, Y の相互情報量は、離散の場合

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

連続の場合

$$I(X,Y) := \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f_{XY}(x,y) \log \frac{f_{XY}(x,y)}{f_X(x)f_Y(y)} dxdy$$

と定義される。しかし、一般の確率変数 X, Y については、ほとんど議論されていないようである。また、分布間の隔たりを示す Kullback-Leibler 情報量に関しても、連続とか離散とか特殊な場合についてだけ定義されている。

本論文では、データマイニングで計算量の低減を目的に使われる、Chow-Liu アルゴリズムの一般化、および木の複雑さを考慮したアルゴリズムの修正版 (正規分布版) を提示する。N 個の確率変数 $X^{(1)}, \cdots, X^{(N)}$ の間の Markov ネットワークを木 (巡回経路をもたないネットワーク) で近似する問題を考える。

巡回経路を生成しない限り相互情報量を最大にする $I(X^{(i)}, X^{(j)})$ を辺として結合していくと、両者の分布の間の Kullback-Leibler 情報量を最小にする木が得られる (Chow-Liu, 1968)。分布の近似ではなく、学習のように n 個の例 $\{(x_{i,1}, \cdots, x_{i,N})\}_{i=1}^n$, $x_{i,j} \in X^{(j)}, j = 1, \cdots, N$ からはじまっても、相対頻度を用いて empirical に相互情報量が定義され、Chow-Liu アルゴリズムが同様に適用される。

しかし、Chow-Liu のオリジナル論文をはじめ、従来は N 個の確率変数が有限個の値をとる場合のみを扱っていた。本論文では、確率空間 $(\Omega, \mathcal{F}, \mu)$ における勝手な N 個の確率変数の場合にどのようになるかを検討する。その際に、一般的な相互情報量の定義が必要となる。

学習の場合、Kullback-Leibler 情報量最小という基準では、尤度 (適合度) 最大という基準で木を選択することになる。MDL 原理を適用して、木の複雑さを考慮した場合でも、やはり相互情報量を修正 (2 個の確率変数の取りうる値の数を考慮) することによって最適解が与えられる (Suzuki, 1993)。しかし、これもやはり、N 個の確率変数が有限個の値をとることを仮定している。本論文では、N 個の確率変数が正規分布にしたがう場合にどのように修正できるかを示す。

2　準備

2.1　無向グラフ

U を有限集合、E を $\mathcal{E} := \{\{X,Y\} \| X, Y \in U, X \neq Y\}$ の部分集合とする。U, E の対 (U, E) を無向グラフと呼ぶ。無向グラフ $G = (U, E)$ について、U を G の頂点集合、その要素を G の頂点と呼ぶ。また E を G の辺集合、その要素を G の辺と呼ぶ。

無向グラフ G に巡回経路が存在しないとき、G は森であるという。また G が森であって連結されているとき、G は木であるという。

*大阪大学大学院理学研究科, 〒 560-0043 大阪府豊中市待兼山町 1-1, e-mail yawn.ikaryaku@gmail.com
Graduate school of Science School of Science,Osaka University, Machikaneyamacho1-1.Toyonakashi.Osaka.560-0043
†大阪大学大学院理学研究科, 〒 560-0043 大阪府豊中市待兼山町 1-1, e-mail suzuki@math.sci.osaka-u.ac.jp
Graduate school of Science School of Science,Osaka University, Machikaneyamacho1-1.Toyonakashi.Osaka.560-0043

2.2 Kullback-Leibler 情報量

$\mathcal{B}$ を実数全体 $\mathbb{R}$ の Borel 集合体として、写像 $X:\Omega\to\mathbb{R}$ が

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

($\mathcal{F}$ 上可測) を満足するとき、X は確率空間 $(\Omega,\mathcal{F},\mu)$ の確率変数という。

2 個の $\mathcal{F}$ 上の測度について、$\nu(A)=0\Longrightarrow\mu(A)=0$ のとき、μ は ν について絶対連続であるといい、$\mu<<\nu$ とかく。そして、任意の $A\in\mathcal{F}$ について、$\mu<<\nu$ のとき、

$$\mu(A)=\int_A fd\nu$$

となる $\frac{d\mu}{d\nu}:=f$ が存在する (Radon-Nikodym の定理)。このとき、

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

を測度 μ の測度 ν に対する Kullback-Leibler(K-L) 情報量とよぶ。

$$D(\mu||\nu)\geq 0$$

$$D(\mu||\nu)=0\Longleftrightarrow\mu=\nu$$

などが容易に導かれる。この性質より K-L 情報量が小さいほど μ は ν に近いと考えられる。X が有限の値をとる $(X(\Omega)<\infty)$ とき、$P(x):=\mu_X(\{x\}),x\in X(\Omega)$, $Q(x):=\nu_X(\{x\}),x\in X(\Omega)$ とおくと、$Q(x)=0\Longrightarrow P(x)=0$ であれば、

$$D(P||Q)=\sum_{x\in X(\Omega)}P(x)\log\frac{P(x)}{Q(x)}$$

それぞれの確率密度関数 f,g が存在する場合、$g(x)dx=0\Longrightarrow f(x)dx=0$ であれば、

$$D(f||g)=\int_{-\infty}^{\infty}f(x)\log\frac{f(x)}{g(x)}dx$$

となる。

確率空間 $(\Omega,\mathcal{F},\mu)$ における 2 個の確率変数 X,Y について、X,Y および X,Y の直積が生成する σ-集合体 $\mathcal{F}_X,\mathcal{F}_Y,\mathcal{F}_{XY}$ に対する測度を μ_X,μ_Y,μ_{XY} とかくとき、積 $\mu_X\mu_Y$ も $\mathcal{F}_{XY}$ に対する測度になる。
そこで、$I(X,Y):=D(\mu_{XY}||\mu_X\mu_Y)$ を X,Y の相互情報量とよぶ。X,Y が有限の値をとる $(X(\Omega),Y(\Omega)<\infty)$ とき、$P_X(x):=\mu_X(\{x\}),x\in X(\Omega)$, $P_Y(y):=\mu_Y(\{y\}),y\in Y(\Omega)$ $P_{XY}(x,y):=\mu_{XY}(\{x\},\{y\}),x\in X(\Omega),y\in Y(\Omega)$ とおくと、

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

確率密度関数 f_X,f_Y,f_{XY} が存在する場合、

$$I(X,Y)=\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}f_{XY}(x,y)\log\frac{f_{XY}(x,y)}{f_X(x)f_Y(y)}dxdy$$

となる。

2.3 Chow-Liu アルゴリズム

n 個の例から推定される確率変数間の条件付独立性を表現する無向グラフを、特に木の範囲で推定する (推定したものを木で近似するとみなしてもよい) ことを考える。そのために、分布 $P_{X^{(1)},...X^{(n)}}(x^{(1)},...,x^{(N)})$, $x^{(1)}\in X^{(1)}(\Omega)$,...,$x^{(N)}\in X^{(N)}(\Omega)$ を

$$\begin{aligned}&\hat{P}_{X^{(1)},...,X^{(N)}}(x^{(1)},...,x^{(N)})\\:=&\prod_{j\in N}P_{X^{(a_j)}|X^{(a_{\lambda_{[j]}})}}(x^{(a_j)},x^{(a_{\lambda_{[j]}})})\end{aligned}$$

(Dendoroid 分布) で近似する問題を考える。ここで、$1\leq\lambda[j]\leq j-1$,$j\neq 1$,$\lambda[1]=0$, $P_{X^{(j)}|X^{(0)}}(\cdot|x^{(0)})=P_{x^{(j)}}(\cdot)$ とし、$(a_1,...,a_N)$ は $(1,...,N)$ の順列であるとする。

補題 1 $U:=\{X^{(i)}|i=1,...,N\}$,
$E\subseteq\mathcal{E}:=\{\{X^{(i)},X^{(j)}\}|i\neq j\}$ をそれぞれ頂点集合、辺集合とする無向グラフ $G=(U,E)$ について、G が木であるならば
$E=\{\{X^{(a_j)},X^{(a_{\lambda[j]})}\}|j=1,...,N\}$、
$1\leq\lambda[j]\leq j-1$,$j\neq 1$,$\lambda[1]=0$, となる $(1,...,N)$ の順列 $(a_1,...,a_N)$ が存在する。

命題 1 (Chow-Liu) *[1]*

$$\begin{aligned}D&\quad(P_{X^{(1)},...,X^{(N)}}|\hat{P}_{X^{(1)},...,X^{(N)}})\\:=&\sum_{x^{(1)}\in X^{(1)}(\Omega),...,x^{(N)}\in X^{(N)}(\Omega)}P_{x^{(1)}...X^{(N)}}(x^{(1)},...,x^{(N)})\\\cdot&\quad\log\frac{P_{x^{(1)}...X^{(N)}}(x^{(1)},...,x^{(N)})}{\prod_{j\in N}P_{X^{(a_j)}|X^{(a_{\lambda_{[j]}})}}(x^{(a_j)},x^{(a_{\lambda_{[j]}})})}\end{aligned}$$

を最小にする $(a_1,...,a_N)$ および $\lambda[k]$、$2\leq k\leq N$ は

$$\begin{aligned}I(i,j)&:=\sum_{x^{(i)}\in X^{(i)}(\Omega),...,x^{(j)}\in X^{(j)}(\Omega)}P_{X^{(i)}X^{(j)}}((x^{(i)},x^{(j)})\\&\cdot\quad\log\frac{P_{X^{(i)}X^{(j)}}(x^{(i)},x^{(j)})}{P_{X^{(i)}}P_{X^{(j)}}}\end{aligned}$$

として、重み最大極大木アルゴリズムに
$U:=\{X^{(i)}|i=1,...,N\},w(X^{(i)},X^{(j)}):=I(X^i,X^j)$
を入力として与えることによって得られる。

証明:

$$
\begin{aligned}
D & \quad (P_{X^{(1)},\cdots,X^{(N)}} | \hat{P}_{X^{(1)},\ldots,X^{(N)}}) \\
= & \sum_{x^{(1)}\in X^{(1)}(\Omega),\ldots,x^{(N)}\in X^{(N)}(\Omega)} P_{X^{(1)}\ldots X^{(N)}}(x^{(1)},\ldots,x^{(N)}) \\
\cdot & \quad \log \frac{P_{X^{(1)}\ldots X^{(N)}}(x^{(1)},\ldots,x^{(N)})}{\prod_{j\in N} P_{X^{(a_j)}|X^{(a_{\lambda_{[j]}})}}(x^{(a_j)}, x^{(a_{\lambda[j]})})} \\
= & -\sum_{a_{\lambda_{[j]}}\to a_j} I(X^{(a_j)}, X^{(a_{\lambda_{[j]}})}) + H(X^{(1)},\cdots,X^{(N)}) \\
- & \sum_{i=1}^{N} H(X^{(i)})
\end{aligned}
$$

$$
\begin{aligned}
H(X^{(1)},\cdots,X^{(N)}) & := \sum_{X^{(1)},\cdots,X^{(N)}} -P_{X^{(1)},\cdots,X^{(N)}}(x^{(1)},\cdots,x^{(N)}) \\
& \cdot \quad \log P_{X^{(1)},\cdots,X^{(N)}}(x^{(1)},\cdots,x^{(N)})
\end{aligned}
$$

$$
H(X^{(i)}) := \sum_{x^{(i)}} -P_{X^{(i)}}(x^{(i)}) \log P_{X^{(i)}}(x^{(i)})
$$

より $D(P||\hat{P})$ の最小化は $\sum_{a_{\lambda_{[j]}}\to a_j} I(X^{(a_{\lambda_j})}, X^{(a_{\lambda_{[j]}})})$ の最大化に等しい。

補題 1 より、これは辺 $\{X^{(a_j)}, X^{(a_{\lambda[j]})}\}$ に割り当てられた重みの合計を最小にする木を見出す問題となる。

n 個の例 $x^n = \{x_1, x_2, \ldots x_n\}$, $x_i \in X(\Omega) = \prod_{j=1}^{N} X^{(j)}(\Omega)$,$i = 1,\ldots,n$ からその条件付独立性を表現する無向グラフを木の範囲で推定をすることを考える。そして、命題 1 の $I(i,j)$ を

$$
I_n(i,j) := \sum_{x\in X^{(i)}, y\in X^{(j)}} c_{i,j}[x,y] \log \frac{c_{i,j}[x,y]}{c_i[x]c_j[y]} \geq 0
$$

に置き換えればよい。ただし $c_i[x]$,$c_{i,j}[x,y]$ を、それぞれ $x_1,\ldots,x_n$ における $X^{(i)} = x \in X^{(i)}(\Omega)$ の頻度、$X^{(i)} = x \in X^{(i)}(\Omega)$、$X^{(j)} = y \in X^{(j)}(\Omega)$ の同時頻度とした。したがって $D(P_{X^{(1)},\ldots,X^{(n)}} | \hat{P}_{X^{(1)},\ldots,X^{(n)}})$ の最小化は

$$
\begin{aligned}
H[\pi](x^n) & := \sum_{x^{(i)}\in X^{(i)}(\Omega)} c_{1,\ldots,N}(x^{(1)},\ldots,x^{(N)}) \\
& \cdot \quad \log \prod \frac{c_{a_{\lambda[j]}}(x^{(a_{\lambda[j]})})}{c_{a_j,a_{\lambda[j]}}(x,^{(a_j)} x^{(a_{\lambda[j]})})}
\end{aligned}
$$

の最小化に相当する。ただし $c_{1,\ldots,N}(x^{(1)},\ldots,x^{(N)})$ を $X^{(1)} = x^{(1)} \in X^{(1)}(\Omega),\ldots,X^{(N)} = x^{(N)} \in X^{(N)}(\Omega)$ が同時に成立する頻度であるとした。

アルゴリズム 1 (相対頻度を用いた Chow-Liu アルゴリズム)

入力 n 個の例 x^n

出力 $H[\pi](x^n)$ を最小とする木 T

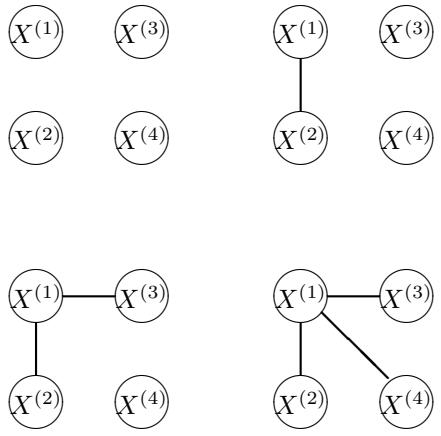
1. $E \leftarrow \{\}$
2. $\mathcal{E} \neq \{\}$ であるかぎり以下を繰り返す。
   (a) $I_n(i,j)$ が最大の $\{X^{(i)}, X^{(j)}\} \in \mathcal{E}$ について、$\mathcal{E} \leftarrow \mathcal{E} - \{X^{(i)}, X^{(j)}\}$
   (b) $(U, E \cup \{X^{(i)}, X^{(j)}\})$ が巡回経路を持たないとき、$E \leftarrow E \cup \{X^{(i)}, X^{(j)}\}$
   (c) $T \leftarrow (U, E)$

例 1 $I_n(i,j)$ の各値が表 1 で与えられているとき

1. $I_n(1,2)$ が最大なので、最初に $X^{(1)}, X^{(2)}$ を結ぶ。
2. それ以外では、$I_n(1,3)$ が最大なので、$X^{(1)}, X^{(3)}$ を結ぶ。
3. それ以外では、$I_n(2,3)$ が最大だが、$X^{(2)}, X^{(3)}$ を結ぶと巡回経路が出来るので、$X^{(2)}, X^{(3)}$ は結ばない。
4. それ以外では、$I_n(1,4)$ が最大なので、$X^{(1)}, X^{(4)}$ を結ぶ。
5. 候補となる辺はあるが、これ以上辺を結ぶと巡回経路が出来るのでアルゴリズムを停止する。

表 1: 各 (i,j) に対する相互情報量

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



2.4 木の複雑さを考慮した Chow-Liu アルゴリズムの一般化

前節では n 個の例 $x_i \in X(\Omega)$、i=1,2, $\cdots$,n から、その条件付独立性を表現する無向グラフを木の範囲で推定することを考えた。
だが上記のアルゴリズムを用いてサンプルからの学習を行う場合、真のグラフが連結でない場合でも出力される木は連結されたものとなる。つまり、興味のある事象とは本来無関係の事象までも、一つの木の要素に組み込んでしまうということが起こりうる。それを解決するため、以下では木 π の構造の簡潔さまでを考慮してモデリングを行う方法を考える。具体的には $H[\pi](x^n)$ の最小化ではなくパラメータ数 $K[\pi]$ を加えた

$$L[\pi](x^n) := H[\pi](x^n) + \frac{K[\pi]}{2}\log n$$

と定義した $L[\pi](x^n)$ を最小化する方法を考える。ここで $H[\pi](x^n)$ は木 π の例 x^n への適合性、$\frac{K[\pi]}{2}\log n$ は木 π の複雑さを表している。[3, pp.106-117]
以下

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

とおく。($\alpha^{(i)} := |X^{(i)}(\Omega)|$) これはエントロピー $H[\pi](x^n)$ の減少分だけではなく、辺を加えることによる木の複雑さの増加分を評価している。

補題 2 $U := \{X^{(i)} | i = 1, ..., N\}$
$E \subseteq \varepsilon := \{\{X^{(i)}, X^{(j)}\} | i \neq j\}$ をそれぞれ頂点集合、辺集合とする無向グラフ $G = (U, E)$ について、G が森であるならば、
$E = \{\{X^{(a_j)}, X^{(a_{\lambda[j]})}\} | j = 1, ..., N\}$、$0 \leq \lambda[j] \leq j-1$、$j = 1, ..., N$、$\lambda[1] = 0$ となる $(1, ..., N)$ の順列 $(a_1, ..., a_N)$ が存在する。

木では $1 \leq \lambda[j] \leq j-1$ となるが、森では $0 \leq \lambda[j] \leq j-1$ となる。

アルゴリズム 2 (パラメータ数を考慮した修正版)
(suzuki,1993)[2]

入力: n 個の例 x^n

出力: $L[\pi](x^n)$ を最小とする森 F

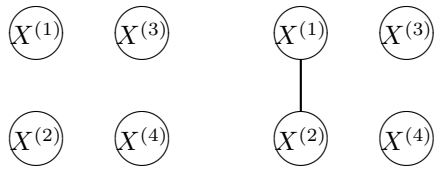
1. $E \leftarrow \{\}$
2. $\mathcal{E} \neq \{\}$ であるかぎり以下を繰り返す。
3. (a) $J_n(i,j)$ が最大の $\{X^{(i)}, X^{(j)}\} \in \mathcal{E}$ について、$\mathcal{E} \leftarrow \varepsilon - \{X^{(i)}, X^{(j)}\}$
   (b) $(U, E \cup \{X^{(i)}, X^{(j)}\})$ が巡回経路を持たず、かつ $J_n(i,j) \geq 0$ のとき、$E \leftarrow E \cup \{X^{(i)}, X^{(j)}\}$
   (c) $F \leftarrow (U, E)$

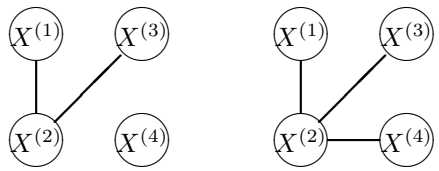
例 2 $I_n(i,j)$ の各値が表 1 で与えられているものとし、$\alpha^{(1)} = 5$,$\alpha^{(2)} = 2$,$\alpha^{(3)} = 3$, $\alpha^{(4)} = 4$ であるとする。

1. 最初に　$J_n(1,2) = 8$ が最大なので、$X^{(1)}, X^{(2)}$ を結ぶ。
2. それ以外では、$J_n(2,3) = 6$ が最大なので、$X^{(2)}, X^{(3)}$ を結ぶ。
3. それ以外では、$J_n(1,3) = 2$ が最大だが、$X^{(1)}, X^{(3)}$ を結ぶと巡回経路が出来るので、　は結ばない。
4. それ以外では、$I_n(2,4) = 1$ が最大なので、$X^{(2)}, X^{(4)}$ を結ぶ。
5. $J_n < 0$ となる辺は連結しない。また連結することによって巡回経路が出来る辺は連結しない。

表 2: 例 2

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





上記のアルゴリズムは n 個の例 x^n から $L[\pi](x^n)$ を最小とする森 F を正しく出力する。

証明: 実際 $X^{(a_j)}$ が $X^{a_{\lambda_{[j]}}}$ に依存するので、

$$K\ [\pi] = \sum_{j=0}^{N}(a_j - 1)(a_{\lambda_{[j]}} - 1)$$

とできる。したがって、$L[\pi](x^n)$ の最小化は、

$$\begin{aligned} & \sum_{j=0}^{N} -I_n(a_j, a_{\lambda_{[j]}}) + \frac{1}{2}\sum_{j=0}^{N}(a_j - 1)(a_{\lambda_{[j]}} - 1)\log n \\ = & -\sum_{j=0}^{N}\{I_n(a_j, a_{\lambda_{[j]}}) - \frac{1}{2}(a_j - 1)(a_{\lambda_{[j]}} - 1)\log n\} \end{aligned}$$

の最小化と同値である。したがって補題 2 より
辺 $(X^{(a_j)}, X^{a_{\lambda_{[j]}}})$ に割り当てられた重みの合計を最小にする森を見出す問題となる。実際、$J_n(i,j) < 0$ なる i,j に関しては、$X^{(i)}, X^{(j)}$ が連結されず、$(X^{(i)}, X^{(j)})$ を辺として加えると $\sum_{j=1}^{n} J_n(a_j, a_{\lambda_{[j]}})$ が増えないので、辺として加える操作は行わないことによって最大化がはかれる。

このアルゴリズムによって最終的に森が木になった場合でも、$|X^{(j)}(\Omega)| \geq 2$ なる $X^{(j)}$ が含まれる場合、$I_n(i,j)$ の順序と $J_n(i,j)$ の順序は一致しないため、アルゴリズム 1 とアルゴリズム 2 では異なる木を出力する場合がある。(例 1、例 2) $J_n(i,j)$ は、辺 $\{X^{(a_j)}, X^{(a_{\lambda[j]})}\}$ を結ぶことによる木表現の例への適合性の増加と、パラメータ数の増加を比べる指標となっている。

3 離散分布を仮定しない、Chow-Liu アルゴリズム

3.1 任意の確率変数の場合への一般化

$X^{(1)}, \cdots, X^{(N)}$ が一般の確率変数の場合においても、Chow-Liu アルゴリズムで Kullback-Leibler 情報量が最小になる。

- 一般的な相互情報量の定義

$\mu_{j|i}$: $X^{(i)}$ を前提とした、$X^{(j)}$ の生成する事象の条件付測度
$\mu_{i,j}$: $X^{(i)}, X^{(j)}$ の生成する事象の測度

$$I(X^{(i)}, X^{(j)}) := \int d\mu_{i,j} \log \frac{d^2\mu_{i,j}}{d\mu_i d\mu_j}$$

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

証明: 無向木の 1 頂点を任意に選び根とする。そして、そこから遠い頂点に向かって、辺に向きをつける。以下では頂点 j の親が頂点 i であることを $i \to j$ と書く。(ただし、i が根の場合 $0 \to i$ と書く)

Radon-Nikodym の定理から、

$$\begin{aligned} \frac{d\mu_{1,\cdots,N}}{d\prod_{i\to j}\mu_{j|i}} &= \frac{d^N\mu_{1,\cdots,N}}{\prod_{i\to j} d\mu_{j|i}} \\ &= \frac{d^N\mu_{1,\cdots,N}}{\prod_{k=1}^{N} d\mu_k} \cdot \frac{\prod_{k=1}^{N} d\mu_k}{\prod_{i\to j} d\mu_{j|i}} \\ &= \frac{d^N\mu_{1,\cdots,N}}{\prod_{k=1}^{N} d\mu_k} \cdot [\prod_{i\to j, i\neq 0} \frac{d^2\mu_{i,j}}{d\mu_i d\mu_j}]^{-1} \end{aligned}$$

であるから、Kullback-Leibler 情報量は

$$\begin{aligned} D(\mu_{1,\cdots,N}||\prod_{i\to j}\mu_{j|i}) &= -\sum_{i\to j, i\neq 0} I(X^{(i)}, X^{(j)}) \\ &+ \int d\mu_{1,\cdots,N} \log\{\frac{d^N\mu_{1,\cdots,N}}{\prod_{k=1}^{N} d\mu_k}\} \end{aligned}$$

と書けて、$I(X^{(i)}, X^{(j)})$ の大きな順に辺を結ぶことによって Kullback-Leibler 情報量が最小の木を得ることが出来る。

3.2 確率変数が正規分布にしたがう場合の修正版

ここまで確率変数が離散型の場合に限って考えていたが、以下では前述のアルゴリズム 1 を正規分布に従う確率変数の条件付独立性の表現に対して実際に適用した場合を考える。

$X^{(i)} \sim N(0, \sigma^2)$ の確率密度関数を

$$f_{X^{(i)}}(x^{(i)}) := \frac{1}{\sqrt{2\pi\sigma_{ii}}} e^{-\frac{x^{(i)2}}{2\sigma_{ii}}}$$

とおく。また $X^{(i)}, X^{(j)}$ の同時確率密度を

$$f_{X^{(i)}X^{(j)}}(x^{(i)}, x^{(j)}) := \frac{1}{2\pi|\Sigma|^{\frac{1}{2}}} e^{-\frac{1}{2}(x^{(i)} \quad x^{(j)})\Sigma^{-1}\begin{pmatrix} x^{(i)} \\ x^{(j)} \end{pmatrix}}$$

とおく。ここで $\Sigma = \begin{pmatrix} \sigma_{ii} & \sigma_{ij} \\ \sigma_{ji} & \sigma_{jj} \end{pmatrix}$ である。このとき相関関数を $\rho_{ij} = \frac{\sigma_{ij}}{\sqrt{\sigma_{ii}\sigma_{jj}}}$ とすると

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

となり $I(i,j)$ の値は相関関数 ρ_{ij} から求めることができる。したがって各 $X^{(i)}$ が正規分布に従うときは、$|\rho_{ij}|$ の大きい順に辺を繋いでいき Chow-Liu アルゴリズムを進めることができる。

$J_n(i,j) = I_n(i,j) - \frac{1}{2}\log n = -\log(1-{\rho_{ij}}^2)^{\frac{1}{2}} - \frac{1}{2}\log n$ とおく

アルゴリズム 3 (正規分布に従うときの修正版)

入力: n 個の例 x^n

出力: $L[\pi](x^n)$ を最小とする森 F

1. $E \leftarrow \{\}$
2. $\mathcal{E} \neq \{\}$ であるかぎり以下を繰り返す。
3. (a) $J_n(i,j)$ が最大の $\{X^{(i)}, X^{(j)}\} \in \mathcal{E}$ について、$\mathcal{E} \leftarrow \varepsilon - \{X^{(i)}, X^{(j)}\}$
   (b) $(U, E \cup \{X^{(i)}, X^{(j)}\})$ が巡回経路を持たず、かつ $J_n(i,j) \geq 0$ のとき、$E \leftarrow E \cup \{X^{(i)}, X^{(j)}\}$
   (c) $F \leftarrow (U, E)$

以上のように確率変数が正規分布にしたがっている場合、例に適合する木は相関関数 ρ から求められる。

3.3 正規分布で、木の複雑さを考慮した Chow-Liu アルゴリズム

- 正規分布で木の複雑さを考慮するとき

$f(\boldsymbol{x}|\Sigma)$: $\boldsymbol{\mu} = 0$, 共分散行列 $\Sigma \in \mathbb{R}^{N\times N}$ の正規分布の確率密度関数

$\hat{\Sigma}(\boldsymbol{x}^n)$: $\boldsymbol{x}^n = (x_1, \cdots, x_n)$ から得られた Σ の最尤推定量

尤度 $H := -\log f(\boldsymbol{x}|\hat{\Sigma}(\boldsymbol{x}^n))dx$

k: Σ における独立なパラメータ数

と置き、

MDL 原理記述長 $L := H + \dfrac{k}{2}\log n \rightarrow$ 最小

となるように

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

とすればよい。

$X^{(i)}, i = 1, ..., N$ が離散確率変数であったときと異なりパラメータ数がどの確率変数についても一定であるために、木の複雑さを考慮しても選択の順序（$I_n(i,j)$ と $J_n(i,j)$ の大小関係の順序）は変化しない。したがって最終的に森が木になった場合にアルゴリズム 1 とアルゴリズム 2 の正規分布に対する修正版は全く同じ木を出力する。

- X_i の分布が、β_i 次元の正規分布に従うとき

共分散行列の次数は $\sum_{i=1}^{N}\beta_i$ であり、独立なパラメータ数 $\sum_{i\rightarrow j}\beta_i\beta_j$ と書ける。

木の範囲で学習するとき $J_n(i,j) := I_n(i,j) - \dfrac{\beta_i\beta_j}{2}\log n$ とすればよい。この場合、木の複雑さを考慮しても選択する順序は異なる。

3.4 正規分布、離散分布が混在している際の Chow-Liu アルゴリズム

【Conjecture】

$X^{(i)}$ は正規分布に従い、$X^{(j)}$ は離散分布に従う確率変数とする。

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

と置く。($\alpha^{(j)} := |X^{(j)}(\Omega)|$) このときアルゴリズム 2 は正規分布、離散分布が混在している場合においても $L[\pi](x^n)$ を最小とする森を出力する。

4 まとめ

確率変数間の条件付独立性を木で表現する手法として、確率変数が有限離散の場合と連続の場合の Chow-Liu アルゴリズムについて考えてきた。どちらの場合であっても与えられた例に適合する木が出力されることが分かったが、Kullback-Leibler(K-L) 情報量 $D(P||Q)$ と相互情報量 $I(i,j)$ の関係から、考慮する確率変数の集合が離散、連続の入り混じったものであっても同様のアルゴリズムが適用できるであろうことが予想される。正規分布での例から分かるように結びつける確率変数のパラメータ数が一定である場合は、Chow-Liu アルゴリズムとその修正版とで確率変数を結び付けていく順序に違いはないが、パラメータ数の異なる確率変数の集合にアルゴリズムを適用した場合、結びつける順序によって出力される木構造の簡潔さに大きな差が生まれるためパラメータ数を考慮した修正版のアルゴリズムは大きな意味を持つ。

参考文献

[1] C.K.CHOW,C.N.LIU: Approximating Discrete Probability Distributions with Dependence Trees.(1968)

[2] J.Suzuki,"A construction of bayesian networks from databases on an MDL principle",Proc 9th UAI,Morgan,Kaufmann,pp.266-273,1993.

[3] 鈴木譲: ベイジアンネットワーク入門［確率的知識情報処理の基礎］(2009)