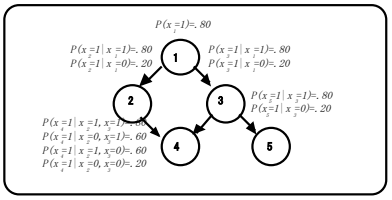
ベイジアン・ネットワークの学習

電気通信大学大学院
情報システム学研究科
植野　真臣

ベイジアン・ネットワークモデル



$$P(x_1, x_2, \cdots, x_N \mid G) = \prod_{i=1}^{N} p(x_i \mid \Pi_i, G)$$

$\Pi_i \subseteq \{x_1, x_2, \cdots, x_{q_i}\}$ は変数iの親ノード集合

パラメータ化とパラメータ推定

ベイジアン・ネットワークのParametrerization
（Spiegelhalter, D.J.,and Lauritzen, S.L. 1990）

Spiegelhalter, D.J.,and Lauritzen, S.L. Sequential updating of conditional Probabilities on directed graphicd structures. Networks 20 (1990), 579-605

今、θ_{ijk} を親ノード変数集合Π_i がj番目のパターンをとったときの

$x_i = k$

となる条件付確率を示すパラメータとする。
このとき、データXを得たときの尤度は、以下のとおりである。

$$L(\Theta_S \mid \mathbf{X}, G) \propto \prod_{k=0}^{r_i-1} \theta_{ijk}^{\,n_{ijk}}$$

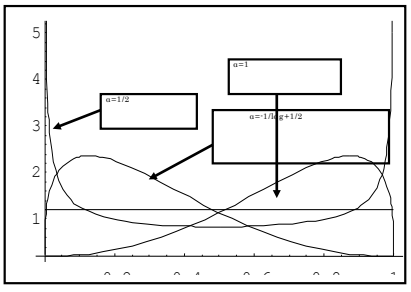
$$\propto \prod_{i=1}^{N} \prod_{j=1}^{q_i} \frac{\left(\sum_{k=0}^{r_i-1} n_{ijk}\right)!}{\prod_{k=0}^{r_i-1} n_{ijk}!} \prod_{k=0}^{r_i-1} \theta_{ijk}^{\,n_{ijk}}$$

Cooper and Herskovits, 1992 A Bayesian methods for the Induction of Probabilistic networks from data, Machine Learning, 9, 309-347

- 多項分布の自然共役分布である以下のディレクレイ分布を事前分布に導入

$$p(\Theta_S \mid G) = \prod_{i=1}^{N} \prod_{j=1}^{q_i} \frac{\Gamma\left(\sum_{k=0}^{r_i-1} \alpha_{ijk}\right)}{\prod_{k=0}^{r_i-1} \Gamma(\alpha_{ijk})} \prod_{k=0}^{r_i-1} \theta_{ijk}^{\,\alpha_{ijj}-1}$$

ディレクレィ分布の周辺分布ベータ分布



事後分布

$$p(\mathbf{X},\Theta_G \mid G) = \prod_{i=1}^{N}\prod_{j=1}^{q_i}\frac{\Gamma(\sum_{k=0}^{r_i-1}(\alpha_{ijk}+n_{ijk}-1))}{\prod_{k=0}^{r_i-1}\Gamma(\alpha_{ijk}+n_{ijk}-1)}\prod_{k=0}^{r_i-1}\theta_{ijk}^{\ \alpha_{ijk}+n_{ijk}-1}$$

$$\propto \prod_{i=1}^{N}\prod_{j=1}^{q_i}\prod_{k=0}^{r_i-1}\theta_{ijk}^{\ \alpha_{ijk}+n_{ijk}-1}$$

EAP推定量

$$\hat{\theta}_{ijk} = \frac{\alpha_{ijk}+n_{ijk}}{\alpha_{ij}+n_{ij}}$$

ただし $\alpha_{ij} = \sum_{k=0}^{r_i-1}\alpha_{ijk}$ $n_{ij} = \sum_{k=0}^{r_i-1}n_{ijk}$

表1

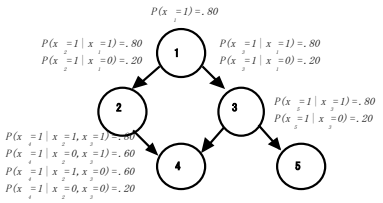
	ノード番号				
	1	2	3	4	5
1	1	0	1	1	1
2	0	0	0	1	0
3	1	1	1	1	1
4	1	1	1	1	0
5	1	1	1	1	0
6	1	1	0	0	0
7	1	0	0	0	0
8	0	0	0	0	0
9	0	0	0	0	0
10	0	1	1	1	1
11	0	0	0	0	0
12	1	1	1	1	1
13	0	0	1	1	1
14	1	1	1	1	1
15	1	0	1	1	1
16	1	1	1	0	0
17	1	1	0	1	0
18	1	1	1	0	0
19	1	1	0	1	0
20	1	1	1	0	1
平均	0.70	0.60	0.60	0.60	0.40

表1より推定した母数推定値

	真の値	$\alpha_{ijk}=0$	$\alpha_{ijk}=1$	$\alpha_{ijk}=1/2$
P(x2=1|x1=1)	0.8	0.8.	0.76	0.78
P(x2=1|x1=0)	0.2	0	0.14	0.08
P(x3=1|x1=1)	0.8	0.8	0.76	0.78
P(x3=1|x1=0)	0.2	0	0.14	0.08
P(x4=1|x2=1, x3=1)	0.8	0.66	0.64	0.65
P(x4=1|x2=1, x3=0)	0.6	0.5	0.5	0.5
P(x4=1|x2=0, x3=1)	0.6	0.5	0.5	0.5
P(x4=1|x2=0, x3=0)	0.2	0.4	0.43	0.41
P(x5=1|x3=1)	0.8	0.67	0.64	0.42
P(x5=1|x3=0)	0.2	0.33	0.35	0.34
真の値との平均自乗誤差		0.0193	0.015	0.028

Learning Bayesian Networks

データから構造を推定する



モデル選択基準

もっとも良いモデルをデータから選択する基準

以下を最小化するモデルを選べばよい

Akaike Information Criterion

AIC = - 2 ln-Likelihood + 2 No.Parameters

Bayesian Information Criterion

BIC = -2 ln-Likelihood + No.Parameters ln(n)

ここで　nはデータ数

AICは期待対数尤度の近似，BICは周辺尤度の近似

AICは漸近的一致性を持たないが，BICは持つ．

AICとBICの条件

- AICとBICは，統計的正則モデルを仮定している。
- 統計的正則モデル：最尤推定量が正規分布に法則収束するモデルを 正則モデルと呼ぶ。このとき，漸近的にフィッシャー情報量行列の固有値がすべて0よりも大きいので漸近展開により，AICやBICが導出される。

ベイジアンネットワークは

- 統計的正則性を持たない。
- 情報科学で用いられる数理モデルは，ほとんど統計的正則性を持たない。
- 理論的には　AICやBICを用いることは問題がある。

ディレクレィ分布の周辺尤度を直接計算

$$p(G \mid X) \propto P(G) \int_{\Theta_S} p(\mathbf{X}, \mid \Theta_G G) p(\Theta_G) d\Theta_G$$

$$= P(G) \prod_{i=1}^{N} \prod_{j=1}^{q_i} \frac{\Gamma(\alpha_{ijk})}{\Gamma\left[\sum_{k=0}^{r_i-1} (\alpha_{ijk} + n_{ijk})\right]} \prod_{k=0}^{r_i-1} \frac{\Gamma(\alpha_{ijk} + n_{ijk})}{\Gamma(\alpha_{ijk})}$$

$$= P(G) \prod_{i=1}^{N} \prod_{j=1}^{q_i} \frac{\Gamma(\alpha_{ijk})}{\Gamma(\alpha_{ij} + n_{ij})} \prod_{k=0}^{r_i-1} \frac{\Gamma(\alpha_{ijk} + n_{ijk})}{\Gamma(\alpha_{ijk})}$$

K2
(Cooper and Herskovits 1992)
α_{ijk}=1（一様分布）のとき

$$p(S \mid \mathbf{X}) \propto p(S) \prod_{i=1}^{N} \prod_{j=1}^{q_i} \frac{(r_i - 1)!}{(n_{ij} + r_i - 1)!} \prod_{k=0}^{r_i-1} n_{ijk}!$$

D.Heckerman, D.Geiger and D.M.Chickering (1995)

- Likelihood equivalence

S_1　S_2 の構造がマルコフ確率構造として同型であるとき

$$p(\Theta_U \mid S_1) = p(\Theta_U \mid S_2)$$

が成り立つこと

n'ijk =1のときの周辺尤度

今、二つの変数x, y について二つの背反するLikelihood equivalenceの構造 $S_{x\to y}$　$S_{y\to x}$ を考える。

$$p(S \mid \mathbf{X}) \propto p(S)\prod_{i=1}^{N}\prod_{j=1}^{q_i}\frac{(r_i-1)!}{(n_{ij}+r_i-1)!}\prod_{k=0}^{r_i-1} n_{ijk}!$$

を用いた場合、

$$p(S_{x\to y} \mid X) \neq p(S_{y\to x} \mid X)$$

となり、Likelihood equivalenceの仮定を満たさない。

BDe Score Metric

(Likelihood equivalent Bayesian Dirichlet scoring)

- Likelihood equivalenceの仮定を満たすScoring Metricの十分条件は

$$P(S)\prod_{i=1}^{N}\prod_{j=1}^{q_i}\frac{\Gamma(\alpha_{ijk})}{\Gamma(\alpha_{ij}+n_{ij})}\prod_{k=0}^{r_i-1}\frac{\Gamma(\alpha_{ijk}+n_{ijk})}{\Gamma(\alpha_{ijk})}$$

ただし $\alpha_{ijk} = \alpha p(x_i = k, \Pi_i = j \mid S)$

α　は事前分布の擬似サンプル数ESSで事前分布の重みでもある

BDeu Score Metric (W.L.Buntine (1991))

- Bdeの一様分布を考えたモデル

$$P(S)\prod_{i=1}^{N}\prod_{j=1}^{q_i}\frac{\Gamma(\alpha_{ijk})}{\Gamma(\alpha_{ij}+n_{ij})}\prod_{k=0}^{r_i-1}\frac{\Gamma(\alpha_{ijk}+n_{ijk})}{\Gamma(\alpha_{ijk})}$$

ただし

$$\alpha_{ijk} = \alpha/(q_i r_i)$$

問：ハイパーパラメータn'を大きくするとエッジは付きやすくなるのか？

$$\hat{\theta}_{ijk} = \frac{\alpha_{ijk}+n_{ijk}}{\alpha_{ij}+n_{ij}}$$

ただし　$\alpha_{ij} = \sum_{k=0}^{r_i-1}\alpha_{ijk}$　$n_{ij} = \sum_{k=0}^{r_i-1} n_{ijk}$

P(X=1|Y=1) P(X=1|Y=0)
真の条件付き確率

P(X=1|Y=1) P(X=1|Y=0)
ハイパーパラメータαを大きくするとその推定値は一様分布に近づく

予測

- ESSを大きくするとエッジはつきにくくなる

Ueno2010

$$logp(x|\alpha,G)=\sum_{i=1}^{N}\sum_{j=1}^{q_i}\sum_{k=1}^{r_i}(\alpha_{ijk}+n_{ijk})log\frac{\alpha_{ijk}+n_{ijk}}{\alpha_{ij}+n_{ij}}$$
$$-\frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{q_i}\sum_{k=1}^{r_i}\frac{r_i-1}{r_i}\log\left(1+\frac{n_{ijk}}{\alpha_{ijk}}\right)+\boldsymbol{O}(1)$$

定理36 (AICの事前知識表現：Ueno(2010))

$For\ \forall i,\forall j,\forall k,\alpha_{ijk}=\frac{1}{3}n_{ijk}$ のとき，対数周辺尤度スコアはAICにより$\boldsymbol{O}(1)$で近似される.

AICではあらかじめ真のモデルがわかっているのと同じ

定理37 (BICの事前知識表現：Ueno(2010))

$For\ \forall i,\forall j,\forall k,\alpha_{ijk}=1.0$ (事前分布が一様分布，K2(Cooper and Herskovits 1992)に一致)のとき，対数周辺尤度スコアはBICにより$\boldsymbol{O}(1)$で近似される.

系2（$\log-$ BDeuのESSによるトレードオフ：Ueno(2010))

$\alpha+n$が十分大きいとき，$\log-$BDeuは以下に近似できる.

$$\log p(x|G)=\alpha\sum_{i=1}^{N}\log r_i+\sum_{i=1}^{N}\sum_{j=1}^{q_i}\sum_{k=0}^{r_i-1}\left(\frac{\alpha}{r_iq_i}+n_{ijk}\right)\log\frac{\frac{\alpha}{r_iq_i}+n_{ijk}}{\frac{\alpha}{q_i}+n_{ij}}$$
$$-\frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{q_i}\sum_{k=0}^{r_i-1}\frac{r_i-1}{r_i}\log\left(1+\frac{r_iq_in_{ijk}}{\alpha}\right)$$

定理38（ESS増大による完全グラフ生成：Ueno(2010))

ESSを大きくしていくと，学習されたベイジアンネットワーク構造のエッジ数は担当増加し，完全グラフ（complete graph）に近づいていく.

定理39（ESS減少によるからグラフ生成：Ueno(2010))

ESSを小さくしていくと，学習されたベイジアンネットワーク構造のエッジ数が単調減少し，空グラフ（empty graph）に漸近的に近づいてい

データへのESS値の最適性

- 真の条件付き確率分布が非一様のとき：
  最適なESS値は比較的小さい値を与えなければならない．なぜならば，式(6.24)右辺において，対数事後分布はエッジを追加しようとするので過学習を起こしやすい．第2項で過学習を抑えるためにはESS値は小さい値に設定しなければならない.
- 真の条件付き確率分布がほぼ一様のとき：
  最適なESS値は比較的大きな値を与えなければならない．なぜならば，真の条件付き確率分布が一様なので式(6.24)右辺において，対数事後分布は真の因果を発見することが難しい．第2項で過小学習（underfitting）を避けるためにESS値は大きい値に設定しなければならない.
- スパース（過疎）なデータのとき：
  n_{ijk}が欠測データもしくは過疎である場合，なるべくESSの影響を抑え，データの学習への影響を最大にしなければならない．式(6.24)よりESS $=1.0$となるとき，データの影響を最大化できる．Castillo Hadi and Solares(1997)　は周辺尤度の分散がn_{ijk}が欠測のとき，ESSの2乗に反比例することを示している．すなわち，この意味でもESS $=1.0$がデータの影響を最大にできることがわかる.

定理40(ESS減少による空グラフ生成：Ueno(2011))

$\alpha<r_iq_i,(i=1,\dots,N)$, nが十分大きいとき，対数周辺尤度は以下に収束する.

$$\log p(x|g,\alpha)=\sum_{i=1}^{N}\sum_{j=1}^{q_i}\sum_{k=0}^{r_i-1}\left(\frac{\alpha}{r_iq_i}+n_{ijk}\right)\log\frac{\frac{\alpha}{r_iq_i}+n_{ijk}}{\frac{\alpha}{q_i}+n_{ij}}$$
$$-\frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{q_i}\left[\frac{r_i-1}{r_i}\sum_{k=1}^{r_i}\log\left(\frac{(r_iq_i)^2n_{ijk}}{2\pi\alpha^2}\right)\right]+\boldsymbol{O}(1)$$

定理41（最適なESSの決定：Ueno(2010)）
BDe(u)は，$\mathrm{ESS} = 1.0$のとき，事後分布の分散を最大化する.

いまおそらく最もよいスコア
事前知識に頑健な基準

定義87(NIP − BIC：Ueno(2011))

$$\mathrm{NIP-BIC} = \sum_{i=1}^{N}\sum_{j=1}^{q_i}\sum_{k=1}^{r_i}\left(\alpha_{ijk}+n_{ijk}\right)\log\frac{\alpha_{ijk}+n_{ijk}}{\alpha_{ij}+n_{ij}} - \frac{1}{2}q_i r_i \log(1+n)$$

完全無情報事前分布

定義88(完全無情報事前分布スコアNIP − BDe：

$$p(x|g,\alpha) = \sum_{g^h\in G} p(g^h)p(x|\alpha,g,g^h) = \sum_{g^h\in G} p(g^h)\left(\prod_{i=1}^{N}\prod_{j=1}^{q_i}\frac{\Gamma\left(\alpha_{ij}^{g^h}\right)}{\Gamma\left(\alpha_{ij}^{g^h}+n_{ij}\right)}\prod_{k=1}^{r_i}\frac{\Gamma\left(\alpha_{ijk}^{g^h}+n_{ijk}\right)}{\Gamma\left(\alpha_{ijk}^{g^h}\right)}\right)$$

ここで$\Sigma_{g^h\in G}$はすべての可能な構造候補に対する和，$p(g^h)$は構造候補の事前分布でここでは一様分布を仮定している.

構造の探索アルゴリズム

変数の数nに対して、構造の候補数は以下のように増える。

$$f(n) = \sum_{i=1}^{n}(-1)^{i+1}\begin{bmatrix} n \\ i \end{bmatrix}2^{i(n-i)}f(n-i)$$

例えば、n=2のとき、構造数は3、n=3のとき、構造数は25、n=5で29000,n=10で4.2×10^{18}

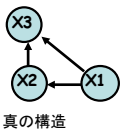
探索問題は、指数爆発する。。なんらかの工夫が必要。

アルゴリズム

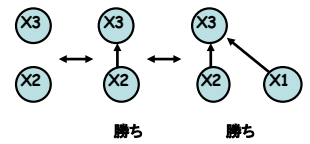
1. 変数間の順序を決める。X1>x2>・・>xn
2. Xnのすべての親ノードパターンを変えながら、情報量基準でどのパターンが最適かを検索。親ノードパターンは、親ノード数をｍ個に制限するのが普通である。
3. Xn-1について、1と同じ手続きを行い、X1まで繰り返す。
4. 最もよい値を置いておく
5. すべての順序について1－4を繰り返す

例

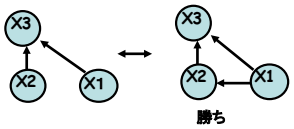
- X1>X2>x3から始める



真の構造



勝ち　勝ち



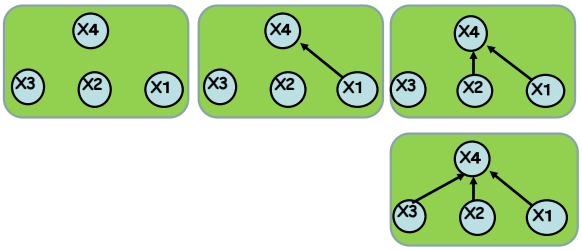
勝ち

- X1>X3>X2
- X2>X3>X1
- X2>X1>X3
- X3>X2>X1
- X3>X1>X2

について同様のことを行い，最もスコアの高かった構造を推定値とする

親探しアルゴリズムが重要

- 順序を所与として，
- 変数Xiの親ノード候補から、いかに早く親ノードをみつけてくるかのためのアルゴリズム
- 例　X4>X3>X2>X1



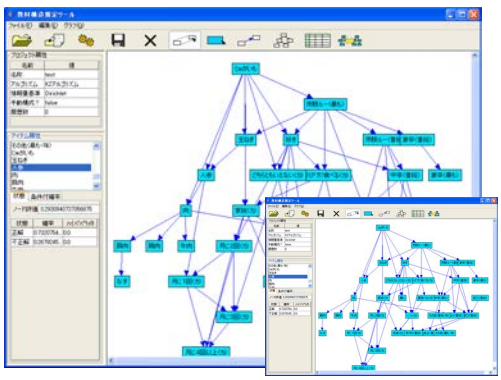
親ノード探索に用いられるアルゴリズム

- 動的計画法DP　O(N2^N)
- A＊探索
- 幅優先探索と分枝限定法（BFBnB）

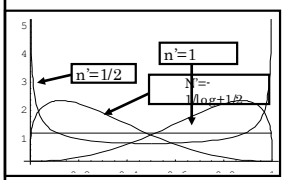
表 6.5　実験結果

データセット			計算時間〔秒〕			必要容量〔B（バイト）〕	
データセット	n	N	DP	BFBnB	A^*	DP	BFBnB
wine	14	178	1	0	0	1.16E+07	2.72E+05
houseVotes	17	435	7	5	3	4.81E+07	4.39E+06
hepatitis	20	126	27	9	6	3.79E+08	2.73E+07
segment	20	2 310	44	28	42	3.79E+08	3.67E+07
meta	22	528	52	57	41	1.67E+09	1.55E+08
import	22	205	123	54	55	1.67E+09	1.52E+08
parkinsons	23	195	297	103	130	3.48E+09	2.63E+08
sensorareadings	25	5 456	12 747	3 061	OM	1.51E+10	1.30E+09
autos	26	159	2 737	1 184	OM	3.15E+10	2.19E+09
flag	29	194	41 733	12 935	OM	2.81E+11	1.55E+10
wdbc	31	569	OD	93 682	OM	OD	6.86E+10
epgenetic	33	72 228	OD	570 760	OM	OD	2.74E+11

表 6.5 の OM はメモリオーバーで探索が打ち切られたことを表し，OD はハードディスクスペースに入らずに探索が打ち切られたことを表す．



性質(1) すべてのパラメータのディレクレ分布のハイパーパラメータを設定できる



例えば、事前に因果関係があると考えられるアークの設定やないと考えられるアークの設定も事前分布の設定により矛盾なく行える

様々な情報量基準を表現できるし、それらをMIXした基準も作り出せる

本手法の特徴 （2）

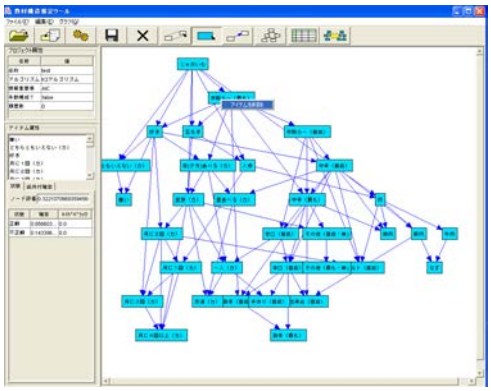
- ベイジアンネットワークモデルの各ノードの価値を定義し、評価しながら、望ましいノード集合を探索する手法

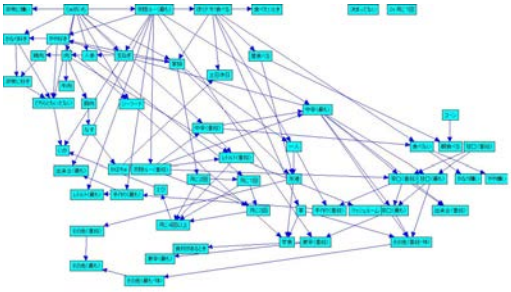
ノードの価値(Expected Value of Node Information)
植野(1993,1994,1996、1999)

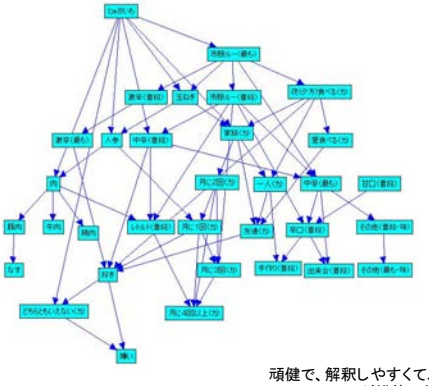
$$EVNIN = \sum_{i=1}^{n} 2^{q_i} p(x_1,\cdots x_n \mid x_i)\log p(x_1,\cdots x_n \mid x_i)$$

$$- \sum_{i=1}^{n} 2^{q_i} p(x_1,\cdots x_n)\log p(x_1,\cdots x_n)$$

変数評価機能：KeyNode







頑健で、解釈しやすくて、予測効率の高い
Causal Modelが構築できる