ベイジアンネットワークと機械学習

植野真臣
電気通信大学
情報理工学研究科　情報数理プログラム

今後のスケジュール（予定）

4月13日　授業の概要とガイダンス
4月20日　ベイズの定理
4月27日　ベイズはどのように誕生したか？
5月11日　ベイズはコンピュータ、人工知能の父である！！
5月18日　アランチューリングとベイズ
5月25日　ビリーフとベイズ
6月1日　　尤度と最尤推定(1)
6月8日　　尤度と最尤推定(2)
6月15日　ベイズ推定と事前分布(1)
6月22日　ベイズ推定と事前分布（２)
6月 29日　国際会議でオンデマンド授業：ベイズ推定と事前分布（3)
7月6日　　データサイエンス：ルービン因果推論
7月13日　ベイジアンネットワークと機械学習
7月27日　テストとデータサイエンス
8月3日　　テストと総括

1. グラフィカルモデル

定義60

X,Y,Zが無向グラフGの互いに排他なノード集合であるとする．もし，XとYの各ノード間の全ての路がZの少なくとも一つのノードを含んでいるとき，ZはXとYを分離する，といい，$I(X,Y|Z)_{\mathrm{G}}$と書く．これは，グラフ上での条件付き独立性を表現する．一方，真の条件付き独立性，すなわち，XとYとZを所与として条件付き独立であるとき，$I(X,Y|Z)_{\mathrm{M}}$と書く．

２．パーフェクトマップ

定義61 グラフGは，ノードに対応する変数間すべての真の条件付き独立性がグラフでの表現に一致し，さらにその逆が成り立つとき，Gをパーフェクトマップ(perfect map)といい，P-mapと書く.

$$I(X,Y|Z)_M \Rightarrow I(X,Y|Z)_G$$

しかし，すべての確率モデルに対応するグラフが存在するわけではない．例えば，四つの変X_1, X_2, Y_1, Y_2が以下の真の条件付き独立性をもっているとする.

$$I(X_1,X_2|Y_1,Y_2)_M, \quad I(X_2,X_1|Y_1,Y_2)_M,$$
$$I(Y_1,Y_2|X_1,X_2)_M, \quad I(Y_2,Y_1|X_1,X_2)_M.$$

問題

- パーフェクト　マップは　グラフィカルモデリングには，厳しすぎる制約！！

３．Iマップ

定義62　グラフGは,グラフでの条件付き独立性のすべての表現が真の条件付き独立性に一致しているとき，Gをインデペンデントマップ(independent map) といい，I-mapと書く.

$$I(X, Y|Z)_G \Rightarrow I(X, Y|Z)_M.$$

I-mapの定義では，完全グラフを仮定するとグラフ上に$I(X, Y|Z)_G$が存在しないので，どんな場合でもI-mapを満たしてしまう.

最小Iマップ

そこで，以下の最小I-mapが重要となる。

定義63　グラフGがI-mapで，かつ，一つでもエッジを取り除くとそれがI-mapでなくなってしまうとき，グラフGは**最小I-map**(minimal I-map)と呼ばれる.

問題

グラフ表現$I(X,Y|Z)_G$と確率表現$I(X,Y|Z)_M$が必ずしも対応しない。

問題

グラフ表現$I(X,Y|Z)_G$と確率表現$I(X,Y|Z)_M$が必ずしも対応しない。

⇓

グラフ表現$I(X,Y|Z)_G$に対応する関係性を表現する手法はないのか？

問題

グラフ表現$I(X,Y|Z)_G$と確率表現$I(X,Y|Z)_M$が必ずしも対応しない。

⇓

グラフ表現$I(X,Y|Z)_G$に対応する関係性を表現する手法はないのか？

⇓

D分離の導入

4. D分離

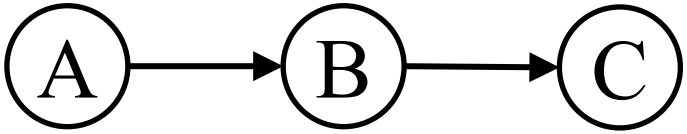


図1　逐次結合

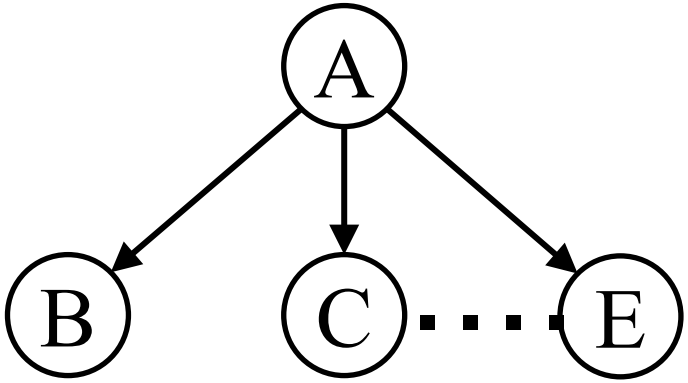


図2　分岐結合

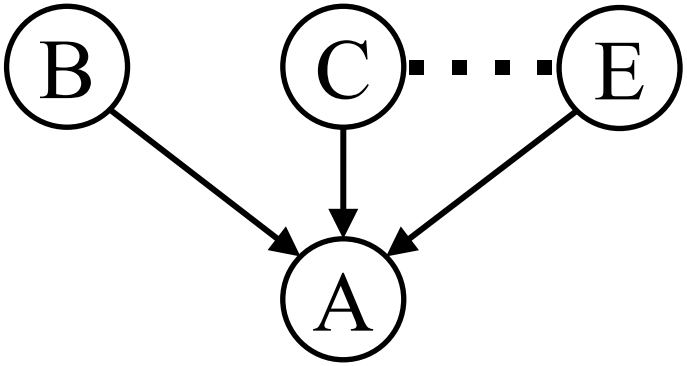
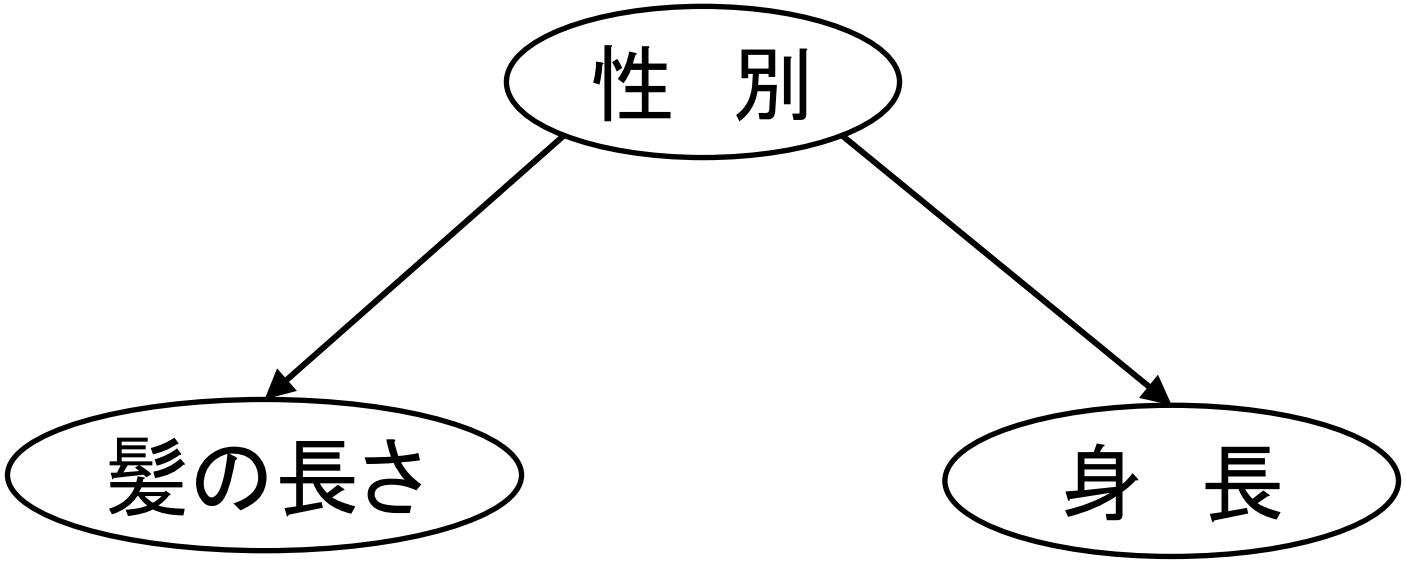


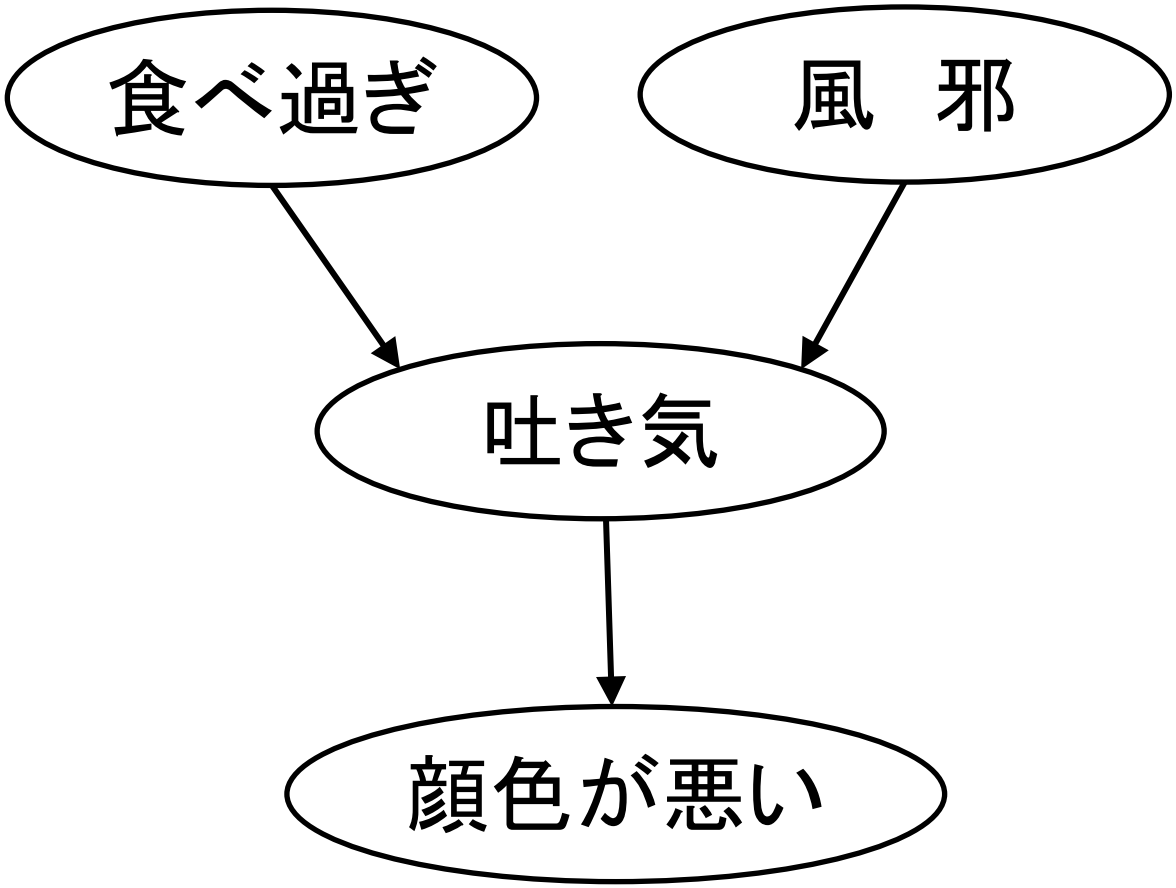
図3　合流結合

分岐結合の例



成人についての性別と髪の長さ，身長との因果ネットワーク

合流結合の例



合流結合のエビデンス

変数B,C,…,Eは，AもしくはAの子孫についての証拠を得ない場合，「**Aを介してd分離である**」と呼ぶ．変数についての証拠はその状態の確からしさの記述でもある．もし，その変数の値が観測されていた場合，「変数がインスタンス化されている」と呼び，その値を「**エビデンス**(evidence)」と呼ぶ．

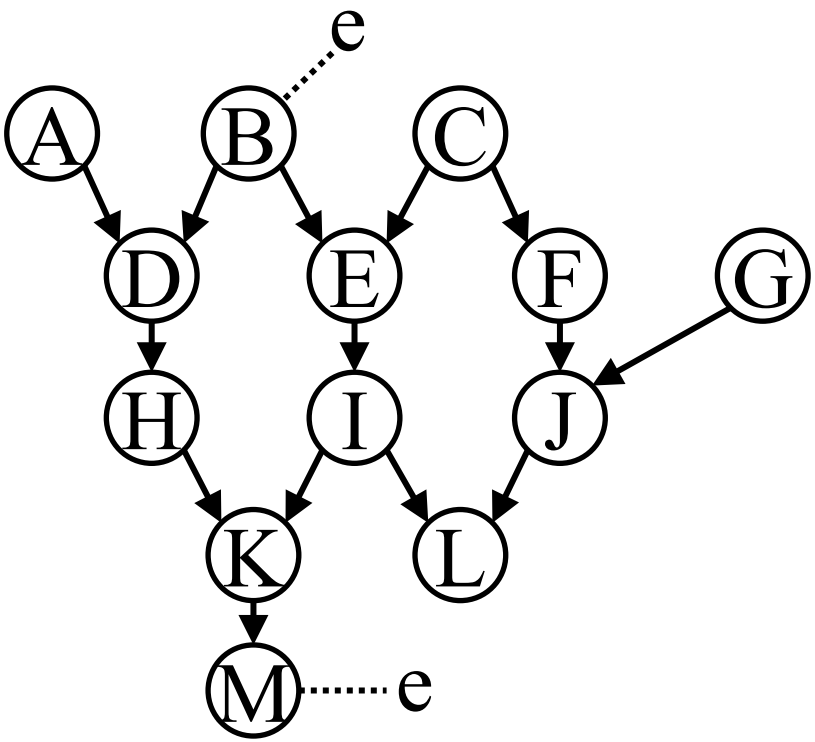
D分離

定義64　因果ネットワークにおける二つの変数AとBは，AとBのすべてのパスに存在する以下のような変数V (AとBを分ける) があるとき，**d分離** (d-separate) である.

・逐次結合もしくは分岐結合でVがインスタンス化されているとき，または，

・合流結合でVもしくはVの子孫がインスタンス化されていないとき，

例

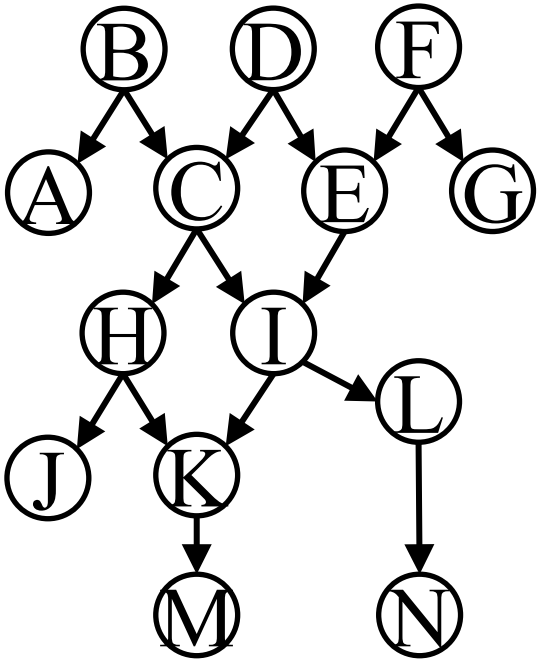


BとMがインスタンス化された
因果ネットワーク

定理(Darwiche 2009)　ある因果ネットワークにおいて，二つのノード集合XとYが，ノード集合Zによってd分離されることは，以下の枝刈り(pruning)によって得られる新しい因果ネットワークにおいて，XとYが分離されていることと同値である.

・$X \cup Y \cup Z$に属していない全てのリーフノード (leaf nodes) を削除する.

例

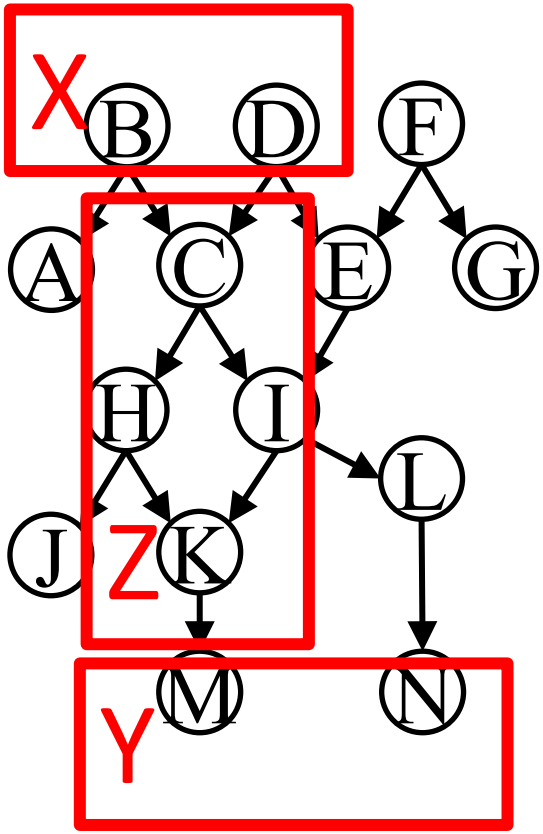


図における因果ネットワークで，$X = \{B, D\}$, $Z = \{C, H, I, K\}$, $Y = \{M, N\}$とする．

$X \cup Y \cup Z = \{B, C, D, H, I, K, M, N\}$であり，それ以外のノードを削除する．Zの要素からのアークをすべて削除すると，XとYは完全に分離される．

すなわち，XとYがノード集合Zを所与としてd分離されている．

例

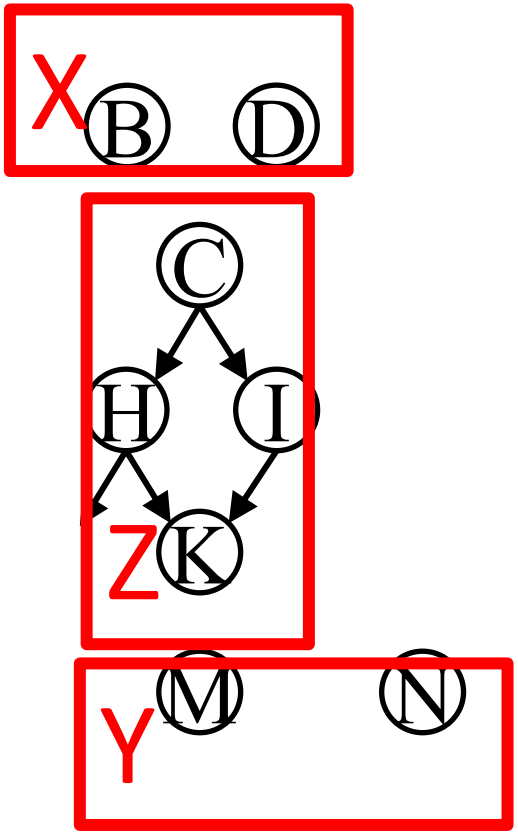


図における因果ネットワークで，$X = \{B, D\}$, $Z = \{C, H, I, K\}$, $Y = \{M, N\}$とする．

$X \cup Y \cup Z = \{B, C, D, H, I, K, M, N\}$であり，それ以外のノードを削除する．Zの要素からのアークをすべて削除すると，XとYは完全に分離される．

すなわち，XとYがノード集合Zを所与としてd分離されている．

例



図における因果ネットワークで，$X = \{B, D\}$, $Z = \{C, H, I, K\}$, $Y = \{M, N\}$とする．

$X \cup Y \cup Z = \{B, C, D, H, I, K, M, N\}$であり，それ以外のノードを削除する．Zの要素からのアークをすべて削除すると，XとYは完全に分離される．

すなわち，XとYがノード集合Zを所与としてd分離されている．

５．マルコフ　ブランケット

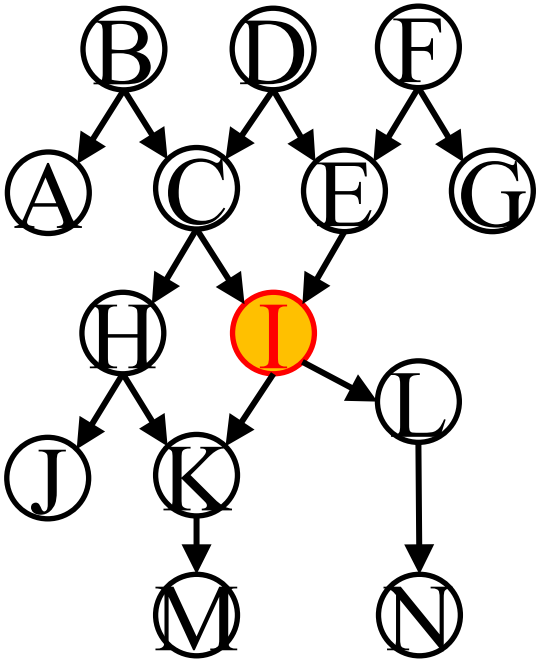
定義65 変数Aのマルコフブランケット (Markov blanket) とは，Aの親集合，子集合，Aと子を共有している変数集合の和集合より成り立つ．

マルコフブランケットとD分離

ノート2　Aについてのマルコフブランケットがすべてエビデンスがある場合，Aはネットワーク中の残りの変数すべてとd分離である.

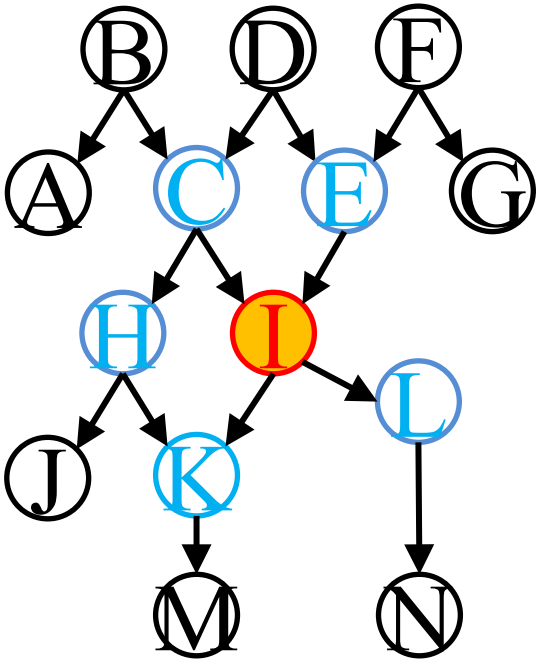
例

図において，Iについてのマルコフブランケットは？

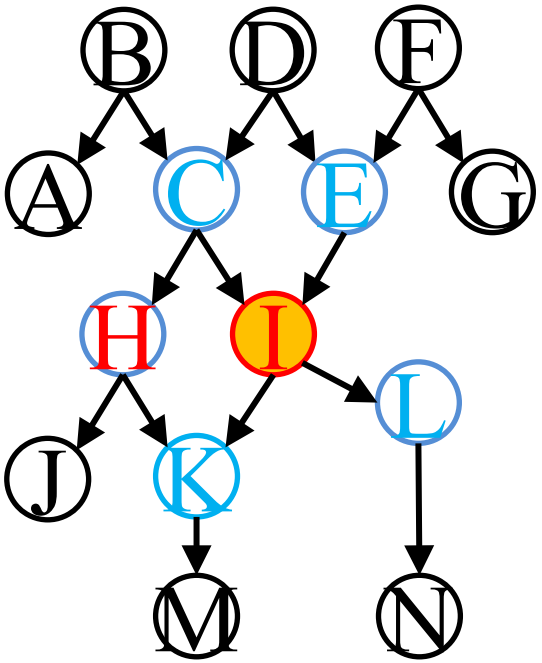


例

図において，Iについてのマルコフブランケットは(C,E,H,K,L)となる.



例



注：Iの近傍の変数のみが所与の場合，JはIとはd分離でないことに注意してほしい．なぜならば，この場合，Iはインスタンス化されないのであるが，Kを所与として合流結合である親Hは Iから影響を受け,Jにも影響を与える.

D分離の表記

定義66　X,Y,ZがグラフGのたがいに排他なノード集合であるとする．XとYがZによってd分離されるとき，$I(X,Y|Z)_d$と書く．

D分離と最小Iマップ

定理19　$I(X,Y|Z)_d$と$I(X,Y|Z)_G$は同値である.

すなわち,　　$I(X,Y|Z)_d \Leftrightarrow I(X,Y|Z)_G$.

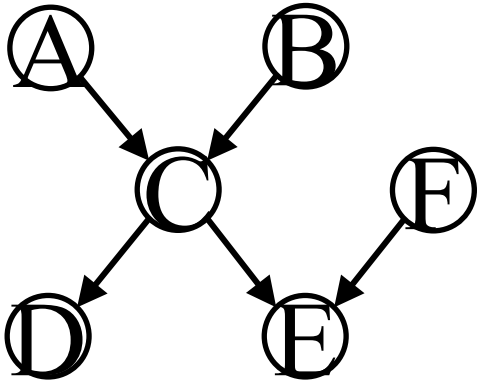
このように，d分離はグラフ上で表現するには都合のよい概念であり，この仮定がベイジアンネットワークのモデルそのものである．真の条件付き独立性すべては表現できておらず，最小I-mapを仮定していることになる.

モラルグラフを用いたD分離の定理

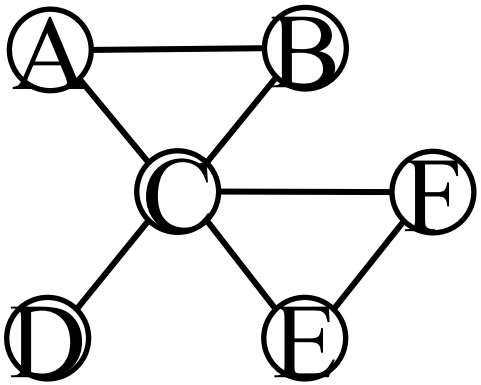
定義67　X,Y,Zがグラフgのたがいに排他なノード集合であるとする．X,Y,Zを含む最小のモラルグラフで，ZがXとYを分離するとき，ZはXとYをd分離する．

例

左図の有向グラフを考える．これを無向化し，モラル化したグラフは右図である．例えば，EはCとFをd分離しない，CはDとFをd分離する，などがわかる．



有向グラフ



モラルグラフ

６．ベイジアンネットワークモデル

定義68　N個の変数集合$x = \{x_1, x_2, \ldots, x_N\}$をもつベイジアンネットワークは，(G,Θ) で表現される．

・Gはxに対応するノード集合によって構成される**非循環有向グラフ**(directed acyclic graph，**DAG**)ネットワーク構造と呼ばれる．

・Θは，Gの各アークに対応する条件付き確率パラメータ集合$\{p(x_i|\Pi_i, \mathrm{G})\}$，$(i = 1, \ldots, N)$である．ただし，$\Pi_i$は変数$x_i$の親変数集合を示している．

同時確率分布

定理20　変数集合$x = \{x_1, x_2, \ldots, x_N\}$をもつベイジアンネットワークの同時確率分布$p(x)$は以下で示される.

$$p(x|\mathrm{G}) = \prod_i p(x_i|\Pi_i, \mathrm{G}))$$

ここで，Gは最小I-mapを示している.

演習

- 定理20を証明せよ.

逐次結合のd分離

[逐次結合の場合] AがBを通じてCに逐次結合している場合を考えよう. このとき,

$p(C|B,A) = p(C|B)$を示せばよい. 定理20より,

$$p(A,B,C) = p(A)p(B|A)p(C|B) = p(A,B)p(C|B).$$

よって

$$p(C|B,A) = \frac{p(A,B,C)}{p(A,B)} = p(C|B)$$

となる.

分岐結合のd分離

[分岐結合の場合] Aの独立な二つの子がB,Cであるとする. このとき,

$p(C|A,B) = p(C|A)$を示せばよい. 定理20より,

$$p(A,B,C) = p(A)p(B|A)p(C|\mathrm{A}) = p(A,B)p(C|\mathrm{A}).$$

よって

$$p(C|A,B) = \frac{p(A,B,C)}{p(A,B)} = p(C|A)$$

となる.

合流結合のd分離

[合流結合の場合] AとBをCの親としよう. いま, $p(A|B) = p(A)$を示せばよい. 定理20より,

$$p(A, B, C) = p(\mathrm{A})p(\mathrm{B})p(C|B, A).$$

Cについて周辺化し,

$$p(A, B) = \sum_C p(\mathrm{A})p(\mathrm{B})p(C|B, A)\,.$$

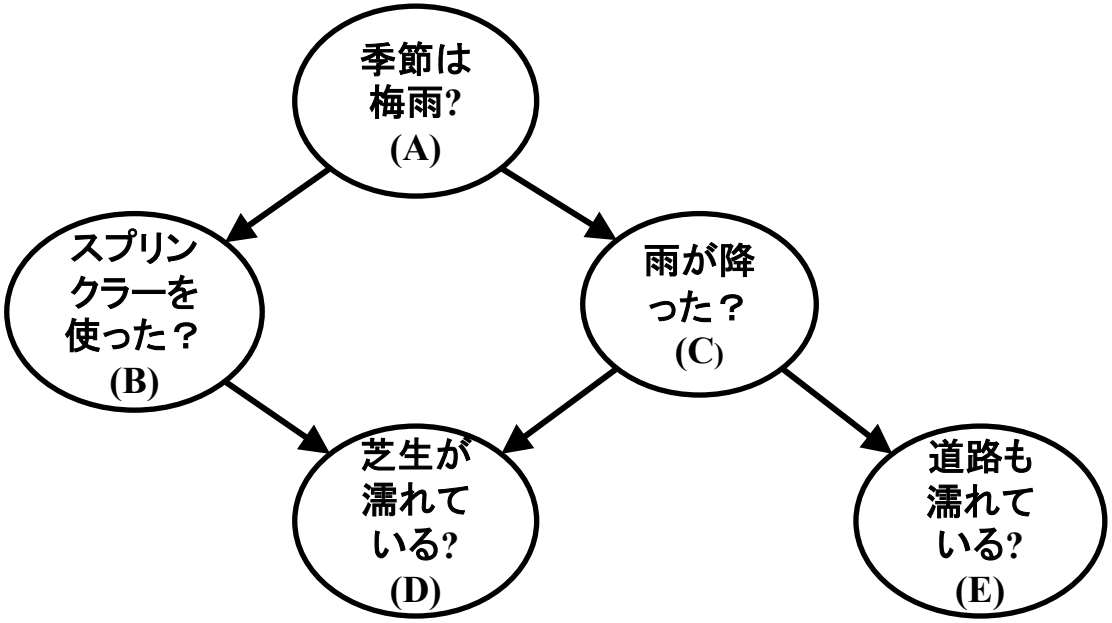
ここで$\sum_C$ は$p(C|B, A)$にのみかかっているので,

$$\sum_C p(\mathrm{A})p(\mathrm{B})p(C|B, A) = p(\mathrm{A})p(\mathrm{B}) \sum_C p(C|B, A)\,.$$

条件付き確率表で列の和は1になるので, $\sum_C p(C|B, A)$は1となる. したがって

$p(A, B) = p(A)p(B)$となるのである.

7. CPT(Conditional probabilities tables)



A	p(A)
真	0.6
偽	0.4

A	B	p(B|A)
真	真	0.2
真	偽	0.8
偽	真	0.75
偽	偽	0.25

A	C	p(C|A)
真	真	0.8
真	偽	0.2
偽	真	0.1
偽	偽	0.9

B	C	D	p(E|C)
真	真	真	0.95
真	真	偽	0.05
真	偽	真	0.9
真	偽	偽	0.1
偽	真	真	0.8
偽	真	偽	0.2
偽	偽	真	0.0
偽	偽	偽	1.0

C	E	p(E|C)
真	真	0.7
真	偽	0.3
偽	真	0.0
偽	偽	1.0

図3.12　ベイジアンネットワークのCPT1[1)]

演習：以下の確率を求めよ

$$p(A = 1, B = 1, C = 1, D = 1, E = 1 | G)$$

8.同時確率分布表 : joint probability distribution table, JPDT

表3.1　図3.2の同時確率分布表 : JPDT

A	B	C	D	E	p(A,B,C,D,E)	A	B	C	D	E	p(A,B,C,D,E)
1	1	1	1	1	0.06384	0	1	1	1	1	0.01995
1	1	1	1	0	0.02736	0	1	1	1	0	0.00855
1	1	1	0	1	0.00336	0	1	1	0	1	0.00105
1	1	1	0	0	0.00144	0	1	1	0	0	0.00045
1	1	0	1	1	0.0	0	1	0	1	1	0.0
1	1	0	1	0	0.02160	0	1	0	1	0	0.24300
1	1	0	0	1	0.0	0	1	0	0	1	0.0
1	1	0	0	0	0.00240	0	1	0	0	0	0.02700
1	0	1	1	1	0.21504	0	0	1	1	1	0.00560
1	0	1	1	0	0.09216	0	0	1	1	0	0.00240
1	0	1	0	1	0.05376	0	0	1	0	1	0.00140
1	0	1	0	0	0.02304	0	0	1	0	0	0.00060
1	0	0	1	1	0.0	0	0	0	1	1	0.0
1	0	0	1	0	0.0	0	0	0	1	0	0.0
1	0	0	0	1	0.0	0	0	0	0	1	0.0
1	0	0	0	0	0.09600	0	0	0	0	0	0.09000

9. ベイジアンネットワークの学習
Ueno (2010, 2011) UAI

予測を最大化する予測分布は

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

ここで、

$\alpha_{ijk} = 1/K$

K: モデルのパラメータ数

n_{ijk}：変数 i が j 番目の親ノードパターンを条件として k をとる頻度

10.背景

世界中のすべての変数の同時確率分布を知ればなんでも推論できる！！

問題

- 変数が増えると状態数が指数的に増える
- すべて2値しかとらなくても　１０変数で１０２４パターンの同時確率を推定しないといけない。
- １００変数で

  1267650000000000000000000000000

パターンの同時確率を推定しなければならない。

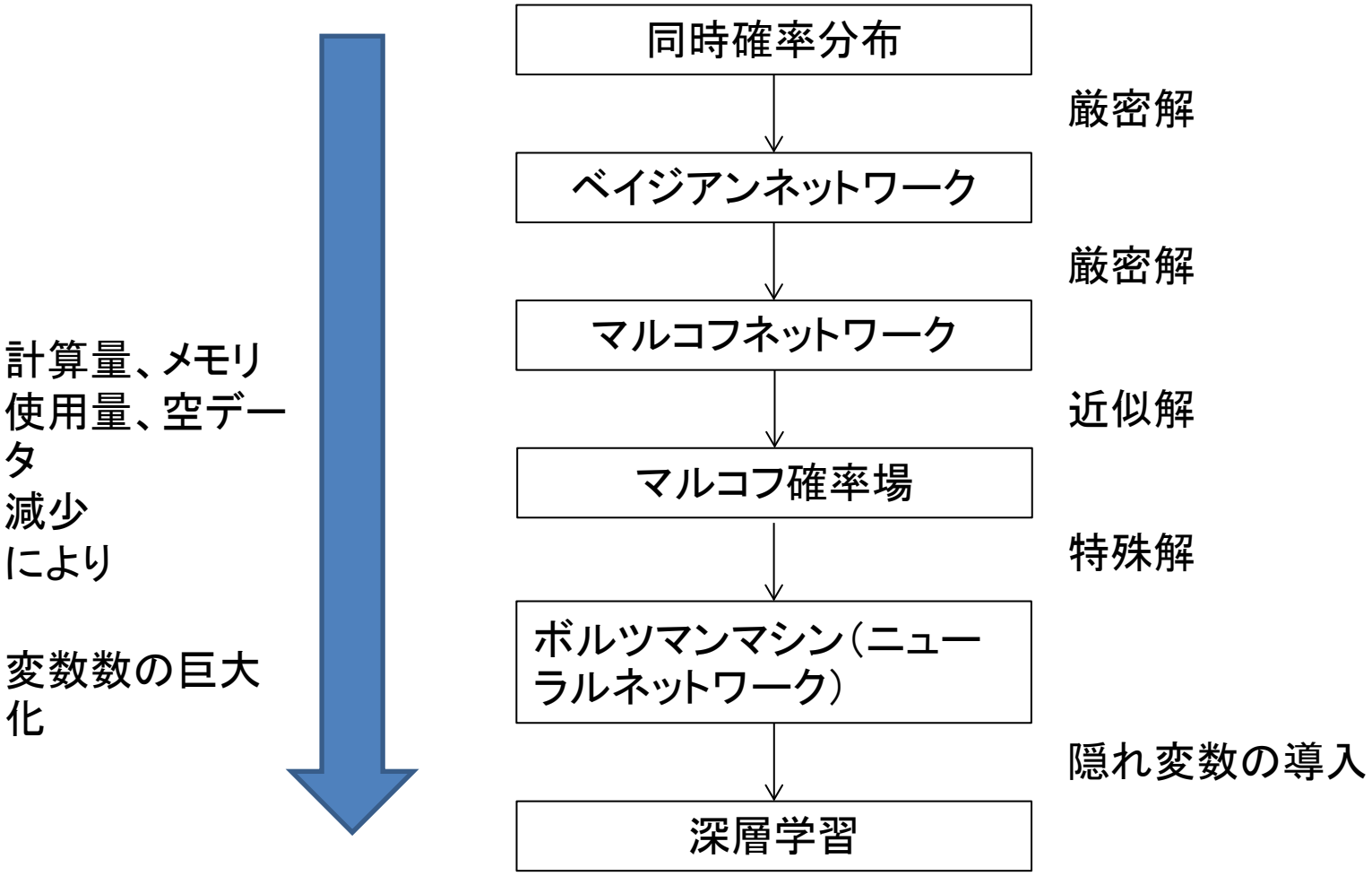
二つの問題

- 計算量が指数的に爆発する。
- データ数よりパターン数のほうが多くなってしまうと、各パターンを推定するためのデータが0になるものが大量発生。

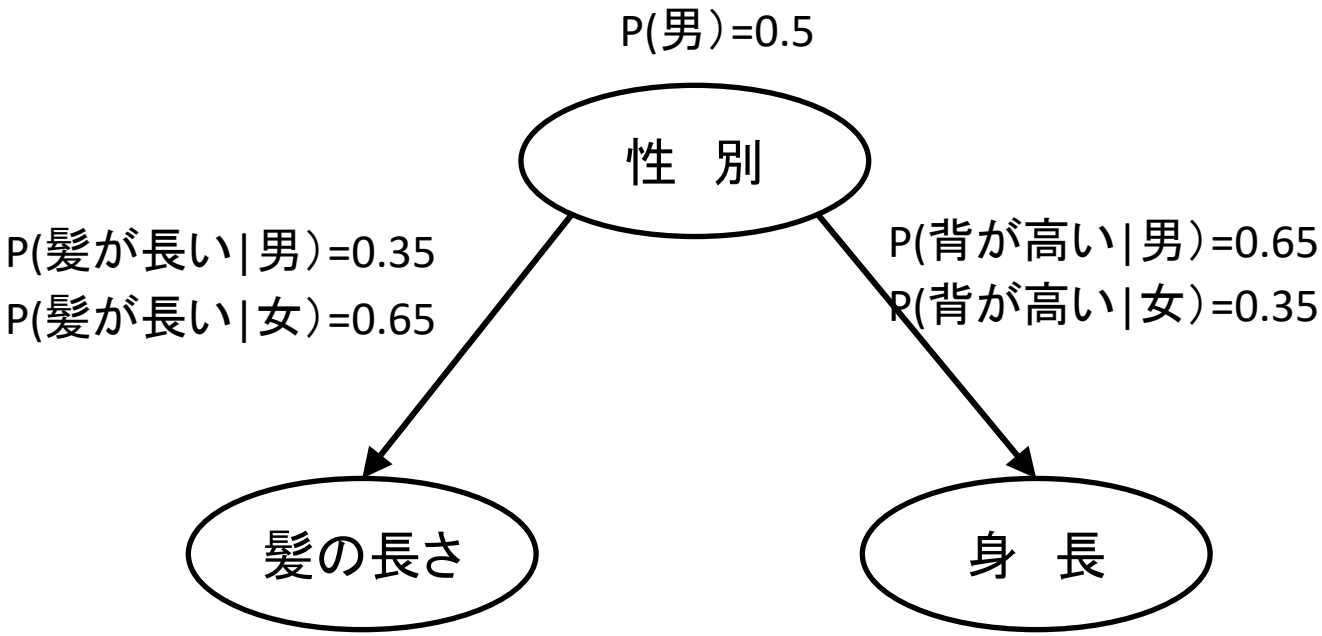
（大量のデータがあっても空データだらけになる）

ビッグデータ問題の課題は、スパースデータ（空データの増加）と計算量(メモリ、計算速度）

同時確率分布推定のための数理モデル



ベイジアン・ネットワーク



$$p(\text{性別、髪の長さ、身長}\,|\,G) =$$
$$p(\text{性別})\,p(\text{髪の長さ}\,|\,\text{性別})\,p(\text{身長}\,|\,\text{性別})$$

同時確率パラメータは2^3-1=7個

条件付き確率パラメータは5個

ベイジアンネットワーク

- 確率構造が非循環有向グラフであれば、

同時確率分布が条件付確率の積に因数分解できることが数学的に証明できる。

IMAP(同時確率分布を表現できる構造)の中で最小パラメータ数の構造が推定される。

- 確率有向グラフが確率因果構造が対応し、解釈性、説明性が高い。

定式化

定義　N個の変数集合$x = \{x_1, x_2, \ldots, x_N\}$をもつベイジアンネットワークは，(G,Θ) で表現される.

・Gはxに対応するノード集合によって構成される**非循環有向グラフ**(directed acyclic graph，**DAG**)ネットワーク構造と呼ばれる.

・Θは，Gの各アークに対応する条件付き確率パラメータ集合$\{p(x_i|\Pi_i, \mathrm{G})\}$，$(i = 1, \ldots, N)$である．ただし，$\Pi_i$は変数$x_i$の親変数集合を示している.

同時確率分布

定理　変数集合$x = \{x_1, x_2, \ldots, x_N\}$をもつベイジアンネットワークの同時確率分布$p(x)$は以下で示される.

$$p(x|\mathrm{G}) = \prod_i p(x_i|\Pi_i, \mathrm{G}))$$

ここで，Gは確率構造を示している.

ベイジアンネットワーク学習
ディレクレィ分布の周辺尤度最大化

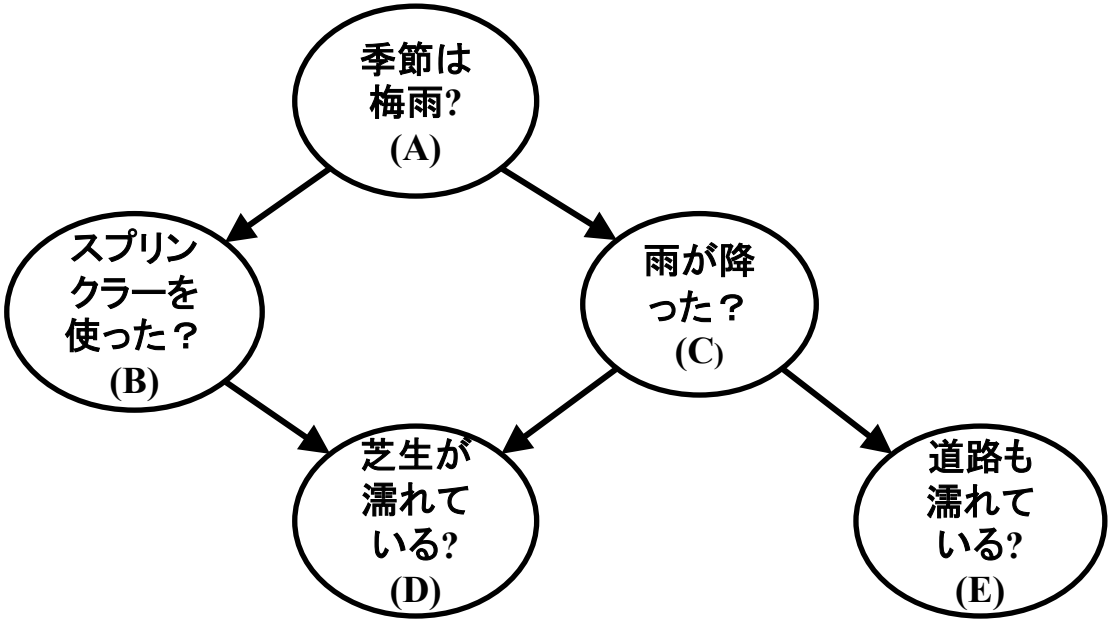
$$p(G \mid X) \propto P(G) \int_{\Theta_S} p(\mathbf{X}, \mid \Theta_G G)\, p(\Theta_G) d\Theta_G$$

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

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

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

CPT(Conditional probabilities tables)



A	p(A)
真	0.6
偽	0.4

A	B	p(B|A)
真	真	0.2
真	偽	0.8
偽	真	0.75
偽	偽	0.25

A	C	p(C|A)
真	真	0.8
真	偽	0.2
偽	真	0.1
偽	偽	0.9

B	C	D	p(E|B, C)
真	真	真	0.95
真	真	偽	0.05
真	偽	真	0.9
真	偽	偽	0.1
偽	真	真	0.8
偽	真	偽	0.2
偽	偽	真	0.0
偽	偽	偽	1.0

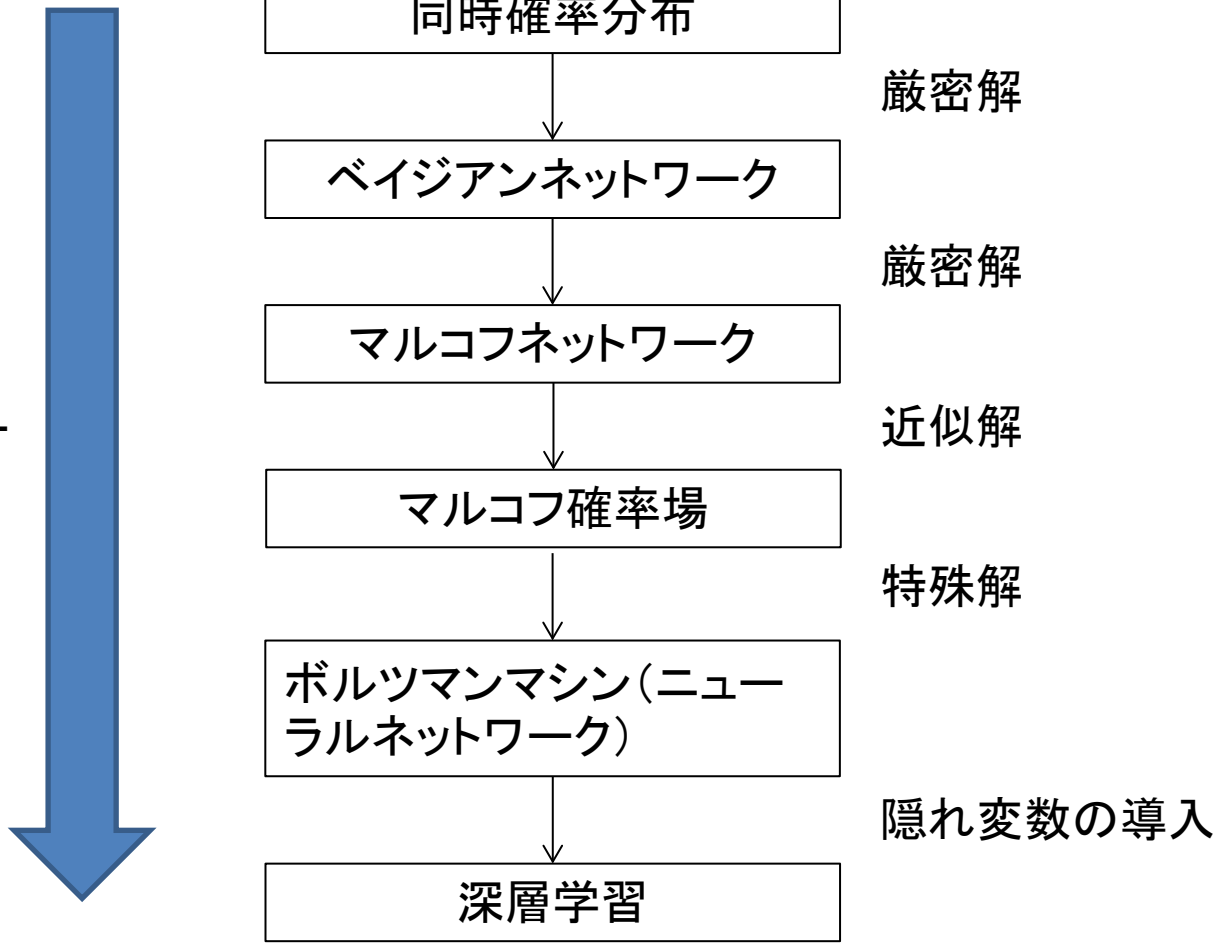
C	E	p(E|C)
真	真	0.7
真	偽	0.3
偽	真	0.0
偽	偽	1.0

図1　ベイジアンネットワークのCPT1[1)]

ベイジアンネットワークの問題

- 計算量の多さ

解決のための数理モデル



マルコフネットワーク

- 無向グラフ構造

利点

- 非循環有向グラフ構造の仮定がいらない。

欠点

- ベイジアンネットワークで表現できない構造を表現できるが逆にベイジアンネットワークでできて表現できない構造もある
- 構造の学習ができない
- パラメータ数が多い

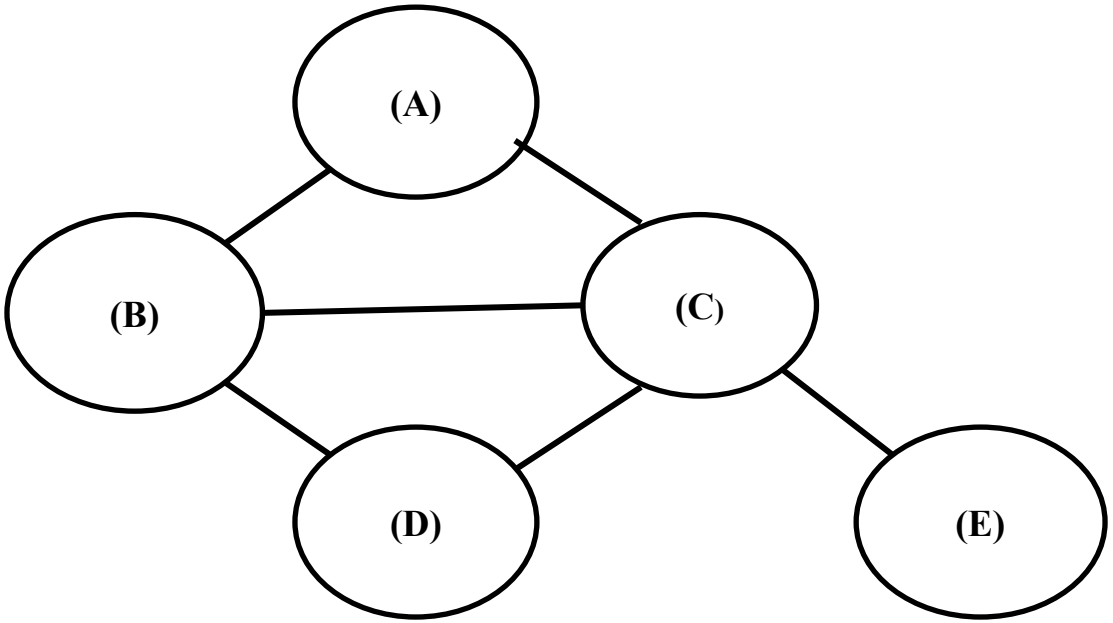
マルコフネットワークの同時確率分布のクリーク分解定理

- $P(x_1, x_2, \cdots, x_N|G) = \frac{1}{Z(\theta)} \prod_{c \in C} \phi_c(x_c|\theta_c)$
- をギブス分布（Gibbs Distribution)と呼ぶ。

Cはクリーク集合

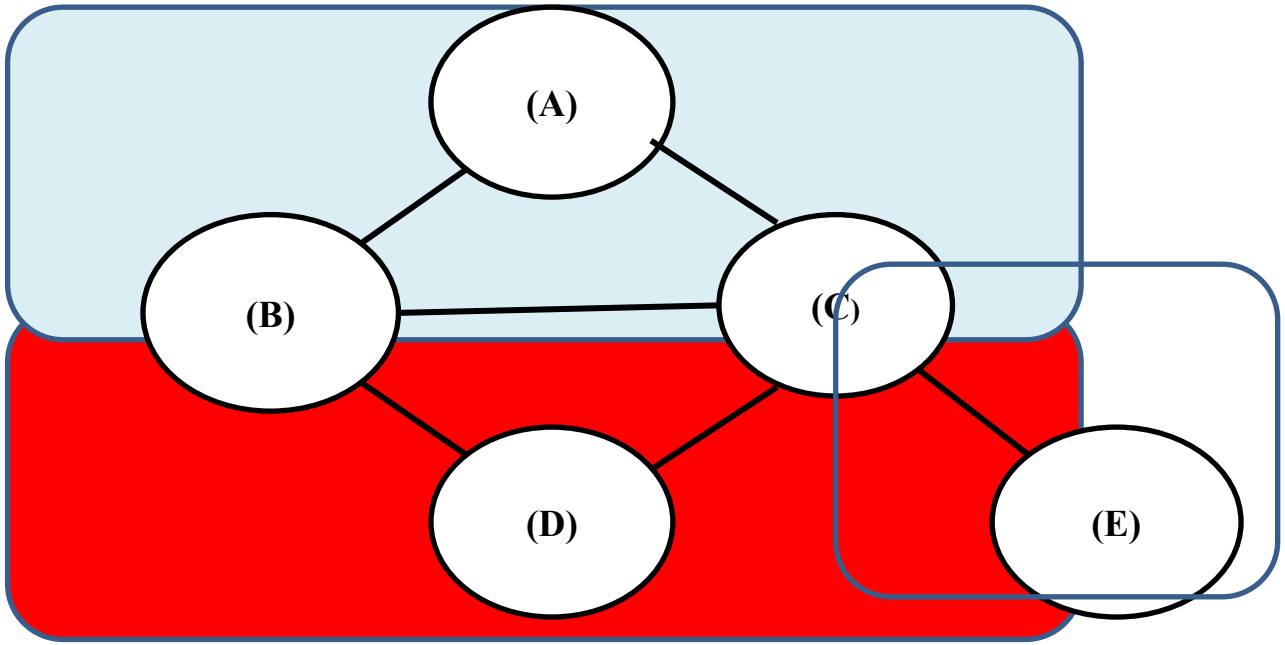
クリーク

- 頂点の部分集合が完全グラフ(すべての頂点間に辺がある)である場合



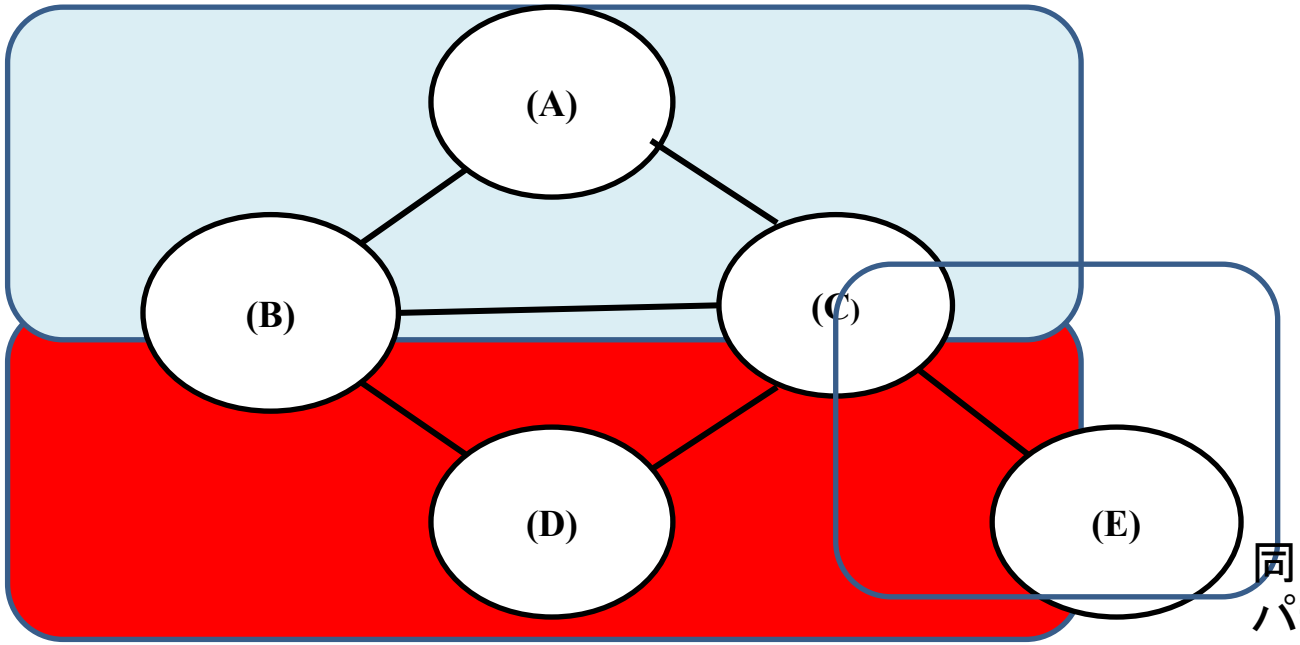
クリーク

- 頂点の部分集合が完全グラフ(すべての頂点間に辺がある)である場合



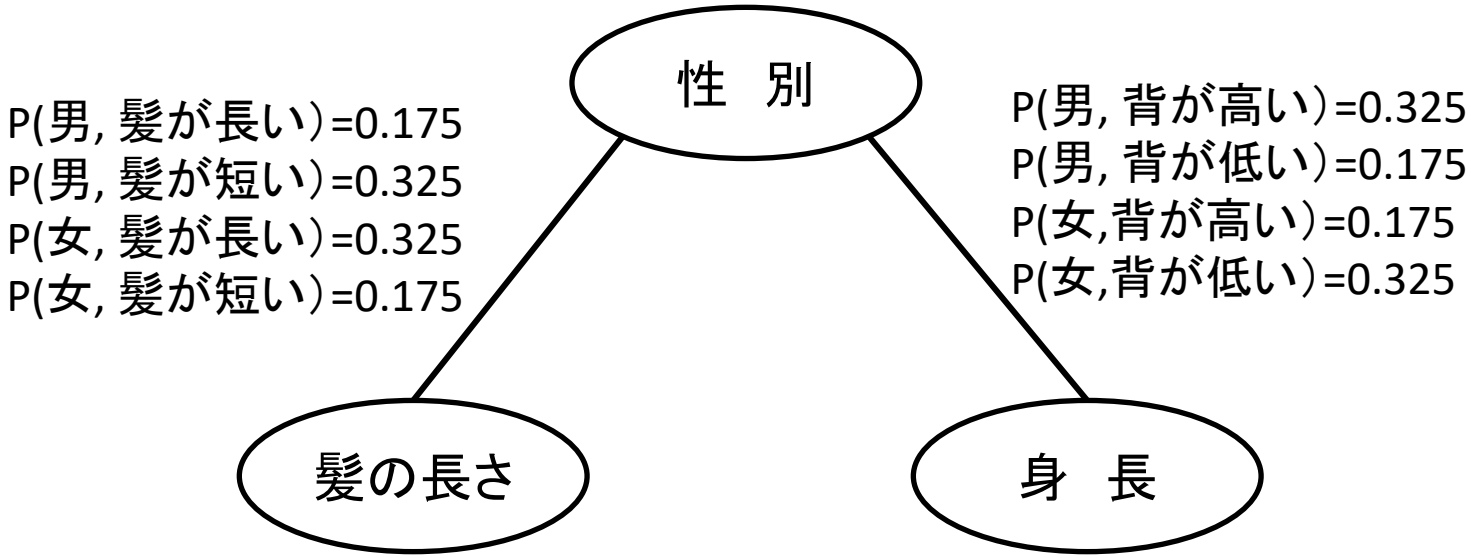
計算例　2値の場合

- $P(x_A, x_B, \cdots, x_E|G) = \frac{1}{Z(\theta)} p(x_A, x_B, x_c) p(x_B, x_c, x_D)\, p(x_c, x_E)$



同時確率分布の
パラメータ数
31
⇛
7+7+3=17

マルコフ・ネットワーク



同時確率パラメータは
2^3-1=7個

パラメータは6個

マルコフネットワークの問題

- パラメータ数はクリークの大きさに対して指数的に増え計算量が増えてしまうのでベイジアンネットワークより計算量が大きい。

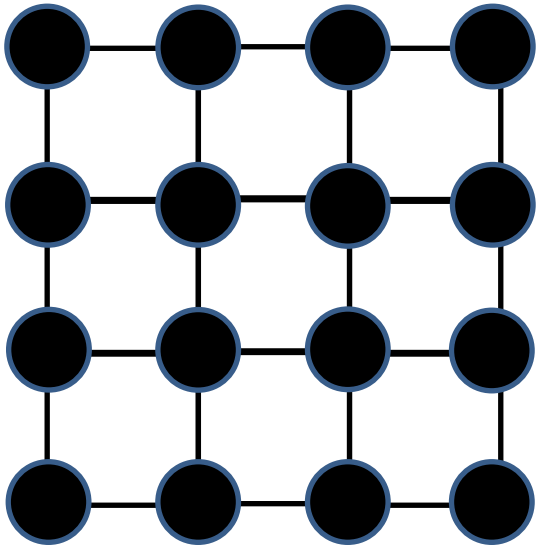
- 数学的厳密性を緩和して計算が簡易なモデルの必要性

マルコフ確率場

$$P(x_1, x_2, \cdots, x_N|G) = \frac{1}{Z(\theta)} \prod_i \phi\ (x_i)$$

$$\prod_{(i,j)} \phi\ (x_i, x_j)$$

- クリークをまともに計算せず、
グラフの辺ごとに分離して
同時確率分布を近似する。
- 画像処理などで用いられる。



マルコフネットワークに戻ろう
Log-Linear モデル

マルコフネットワークのファクターを

$$\phi_c(x_c|\theta_c)=\exp(-E(x_c|\theta_c))$$

と定義する。

ここで、$E(x_c|\theta_c) = -\log(\phi_c(x_c|\theta_c)) > 0$はクリークcのエネルギー関数と呼ばれる。

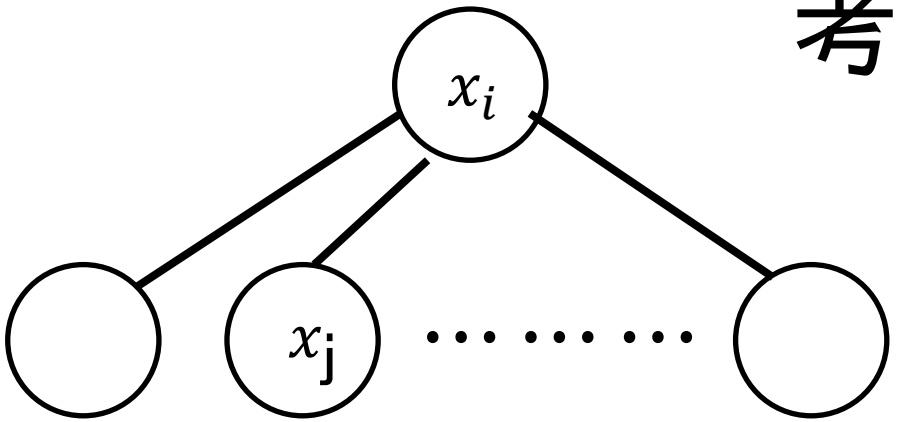
すなわち

- $P(x_1, x_2, \cdots, x_N|G) = \frac{1}{Z(\theta)}\exp(-\sum_c E(x_c|\theta_c))$

マルコフ確率場のLog-Linear モデル表現

$$P(x_1, x_2, \cdots, x_N|G)$$
$$= \frac{1}{Z(\theta)} \exp(-\sum_i E\ (X_i) - \sum_{(i,j)} E(X_i, X_j))$$

x_iは二値しかとらずに以下の構造を考える



$$P(x_i = 1|x_1, x_2, \cdots, x_N, G) = \frac{1}{Z(\theta)} \exp\left(-\sum_i E\ (x_i = 1) - \sum_{(i,j)} E(x_i = 1, x_j)\right)$$

$$= \frac{\exp\left(-\sum_i E\ (x_i = 1) - \sum_{(i,j)} E(x_i = 1, x_j)\right)}{\exp\left(-\sum_i E\ (x_i = 1) - \sum_{(i,j)} E(x_i = 1, x_j)\right) + \exp\left(-\sum_i E\ (x_i = 0) - \sum_{(i,j)} E(x_i = 0, x_j)\right)}$$

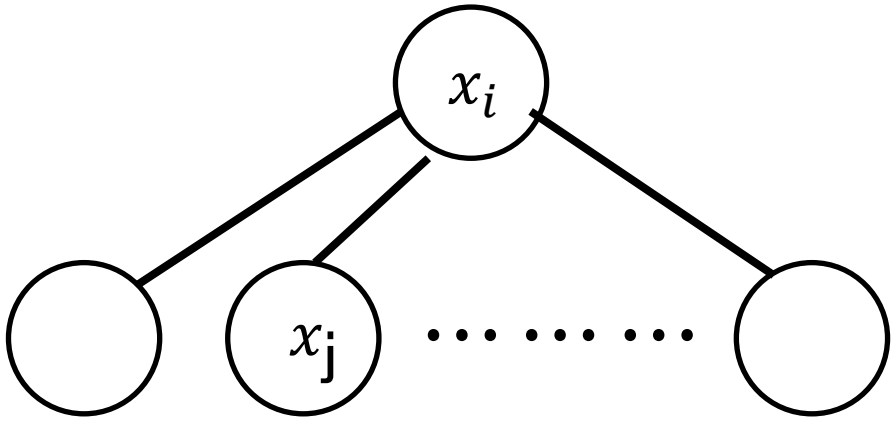
$$= \frac{1}{1 + \exp\left(\sum_i E\ (x_i = 1) - \sum_i E\ (x_i = 0) + \sum_{(i,j)} E(x_i = 1, x_j) - \sum_{(i,j)} E(x_i = 0, x_j)\right)}$$

$-\sum_i E\ (x_i = 1) + \sum_i E\ (x_i = 0) = b_i$,$\sum_{(i,j)} E(x_i = 1, x_j) - \sum_{(i,j)} E(x_i = 0, x_j) = w_{ij} x_j$とおくと

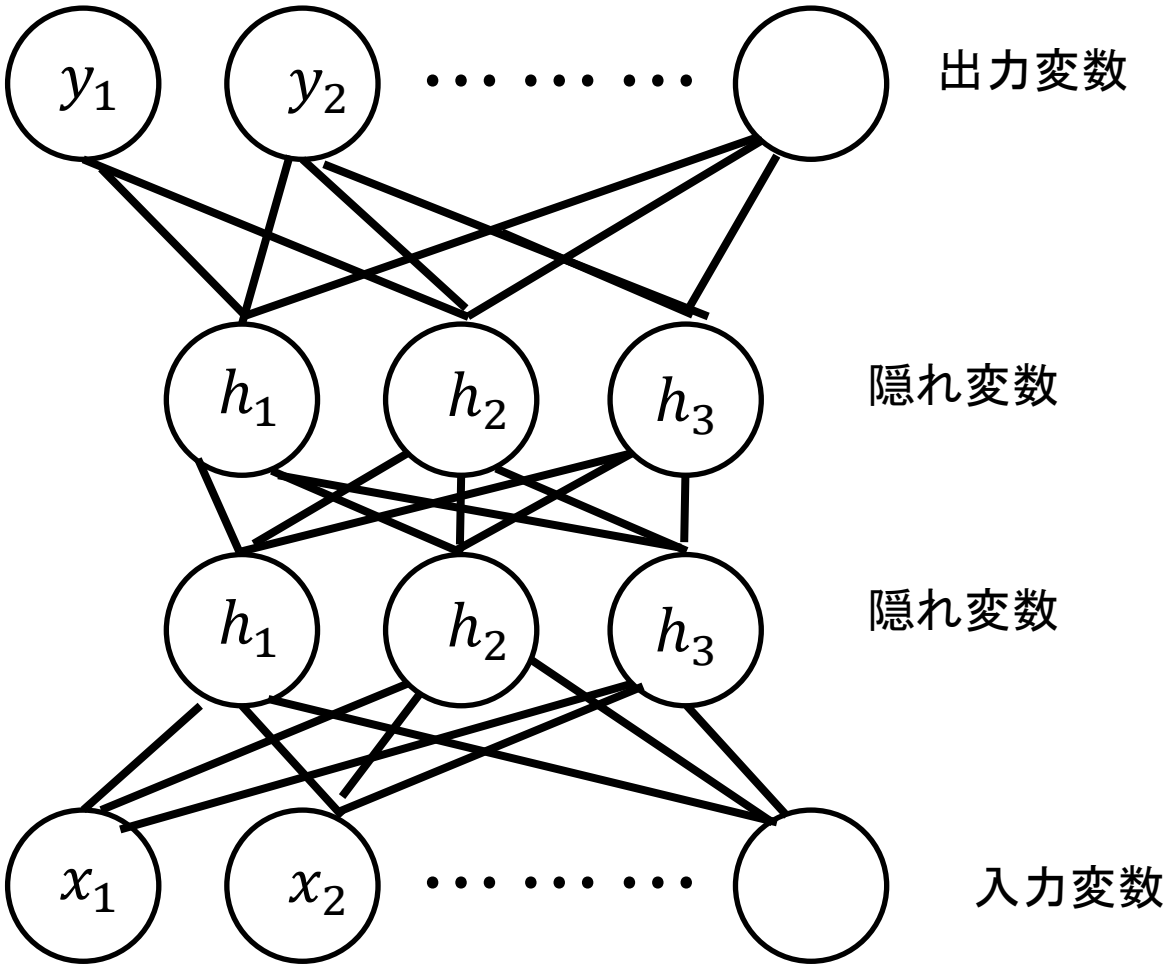
ニューラルネットワークのユニット

$-\sum_i E\ (x_i = 1) + \sum_i E\ (x_i = 0) = b_i,$
$\sum_{(i,j)} E(x_i = 1, x_j) - \sum_{(i,j)} E(x_i = 0, x_j) = w_{ij} x_j$
とおくと

$$P(x_i = 1 | x_1, x_2, \cdots, x_N, G) = \frac{1}{1 + \exp\left(-\sum_j w_{ij} x_j - b_i\right)}$$



深層学習モデル
（ディープラーニング）



隠れ変数の役割

- 隠れ変数を積分消去すると全変数間に辺が引かれた完全グラフ構造となる。
- 計算不可能な複雑な構造を　隠れ変数を導入することにより、単純で計算可能な階層構造に変換している。
- 隠れ変数はモデル平均のような役割を行う。

その後

- 深層学習モデルは　大爆発し、現在のAI時代を導く！！

当時の私の仮説

ベイジアンネットワークは一致性を持って同時確率分布を推定できるので　近似である深層学習より　少ない変数数であれば　予測精度が高い。

しかし現実は

- 10程度の少数変数しか持たない場合でもBayesian networkは　深層学習に予測精度で勝てない！！

- Bayesian networkは　すべての変数の同時確率分布を求めようとしている。
- 深層学習は　特徴量を所与として各変数を一つずつ予測している。

11. 主なアイデア

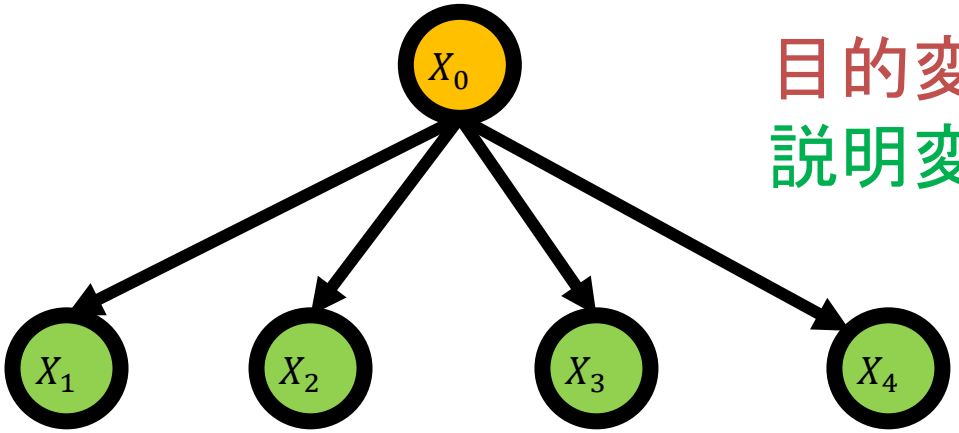
- Bayesian networkは　すべての変数の同時確率分布を求めようとしている。深層学習は　特徴量を所与として各変数を一つずつ予測している。
- Bayesian networkも所与とする特徴量の変数は予測する必要がないので、特徴量となる変数を条件として、一つの目的変数の予測を最適化する手法を提案できないか？

- Shouta Sugahara, Koya Kato, James Cussens, Maomi Ueno: Learning Bayesian Network Classifiers to Minimize Class Variable Parameters, Journal of Machine Learning Research, 27(21):1−41, (2026)
- Shouta Sugahara, Koya Kato and Maomi Ueno: Learning Bayesian Network Classifiers to Minimize the Class Variable Parameters. Proceedings of the AAAI Conference on Artificial Intelligence, 38(18), 20540-20549. (2024)

実は古くからBayesian networkの分類精度が悪いことへの批判は存在した！！

周辺尤度で学習したベイジアンネットワークの分類精度が，
単純な構造をとるNaive Bayesより劣ることが
多々ある．

Naïve Bayesの例



目的変数：X_0

説明変数：$X_1, \cdots, X_4$

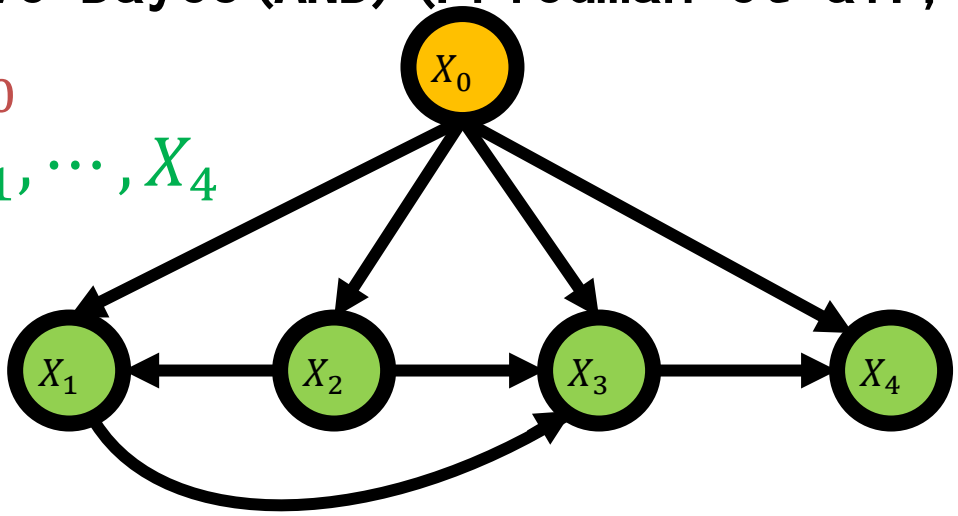
N. Friedman, D. Geiger, and M. Goldszmidt, “Bayesian Network Classifiers,” Machine Learning, vol.29, no.2, pp.131–163, 1997.

Augmented Naïve Bayes Classifiers (ANB)

目的変数から全説明変数にエッジが引かれている構造を，
Augmented Naïve Bayes(ANB)(Friedman et al.,1997)と呼ぶ.

目的変数：X_0

説明変数：$X_1, \cdots, X_4$



目的変数の親変数が0で，全ての説明変数を子変数とするため，先述した分類精度低下の問題が生じない.

N. Friedman, D. Geiger, and M. Goldszmidt, “Bayesian Network Classifiers,” Machine Learning, vol.29, no.2, pp.131–163, 1997.

ANBの厳密解(Sugahara and Ueno 2021)

分類予測精度では　深層学習には負けてしまった。

ANBの厳密解(Sugahara and Ueno 2021)

- ANB構造を制約としてパラメータ数最小のI-map ANBが得られることは保証する
- データがBayesian networkの下位モデルに従っていれば、目的変数の予測確率は漸近一致性を持つという定理を証明。(菅原氏の修士論文)

- Shouta Sugahara, Maomi Ueno: Exact Learning Augmented Naive Bayes Classifier. Entropy 2021, 23, 1703.https://doi.org/10.3390/e23121703,2021

12.BNOC: Bayesian Network Optimal Classifier

- 目的変数の分類確率の推定に関係する全てのパラメータ数（Number of Class variable Parameters: NCP）を最小化するBayesian network

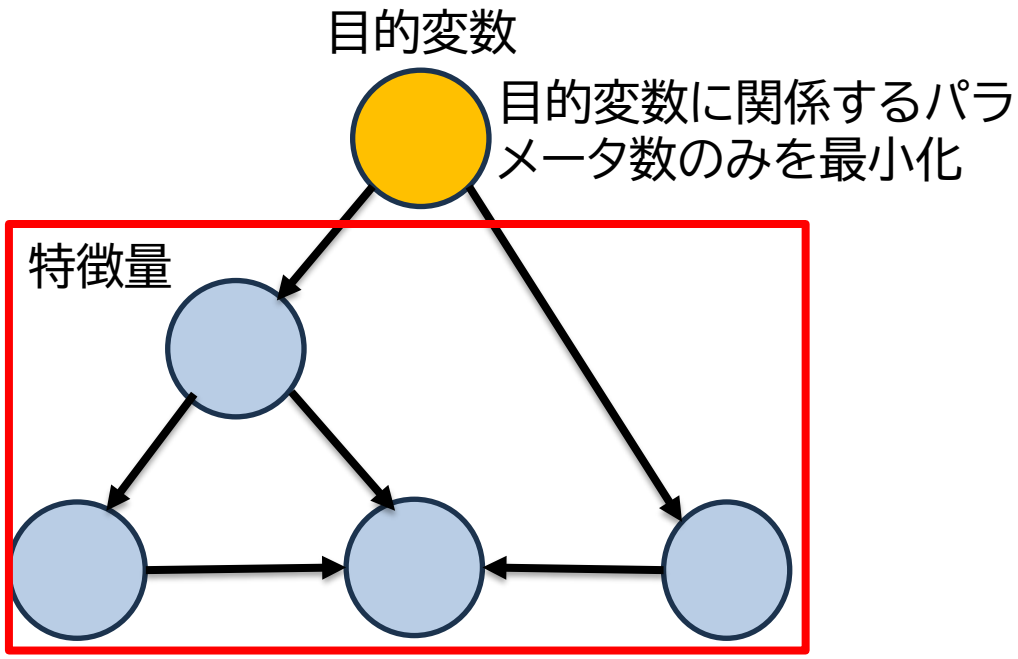
$$NCP(G) = \sum_{i=0}^{n} NCP_i(\mathbf{Pa}_{X_i}^G),$$

$$NCP_i(\mathbf{Pa}_{X_i}^G) = \begin{cases} (r_i - 1)q_i & (i = 0 \ \vee X_0 \in \mathbf{Pa}_{X_i}^G) \\ 0 & (\text{その他の場合}) \end{cases}$$

Shouta Sugahara, Koya Kato, James Cussens, Maomi Ueno: Learning Bayesian Network Classifiers to Minimize Class Variable Parameters, Journal of Machine Learning Research, 27(21):1−41, (2026)

Shouta Sugahara, Koya Kato and Maomi Ueno: Learning Bayesian Network Classifiers to Minimize the Class Variable Parameters. Proceedings of the AAAI Conference on Artificial Intelligence, 38(18), 20540-20549. (2024)

BNOC



定理１

各変数順序を所与として周辺尤度を最大にする各構造の中で NCPを最小とする構造が、すべての構造で目的変数確率の推定値の一致性を満たしてNCPを最小にできる構造である。

→

BNOCは　Bayesian networkではなくなってしまった。

定理2

BNOCは　真のモデルに関わらず真の分類確率に漸近的に一致する.

予測精度比較

No.	Dataset	Variables	Sample size	SPP	*RF*	*DL*	*BNOC-BFS*	*PCMCE*
13	Heart	14	270	1.22×10^{-3}	0.8296	0.8370	0.8074	0.8037
14	HTRU2	9	17898	1.56×10^{-3}	0.9797	0.9797	0.9783	0.9788
15	Congressional Voting Records	17	232	1.77×10^{-3}	0.9567	0.9567	0.9655	0.9569
16	Solar Flare	11	1389	3.72×10^{-3}	0.8294	0.8409	0.8431	0.8431
17	Glass	10	214	6.63×10^{-3}	0.6310	0.6602	0.6262	0.6946
18	Contraceptive Method Choice	10	1473	2.66×10^{-2}	0.4772	0.4983	0.4623	0.4912
19	Hayes-Roth	5	132	2.29×10^{-1}	0.8088	0.7571	0.8333	0.8258
20	Balance Scale	5	625	3.33×10^{-1}	0.8287	0.9808	0.9152	0.9904
21	Lenses	5	24	3.33×10^{-1}	0.8167	0.7667	0.8750	0.8750
22	Iris	5	150	6.17×10^{-1}	0.9333	0.9467	0.9467	0.9333
23	LED7	8	3200	2.50×10^{0}	0.7269	0.7344	0.7316	0.7334
24	Banknote authentication	5	1372	2.72×10^{0}	0.9432	0.9432	0.9410	0.9406
Average accuracy for datasets with large SPP (Nos. 13-24)					0.8134	0.8251	0.8271	0.8389

提案手法

3つの構造によるシミュレーション

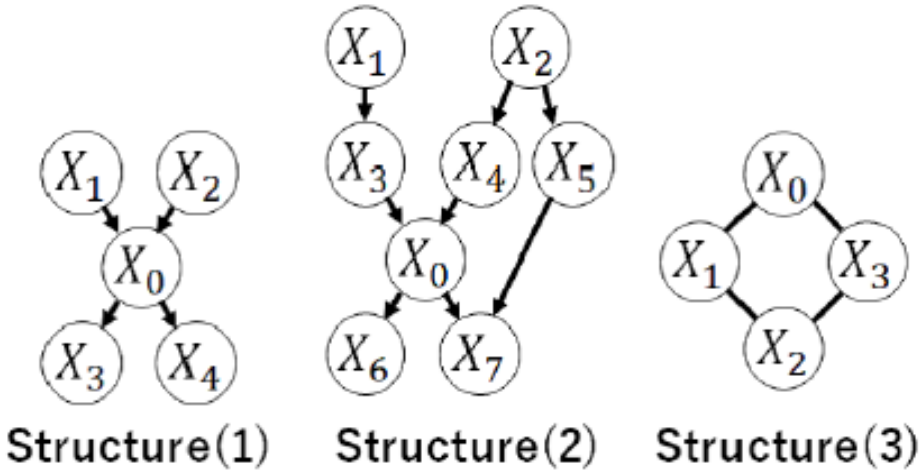


Figure 3: Structures: (1) the Cancer network, (2) the Asia network, and (3) a Markov network with a cycle.

真の目的変数の確率と推定値とのカルバックライブラー距離

Network	Sample size	*RF*	*DL*	*BNOC-BFS*	*PCMCE*
Cancer (Structure (1))	100	8.49×10^{-2}	5.29×10^{-2}	2.70×10^{-2}	2.88×10^{-2}
	1,000	5.33×10^{-2}	4.85×10^{-2}	2.67×10^{-2}	2.63×10^{-2}
	10,000	1.88×10^{-3}	8.38×10^{-3}	1.27×10^{-3}	1.45×10^{-3}
	100,000	2.73×10^{-4}	5.11×10^{-3}	8.46×10^{-5}	3.80×10^{-5}
Asia (Structure (2))	100	1.23×10^{-1}	1.51×10^{-1}	4.40×10^{-2}	4.37×10^{-2}
	1,000	1.82×10^{-1}	2.31×10^{-1}	6.38×10^{-2}	6.35×10^{-2}
	10,000	5.62×10^{-2}	3.11×10^{-2}	2.46×10^{-2}	2.40×10^{-2}
	100,000	3.26×10^{-2}	1.92×10^{-2}	2.72×10^{-4}	5.88×10^{-4}
Markov net (Structure (3))	100	2.73×10^{-1}	1.16×10^{-1}	8.18×10^{-2}	8.16×10^{-2}
	1,000	1.38×10^{-1}	7.86×10^{-2}	6.63×10^{-2}	6.65×10^{-2}
	10,000	1.30×10^{-1}	1.36×10^{-1}	4.43×10^{-4}	4.42×10^{-4}
	100,000	1.17×10^{-1}	1.44×10^{-1}	7.94×10^{-5}	7.94×10^{-5}

Table 8: Average KLDs between the true probability distribution of the class variable and those using *RF*, *DL*, *BNOC-BFS*, and *PCMCE*.

13. 整数計画法によるBNOCの大規模化

メリット

1. 第一ステップ手順(ii)，第二ステップの探索を一つの目的関数で同時に実施→効率的な探索を実現
2. 最大親変数数をdに制限にしたとき，空間計算量を$O\left(n\sum_{j=0}^{d}\binom{n-1}{j}\right)$に削減可能.

13.1. 提案手法の目的関数

$$\text{maximize} \quad \sum_{X,\mathbf{W}} Score_X(\mathbf{W}) I(\mathbf{W} \to X) - \gamma \times \sum_{X,\mathbf{W}} NCP_X(\mathbf{W}) I(\mathbf{W} \to X)$$

周辺尤度の和　　NCPの大きさ

γ：チューニングパラメータ
$Score_X(\mathbf{W})$：変数集合$\mathbf{W}$が変数Xの親であるときの周辺尤度
$I(\mathbf{W} \to X)$：BNCにおいて$\mathbf{W}$がXの親であるなら1, そうでなければ0

13.2. 提案手法の制約

目的変数が親変数を持たないBNCであるための制約

1. 各変数$X \in \mathrm{V}$の親変数集合Wがただ一つである

$$\forall X : \sum_{\mathbf{W}} I(\mathbf{W} \to X) = 1$$

2. 構造に循環を含まない

$$\forall C \subseteq \mathbf{V} : \sum_{X \in C} \sum_{\mathbf{W}:\mathbf{W} \cap C = \emptyset} I(\mathbf{W} \to X) \geq 1$$

3. 目的変数が親変数を持たない

$$I(\emptyset \to X_0) = 1$$

13.3. 全体の定式化

$$\text{maximize} \quad \sum_{X,\mathbf{W}} Score_X(\mathbf{W}) I(\mathbf{W} \to X) - \gamma \times \sum_{X,\mathbf{W}} NCP_X(\mathbf{W}) I(\mathbf{W} \to X)$$

$$\text{subject to} \quad \forall X : \sum_{\mathbf{W}} I(\mathbf{W} \to X) = 1$$

$$\forall C \subseteq \mathbf{V} : \sum_{X \in C} \sum_{\mathbf{W}:\mathbf{W} \cap C = \emptyset} I(\mathbf{W} \to X) \geq 1$$

$$I(\emptyset \to X_0) = 1$$

$$\forall X, \mathbf{W} : I(\mathbf{W} \to X) \in \{0, 1\}$$

13.4. 提案手法の時間計算量

$$O\left(2^{n\sum_{j=0}^{d}\binom{n-1}{j}}\right)$$

(例)最大親変数数を3に制限したときの時間計算量

AAAI 2024 → $O(n2^n)$

提案手法→ $O\left(2^{n^4}\right)$

時間計算量は増

13.5. 提案手法の空間計算量

$$O\left(n\sum_{j=0}^{d}\binom{n-1}{j}\right)$$

(例)最大親変数数を3に制限したときの空間計算量

AAAI2024 の手法 $\rightarrow O(2^n)$

提案手法$\rightarrow O(n^4)$

空間計算量は減

13.6. 小規模ネットワークの評価実験

- 比較手法:
  - Naive Bayes
  - gGBN : 周辺尤度を用いて近似学習したGBN
  - GBN : 周辺尤度を用いて厳密学習したGBN
  - ANB : ANB構造を制約とした学習手法
  - 幅優先探索によるNCP最小化(以下 AAAI2024 幅優先と表記)
  - 深さ優先分枝限定法によるNCP最小化(以下 AAAI2024 深さ優先と表記)
  - 提案手法(ハイパーパラメータは0.05、IBN CPLEX)
- 実データ:
  UCIレポジトリデータベースに登録されているベンチマークデータセット
- 実験手順:
  各手法，各データセットに対して，10分割交差検証によるテストデータの平均一致率を求めて分類精度とした.

13.7.小規模ネットワークでの実験結果

No.	Dataset	Variables	Sample size	SPP	Naive-Bayes	gGBN	GBN	ANB	AAAI2024 幅優先	AAAI2024 深さ優先	提案手法 $\gamma = 0.05$
1	Lymphography	19	148	1.63×10^{-7}	0.8378	0.7905	0.7500	0.8108	0.7635	0.7838	0.7905
2	Breast Cancer Wisconsin	10	683	3.42×10^{-7}	0.9751	0.7094	0.9751	0.9751	0.9737	0.9737	0.9737
3	Hepatitis	20	80	7.63×10^{-5}	0.8500	0.8500	0.6125	0.5750	0.8250	0.8000	0.8250
4	Zoo	17	101	1.03×10^{-4}	0.9802	0.9505	0.9228	0.9406	0.9505	0.9208	0.9703
5	Australian	15	690	2.97×10^{-4}	0.8290	0.8420	0.8507	0.8203	0.8493	0.8449	0.8580
6	Vehicle	19	846	8.07×10^{-4}	0.4314	0.5461	0.5898	0.6217	0.6050	0.5843	0.5946
7	Breast Cancer	10	277	8.33×10^{-4}	0.7364	0.7058	0.7256	0.6968	0.7076	0.7365	0.7184
8	Image Segmentation	19	2310	1.26×10^{-3}	0.7290	0.8026	0.8255	0.8273	0.8264	0.8320	0.8264
9	Congressional Voting Records	17	232	1.77×10^{-3}	0.9095	0.9741	0.9655	0.9483	0.9698	0.9655	0.9612
10	Heart	14	270	2.44×10^{-3}	0.8259	0.8222	0.8370	0.8037	0.8222	0.8333	0.8222
11	Solar Flare	11	1389	3.72×10^{-3}	0.7804	0.8431	0.8431	0.8215	0.8431	0.8409	0.8398
12	Wine	14	178	7.24×10^{-3}	0.9270	0.9045	0.9270	0.9270	0.9494	0.9494	0.9494
13	Letter	17	20000	1.17×10^{-2}	0.4466	0.5761	0.6434	0.6434	0.6290	0.6303	0.6237
14	Pendigits	17	10992	1.68×10^{-2}	0.8032	0.9253	0.9342	0.9332	0.9368	0.9373	0.9314
15	Contraceptive Method Choice	10	1473	5.99×10^{-2}	0.4671	0.4440	0.4792	0.4481	0.4616	0.4396	0.4742
16	Glass	10	214	6.97×10^{-2}	0.5514	0.4626	0.5888	0.6355	0.5794	0.6036	0.6122
17	Hayes-Roth	5	132	2.29×10^{-1}	0.8333	0.7525	0.6212	0.7879	0.8333	0.8333	0.8333
18	Balance Scale	5	625	3.33×10^{-1}	0.9152	0.9152	0.9152	0.9152	0.9152	0.9152	0.9152
19	Lenses	5	24	3.33×10^{-1}	0.7083	0.8333	0.8333	0.7500	0.8750	0.8750	0.8750
20	EEG	15	14980	4.57×10^{-1}	0.5778	0.6732	0.7246	0.7212	0.7155	0.7135	0.7115
21	LED7	8	3200	2.50×10^{0}	0.7294	0.7297	0.7303	0.7303	0.7316	0.7325	0.7275
22	Iris	5	150	3.13×10^{0}	0.7133	0.8133	0.8267	0.8156	0.8200	0.8200	0.8133
23	HTRU2	9	17898	3.50×10^{1}	0.8966	0.9092	0.9141	0.9141	0.9140	0.9140	0.9141
24	Banknote authentication	5	1372	4.29×10^{1}	0.8433	0.8819	0.8812	0.8812	0.8819	0.8819	0.8812
	average				0.7624	0.7774	0.7882	0.7893	0.8070	0.8067	0.8101

13.8.大規模ネットワークでの評価実験

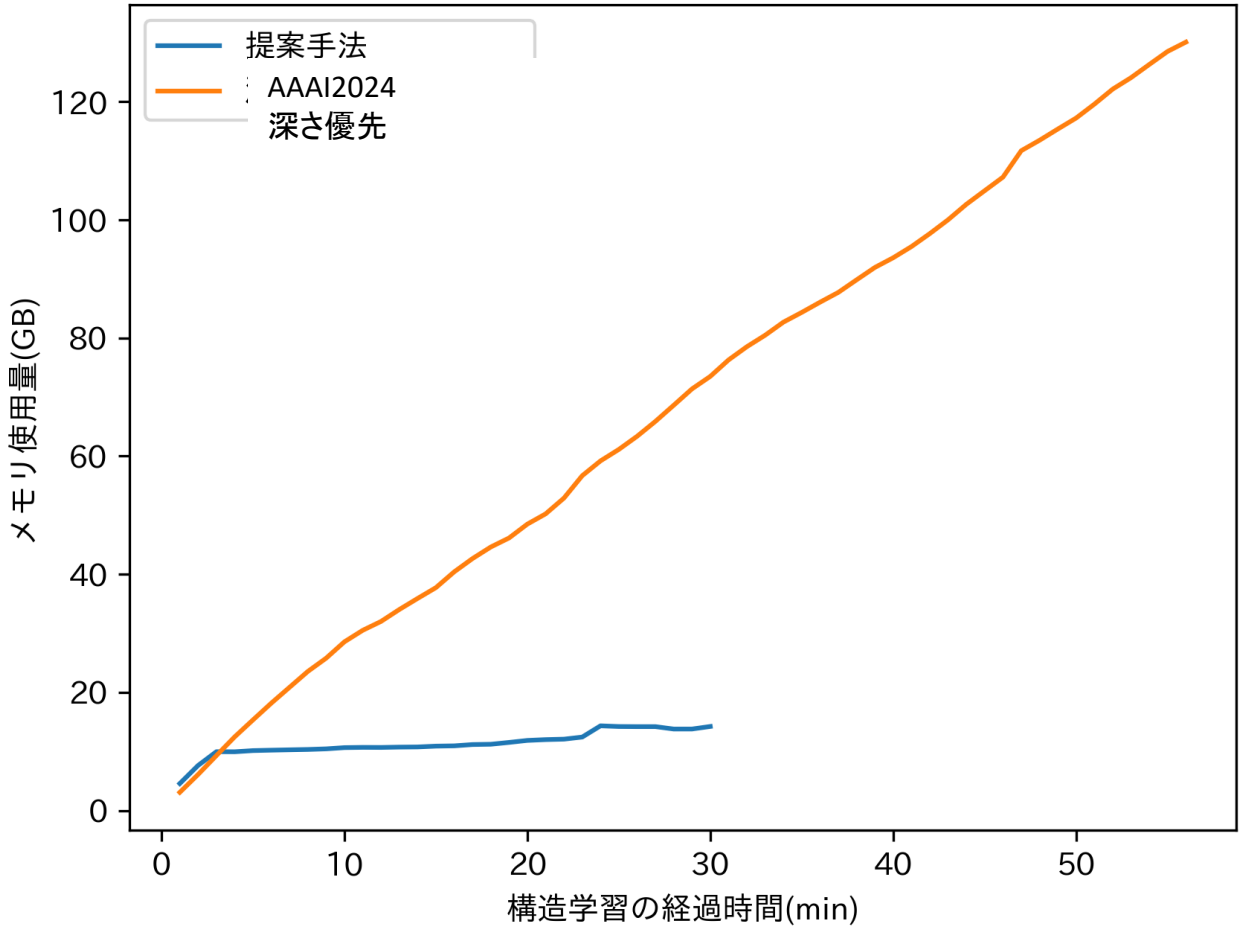
- 比較手法:
  - Naive Bayes
  - 深さ優先分枝限定法によるNCP最小化(以下AAAI2024 深さ優先と表記)
  - 提案手法(ハイパーパラメータは0.05, IBN CPLEX)
- 実データ:
  UCIレポジトリデータベースに登録されているベンチマークデータセット
- 実験手順:
  - 各手法，各データセットに対して，10分割交差検証によるテストデータの平均一致率を求めて分類精度とした.
  - AAAI2024の手法と提案手法は構造学習中のメモリ使用量を測定した.
  - 最大親変数数を3にした.

13.9.大規模ネットワークでの実験結果

No.	Dataset	Variables	Sample size	Naive-Bayes	AAAI2024 深さ優先	提案手法 $\gamma = 0.05$
25	Phishing	31	11055	0.9276	0.9337*	0.9433
26	Postures	31	23906	0.8290	0.8727*	0.8760
27	connect-4	43	67557	0.7058	0.6751*	0.7138
28	PAMAP2	53	174915	0.6862	0.8266*	0.8430
29	spam	58	4601	0.8794	0.9063*	0.9107
	average			0.8056	0.8429	0.8574

※ "*"は, メモリオーバによって打ち切りが発生し, それまでに得た最もNCPの小さい構造を用いたときの分類精度に付与されている.

13.10.大規模ネットワークでのメモリ使用量(58変数)



13.11大規模ネットワークでのメモリ使用量 (58変数)

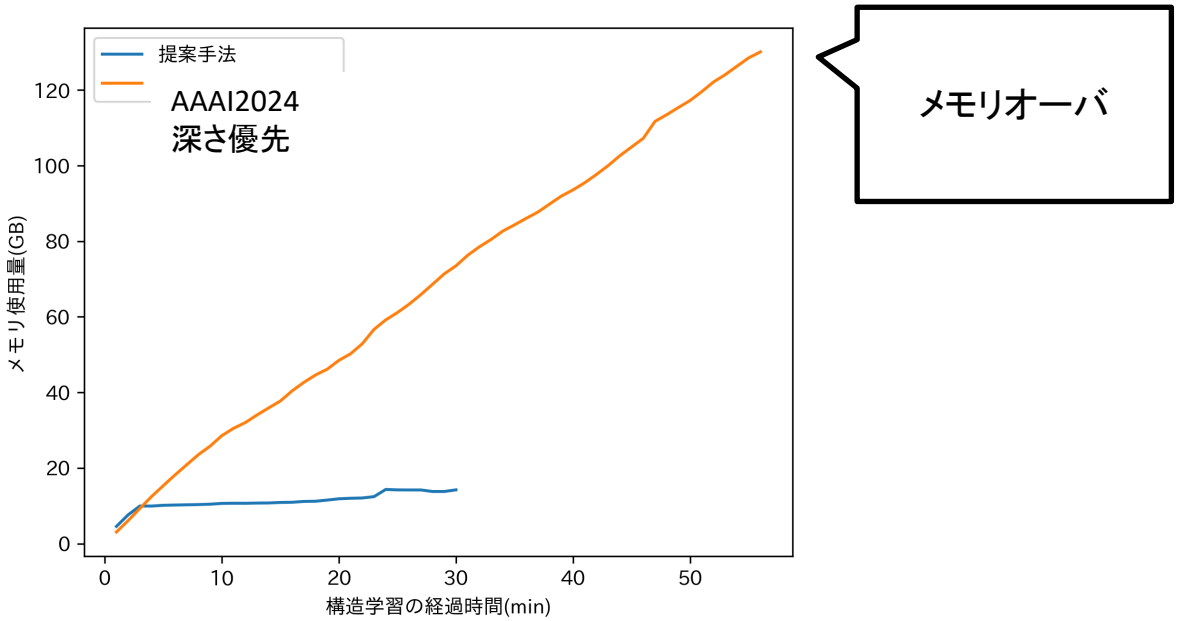


図4: 5番の構造学習中のメモリ使用量

提案手法はAAAI2024の手法と比較して構造学習中のメモリ使用量が少ない.

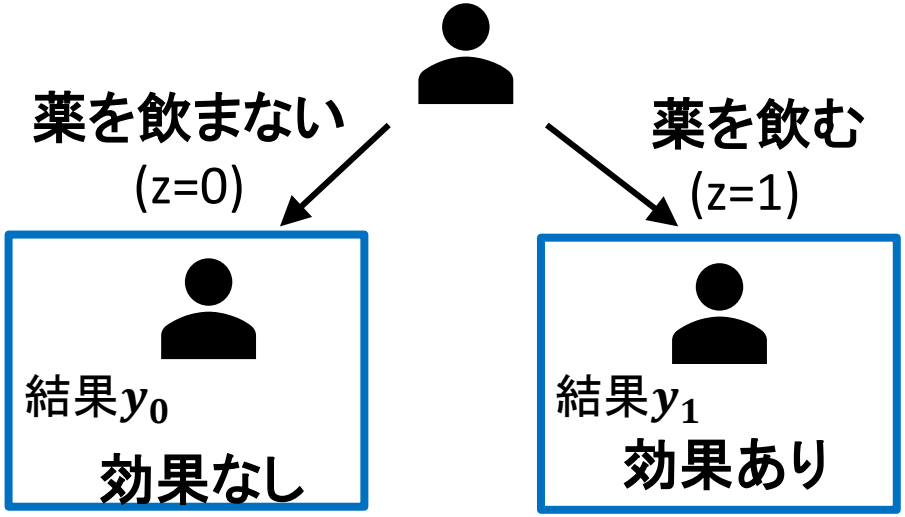
- たくさんのBDOCを組み合わせれば深層学習より精度と解釈性が高い機械学習が可能かもしれない

14. Bayesian network optimal classifierを用いた傾向スコア推定による因果推論

Rubin因果モデル

反事実の仮定をおいたRubin因果モデル[1]は, 介入の有無による結果を比較して因果推論を行う統計的手法

例) 頭痛薬の効果があるか確かめたい
介入z:頭痛薬を投与する
y_1:薬を飲んだ時の効き目
y_0 :飲まなかったときの効き目



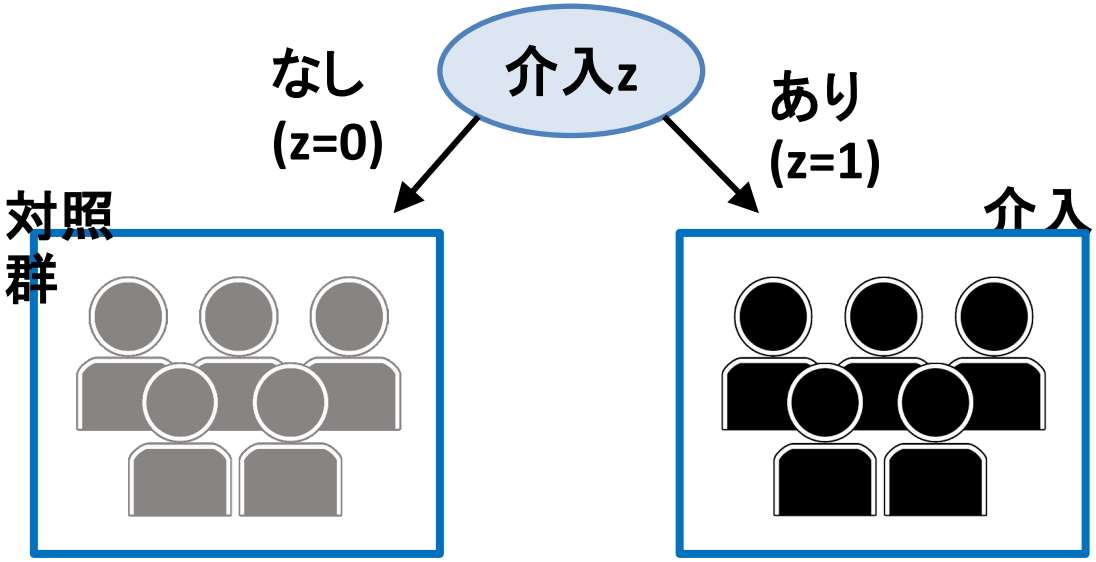
因果効果$= y_1 - y_0$

しかし, 同一の対象では介入の有無を同時に観測できない

[1] Rubin, D. B. 1974. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology* 66(5):688–701.

無作為化比較試験

対象をランダムに介入群と対照群に分けて介入の効果を評価する方法



無作為化比較試験では, $(y_1, y_0) \perp z$が成り立つため,

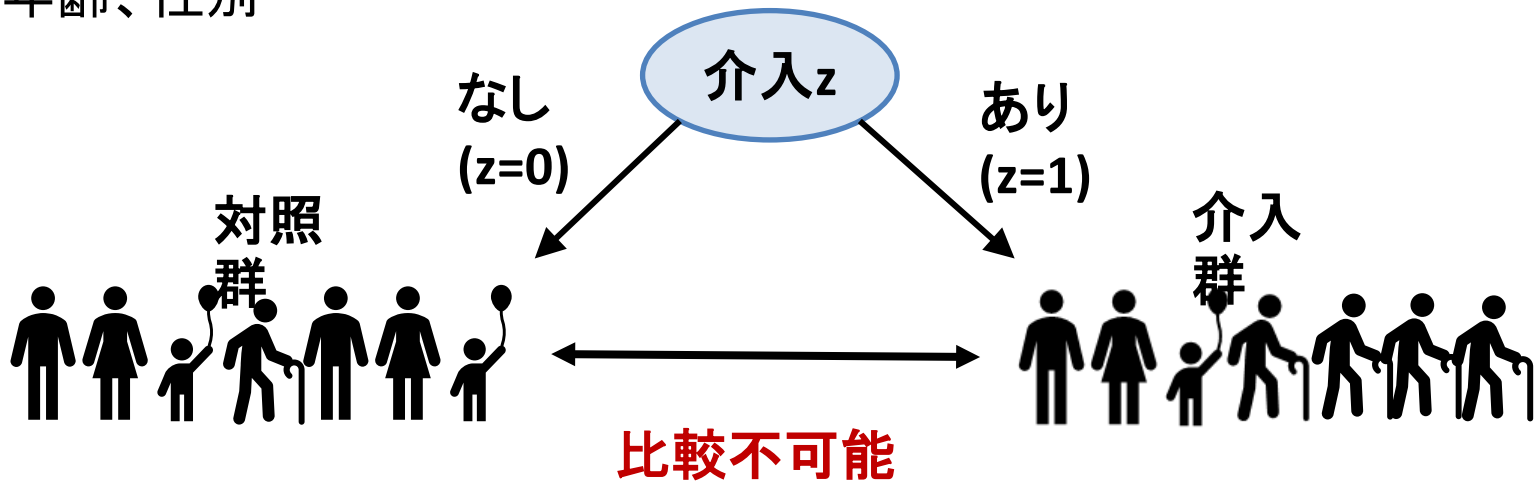
$$E(y_j) = E(y_j \mid z = j), (j = 1, 0)$$

となるので, ***因果効果または平均介入効果*(Average Treatment Effect : ATE)は**

$$\boldsymbol{ATE = E(y_1 - y_0) = E(y_1|z = 1) - E(y_0|z = 0)}$$

無作為化比較試験ができない場合

共変量x：対象の特徴
例）年齢、性別



共変量xに偏りがあり，平均介入効果を推定できない

$$ATE \neq E(y_1 \mid z = 1) - E(y_0 \mid z = 0)$$

傾向スコア

Rosenbaum & Rubin [2] は傾向スコアを用いた共変量調整法を提案した.

傾向スコア$e(x)$: 共変量xを所与としたときの対象が介入を受ける確率

$$e(x) = p(z = 1|y_1, y_0, x) = p(z = 1|x)$$

このとき $E(y_j \mid x) = E(y_j \mid e(x)), (j = 1,0)$

$$(y_1, y_0) \perp z \mid e(x), \quad 0 < p(z = 1 \mid e(x)) < 1$$

この条件のもと, $E(y_j \mid e(x)) = E(y_j \mid e(x), z = j), (j = 1,0)$

となり, $\mathrm{ATE} = E_x(E(y_1 \mid e(x), z = 1) - E(y_0 \mid e(x), z = 0))$

が成り立つため, 逆確率重み付け法(Inverse probability weighting : IPW)でATEを求められる[3].

$$\mathrm{ATE} = \frac{1}{N}\sum_i^N \frac{y_{1i}}{e(x_i)} z_i - \frac{1}{N}\sum_i^N \frac{y_{0i}}{1 - e(x_i)}(1 - z_i)$$

Nは対象全体の数, y_{1i}は対象iが介入を受けた時の結果変数, y_{0i}は対象iが介入を受けなかった時の結果変数, x_iは対象iの共変量, z_iは対象iの介入の有無

[2] Rosenbaum, P. R., and Rubin, D. B. 1983. The central role of the propensity score in observational studies for causal effects. *Biometrika* 70(1):41–55.
[3]K. Hirano, G.W. Imbens, and G. Ridder, “Efficient estimation of average treatment effects using the estimated propensity score,” Econometrica, vol.71, no.4, pp.1161–1189, 2003.

ロジスティック回帰を用いた傾向スコア推定

傾向スコアの推定にはロジスティック回帰[2]が最も用いられる

しかし

傾向スコアのlogitが共変量の線形関数で表現できない場合,
傾向スコアの推定に一致性がなくなる[3]

$$\mathrm{logit}\big(\hat{e}(x)\big) = \log\frac{\hat{e}(x)}{1-\hat{e}(x)} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + x_n\beta_n$$

ここで$\hat{e}(x)$は介入を受ける確率(傾向スコア)の推定値, β_0は定数項, $\beta_1, \beta_2, \ldots, \beta_n$ は各説明変数の回帰係数, $x_1,\ x_2, \ldots, x_n$ は説明変数である.

因果効果推定の精度は傾向スコア推定の精度に大きく依存する[4,5] ため,
この手法は適さない

[3] Sant'Anna, P. H., Song, X., and Xu, Q. Covariate distribution balance via propensity scores. *Journal of Applied Econometrics*, 37(6):1093–1120, 2022.
[4] Hainmueller, J. Entropy balancing for causal effects: A multivariate reweighting method to produce balanced samples in observational studies. *Political analysis*, 20 (1):25–46, 2012.
[5] Imai, K. and Ratkovic, M. Covariate balancing propensity score. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76(1):243–263, 2014.

機械学習を用いた傾向スコア推定

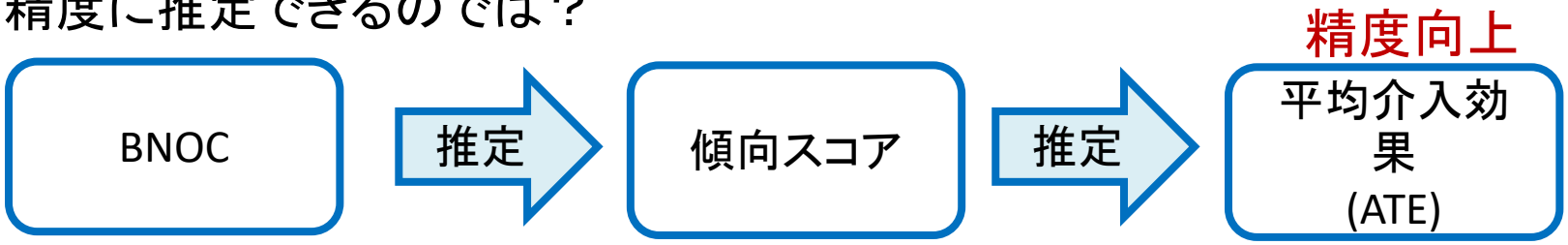
- 勾配ブースティング[6]
- ニューラルネットワーク[7]
- 決定木[7,8]

などの手法が知られているが, 漸近的に真の確率を推定する保証がない

[6] McCaffrey, D. F.; Ridgeway, G.; and Morral, A. R. 2004. Propensity score estimation with boosted regression for evaluating causal effects in observational studies. *Psychological methods*, 9(4): 403.
[7] Setoguchi, S.; Schneeweiss, S.; Brookhart, M. A.; Glynn, R. J.; and Cook, E. F. 2008. Evaluating uses of data mining techniques in propensity score estimation: a simulation study. *Pharmacoepidemiology and drug safety*, 17(6): 546– 555.
[8] Westreich, D.; Lessler, J.; and Funk, M. J. 2010. Propensity score estimation: machine learning and classification methods as alternatives to logistic regression. *Journal of clinical epidemiology*, 63(8).

ベイジアンネットワーク分類器を用いた傾向スコア推定

BNOCで傾向スコアを利用して平均介入効果(ATE)を高精度に推定できるのでは？



介入zを受けるかどうかを目的変数とし，共変量xと結果変数yを説明変数とした真の分類確率に漸近的に一致するBNOCを利用し，傾向スコアを推定する

BNOCの利点

真のモデルがベイジアンネットワークに従わない場合でも真の傾向スコアを漸近的に推定できる

解釈のしやすさ

因果関係を視覚的に表現することで結果を理解しやすくなる

自動的な変数選択

データから構造を学び、状況に応じた最適なモデルを構築可能

Jobsデータセット

概要

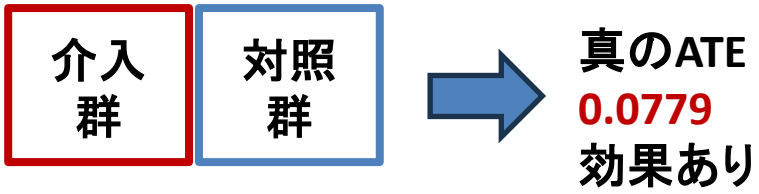
- Jobsデータセット[12]は職業訓練プログラムの効果を評価するための実データセット
- 参加者の雇用状況や収入に関する情報が含まれている

データの内容

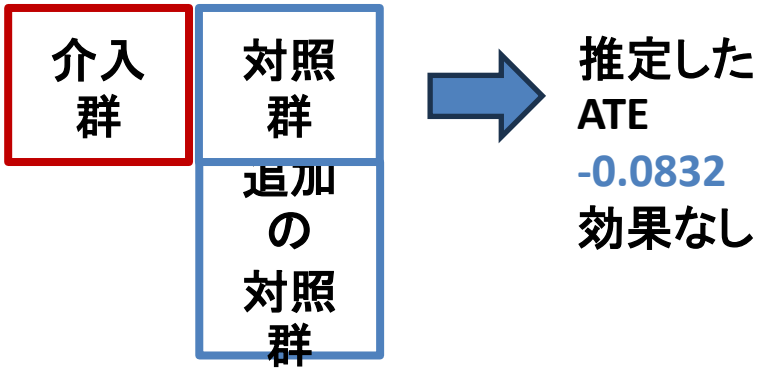
- 共変量：年齢, 最終学歴, 民族(black, hispanic), 婚姻状況, 大学の学位を持っているか, 1975年の収入の7つの変数
- 介入：職業訓練への参加の有無
- 効果：介入後に就職できたかどうか

データの特徴

無作為化比較試験



バイアスのあるデータセットを作成



[12]Rajeev H Dehejia and Sadek Wahba. Causal effects in nonexperimental studies: Reevaluating the evaluation of training programs. Journal of the American statistical Association, 94(448):1053–1062, 1999.

実験手順

1. Jobsデータを訓練データ : テストデータ = 8 : 2 で分ける
2. 各手法を用いて傾向スコアの推定を行う
   訓練データで学習し, テストデータに適用する
3. 2 で推定した傾向スコアを用いてATEの推定を行う
4. 1 ~ 3 を 5 回繰り返す

評価指標

推定したATEに対しBiasとRMSEとMAE

ATE推定の実験結果

	LR	NN	BOOST	GBN	提案手法
Bias	-0.3755	-0.5486	-0.2939	13.0158	0.0060
RMSE	0.3883	0.5530	0.3384	26.5959	0.1877
MAE	0.3755	0.5486	0.2939	13.4092	0.1515

- 提案手法の精度が最も高くなる

Biasの定義

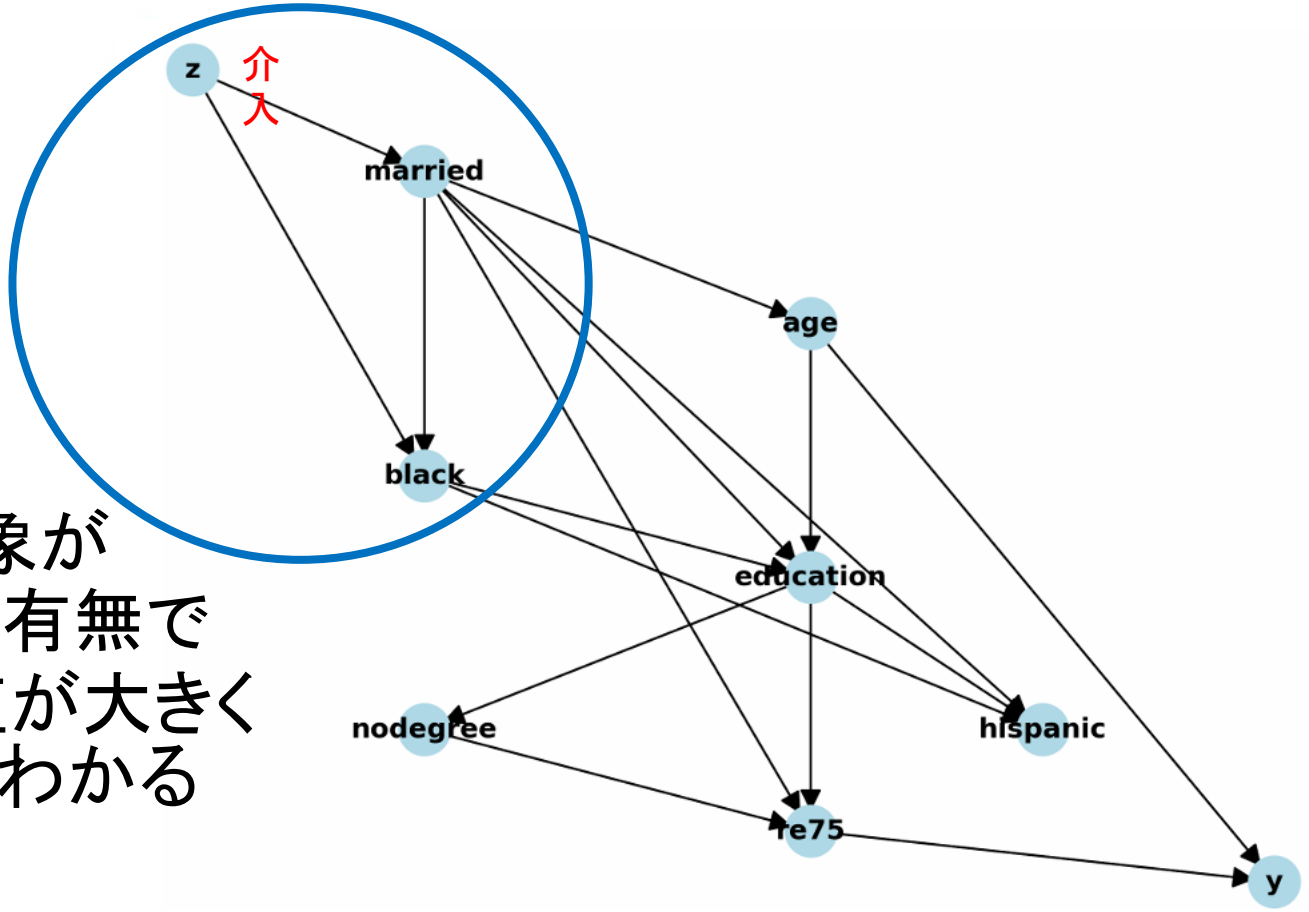
$$Bias = E[\widehat{ATE} - ATE_{true}]$$

より, 以下が成り立つ.

$$E\left[\widehat{ATE}\right] = Bias + E[ATE_{true}] = Bias + 0.0779$$

Bias の結果から, 職業訓練に正の効果があると推定できたのは 提案手法のみである

Jobsデータの確率的因果構造



提案手法は対象が
黒人か, 結婚の有無で
傾向スコアの値が大きく
変化することがわかる

Twinsデータセット

概要

1989年から1991年の米国での全出生記録から得られた実データセット[13]

乳幼児の出生体重による死亡率を評価するために使用される

データの特徴

体重が重い方 z= 1　　軽い方 z= 0

死亡率: 14.92%　　16.91%

真のATEは-0.0248

データの内容

出生体重が2kg未満の同性の双子を選択

共変量：年齢, 性別, 母親の健康状態, 妊娠, 両親の社会経済的背景などの46変数

介入：双子のうち体重が重い方

効果：1年後に死亡しているかどうか

介入の選択

共変量ベクトルをx, 共変量の個数をdとすると,

$$z|\mathbf{x} \sim Bern(Sigmoid(\mathbf{w}^{\mathrm{T}}\mathbf{x} + n))$$

ここで $\mathbf{w}^{\mathrm{T}} \sim \mathcal{U}\left((0.1, 0.1)^{d\times 1}\right), n \sim \mathcal{N}(0, 0.1)$

[13] Douglas Almond, Kenneth Y Chay, and David S Lee. The costs of low birth weight. The Quarterly Journal of Economics, 120(3):1031–1083, 2005.

ATE推定の実験結果

	LR	NN	BOOST	提案手法
Bias	0.0689	0.0498	0.0317	0.0051
RMSE	0.0784	0.0984	0.0439	0.0417
MAE	0.0689	0.0796	0.0393	0.0328

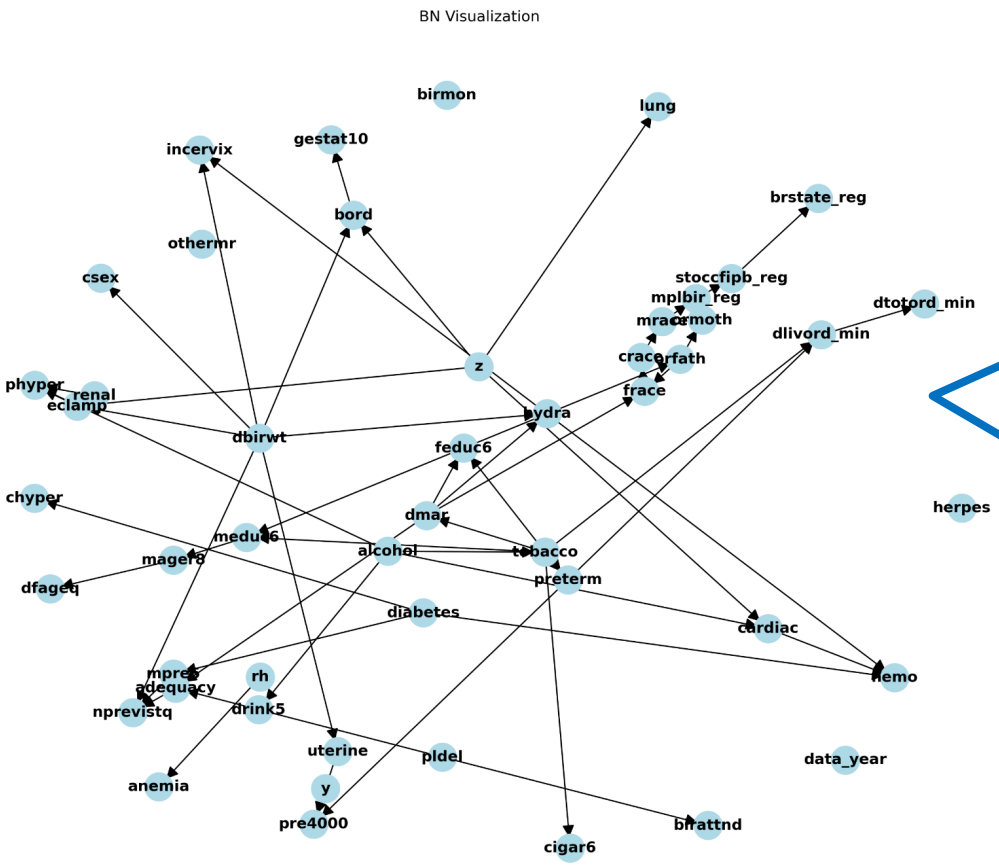
提案手法の精度が
最も高くなる

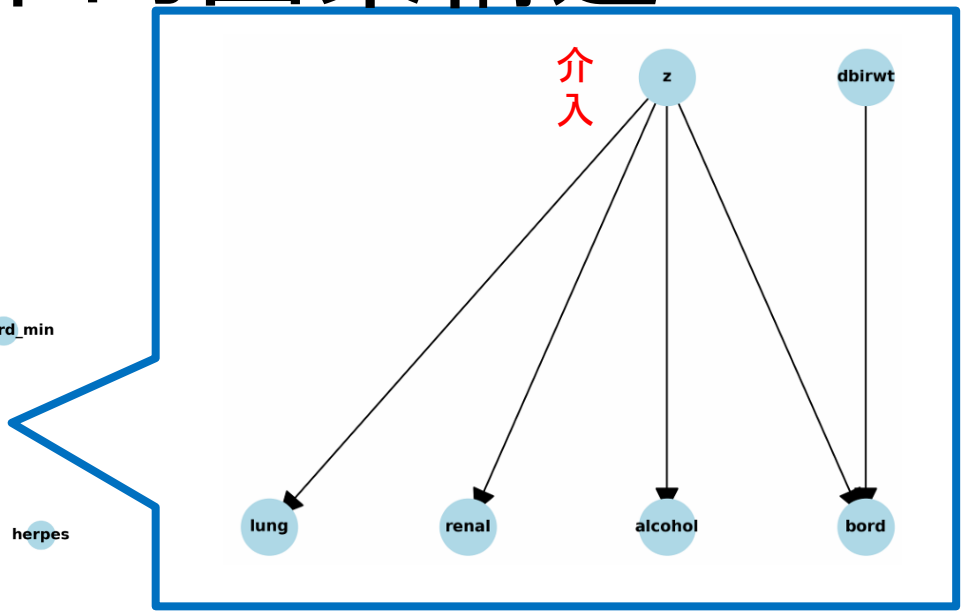
Biasの定義より, 以下が成り立つ.

$$E\left[\widehat{ATE}\right] = Bias + E[ATE_{true}] = Bias - 0.0248$$

Bias の結果から, 推定したATEが負の値をとる, つまり, 体重が重い方が死亡率が低いと推定できたのは提案手法のみである

Twinsデータの確率的因果構造





母親の急性及び慢性の肺疾患(lung),腎臓病(renal), 飲酒 (alcohol), 子供の生まれた順番(bord), が傾向スコアに影響を与えている。

BNOCの欠点と長所

欠点

- 時間計算量の大きさ
- 現状　200以上のネットワークは学習できない

長所

- 自動的に変数選択する。
- 20までの特徴量では予測精度、解釈性が高い。
- 確率予測は圧倒的に強い。確率予測が重要となる分野（意思決定問題や教育問題など）では役立つ？
- 因果推論には傾向スコアの推定、その解釈性で役に立ちそうである。データサイエンスでは深層学習などはそれほど効果的でない。