

目 次

第 1 章	グラフィカルモデルの構造学習	1
1.1	はじめに	1
1.2	Bayesian ネットワークと Markov ネットワーク	1
1.3	事後確率最大の森の構造学習	6
1.3.1	局所スコア	6
1.3.2	独立性の検定	8
1.3.3	Chow-Liu アルゴリズム	10
1.3.4	条件付き独立性の検定	13
1.4	事後確率最大の BN の構造学習	15
1.4.1	大域スコア	15
1.4.2	動的計画法としての定式化	16
1.4.3	最短経路問題としての定式化	20
1.4.4	条件付き独立性と BN の構造推定	22
1.5	MDL 原理の適用	23
1.5.1	スコアの導出	23
1.5.2	計算量の削減	25
1.5.3	相互情報量推定への応用	27
1.6	おわりに	28
	参考文献	29

第 1 章

グラフィカルモデルの構造学習

1.1 はじめに

複数の属性に関する有限個のサンプルから、Bayesian ネットワークや Markov ネットワークといったグラフィカルモデルの構造を学習する問題を考える。各構造に事前確率が与えられていれば、サンプルに基いて事後確率最大の構造を選択することができる。

学術的な研究は 1990 年代の前半に始まって 2000 年くらいまでのうちに、基本的な問題が解決したように思われたが、2010 年以降でも新たな進展が見られている。また、ビックデータ時代の幕開けということもあって、応用面でも広がりを見せている。

本章では、グラフィカルモデルの構造学習の題材のうち、オーソドックスなことだけを取り上げた。ただ、これらのことを系統だって平易に書かれているものは、論文でも単行本でも、英語でも日本語でも皆無であるように思われる。大学院生やビックデータの若手で数学を苦にしない方であれば、一読する意味があるものと思われる。

1.2 Bayesian ネットワークと Markov ネットワーク

有向非巡回グラフ (DAG, directed acyclic graph)、無向グラフ (undirected graph) を用いて、確率変数の間の依存関係をあらわしたものを、それぞれ Bayesian ネットワーク (BN) および Markov ネットワーク (MN) と

2 第1章 グラフィカルモデルの構造学習

よぶ。

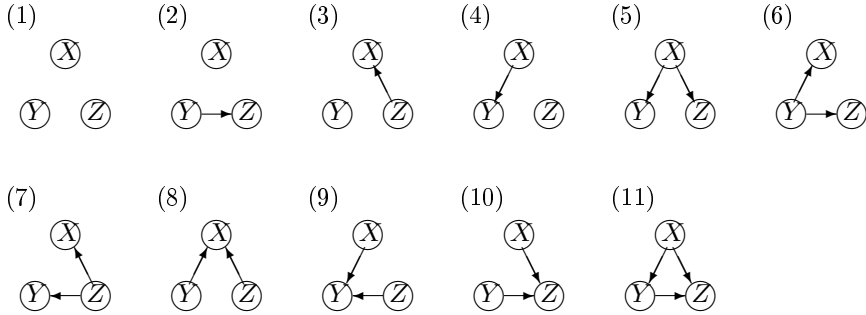


図 1.1 Bayesian ネットワーク

以下では、 $N \geq 2$ として、 $X^{(1)}, \dots, X^{(N)}$ を有限個の値をとる確率変数とする。

まず、BN を分布の因数分解という概念を用いて定義する。簡単のため、 $N = 3$ とし、3 変数を X, Y, Z とおくと、その分布 $P(XYZ)$ は、以下の 11 個のいずれかに因数分解される。

$$P(X)P(Y)P(Z)$$

$$P(X)P(YZ), P(Y)P(ZX), P(Z)P(XY)$$

$$\frac{P(ZX)P(XY)}{P(X)}, \frac{P(XY)P(YZ)}{P(Y)}, \frac{P(ZX)P(XY)}{P(Z)}$$

$$\frac{P(Y)P(Z)P(XYZ)}{P(YZ)}, \frac{P(Z)P(X)P(XYZ)}{P(ZX)}, \frac{P(X)P(Y)P(XYZ)}{P(XY)},$$

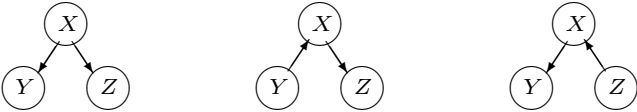
$$P(XYZ)$$

これらは図 1.1 のように、ループをもたない有向グラフ、すなわち DAG であらわされる。

他方、DAG は、 $N = 3$ であれば、25 個ある (3^3 個の有向グラフのうち、時計回り、反時計回りのループを除外する)。しかし、 $P(X)P(Y|X)P(Z|X)$, $P(Y)P(X|Y)P(Z|X)$, $P(Z)P(X|Z)P(Y|X)$ は最終的に $\frac{P(XY)P(XZ)}{P(X)}$ 、

1.2 Bayesian ネットワークと Markov ネットワーク 3

$P(Y)P(Z)P(X|YZ)$ は最終的に $\frac{P(X)P(Z)P(XYZ)}{P(ZX)}$ に因数分解される
 というように、本来は 25 種類あった因数分解のあるもののどうしが同一視
 され、11 個のクラスに分類される (図 1.1)。以下では、これら 11 個の式を
 (1)-(11) とかくものとする。



$$\begin{aligned}
 & \underbrace{P(X)P(Y|X)P(Z|X) = P(Y)P(X|Y)P(Z|X) = P(Z)P(X|Z)P(Y|X)}_{= \frac{P(XY)P(XZ)}{P(X)}} \quad (5) \\
 & \underbrace{P(Y)P(Z)P(X|YZ)}_{= \frac{P(Y)P(Z)P(XYZ)}{P(YZ)}} \quad (8)
 \end{aligned}$$

図 1.2 複数の因数分解が同一視されるケース: (5) 式と (8) 式

他方、MN は以下のように定義される (Hammersley-Clifford)。 V_1, \dots, V_M をそれぞれ、確率変数 $X^{(1)}, \dots, X^{(N)}$ の部分集合で互いに包含関係をもたないものとして、 f_1, \dots, f_M をそれぞれ V_1, \dots, V_M の各値に対応して、正の値を返す関数とする。そして、 $P(X^{(1)}, \dots, X^{(N)})$ がこれらを用いて、 $\frac{1}{K} \prod_{k=1}^M f_k(V_k)$ の形でかけるとき、各 V_k の中の変数のすべての対を無向辺で結んだものをクリークとよび、それらの和集合を辺集合にもつ無向グラフを MN とよぶ。ただし、 K は正規化定数である。たとえば、 $N = 3$ 変数がそれぞれ $\{0, 1\}$ の値を取り、 $V_1 = \{X, Y\}$, $V_2 = \{X, Z\}$, $M = 2$ であれば、

$$P(X = x, Y = y, Z = z) = \frac{f_1(x, y)f_2(x, z)}{K}, \quad x, y, z \in \{0, 1\}$$

$$\begin{aligned}
 K = & f_1(0, 0)f_2(0, 0) + f_1(0, 0)f_2(0, 1) + f_1(0, 1)f_2(0, 0) + f_1(0, 1)f_2(0, 1) \\
 & + f_1(1, 0)f_2(1, 0) + f_1(1, 0)f_2(1, 1) + f_1(1, 1)f_2(1, 0) + f_1(1, 1)f_2(1, 1)
 \end{aligned}$$

4 第1章 グラフィカルモデルの構造学習

となる。この場合、 $\{X, Y\}, \{X, Z\}$ がクリークである。 $N = 3$ であれば、クリークが $\{X\}$ と $\{Y\}$ と $\{Z\}$ 、 $\{Y, Z\}$ 、 $\{Z, X\}$ 、 $\{X, Y\}$ 、 $\{Z, X\}$ と $\{X, Y\}$ 、 $\{X, Y\}$ と $\{Y, Z\}$ 、 $\{Y, Z\}$ と $\{Z, X\}$ 、 $\{X, Y, Z\}$ となる分布をそれぞれ (12)-(19) と書くと、それぞれの MN は図 1.3 のように書ける。

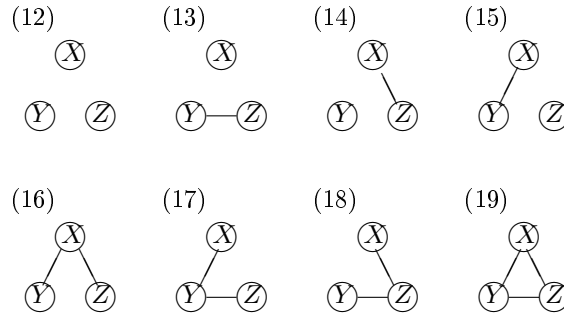


図 1.3 Markov ネットワーク

ここで、確率変数 X, Y について、 $P(XY) = P(X)P(Y)$ であれば、 X, Y が独立であるといい、 $X \perp\!\!\!\perp Y$ とかく。この概念を一般化して、確率変数 X, Y, Z について、 $P(Z = z) \neq 0$ なる各 z について、

$$P(XY|Z = z) = P(X|Z = z)P(Y|Z = z)$$

が成立するとき、 X, Y が Z のもとで条件付き独立であるといい、 $X \perp\!\!\!\perp Y|Z$ とかく (一般の確率変数の集合 $\mathbf{X}, \mathbf{Y}, \mathbf{Z}$ については、 $\mathbf{X} \perp\!\!\!\perp \mathbf{Y}|\mathbf{Z}$ とかく)。ただし、要素が 1 個しかない場合は、たとえば、 $\{X\} \perp\!\!\!\perp \{Y\}$ や $\{X\} \perp\!\!\!\perp \{Y, Z\}|\{W\}$ ではなく $X \perp\!\!\!\perp Y$ や $X \perp\!\!\!\perp \{Y, Z\}|W$ と記述することが多い。たとえば、(1)-(19) の各式で成立する条件付き独立性は、表 1.1 のようになる。

本章の冒頭では、確率変数の間の依存関係という曖昧な表現を用いたが、BN も MN も本来は条件付き独立性をあらわすグラフィカルモデルとして定義される (第 1 章「Bayesian ネットワークの基礎」参照)。BN の因数分解による定義、MN の Hammersley-Clifford による定義も、条件付き独立性による定義と同値であることが証明されている [8]。

(8)-(10) の表現する条件付き独立性は、BN で表現できても MN では表現

表 1.1 (1)-(19) で成立する条件付き独立性

分布	成立している条件付き独立性
(1)(12)	$Y \perp\!\!\!\perp Z, Z \perp\!\!\!\perp X, X \perp\!\!\!\perp Y, \{YZ\} \perp\!\!\!\perp X, \{ZX\} \perp\!\!\!\perp Y, \{XY\} \perp\!\!\!\perp Z$
(2)(13)	$X \perp\!\!\!\perp Y, X \perp\!\!\!\perp Z, \{YZ\} \perp\!\!\!\perp X$
(3)(14)	$Y \perp\!\!\!\perp Z, Y \perp\!\!\!\perp X, \{ZX\} \perp\!\!\!\perp Y$
(4)(15)	$Z \perp\!\!\!\perp X, Z \perp\!\!\!\perp Y, \{XY\} \perp\!\!\!\perp Z$
(5)(16)	$Y \perp\!\!\!\perp Z X$
(6)(17)	$Z \perp\!\!\!\perp X Y$
(7)(18)	$X \perp\!\!\!\perp Y Z$
(8)	$Y \perp\!\!\!\perp Z$
(9)	$Z \perp\!\!\!\perp X$
(10)	$X \perp\!\!\!\perp Y$
(11)(19)	

できない。たとえば、(10) では、 $X \perp\!\!\!\perp Y$ は成立するが、 $Y \perp\!\!\!\perp Z$ および $Z \perp\!\!\!\perp X$ が成立しない。これは衝突といって、有向辺の先を 2 個以上含む頂点 (Z) で、それらの元 (X, Y) どうしが結ばれていないものが存在していることに起因する。(11) も (10) と同様、 X から Z 、 Y から Z の 2 個の有向辺が Z に向かっているが、 X, Y が結ばれているので、衝突ではない。

N が 4 以上であれば、逆に MN で表現できて、BN で表現できない条件付き独立性が存在する。MN で書いたときに、弧を含まない大きさ 4 以上のループが存在する場合がそれにあたる。たとえば、 $V = \{X, Y, Z, W\}$, $V_1 = \{X, Y\}$, $V_2 = \{Y, Z\}$, $V_3 = \{Z, W\}$, $V_4 = \{W, X\}$ のように、 K を正規化定数として、分布

$$P(X, Y, Z, W) = f_1(X, Y)f_2(Y, Z)f_3(Z, W)f_4(W, X)/K$$

が表現する条件付き独立性 $X \perp\!\!\!\perp Z|\{Y, W\}$, $Y \perp\!\!\!\perp W|\{X, Z\}$ は、BN では表現できない (図 1.4)。

6 第1章 グラフィカルモデルの構造学習

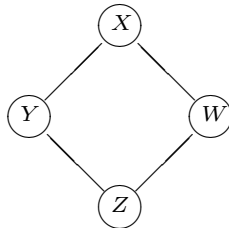


図 1.4 MN で表現できても BN で表現できない条件付き独立性の例

1.3 事後確率最大の森の構造学習

統計学では、条件付き独立性を検定する方法はいくつも提案されていて、サンプルからグラフィカルモデルを学習する際によく用いられている。

ここでは、条件付き独立性の事前確率が与えられた場合に、サンプルから、事後確率を最大にする検定方法を考えてみよう。

1.3.1 局所スコア

未知の2変数 X, Y にしたがって発生した n 対のサンプル $x^n = (x_1, \dots, x_n)$, $y^n = (y_1, \dots, y_n)$ から、 X, Y が独立であるか否かを検定する。以下では、 X, Y はそれぞれ $\{0, 1\}$ の値を取るものとする。独立であることの事前確率を p とおき、系列 $x^n, y^n, (x^n, y^n)$ の確率のかわりをする何らかの $Q^n(X), Q^n(Y), Q^n(X, Y)$ を用意し、

$$pQ^n(X)Q^n(Y) \geq (1-p)Q^n(X, Y) \quad (1.20)$$

が成立すれば X, Y が独立、成立しなければ独立でないという判定をしたい。

そのために、 $X = 1$ の生起する確率を $0 \leq \theta \leq 1$ 、 x^n における $X = 1$ の頻度を $0 \leq c \leq n$ とすれば、系列 $X^n = x^n$ の生起する確率は $\theta^c(1-\theta)^{n-c}$ とかける。このとき、 θ に関する事前分布 (確率密度関数) $w(\theta)$ が存在すると仮定すると、 $\theta^c(1-\theta)^{n-c}w(\theta)$ を $0 \leq \theta \leq 1$ について積分すれば、特定の θ を仮定しない $X^n = x^n$ の生起する確率が計算できる。これを

$$Q^n(X) = \int_0^1 \theta^c(1-\theta)^{n-c}w(\theta)d\theta$$

とすれば、独立である事前確率 p 、 $X = 1$ の確率 θ の事前確率という2段階

1.3 事後確率最大の森の構造学習 7

の事前確率をおく必要があるが、それらのもとの事後確率最大の判定がなされる。

たとえば、 $0 \leq \theta \leq 1$ が一様に分布すると仮定すると、 $w(\theta) = 1$ であるので、

$$Q^n(X) = \int_0^1 \theta^c (1-\theta)^{n-c} d\theta = \frac{c!(n-c)!}{(n+1)!}$$

となる。右側の等号は、部分積分から求まる。

$$J_0 = \int_0^1 (1-\theta)^n d\theta = \frac{1}{n+1} [t^{n+1}]_0^1 = \frac{1}{n+1}$$

$$\begin{aligned} J_c &= \int_0^1 \theta^c (1-\theta)^{n-c} d\theta = \int_0^1 \theta^c \left\{ -\frac{(1-\theta)^{n-c+1}}{n-c+1} \right\}' d\theta \\ &= \left[-\theta^c \frac{(1-\theta)^{n-c+1}}{n-c+1} \right]_0^1 + \frac{c}{n-c+1} \int_0^1 \theta^{c-1} (1-\theta)^{n-c+1} d\theta \\ &= \frac{c}{n-c+1} J_{c-1} = \frac{c}{n-c+1} \cdot \frac{c-1}{n-c+2} \cdots \frac{1}{n} \cdot J_0 = \frac{c!(n-c)!}{(n+1)!} \end{aligned}$$

この値は、系列 x^n を前から順に見ていき、 $i-1$ 番目までで $X=0, 1$ がそれぞれ $c_{i-1}(0), c_{i-1}(1)$ 回であれば、 $X=0$ のとき $\frac{c_{i-1}(0)+1}{i+1}$ 、 $X=1$ のとき $\frac{c_{i-1}(1)+1}{i+1}$ をかけていくと求まる量である。たとえば、 $x^5 = (1, 0, 1, 1, 0)$ であれば、

$$\frac{1}{2} \cdot \frac{1}{3} \cdot \frac{2}{4} \cdot \frac{3}{5} \cdot \frac{2}{6} = \frac{1}{60}$$

となる。

証明は省略するが、一般には、 X が $0, 1, \dots, \alpha-1$ のいずれかの値をとるとき、最初に加える値を $1, \dots, 1$ とせず、 $a(0), \dots, a(\alpha-1)$ とすれば、 $X = x_i$ の確率を

$$\frac{c_{i-1}(x_i) + a(x_i)}{i-1 + \sum_x a(x)}$$

8 第1章 グラフィカルモデルの構造学習

で予測でき、 $Q^n(X)$ および $w(\theta)$, $\theta = (\theta_0, \dots, \theta_{\alpha-1})$ が以下のようになることが知られている [5]。

$$Q^n(X) = \frac{\Gamma(\sum_x a(x)) \prod_x \Gamma(c_n(x) + a(x))}{\prod_x \Gamma(a(x)) \cdot \Gamma(n + \sum_x a(x))} \quad (1.21)$$

$$w(\theta) = K \prod_x \theta_x^{a(x)-1}$$

ただし、 K は正規化定数である。また、 $\Gamma(\cdot)$ は Gamma 関数 $\Gamma(z) = \int_0^\infty t^{z-1} e^{-t} dt$ で、 $z\Gamma(z) = \Gamma(z+1)$ などが成立する (自然数 n に対し、 $\Gamma(n+1) = n!$)。

1.3.2 独立性の検定

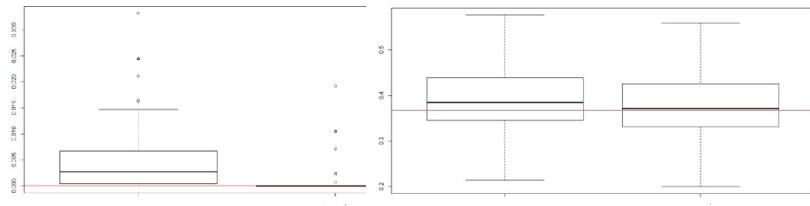


図 1.5 相互情報量の推定。左 2 個が独立な場合、右 2 個が独立でない場合。それぞれで左が $I^n(X, Y)$ (最尤法)、右が $J^n(X, Y)$ (Bayes)。いずれも、最尤法は大きな値をとる傾向がある。

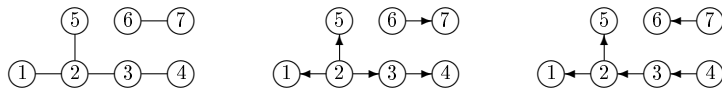


図 1.6 森 (V, E) , $V = \{1, 2, 3, 4, 5, 6, 7\}$, $E = \{\{1, 2\}, \{2, 3\}, \{2, 5\}, \{3, 4\}, \{6, 7\}\}$ (左) を、2 と 6 を根に選んだ場合 (中) と 4 と 7 を根に選んだ場合 (右)

$Q^n(Y)$, $Q^n(X, Y)$ の構成も同様である。たとえば、 $X = x$ の頻度 $c_n(x)$ を $c_n(x, y)$ に、 $X = x$ の定数 $a(x)$ を $a(x, y)$ にかえれば、 $Q^n(X, Y)$ が得ら

1.3 事後確率最大の森の構造学習 9

れる。このように $Q^n(\cdot)$ を構成すると、十分大きな n に対して、

$$(1.20) \iff X \perp\!\!\!\perp Y \quad (1.22)$$

が確率 1 で成立する [12]。また、 $p = 0.5$ としたときの (1.20) の両辺の比についての関数

$$J^n(X, Y) := \frac{1}{n} \log \frac{Q^n(X, Y)}{Q^n(X)Q^n(Y)} \quad (1.23)$$

は、相互情報量

$$\begin{aligned} I(X, Y) &:= H(X) + H(Y) - H(X, Y) \\ &= \sum_x \sum_y P(X = x, Y = y) \log \frac{P(X = x, Y = y)}{P(X = x)P(Y = y)} \end{aligned}$$

の推定量とみなすこともできる。ただし、

$$H(X) := \sum_x -P(X = x) \log P(X = x)$$

$$H(Y) := \sum_y -P(Y = y) \log P(Y = y)$$

$$H(X, Y) := \sum_x \sum_y -P(X = x, Y = y) \log P(X = x, Y = y)$$

は、それぞれ $X, Y, (X, Y)$ のエントロピーである。すなわち、十分大きな n に対して、

$$J^n(X, Y) \leq 0 \iff X \perp\!\!\!\perp Y \quad (1.24)$$

が確率 1 で成立する推定量になっている [12]。

無論、

$$I^n(X, Y) := \sum_x \sum_y \frac{c_n(x, y)}{n} \log \frac{c_n(x, y)/n}{c_n(x)/n \cdot c_n(x, y)/n} \quad (1.25)$$

も、大きな n で真の相互情報量 $I(X, Y)$ に収束するが、(1.24) が成立しない。実際、 X, Y が独立であっても、正の確率で $I^n(X, Y) > 0$ となる。 $\alpha = 2$ で

10 第1章 グラフィカルモデルの構造学習

独立の X, Y 、非独立の X, Y で実験してみると、相互情報量の推定量として、 $I^n(X, Y)$ は $J^n(X, Y)$ より大きめな値をとっている (図 1.5)。これは、 $I^n(X, Y)$ が最尤法によって推定しているので、過学習をおこしているためと考えられる。

1.3.3 Chow-Liu アルゴリズム

以下では、相互情報量の推定に基いて、 N 変数 $X^{(1)}, \dots, X^{(N)}$ についてのループをもたない無向グラフ、すなわち森 (forest) の構造を学習する問題について考えてみたい。

まず、頂点集合 $V = \{1, \dots, N\}$ 、辺集合 $E = \{\{i, j\} | i, j \in V, i \neq j\}$ をもつ森 (V, E) に、分布の因数分解を対応付ける。

$$P'(X^{(1)}, \dots, X^{(N)}) := \prod_{i \in V} P(X^{(i)}) \prod_{\{i, j\} \in E} \frac{P(X^{(i)}, X^{(j)})}{P(X^{(i)})P(X^{(j)})} \quad (1.26)$$

たとえば、森が図 1.6 左のように与えられれば、

$$\begin{aligned} & P(X^{(1)})P(X^{(2)})P(X^{(3)})P(X^{(4)})P(X^{(5)})P(X^{(6)})P(X^{(7)}) \cdot \frac{P(X^{(1)}, X^{(2)})}{P(X^{(1)})P(X^{(2)})} \\ & \cdot \frac{P(X^{(2)}, X^{(3)})}{P(X^{(2)})P(X^{(3)})} \cdot \frac{P(X^{(2)}, X^{(5)})}{P(X^{(2)})P(X^{(5)})} \cdot \frac{P(X^{(3)}, X^{(4)})}{P(X^{(3)})P(X^{(4)})} \cdot \frac{P(X^{(6)}, X^{(7)})}{P(X^{(6)})P(X^{(7)})} \end{aligned}$$

となるが、これは、たとえば 2 と 6 を根として選べば (図 1.6 中)、

$$P(X^{(2)})P(X^{(1)}|X^{(2)})P(X^{(3)}|X^{(2)})P(X^{(5)}|X^{(2)})P(X^{(4)}|X^{(3)})P(X^{(6)})P(X^{(7)}|X^{(6)})$$

と一致し、4 と 7 を根として選べば (図 1.6 右)、

$$P(X^{(4)})P(X^{(3)}|X^{(4)})P(X^{(2)}|X^{(3)})P(X^{(5)}|X^{(2)})P(X^{(1)}|X^{(2)})P(X^{(7)})P(X^{(6)}|X^{(7)})$$

と一致する。

そして、分布 $P(X^{(1)}, \dots, X^{(N)})$ が任意に与えられたときに、それを (1.26) の形の分布に近似することを考える。まず、分布 P, P' の間の Kullback-Leibler 情報量 $D(P||P')$ が

$$\begin{aligned}
D(P||P') &= \sum_{x^{(1)}, \dots, x^{(N)}} P(x^{(1)}, \dots, x^{(N)}) \log \frac{P(x^{(1)}, \dots, x^{(N)})}{P'(x^{(1)}, \dots, x^{(N)})} \\
&= -H(1, \dots, N) + \sum_{i \in V} H(i) - \sum_{\{i, j\} \in E} I(i, j)
\end{aligned}$$

とかけることに注意する。ここで、

$$\begin{aligned}
H(1, \dots, N) &:= \sum_{x^{(1)}, \dots, x^{(N)}} -P(X^{(1)} = x^{(1)}, \dots, X^{(N)} = x^{(N)}) \\
&\quad \log P(X^{(1)} = x^{(1)}, \dots, X^{(N)} = x^{(N)}) \\
H(i) &:= \sum_{x^{(i)}} -P(X^{(i)} = x^{(i)}) \log P(X^{(i)} = x^{(i)})
\end{aligned}$$

$$I(i, j) := \sum_{x^{(i)}, x^{(j)}} P(X^{(i)} = x^{(i)}, X^{(j)} = x^{(j)}) \log \frac{P(X^{(i)} = x^{(i)}, X^{(j)} = x^{(j)})}{P(X^{(i)} = x^{(i)})P(X^{(j)} = x^{(j)})}$$

は、 $X^{(1)}, \dots, X^{(N)}$ のエントロピー、 $X^{(i)}$ のエントロピー、 $X^{(i)}, X^{(j)}$ の相互情報量である。したがって、 $D(P||P')$ を最小にするには、相互情報量の和 $\sum_{\{i, j\} \in E} I(i, j)$ を最大にすれば良い。そのために、 $i, j \in V, i \neq j$ に非負の重み $w(i, j) = w(j, i)$ に対して、その重みの和を最大にする木 (すべての頂点が連結された森) を求める方法 (Kruskal のアルゴリズム) を適用する。すなわち、重みが最大の2辺を結ぶことによってループができないかぎり結合し、ループができる場合その2頂点は結合せしない。最終的に、 $N(N-1)/2$ 個のすべての頂点の対に関してこの操作を行う。

この重みとして、相互情報量 $I(i, j)$ を適用して、 $D(P||P')$ を最小にする木を見出したのが、Chow-Liu アルゴリズム (1968)[4] である。図 1.7 において、 $I(1, 2) > I(1, 3) > I(2, 3) > I(1, 4) > I(3, 4) > I(2, 4) > 0$ であれば、 $\{1, 2\}$, $\{1, 3\}$ と辺を結ぶが、3 番目に相互情報量の大きな $\{2, 3\}$ は結ぶとループを生成するので、結ばない。さらに、 $\{1, 4\}$ は結んでもループが生成されないので、結ぶ。また、それ以上結ぶとループが生成されるので、ここで森の生成は完了する。

12 第1章 グラフィカルモデルの構造学習

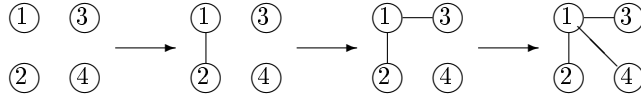


図 1.7 The Chow-Liu アルゴリズム: $I(1, 2) > I(1, 3) > I(2, 3) > I(1, 4) > I(3, 4) > I(2, 4) > 0$

それでは、分布が与えられてなく、サンプルとして、変数 $X^{(1)} \dots, X^{(N)}$ の実現値の n 組

$$\begin{aligned} X^{(1)} &= x_{1,1}, \dots, X^{(N)} = x_{1,N} \\ &\dots, \dots, \dots \\ X^{(1)} &= x_{n,1}, \dots, X^{(N)} = x_{n,N} \end{aligned}$$

が得られたとする。ただし、サンプルに欠損は含まれていないものとする。この場合に、どのようにして、同様の木を生成すればよいだろうか。(1.25) で、相互情報量を推定して、その値に基いて木を生成することは可能である。しかしながら、たとえば、 $N = 2$ で $V = \{1, 2\}$ として、(1.25) を用いると、その 2 変数が独立であってもなくても 2 頂点を結んだ木ができる。

1993 年に Suzuki [11] は以下の方法を提案している。Kruskal の方法は、重み $w(i, j)$ が負であっても問題なく動作する (重みが正の辺だけを結合する)。そして、(1.25) ではなく (1.23) に基いて相互情報量を推定して、それを相互情報量として、Chow-Liu アルゴリズムを適用すると、(1.23) は負の値を取り、ループができなくても結合しないことがあるが、その場合 (1.23) より、それはその 2 変数が独立であることを意味する。すなわち、大きなサンプル数 n で、ある 2 頂点が結合されることとその 2 変数が独立でないことが必要十分の関係になる。もっと本質的には、(1.23) を最大にする辺を逐次選択するということは、(1.26) に対応して、

$$R^n(E) := \prod_{i \in V} Q^n(X^{(i)}) \prod_{\{i,j\} \in E} \frac{Q^n(X^{(i)}, X^{(j)})}{Q^n(X^{(i)})Q^n(X^{(j)})}$$

を最大にする森 (V, E) を選択していることにほかならない。実際、

$$-\frac{1}{n} \log R^n(E) = \sum_{i \in V} -\frac{1}{n} \log Q^n(X^{(i)}) - \sum_{\{i,j\} \in E} J^n(X^{(i)}, X^{(j)})$$

の右辺第1項は E によらず一定で、 $-\frac{1}{n} \log R^n(E)$ の最小化と $J^n(X^{(i)}, X^{(j)})$ の和の最大化は一致している。したがって、各森が等確率で生起するという事前確率をおけば、サンプルのもとで事後確率を最大にする森を選択しているといえる。

$I^n(X^{(i)}, X^{(j)})$ の和を最小化 (最尤法) するのと $J^n(X^{(i)}, X^{(j)})$ の和を最小化 (事後確率最大) するのでは、

1. 後者で生成された森は、必ずしも木でない
2. 後者の辺集合が、必ずしも前者の辺集合の部分集合にはなっていない

という特徴がある。R パッケージ `bnlearn` [9] のデータセット `Asia` を適用した最尤、事後確率最大で得られた森を図 1.8 に示す。

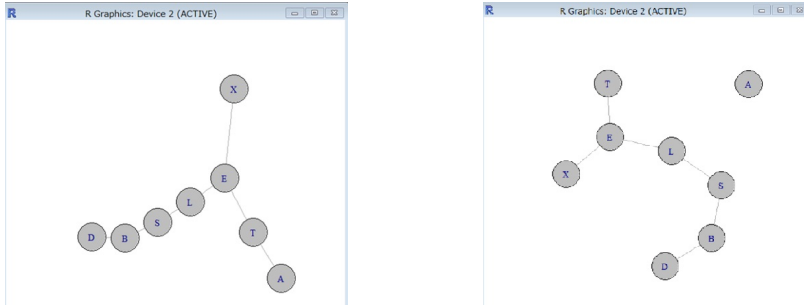


図 1.8 R パッケージ `bnlearn` のデータセット `Asia` を適用。左が最尤、右が事後確率最大で得られた森。最尤法を適用した場合、過学習を検知できないので必ず木になる。

1.3.4 条件付き独立性の検定

同様のことは、未知の3変数 X, Y, Z にしたがって発生した n 対のサンプル $x^n = (x_1, \dots, x_n)$, $y^n = (y_1, \dots, y_n)$, $z^n = (z_1, \dots, z_n)$ から、 X, Y が Z のもとで条件付き独立であるか否かを検定する場合にも一般化できる。以下では、 X, Y, Z はそれぞれ α, β, γ 通りの値を取るものとする。条件付き独立で

14 第1章 グラフィカルモデルの構造学習

あることの事前確率を p とおき、 $Q^n(Z), Q^n(X, Z), Q^n(Y, Z), Q^n(X, Y, Z)$ を前述のように構成し、

$$pQ^n(X, Z)Q^n(Y, Z) \geq (1 - p)Q^n(X, Y, Z)Q^n(Z) \quad (1.27)$$

が成立すれば X, Y が Z のもとで条件付き独立であるという判定をするようにする。すると、十分大きな n に対して、

$$(1.27) \iff X \perp\!\!\!\perp Y|Z$$

が確率 1 で成立する [13]。

また、

$$J^n(X, Y|Z) := \frac{1}{n} \log \frac{Q^n(X, Y, Z)Q(Z)}{Q^n(X, Z)Q^n(Y, Z)}$$

は、条件付き相互情報量

$$\begin{aligned} I(X, Y|Z) &:= H(X, Z) + H(Y, Z) - H(X, Y, Z) - H(Z) \\ &= \sum_x \sum_y P(X = x, Y = y, Z = z) \log \frac{P(X = x, Y = y, Z = z)P(Z = z)}{P(X = x, Z = z)P(Y = y, Z = z)} \end{aligned}$$

の推定量とみなすこともできる。ただし、

$$H(X, Z) := \sum_x \sum_z -P(X = x, Z = z) \log P(X = x, Z = z)$$

$$H(Y, Z) := \sum_y \sum_z -P(Y = y, Z = z) \log P(Y = y, Z = z)$$

$$H(X, Y, Z) := \sum_x \sum_y \sum_z -P(X = x, Y = y, Z = z) \log P(X = x, Y = y, Z = z)$$

は、それぞれ $(X, Z), (Y, Z), (X, Y, Z)$ のエントロピーである。すなわち、十分大きな n に対して、

$$J^n(X, Y|Z) \leq 0 \iff X \perp\!\!\!\perp Y|Z$$

が確率 1 で成立する推定量になっている [12]。

1.4 事後確率最大の BN の構造学習

N 変数 $X^{(1)}, \dots, X^{(N)}$ に関する n 組のサンプルから、最も事後確率の高い BN を求めることを考える。ただし、以下では簡単のため、断らないかぎり、すべての構造に等しい事前確率が割り振られているものとする。

まず、サンプルとして、変数 $X^{(1)} \dots, X^{(N)}$ の実現値の n 組

$$\begin{aligned} X^{(1)} &= x_{1,1}, \dots, X^{(N)} = x_{1,N} \\ &\dots, \dots, \dots \\ X^{(1)} &= x_{n,1}, \dots, X^{(N)} = x_{n,N} \end{aligned}$$

が得られたとする。ただし、サンプルに欠損は含まれていないものとする。

1.4.1 大域スコア

まず、前節で定義した $Q^n(X), Q^n(Y, Z)$ などの局所スコア (local score) を計算する。記法が煩雑になるのを防ぐため、混乱のない限り、 $Q^n(\{X\})$ や $Q^n(\{Y, Z\})$ と書かず、 $Q^n(X), Q^n(Y, Z)$ のように書くものとする。 N 変数であれば、局所スコアは $2^N - 1$ 個存在する。そして、 $N = 3$ であれば、(1)-(11) に対応して、

$$\begin{aligned} &Q^n(X)Q^n(Y)Q^n(Z) \\ &Q^n(X)Q^n(Y, Z), Q^n(Y)Q^n(Z, X), Q^n(Z)Q^n(X, Y) \\ &\frac{Q^n(Z, X)Q^n(X, Y)}{Q^n(X)}, \frac{Q^n(X, Y)Q^n(Y, Z)}{Q^n(Y)}, \frac{Q^n(Z, X)Q^n(X, Y)}{Q^n(Z)}, \\ &\frac{Q^n(Y)Q^n(Z)Q^n(X, Y, Z)}{Q^n(Y, Z)}, \frac{Q^n(Z)Q^n(X)Q^n(X, Y, Z)}{Q^n(Z, X)}, \frac{Q^n(X)Q^n(Y)Q^n(X, Y, Z)}{Q^n(X, Y)}, \\ &Q^n(X, Y, Z) \end{aligned}$$

の各値を求め、その値 (事後確率に比例する) を最大にする構造を選択する。これら 11 個のような値を大域スコア (global score) という。各 N で、 2^N を超える個数の大域スコアが存在する。

以下では、局所スコアから大域スコアを求める方法を述べる (Silander and Myllymaki, 2006 [7])。

16 第1章 グラフィカルモデルの構造学習

BNの構造を学習するには、各変数が依存する他の変数の集合(親集合)を求める必要がある。図1.1で、(2)の Z の親集合は $\{Y\}$ 、(5)の Y の親集合は $\{X\}$ 、(8)の X の親集合は $\{Y, Z\}$ になる。そして、親集合は、条件付き独立性の検定によって求めることができる。親集合は、狭義にはその変数が依存する他の変数全てであるが、実際にはある範囲の親集合の中で、その変数が依存する変数の集合を求めるという処理が必要となる。すなわち、 $X \notin U \subseteq V$ なる各 (U, V) について、条件付き局所スコア

$$Q^n(X|U) := \frac{Q^n(X, U)}{Q^n(U)}$$

が定義できる。そして、 $X \in S \subseteq V$ なるすべての (X, S) について、 $Q^n(X|U)$ を最大にする $U \subseteq S \setminus \{X\}$ を、 X の S に制限した親集合といい、 $\pi_S(X)$ と書くものとする。記法が煩雑になるのを防ぐため、混乱のない限り、 $\pi_{\{Y\}}(Y)$ や $\pi_{\{Y, Z\}}(Z)$ と書かず、 $\pi_Y(Y)$ 、 $\pi_{YZ}(Z)$ のように書くものとする。たとえば、 $V = \{X, Y, Z, W\}$ 、 $S = \{X, Y, Z\}$ であれば、

$$Q^n(X), Q^n(X|\{Y\}), Q^n(X|\{Z\}), Q^n(X|\{Y, Z\})$$

のどれが最適であるかによって、 $\pi_{XYZ}(X) = \{\}, \{Y\}, \{Z\}, \{Y, Z\}$ が定まる。

1.4.2 動的計画法としての定式化

最初に、順序グラフという概念を説明しておきたい。 N 個の変数の集合を V とし、その 2^N 個の部分集合をラベルに持つ有向グラフで、 $X \notin U \subseteq V$ なる各 (X, U) に対して、ラベル U の頂点からラベル $S = U \cup \{X\}$ に向けて有向辺をひいたものを、その N 変数の順序グラフ(ordred graph)と呼ぶ。 $V = \{X, Y, Z, W\}$ に対する順序グラフは、図1.11のようになる。

X の S に制限した親集合をすべて効率よく求める方法を、図1.9を例として説明する。各頂点には、 $S \setminus \{X\}$ の値が記載されていて、最下段の $\{\}$ から最上段の $\{Y, Z, W\}$ にむかって上方に進んでいくものとする。

まず、 $\pi_{\{X\}}(X) = \{\}$ とする。次に $Q^n(X|\{Y\})$ と $Q^n(X|\{\}) (= Q^n(X))$ を比較して、前者が大きければ $\pi_{XY}(X) = \{Y\}$ 、後者が大きければ $\pi_{XY} =$

$\{\}$ となる。同様にして、 $\pi_{XZ}(X)$ 、 $\pi_{XW}(X)$ が求まる。そして、 $Q^n(X|Y, Z)$ 、 $Q^n(X|\pi_{XY}(X))$ 、 $Q^n(X|\pi_{XZ}(X))$ の3者を比較して、 $\pi_{XYZ}(X) = \{Y, Z\}$ 、 $\pi_{XY}(X)$ 、 $\pi_{XZ}(X)$ のいずれであるかがきまる。同様にして、 $\pi_{XYW}(X)$ 、 $\pi_{XZW}(X)$ が求まる。最後に、 $Q^n(X|Y, Z, W)$ 、 $Q^n(X|\pi_{XYZ}(X))$ 、 $Q^n(X|\pi_{XYW}(X))$ 、 $Q^n(X|\pi_{XZW}(X))$ の4者を比較して、 $\pi_{XYZW}(X) = \{Y, Z, W\}$ 、 $\pi_{XYZ}(X)$ 、 $\pi_{XYW}(X)$ 、 $\pi_{XZW}(X)$ のいずれであるかがきまる。

再帰的に書くと、 $X \in S$, $U \cup \{X\} = S$ として、

$$Q^n(X|\pi_S(X)) = \max\{Q^n(X|U), \max_{Y \in U} Q^n(X|\pi_{S \setminus \{Y\}}(X))\}$$

となる。同様にして、 $\pi_S(Y)$, $Y \in S$ 、 $\pi_S(Z)$, $Z \in S$ 、 $\pi_S(W)$, $W \in S$ を求めることができる。

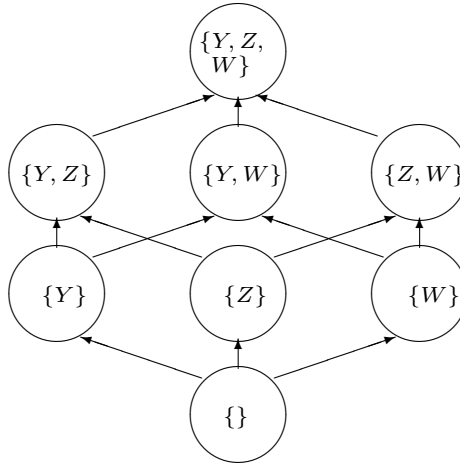


図 1.9 $\{\}$ から $\{Y, Z, W\}$ までの順序グラフ: $X \in S \subseteq V$ なる各 S について、 $\pi_S(X)$ を求める

順序グラフで、 j 個の要素からなる V の部分集合 $U = S \setminus \{X\}$ をラベルとする頂点では、 $Q^n(X|U)$ を $Q^n(X|\pi_{S \setminus \{Y\}}(X))$ と比較する (後者は j 個ある)。したがって、比較回数は

$$\sum_{j=1}^{N-1} j \binom{N-1}{j} = \sum_{j=1}^{N-1} (N-1) \binom{N-2}{j-1} = (N-1)2^{N-2}$$

18 第1章 グラフィカルモデルの構造学習

と評価される。これをすべての $X \in V$ に対して行うので、 $O(N^2 2^N)$ の計算が必要である。順序グラフの頂点は 2^{N-1} あるので、局所スコアが長期記憶に格納されていて、 X 以外の Y, Z, W に関して $\pi_S(Y), \pi_S(Z), \pi_S(W)$ を求める前に、結果を長期記憶に退避させることを前提にすると、 $O(2^N)$ のメモリを必要とする。

次に、大域スコアを最大値とする因数分解を求めたい。ここで、 $\pi_V(X), \pi_V(Y), \pi_V(Z), \pi_V(W)$ を求めて、それぞれ X, Y, Z, W から有向辺を結ぶと、ループができる可能性がある。

しかし、たとえば、 X, Y, Z, W が $X \rightarrow Y \rightarrow Z \rightarrow W$ の順序である (順序の後のものから前のものに有向辺がない) ことがわかっていれば、

$$Q^n(X|\pi_X(X))Q^n(Y|\pi_{X,Y}(Y))Q^n(Z|\pi_{X,Y,Z}(Z))Q^n(W|\pi_{X,Y,Z,W}(W))$$

が、 $Y \rightarrow W \rightarrow X \rightarrow Z$ の順序であることがわかっていれば、

$$Q^n(Y|\pi_Y(Y))Q^n(W|\pi_{Y,W}(W))Q^n(X|\pi_{X,Y,W}(X))Q^n(Z|\pi_{X,Y,Z,W}(Z))$$

がループをもたない有向グラフ (DAG) の中で、大域スコアが最大値となることがわかる。そして、たとえば、 $X \rightarrow Y \rightarrow Z \rightarrow W$ であって、 $\pi_{XY}(Y) = \{\}$, $\pi_{XYZ}(Z) = \{X\}$, $\pi_{XYZW}(W) = \{Y, Z\}$ であれば、図 1.10 のような BN が得られる。

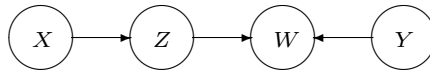
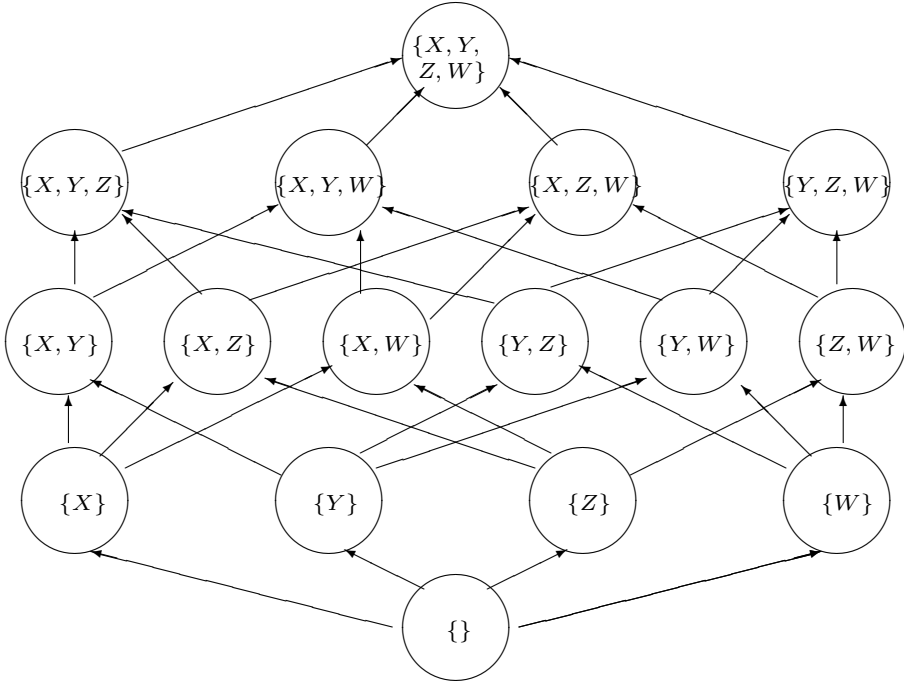


図 1.10 $\pi_{XY}(Y) = \{\}$, $\pi_{XYZ}(Z) = \{X\}$, $\pi_{XYZW}(W) = \{Y, Z\}$ であって、 $X \rightarrow Y \rightarrow Z \rightarrow W$ の順序のときの BN

しかし、 N 個の変数の順序は $N!$ 個だけあるので、実際には、図 1.11 にあるような順序グラフを用いる。以下では、 V の部分集合 S に対する大域スコアを $R^n(S)$ とかくものとする。また、記法が煩雑になるのを防ぐため、 $R^n(\{X\})$ や $R^n(\{Y, Z\})$ と書かず、 $R^n(X)$, $R^n(Y, Z)$ のように書くものとする。

$V = \{X, Y, Z, W\}$ では以下のようなになる。最初に $R^n(X) = Q^n(X)$, $R^n(Y) = Q^n(Y)$, $R^n(Z) = Q^n(Z)$, $R^n(W) = Q^n(W)$ とおく。そして、

図 1.11 $\{\}$ から $\{X, Y, Z, W\}$ までの順序グラフ

$Q^n(X, Y)$, $Q^n(X|\pi_{XY}(X))R^n(Y)$, $Q^n(Y|\pi_{XY}(Y))R^n(X)$ の 3 者の最大値を求め、それを $R^n(X, Y)$ とおく。同様にして、 $R^n(X, Z)$, $R^n(X, W)$, $R^n(Y, Z)$, $R^n(Y, W)$, $R^n(Z, W)$ を求める。次に、 $Q^n(X, Y, Z)$, $Q^n(X|\pi_{XYZ}(X))R^n(Y, Z)$, $Q^n(Y|\pi_{XYZ}(Y))R^n(X, Z)$, $Q^n(Z|\pi_{XYZ}(Z))R^n(X, Y)$ の 4 者の最大値を求め、それを $R^n(X, Y, Z)$ とおく。同様にして、 $R^n(X, Y, W)$, $R^n(X, Z, W)$, $R^n(Y, Z, W)$ を求める。最後に、 $Q^n(X, Y, Z, W)$, $Q^n(X|\pi_{XYZW}(X))R^n(Y, Z, W)$, $Q^n(Y|\pi_{XYZW}(Y))R^n(X, Z, W)$, $Q^n(Z|\pi_{XYZW}(Z))R^n(X, Y, W)$, $Q^n(W|\pi_{XYZW}(W))R^n(X, Y, Z)$, の 5 者の最大値を求め、その値 $R^n(X, Y, Z, W)$ が大域スコアの最大値となる。そして、 $\{\}$ から $\{X, Y, Z, W\}$ に至る経路が、変数 X, Y, Z, W の順序を与える。

一般には、 $S \subseteq V$ に対して、 $R^n(X) = Q^n(X)$, $X \in V$ として、大域スコ

20 第1章 グラフィカルモデルの構造学習

アの最大値は再帰的に以下のようにして求まる。

$$R^n(S) := \max\{Q^n(S), \max_{X \in S} Q^n(X|\pi_S(X))R^n(U)\}, U \cup \{X\} = S$$

$X^{(1)}, \dots, X^{(N)}$ のそれぞれが $\{\}$ から V に向かって、 X_1, \dots, X_N と順序付けられたとし、 $V_j := V \setminus \{X_{j+1}, \dots, X_N\}$ 、

$$\pi_N := \pi_{V_N}(X_N), \pi_{N-1} := \pi_{V_{N-1}}(X_{N-1}), \dots, \pi_1 := \pi_{V_1}(X_1) = \{\}$$

とおくとき、各 $j = 1, 2, \dots, N$ で、 π_j の各ノードから X_j に有向辺をひいて、大域スコア最大

$$R^n(V) = Q^n(X_N|\pi_N) \cdot Q^n(X_{N-1}|\pi_{N-1}) \cdots Q^n(X_2|\pi_2)Q^n(X_1)$$

のBNが得られる。そして、このときの因数分解は、

$$P(X_N|\pi_N)P(X_{N-1}|\pi_{N-1}) \cdots P(X_2|\pi_2)P(X_1)$$

したがって、比較回数は

$$\sum_{j=1}^N j \binom{N}{j} = \sum_{j=1}^N N \binom{N-1}{j-1} = N2^{N-1}$$

と計算でき、 $O(N2^N)$ の計算が必要である。順序グラフの頂点は 2^N あるので、 $O(2^N)$ のメモリを必要とする。

1.4.3 最短経路問題としての定式化

後半の処理は、 $X \notin U \subseteq V, U \cup \{X\} = S$ なる各 (U, S) について、

$$d(U, S) = -\log Q^n(X|\pi_S(X)) \quad (1.28)$$

の合計を最小にする経路を求める問題であったが、順序グラフの有向辺 (U, S) の間の距離をそのようにおいた最短経路問題として定式化できる。

以下では、 $g(S)$ を $\{\}$ から S に至る経路での $d(\cdot, \cdot)$ の和の最小値、 $h(S)$ を S から V に至る経路での $d(\cdot, \cdot)$ の和の下界 (ヒューリスティック) とする。

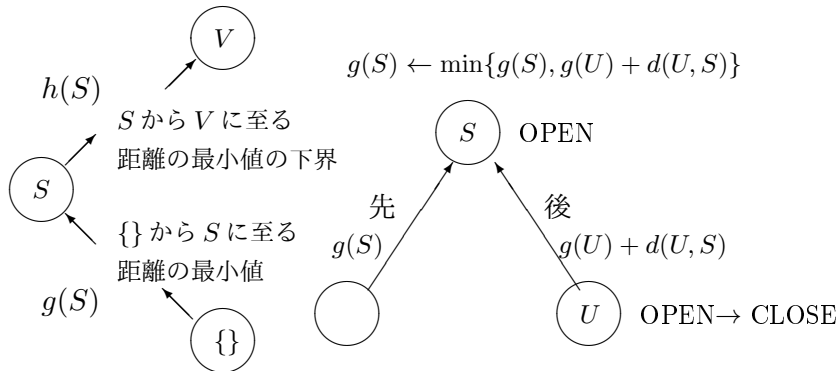


図 1.12 A* アルゴリズム

下界を計算する情報がない場合、 $h(S) = 0$ とおくものとする。最初、順序グラフにおいて $\{ \}$ に OPEN、他の頂点には CLOSE のラベルが貼られているものとする。次に、頂点 $\{ \}$ を CLOSE とし、 N 個の頂点 $\{X\}$, $X \in V$ を OPEN とする。それ以降は、OPEN となっている頂点の中で $f(U)$ の最も小さな頂点 U を選択し、それを CLOSE、 U から有向辺 (U, S) として出ている各頂点 S を OPEN とし、 $g(U) + d(U, S)$ を $g(S)$ の値とする。しかし、OPEN しようとした頂点がすでに OPEN となっている場合には、すでにもっている $g(S)$ の値と、 $g(U) + d(U, S)$ のうちで小さい方の値を $g(S)$ とする。各 $S \subseteq V$ は、 $g(S) = g(U) + d(U, S)$ となった U の値を保持する。複数の U が同じ $g(S)$ の最小値を保つ場合には、どちらの U の値を保持してもよい。このようにして、最終的に頂点 V が OPEN から CLOSE になった時点で、処理が終了する。 $h(S)$ が S から V に至る経路での $d(U, S)$ の和の最小値の下界であるので、OPEN から CLOSE になる頂点は $g(V)$ 以下の値をもつ (図 1.12)。

このような方法は、A* アルゴリズムとよばれ、不要な頂点のコストを計算しないので、計算やメモリが節約される。特に、下界 $h(S)$ が実際の値に近ければ、不要な頂点を OPEN することが少なくなるので、処理時間が短くなる。

しかし、ヒューリスティック $h(S)$ を計算するため、計算のオーバーヘッドに時間がかかり、データセットによっては、全体の実行時間が多くなる場合もある。ヒューリスティックの例として、まだ選択されていない変数 X

22 第1章 グラフィカルモデルの構造学習

の

$$h(S) = \sum_{X \notin S} -\log Q^n(X|\pi_V(X))$$

の値を計算するなど (単純ヒューリスティック) がある。実際、 S から V への $d(\cdot, \cdot)$ の合計の最小値 $h^*(S)$ は、 $S = \{X_1, \dots, X_m\}$ として、

$$\begin{aligned} & -\log Q^n(X_{m+1}|\pi_{S \cup \{X_{m+1}\}}(X_{m+1})) - \log Q^n(X_{m+2}|\pi_{S \cup \{X_{m+1}, X_{m+2}\}}(X_{m+2})) \cdots \\ & -\log Q^n(X_{N-1}|\pi_{V \setminus \{X_N\}}(X_{N-1})) - \log Q^n(X_N|\pi_V(X_N)) \end{aligned}$$

の形になるが、任意の $X \in T \subseteq V$ について、 $Q^n(X|\pi_T(X)) \leq Q^n(X|\pi_V(X))$ となり、単純ヒューリスティックは $h^*(S)$ の下界となる。

1.4.4 条件付き独立性と BN の構造推定

他方、各 $X \in S \subseteq V$ について、

$$K^n(X, S) := \frac{Q^n(X|\pi_S(X))}{Q^n(X|S \setminus \{X\})}$$

とおくと、

$$\begin{aligned} R^n(V) &= Q^n(X_N|\pi_N) \cdots Q^n(X_1|\pi_1) \\ &= Q^n(V) \cdot \frac{Q^n(X_N|\pi_N)Q^n(V \setminus \{X_N\})}{Q^n(V)} \cdot \frac{Q^n(X_{N-1}|\pi_{N-1})Q^n(V \setminus \{X_{N-1}, X_N\})}{Q^n(V \setminus \{X_N\})} \\ &\quad \cdots \frac{Q^n(X_2|\pi_2)Q^n(X_1)}{Q^n(X_1, X_2)} \cdot \frac{Q^n(X_1|\pi_1)}{Q^n(X_1)} \\ &= Q^n(V) \cdot K^n(X_N, V_N) \cdots K^n(X_1, V_1) \end{aligned}$$

とかける。したがって、 $U \cup \{X\} = S$ なる各 (U, S) について定義される (1.28) の和を最小化することと、 $K^n(X, S)$ の和を最小化することは等価であり、前節で定義した条件付き情報量の推定量 $J^n(\cdot, \cdot)$ を用いると、 $J^n(X, S \setminus (\{X\} \cup \pi_S(X))|\pi_S(X), \cdot)$ の和を最大化することと等価である。このことは、 S を $\{ \}$ から V まで変化させ、 X の S における親集合 $\pi_S(X)$ を求め、その条件付き相互情報量の推定量の和を最大化していることにほかな

らない。そして、そのようにして得られた $\{(Y, X) | Y \in \pi_S(X)\}$ の和を辺集合とする BN が求まるのである。

たとえば、事前確率はすべて等しいとして、(1) と (4)、もしくは (10) と (11) で事後確率を比較することは、(1.20) の独立性の検定に帰着される。また、(7) と (11) で事後確率を比較することは、(1.27) の条件付き独立性の検定に帰着される。

1.5 MDL 原理の適用

MDL(minimum description length, MDL[6]) は、サンプルが与えられると、それをある規則とその例外という 2 段階で記述し、その記述の長さが最も短くなる規則を、真の規則とする学習の原理である。AIC (Akaike Information Criterion) などと同様、モデルを選択するための情報量基準のひとつである¹⁾。

BN の構造学習は、大域スコアに事前確率をかけた値を最大にするが、ここでは、事前確率を等しいとしたときに、大域スコアに対数を施し、その値を記述長として、その値を最小にする方法 (Suzuki 1993 [11]) を紹介する。

1.5.1 スコアの導出

$-\log Q^n(X)$ は、 $a = 0.5$ とおいて、Stirling の公式を用いると、

$$L^n(X) = nH^n(X) + \frac{k(X)}{2} \log n$$

にある定数を加えたかたちでかける。ただし、第 1 項は経験的エントロピーとよばれる量で、

$$\sum_x -\frac{c_n(x)}{n} \log \frac{c_n(x)}{n}$$

で定義される。第 2 項の $k(X)$ は、独立な確率パラメータの個数と解釈され、 $X = 0, 1, \dots, \alpha - 1$ の各確率の値の和が 1 であるために、 α 個の中の $\alpha - 1$

¹⁾ BIC(Bayesian Information Criterion) や事後確率最大と本質的な差異はないと見る場合もある

24 第1章 グラフィカルモデルの構造学習

個の値をを指定すればよく、 $k(X) = \alpha - 1$ となる。

実際、(1.21) で $\alpha = 2$, $a(x) = 0.5$ のとき、 (x_1, \dots, x_n) における 1 の頻度を c とすれば、

$$Q^n(X) = \frac{\Gamma(n - c + 0.5)\Gamma(c + 0.5)}{\Gamma(0.5)^2\Gamma(n + 1)}$$

これに、Stirling の公式:

$$n! = \sqrt{2\pi n} \left(\frac{n}{e}\right)^n e^{\lambda_n}$$

なる $(12n + 1)^{-1} < \lambda_n < (12n)^{-1}$ が存在すること、およびその変種 [2]

$$\sqrt{2e} \left(\frac{c}{e}\right)^c \leq \Gamma(c + 0.5) \leq \sqrt{2\pi} \left(\frac{c}{e}\right)^c$$

$$\sqrt{2e} \left(\frac{n - c}{e}\right)^{n - c} \leq \Gamma(n - c + 0.5) \leq \sqrt{2\pi} \left(\frac{n - c}{e}\right)^{n - c}$$

を用いると、

$$\frac{1}{12n + 1} \leq -\log Q^n(X) - nH^n(X) - \frac{1}{2} \log n \leq \log \frac{\pi}{e} + \frac{1}{12n}$$

となり、上からも下からも定数でおさえられることがわかる。同様に、一般の α についても、

$$\left| -\log Q^n(X) - nH^n(X) - \frac{\alpha - 1}{2} \log n \right| \quad (1.29)$$

が n によらないある定数以下になることが示される。

また、たとえば、

$$L^n(X)$$

$$L^n(X|Y) = L^n(X, Y) - L^n(Y)$$

$$L^n(X|Z) = L^n(X, Z) - L^n(Z)$$

$$L^n(Z|X, Y) = L^n(X, Y, Z) - L^n(Y, Z)$$

の4個の記述長(の差)を計算することで、親集合が $\pi(X) = \{\}, \{Y\}, \{Z\}, \{Y, Z\}$ のいずれかであるかの結論が得られる。その場合、それぞれの記述長は、

$$\sum_s n_s H_s^n(X) + \sum_s \frac{k(X)}{2} \log n_s$$

となる。ただし、 $H_s^n(X)$ は $\pi(X) = s$ となるサンプルについての経験的エントロピー、 n_s は $\pi(X) = s$ となるサンプルの個数である。しかし、この値を評価するのに、その上界

$$L^n(X|\pi(X)) = nH^n(X|\pi(X)) + \frac{k(X|\pi(X))}{2} \log n$$

が用いられることが多い[11]。ただし、 $H^n(X|\pi(X)) := \sum_s \frac{n_s}{n} H_s^n(X)$ 、

$k(X)$ にその $\pi(X)$ の取りうる値を乗じた整数を $k(X|\pi(X))$ とおいた。たとえば、 X, Y, Z が α, β, γ 個の値を取る場合、

$$k(X|\{Y\}) = (\alpha - 1)\beta, \quad k(X|\{Y, Z\}) = (\alpha - 1)\beta\gamma$$

となる。この近似誤差は、特にサンプル数 n と比較して、 $k(X|\pi(X))$ が大きい場合に、大きくなる。

1.5.2 計算量の削減

しかし、それでもなお、BNの構造選択でMDL原理を用いるメリットがある。そのひとつは、MDLにかえて、AIC ($\log n$ を2に置き換える) や他の情報量基準を用いることができる点である。そのようにして、サンプルの規則性とその例外(雑音とみなされる)のバランスを変えることができる。

もう一つのメリットは、最適な構造を求めるための計算量が削減されることである。変数 X の親集合を探す場合に、 $\pi(X)$ に $Z \notin \{X\} \cup \pi(X)$ をくわえたときに、

$$nH^n(X|\pi(X)) + \frac{k(X|\pi(X))}{2} \log n \leq \frac{k(X|\pi(X) \cup \{Z\})}{2} \log n$$

すなわち、

$$nH^n(X|\pi(X)) \leq \frac{k(X|\pi(X) \cup \{Z\}) - k(X|\pi(X))}{2} \log n \quad (1.30)$$

26 第1章 グラフィカルモデルの構造学習

であれば、

$$nH^n(X|\pi(X)) + \frac{k(X|\pi(X))}{2} \log n \leq nH^n(X|\pi(X) \cup \{Z\}) + \frac{k(X|\pi(X) \cup \{Z\})}{2} \log n$$

の右辺を計算しなくてもその不等式が成立することがわかるが、 $\pi(X) \cup \{Y\} \subseteq T \subseteq V$ なる T に対しても、 $k(X|\pi(X) \cup \{Z\}) \leq k(X|T)$ と (1.30) より、

$$nH^n(X|\pi(X)) + \frac{k(X|\pi(X))}{2} \log n \leq nH^n(X|T) + \frac{k(X|T)}{2} \log n$$

が成立することがわかる。

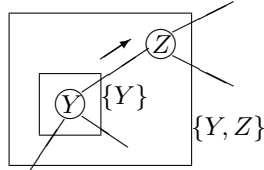


図 1.13 分枝限定法: X の親集合として、 $\{Y\}$, $\{Y, Z\}$ もしくは、それらを含む集合を選択するのか

Suzuki 1996 [10] は、その事実を用いて、分枝限定法 (図 1.13) で探索の計算を削減できることを示した。たとえば、 $\pi(X) = \{Y\}$ で、探索の途中であらたに Z を親集合に含めるかどうか判断する場合に、

$$nH^n(X|\pi(X)) \leq \frac{(\alpha - 1)\beta(\gamma - 1)}{2} \log n \quad (1.31)$$

であれば、 $\{Y, Z\}$ を含む V の部分集合は、 X の最適な親集合にはなれず、深さ優先の探索で、それ以上深い探索は削減することができる。

他方、 $H^n(X) \leq \log \alpha$ と同様に、 $H^n(X|\pi(X)) \leq \log \alpha$ が成立する。また、 $\gamma \geq 2$ であるので、 $\beta \geq \frac{2n}{\log n}$ であれば、

$$\beta \geq \frac{2n \log \alpha}{(\alpha - 1) \log n}$$

がいえて、(1.31) が成立する。さらに、 Y が 1 変数ではなく、複数 (L 変数) の集合で、それらが β 通りの値をとれば、 $\beta \geq 2^L$ が成立する。したがって、 $L \geq \log(\frac{2n}{\log n})$ であれば、(1.31) が成立する。このことから、 $\log(\frac{2n}{\log n})$ を越える個数の親集合は、MDL 原理の意味で最適にはならず、深さ優先探索の途中で棄却される。

したがって、MDL 原理を適用した場合、最適な親集合の候補は、 $\sum_{j=0}^L \binom{N}{j} \leq$

$N^L + 1$ 個に絞られる。特に、スパースな仮定をおくなど、 n が定数であれば L も定数になり、親集合を求める計算は N の多項式時間で完了し、順序グラフ (たとえば図 1.11) の深さが L までのノードで情報を保持すればよく、メモリも N の多項式でおさえることができる。

1.5.3 相互情報量推定への応用

最後に、もうひとつ MDL 原理を適用するメリットとして、見通しのよい解析が得られるということがあげられる。たとえば、(1.29) と同様に、

$$| -\log Q^n(Y) - nH^n(Y) - \frac{\beta-1}{2} \log n |$$

$$| -\log Q^n(X, Y) - nH^n(X, Y) - \frac{\alpha\beta-1}{2} \log n |$$

がそれぞれ定数で抑えられることと、

$$I^n(X, Y) = H^n(X) + H^n(Y) - H^n(X, Y)$$

$$J^n(X, Y) = \frac{1}{n} \log \frac{Q^n(X, Y)}{Q^n(X)Q^n(Y)}$$

を用いると、

$$J^n(X, Y) = I^n(X, Y) - \frac{(\alpha-1)(\beta-1)}{2n} \log n$$

が得られる。すなわち、 $I^n(X, Y)$ は、上式の第 2 項の分だけ大きな値を推定していて、 X, Y のそれぞれの取りうる値 α, β が大きいほど、データがモ

28 第1章 グラフィカルモデルの構造学習

デルに適合しやすく、ペナルティが大きくなっているという解釈が得られる [11]。

1.6 おわりに

以上、Bayesian ネットワークと Markov ネットワーク、事後確率最大の森の構造学習、事後確率最大の BN の構造学習、MDL 原理の適用に関して基本的なことを説明した。

この他、条件付き独立性の検定結果を用いる方法、離散と連続の変数が混在する場合、Markov ネットワークの構造学習など、紙面の都合で説明できなかったこともたくさんある。

ただ、Bayes による構造学習に関しては、この章を理解しただけで十分すぎるくらいのレベルになっているものと思われる。

参考文献

- [1] H. Akaike, "Information theory and an extension of the maximum likelihood principle, *2nd International Symposium on Information Theory*, Budapest, Hungary (1973).
- [2] N. Batir, "Inequalities for the Gamma function", *Archiv Der Mathematik*, 91(6):554-563, Dec (2008)
- [3] C. P. de Campos, Q. Ji, "Efficient Structure Learning of Bayesian Networks using Constraints", *Journal of Machine Learning Research*, vol. 12, no. Mar, pages 663-689, (2011).
- [4] C. K. Chow and C. N. Liu. "Approximating discrete probability distributions with dependence trees". *IEEE Transactions on Information Theory*, IT-14(3):462-467, May (1968).
- [5] R.E. Krichevsky and V.K. Trofimov. "The Performance of Universal Encoding", *IEEE Trans. Information Theory*, Vol. IT-27, No. 2, pp. 199-207 (1981)
- [6] J. Rissanen, "Modeling by shortest data description," *Automatica* 14: 465-471 (1978).
- [7] T. Silander, P. Myllymaki, "A Simple Approach for Finding the Globally Optimal Bayesian Network Structure". *Uncertainty in Artificial Intelligence* (2006).
- [8] J. Pearl, *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference (Representation and Reasoning)*, Morgan Kaufmann Pub, 2nd edition (1988)
- [9] Marco Scutari, *Package 'bnlearn'*, <https://cran.r-project.org/web/packages/bnlearn/bnlearn.pdf> (2015).
- [10] J. Suzuki, "Learning Bayesian Belief Networks Based on the Minimum

30 参考文献

- Description Length Principle: An Efficient Algorithm Using the B & B Technique". *International Conference on Machine Learning*, pages 462-470 (1996)
- [11] J. Suzuki, "A Construction of Bayesian Networks from Databases on an MDL Principle". The Ninth Conference on Uncertainty in Artificial Intelligence, Washington D. C., 266-273 (1993).
 - [12] J. Suzuki, "The Bayesian Chow-Liu Algorithm", in the proceedings of *The Sixth European Workshop on Probabilistic Graphical Models*, Granada, Spain (2012).
 - [13] J. Suzuki, "Consistency of Learning Bayesian Network Structures with Continuous Variables: An Information Theoretic Approach", *Entropy* 2015, vol. 17, no. 8, page 5752-5770 (2015).
 - [14] C. Yuan and B. Malone, "Learning Optimal Bayesian Networks: A Shortest Path Perspective", *Journal of Machine Learning Research*, vol. 48, pages 23-65 (2013).