

9. ベイジアンネットワークと 他の機械学習モデルとの関係

植野真臣

電気通信大学

情報理工学研究所 情報数理プログラム

今後のスケジュール（予定）

- 4月8日 授業の概要とガイダンス
- 4月15日 ベイズの定理
- 4月22日 ベイズはどのように誕生したか？
- 5月13日 ベイズはコンピュータ、人工知能の父である！！
- 5月20日 ビリーフとベイズ
- 5月27日 6月3日 6月10日 尤度と最尤推定
- 6月17日 6月24日 7月1日 ベイズ推定と事前分布、階層ベイズ、因果推論
- 7月8日 自宅でオンデマンド授業
- 7月15日 自宅でオンデマンド授業
- 7月22日 7月29日 確率的グラフィカルモデル、ベイジアンネットワークと機械学習
- 8月5日 テストと総括

1. 本日の目標

- ベイジアンネットワークと他の機械学習モデルとの関係を理解する

2. ビック・データ時代

- 90年代—00時代 インフラストラクチャー時代
いかにデータを蓄えるか、データを蓄えさせる時代
- 有り余るデータとその有効活用が課題：大量のデータから高精度に高度な処理ができる手法→次世代の企業コンピテンシー
- 簡単にまねできない高精度な処理技法が注目される時代に突入
- 総合格闘技→数理統計、アルゴリズム、データベースの統合的技術

3. 古典的人工知能(ルール)

- 古典的AI (アリストテレス)
- 論理推論、 IF then rule
- 人は死ぬ、ソクラテスは人である、ソクラテスは死ぬ

▪

古典的AIの問題

- 当たり前のことしか推論できない
- 不確実な現象を推論できない
- 例外が多い
- 学習ができない

4. 確率推論では

- より一般化した表現
- **同時確率分布**

例：性別、髪の高さ、背の高さの同時確率分布

データより性別、髪の高さ、背の高さの同時確率分布が以下であることが分かっているとします。

$$P(\text{男、髪短い、背高い})=0.2$$

$$P(\text{男、髪長い、背高い})=0.125$$

$$P(\text{男、髪長い、背低い})=0.05$$

$$P(\text{男、髪短い、背低い})=0.125$$

$$P(\text{女、髪短い、背高い})=0.05$$

$$P(\text{女、髪長い、背高い})=0.125$$

$$P(\text{女、髪長い、背低い})=0.2$$

$$P(\text{女、髪短い、背低い})=0.125$$

$$P(\text{男})=0.5, P(\text{女})=0.5$$

同時確率分布からの確率推論

- 「その人は髪が短い」ことがわかった

$$P(\text{男、髪短い、背高い})=0.2$$

$$P(\text{男、髪短い、背低い})=0.125$$

$$P(\text{女、髪短い、背高い})=0.05$$

$$P(\text{女、髪短い、背低い})=0.125$$

$$\text{男の確率}=0.325/0.5=0.65$$

同時確率分布からの確率推論

- さらに「その人は背が高い」ことがわかった

$$P(\text{男、髪短い、背高い})=0.2$$

$$P(\text{女、髪短い、背高い})=0.05$$

$$\text{男の確率}=0.2/0.25=0.8$$

確率推論の数学的定式化

データ x_d が得られたときの x_i の確率は

$$p(x_i | x_d) = \sum_{j \neq i} p(x_1, x_2, \dots, x_N | x_d)$$

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！！

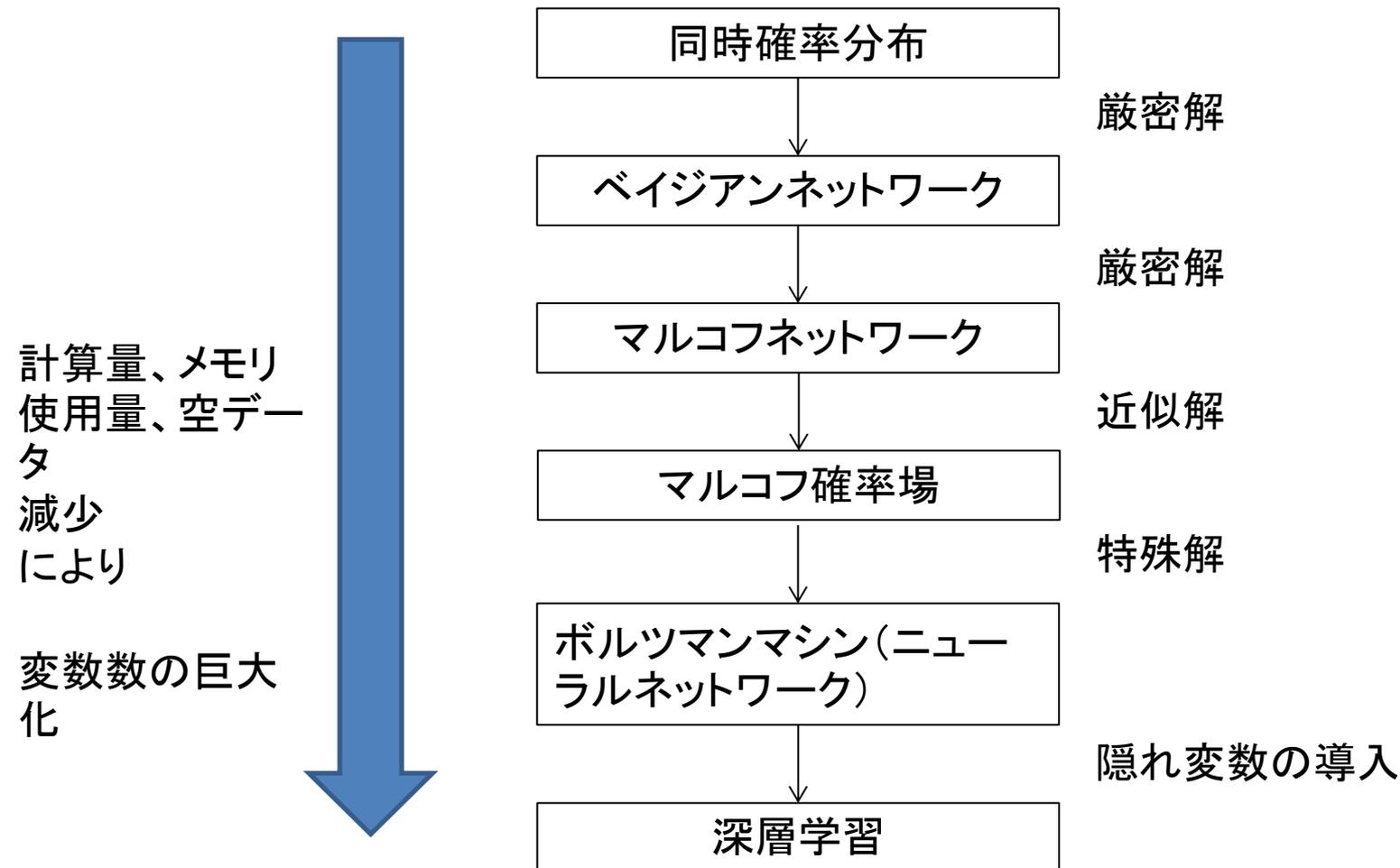
5. 二つの問題

- 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)

6. 解決のための数理モデル

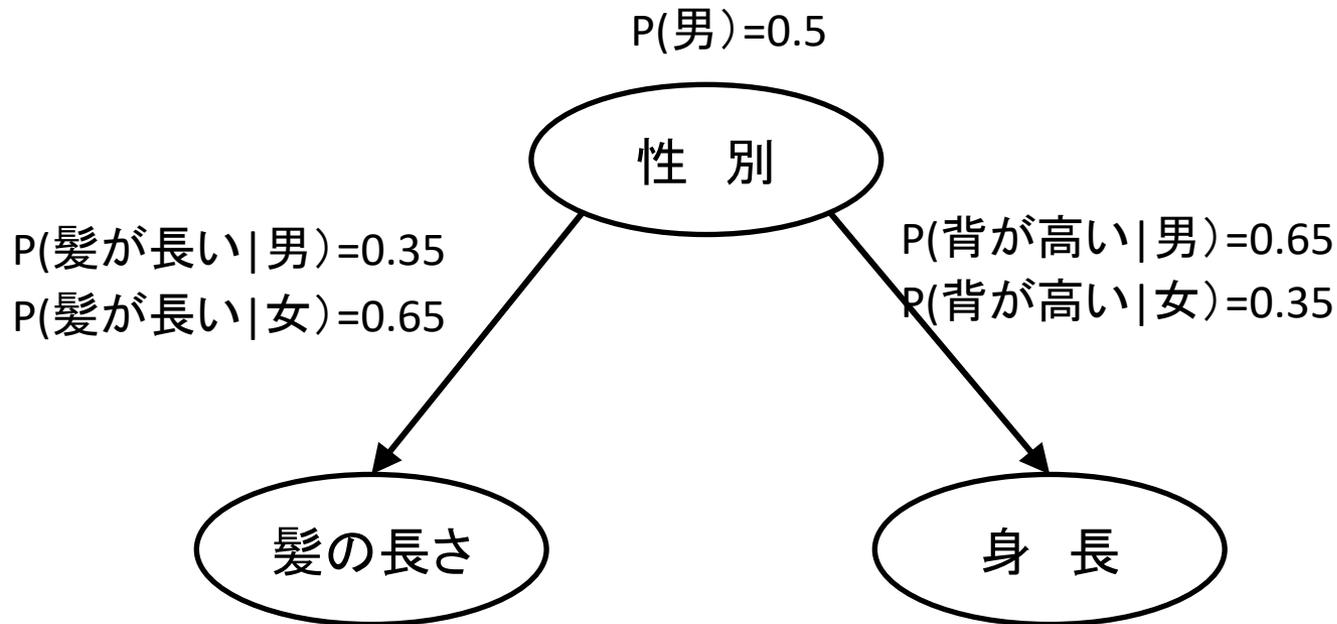


7. グラフィカルモデル

定義

X, Y, Z が無向グラフ G の互いに排他的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, $I(X, Y|Z)_G$ 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, $I(X, Y|Z)_M$ と書く.

8. ベイジアン・ネットワーク



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

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。
- 確率有向グラフが確率因果構造が対応し、ものごとの因果もわかる！！

現在考えられる最もよい同時確率分布の推定値
⇒ 推論の予測精度が最高のはず！！

9 定式化

定義 N 個の変数集合 $x = \{x_1, x_2, \dots, x_N\}$ 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 $\{p(x_i | \Pi_i, G)\}$, $(i = 1, \dots, N)$ である. ただし,

Π_i は変数 x_i の親変数集合を示している.

10. 同時確率分布

定理 変数集合 $x = \{x_1, x_2, \dots, x_N\}$ 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.

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

ここで, G は確率構造を示している.

11. ベイジアンネットワーク学習 ディレクレイ分布の周辺尤度を直接計算

$$\begin{aligned} p(G | X) &\propto P(G) \int_{\Theta_S} p(\mathbf{X}, \Theta_G | G) p(\Theta_G) d\Theta_G \\ &= P(G) \prod_{i=1}^N \prod_{j=1}^{q_i} \frac{\Gamma(\alpha_{ijk})}{\Gamma\left[\sum_{k=0}^{r_i-1} (\alpha_{ijk} + n_{ijk})\right]} \prod_{k=0}^{r_i-1} \frac{\Gamma(\alpha_{ijk} + n_{ijk})}{\Gamma(\alpha_{ijk})} \\ &= P(G) \prod_{i=1}^N \prod_{j=1}^{q_i} \frac{\Gamma(\alpha_{ijk})}{\Gamma(\alpha_{ij} + n_{ij})} \prod_{k=0}^{r_i-1} \frac{\Gamma(\alpha_{ijk} + n_{ijk})}{\Gamma(\alpha_{ijk})} \end{aligned}$$

$\alpha_{ijk} = \alpha / (q_i r_i)$ $\alpha = 1.0$ が漸近的には最適

12. ベイジアンネットワークの学習

未知のデータへの予測を最大化する構造は

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

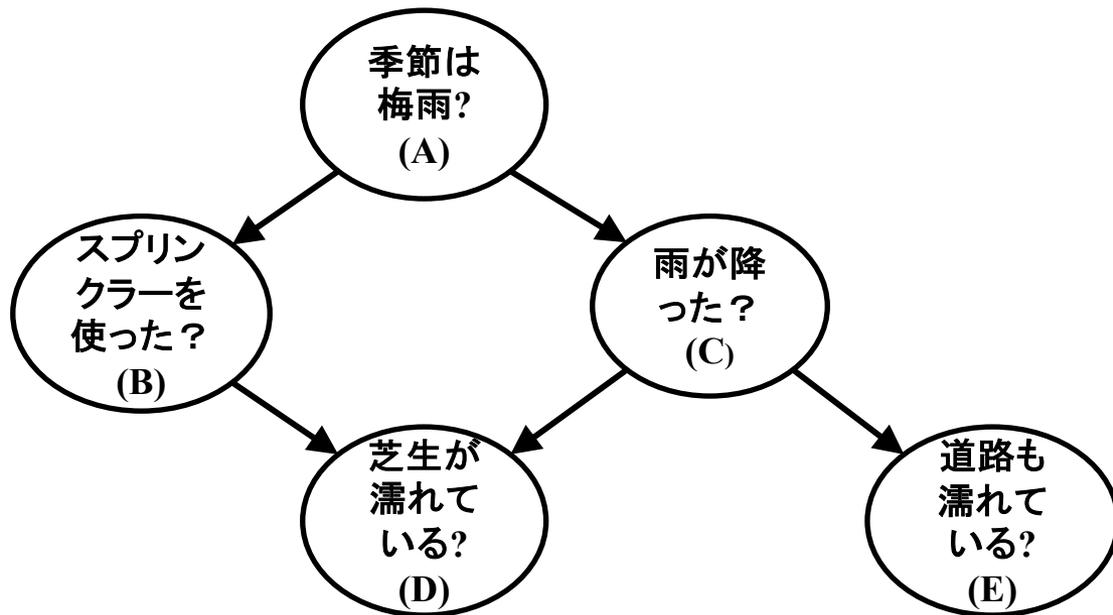
ここで、

$$\alpha_{ijk} = 1/K$$

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

13. 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1 ベイジアンネットワークのCPT1¹⁾

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

表1 図1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

15. 周辺確率の計算効率化

例えば, 図1について周辺確率 $p(E = 1|G)$ を求める。

$$\sum_D \sum_C \sum_B \sum_A p(E|C)p(D|B, C)p(A)p(B|A)p(C|A) =$$

$$\sum_D \sum_C p(E|C) \sum_B p(D|B, C) \sum_A p(A)p(B|A)p(C|A) = 0.364$$

16. 周辺消去アルゴリズム

アルゴリズム (周辺事前確率のための変数消去アルゴリズム)

• Input: ベイジアンネットワーク $\{G, \Theta\}$, ベイジアンネットワークでのクエリ (query) 変数集合 Q

• Output: 周辺確率 $p(Q|G)$

1. Main
2. $S \leftarrow$ CPTの値
3. For $i=1$ to N do
4. $\varphi \leftarrow \prod_k \varphi_k$, ここで φ_k は Q に含まれないノード i に関する (を含む) S に属する条件付き確率
5. $\varphi_i \leftarrow \sum_i \varphi$
6. S のすべての φ_k を φ_i によって置き換える.
7. end for
8. return $\prod_{\varphi \in S} \varphi$
9. end procedure

例

変数消去の順を $A \rightarrow B \rightarrow C \rightarrow D \rightarrow E$ の順としたとき、計算のステップは以下のようになる。

1. (step1) $i = 1$

消去する変数 A に関する変数は B, C なので φ_k は以下の通り。

$$\varphi \leftarrow \prod_k \varphi_k = p(A)p(B|A)p(C|A) \quad (1)$$

$$= p(A, B, C) \quad (2)$$

具体的なポテンシャルは表1のようになる。

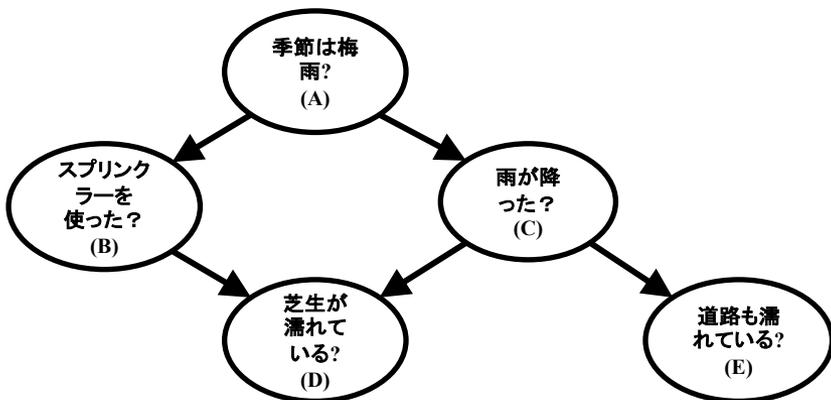


表1: $p(A, B, C)$

A	B	C	$p(A, B, C)$
真	真	真	$0.6 \times 0.2 \times 0.8 = 0.096$
真	真	偽	$0.6 \times 0.2 \times 0.2 = 0.024$
真	真	真	$0.6 \times 0.8 \times 0.8 = 0.384$
真	偽	偽	$0.6 \times 0.8 \times 0.2 = 0.096$
真	偽	真	$0.4 \times 0.75 \times 0.1 = 0.03$
真	真	偽	$0.4 \times 0.75 \times 0.9 = 0.03$
真	偽	真	$0.4 \times 0.25 \times 0.1 = 0.01$
真	偽	偽	$0.4 \times 0.25 \times 0.9 = 0.09$
偽	真	真	

次に, Aを消去したポテンシャルは以下のようになる.

$$\varphi_1 \leftarrow \sum_A \varphi = \sum_A p(A, B, C) \quad (3)$$

$$= p(B, C) \quad (4)$$

具体的なポテンシャルは表2のようになる.

表1: p(A,B,C)

A	B	C	p(A,B,C)
真	真	真	$0.6 \times 0.2 \times 0.8 = 0.096$
真	真	偽	$0.6 \times 0.2 \times 0.2 = 0.024$
真	偽	真	$0.6 \times 0.8 \times 0.8 = 0.384$
真	偽	偽	$0.6 \times 0.8 \times 0.2 = 0.096$
偽	真	真	$0.4 \times 0.75 \times 0.1 = 0.03$
偽	真	偽	$0.4 \times 0.75 \times 0.9 = 0.03$
偽	偽	真	$0.4 \times 0.25 \times 0.1 = 0.01$
偽	偽	偽	$0.4 \times 0.25 \times 0.9 = 0.09$

表2: p(B,C)

B	C	p(B,C)
真	真	$0.096 + 0.03 = 0.126$
真	偽	$0.024 + 0.27 = 0.294$
偽	真	$0.384 + 0.01 = 0.394$
偽	偽	$0.096 + 0.09 = 0.186$

例

2. (step2) $i = 2$

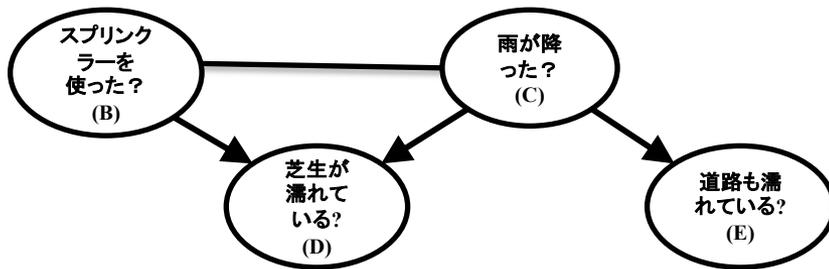
消去する変数Bに関する変数はC,Dなので φ_k は以下の通り.

$$\varphi \leftarrow \prod_k \varphi_k = p(B, C)p(D|B, C) \quad (5)$$

$$= p(B, C, D) \quad (6)$$

具体的なポテンシャルは表3のようになる.

表3: $p(B, C, D)$



B	C	D	$p(B, C, D)$
真	真	真	$0.126 \times 0.95 = 0.1197$
真	真	偽	$0.126 \times 0.05 = 0.0063$
真	偽	真	$0.294 \times 0.9 = 0.2646$
真	偽	偽	$0.294 \times 0.1 = 0.0294$
偽	真	真	$0.394 \times 0.8 = 0.3152$
偽	真	偽	$0.394 \times 0.2 = 0.0788$
偽	偽	真	$0.186 \times 0.0 = 0.0$
偽	偽	偽	$0.186 \times 1.0 = 0.186$

例

次に、Bを消去したポテンシャルは以下のようになる。

$$\varphi_2 \leftarrow \sum_B \varphi = \sum_B p(B, C, D) = p(C, D)$$

具体的なポテンシャルは表4のようになる。

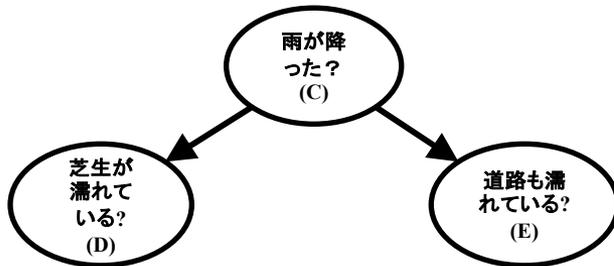


表4:p(C,D)

C	D	p(C,D)
真	真	0.1197 + 0.3152 = 0.4349
真	偽	0.0063 + 0.0788 = 0.0851
偽	真	0.2646 + 0.0 = 0.2646
偽	偽	0.0294 + 0.186 = 0.2154

例

3. (step3) $i = 3$

消去する変数Cに関する変数はD,Eなので φ_k は以下の通り.

$$\begin{aligned}\varphi &\leftarrow \prod_k \varphi_k = p(C, D)p(E|C) \\ &= p(C, D, E)\end{aligned}\tag{9}$$

具体的なポテンシャルは表5のようになる.\tag{10}

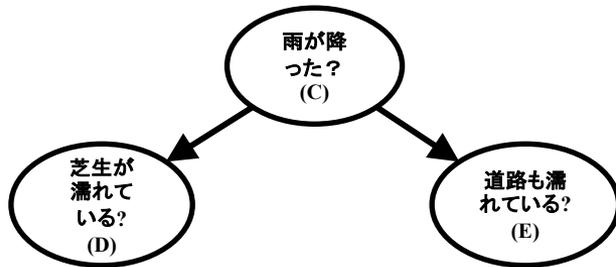


表5: $p(C, D, E)$

C	D	E	$p(C, D, E)$
真	真	真	$0.4349 \times 0.7 = 0.30443$
真	真	偽	$0.4349 \times 0.3 = 0.13047$
真	偽	真	$0.0851 \times 0.7 = 0.05957$
真	偽	偽	$0.0851 \times 0.3 = 0.02553$
偽	真	真	$0.2646 \times 0.0 = 0.0$
偽	真	偽	$0.2646 \times 1.0 = 0.2646$
偽	偽	真	$0.2154 \times 0.0 = 0.0$
偽	偽	偽	$0.2154 \times 1.0 = 0.2154$

例

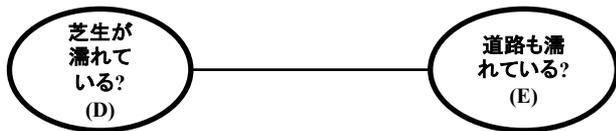
次に、Cを消去したポテンシャルは以下のようになる。

$$\varphi_2 \leftarrow \sum_C \varphi = \sum_C p(C, D, E) \quad (11)$$

$$= p(D, E) \quad (12)$$

具体的なポテンシャルは表6のようになる。

表6:p(D,E)



D	E	p(D,E)
真	真	0.30443 + 0.0 = 0.30443
真	偽	0.13047 + 0.2646 = 0.39507
偽	真	0.05957 + 0.0 = 0.05957
偽	偽	0.02553 + 0.2154 = 0.24093

例

4. (step4) $i = 4$

消去する変数Dに関する変数はEなので φ_k は以下の通り.

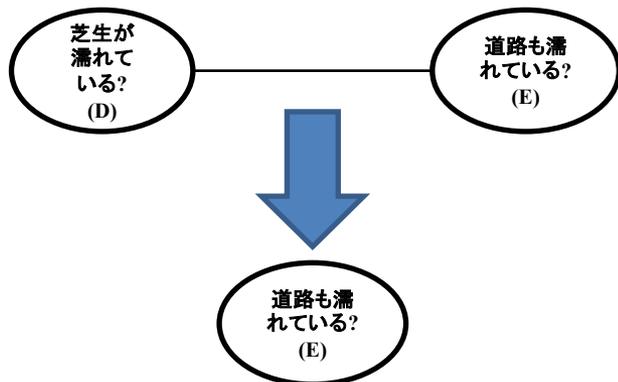
$$\varphi \leftarrow \prod_k \varphi_k = p(D, E)$$

次に, Dを消去したポテンシャルは以下のようにになる.

$$\begin{aligned} \varphi_4 \leftarrow \sum_D \varphi &= p(D, E) \\ &= p(E) \end{aligned}$$

具体的なポテンシャルは表7のようにになる.

表7:p(E)



E	p(D,E)
真	0.30443 + 0.05957 = 0.364
偽	0.39507 + 0.24093 = 0.636

例

5. (step4) $i = 4$

消去する変数はEだがEはクエリQに含まれているため処理を行わない。

6. return

S注に含まれるポテンシャルをすべて掛け合わせたものを出力する。この例では7が出力される。

ここでの変数の周辺化は、ベイジアンネットワークのいくつかの変数がインスタンス化される (エビデンスを得る) 前の事前確率分布について行われるものであり、得られた各変数の周辺確率を周辺事前確率 (marginal prior) と呼び、この操作を事前分布周辺化 (prior marginals) と呼ぶ。それに対して、ベイジアンネットワークでいくつかの変数がインスタンス化 (エビデンスを得る) された場合の各変数の周辺確率を周辺事後確率 (marginal posterior) と呼び、この操作を事後分布周辺化 (posterior marginals) と呼ぶ。

17. エビデンスを得た後の周辺事後確率

エビデンス e を所与とした種変事後確率を計算する場合 (エビデンスを得た場合の変数消去の場合), まず同時周辺確率 (joint marginals) $p(Q, e|G)$ を計算する. そのために, エビデンスに一致しないファクターの値を0にするように, ファクターを以下のように再定義する.

定義70 エビデンス e を所与としたときのファクター $\varphi^e(x)$ は以下のように定義される.

$$\varphi^e(x) \begin{cases} \varphi(x) & (x \text{ が } e \text{ に一致しているとき}) \\ 0 & (\text{上記以外}) \end{cases} .$$

さらに, この変換について以下の分配法則が成り立つ.

定理21 φ_1 と φ_2 が二つの異なるファクターであり, エビデンス e を得たとき,

$$(\varphi_1 \varphi_2)^e = \varphi_1^e \varphi_2^e$$

が成り立つ.

18. 周辺事後確率のための変数消去アルゴリズム

アルゴリズム7 (周辺事後確率のための変数消去アルゴリズム)

• Input: ベイジアンネットワーク $\{G, \Theta\}$, ベイジアンネットワークでのクエリ変数集合 Q , エビデンス e

• Output: 周辺確率 $p(Q, e | G)$

1. Main

2. $S \leftarrow \varphi^e \leftarrow \varphi$

3. For $i=1$ to N do

4. $\varphi \leftarrow \prod_k \varphi_k$, ここで φ_k はノード i に関する (を含む) S に属する φ^e

5. $\varphi_i \leftarrow \prod_i \varphi$

6. S のすべての φ_k を φ_i によって置き換える.

7. end for

8. return $\prod_{\varphi \in S} \varphi$

9. end procedure

19. エビデンスを得た後の周辺事後確率

エビデンス e を所与とした種変事後確率を計算する場合 (エビデンスを得た場合の変数消去の場合), まず同時周辺確率 (joint marginals) $p(Q, e|G)$ を計算する. そのために, エビデンスに一致しないファクターの値を0にするように, ファクターを以下のように再定義する.

定義70 エビデンス e を所与としたときのファクター $\varphi^e(x)$ は以下のように定義される.

$$\varphi^e(x) \begin{cases} \varphi(x) & (x \text{ が } e \text{ に一致しているとき}) \\ 0 & (\text{上記以外}) \end{cases} .$$

さらに, この変換について以下の分配法則が成り立つ.

定理21 φ_1 と φ_2 が二つの異なるファクターであり, エビデンス e を得たとき,

$$(\varphi_1 \varphi_2)^e = \varphi_1^e \varphi_2^e$$

が成り立つ.

20. 周辺事後確率のための変数消去アルゴリズム

アルゴリズム7 (周辺事後確率のための変数消去アルゴリズム)

• Input: ベイジアンネットワーク $\{G, \Theta\}$, ベイジアンネットワークでのクエリ変数集合 Q , エビデンス e

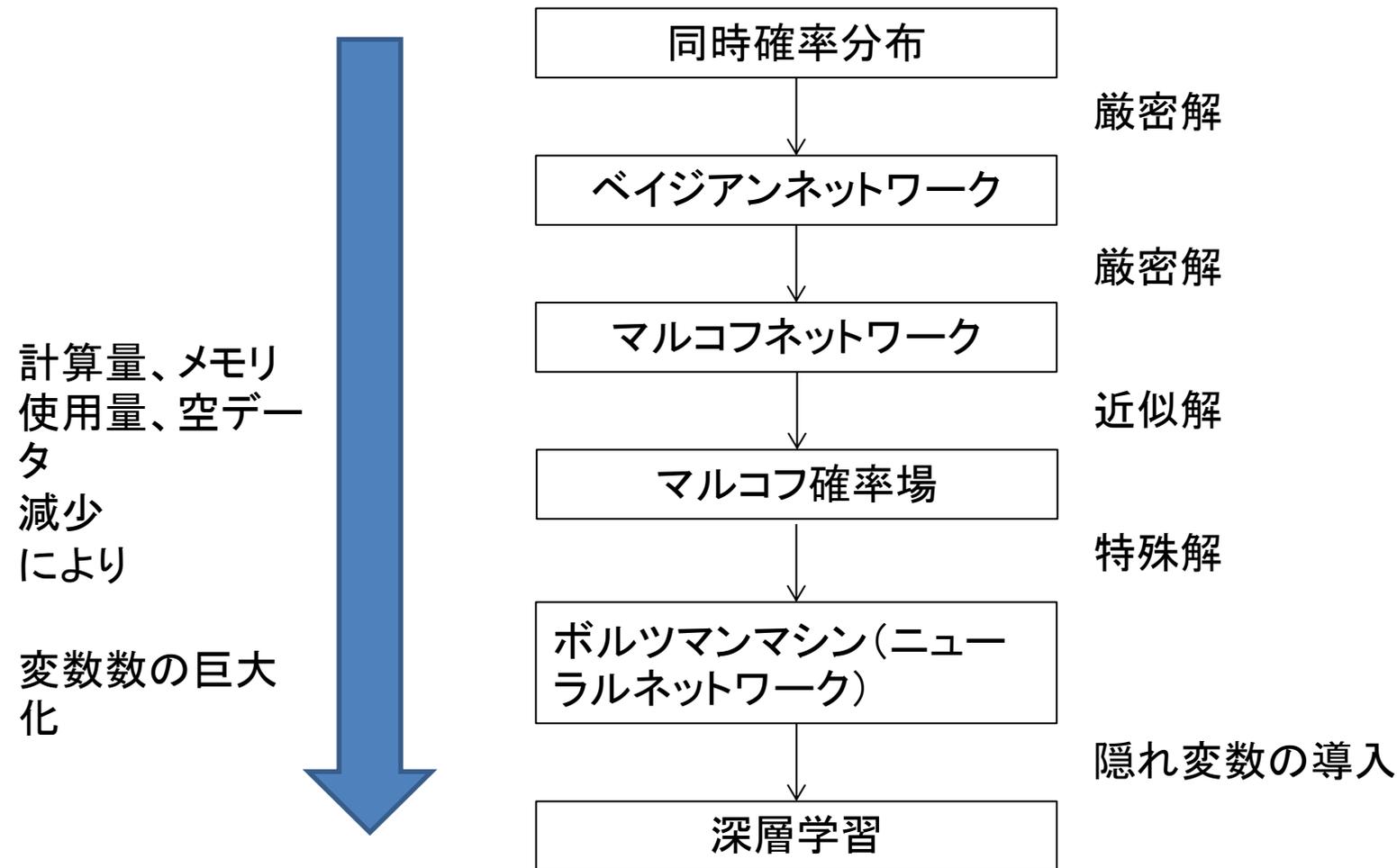
• Output: 周辺確率 $p(Q, e | G)$

1. Main
2. $S \leftarrow \varphi^e \leftarrow \varphi$
3. For $i=1$ to N do
4. $\varphi \leftarrow \prod_k \varphi_k$, ここで φ_k はノード i に関する (を含む) S に属する φ^e
5. $\varphi_i \leftarrow \prod_i \varphi$
6. S のすべての φ_k を φ_i によって置き換える.
7. end for
8. return $\prod_{\varphi \in S} \varphi$
9. end procedure

21. ベイジアンネットワークの問題

- 欠点として計算量の多さ
- 現在 厳密学習では、加藤ら(2024)が2万越え
- 将来的にはこれを克服すれば最強ツールになる！！

22. 解決のための数理モデル



23. 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ぬ。

欠点

- ベイジアンネットワークから変換できるがその逆は不可
- 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

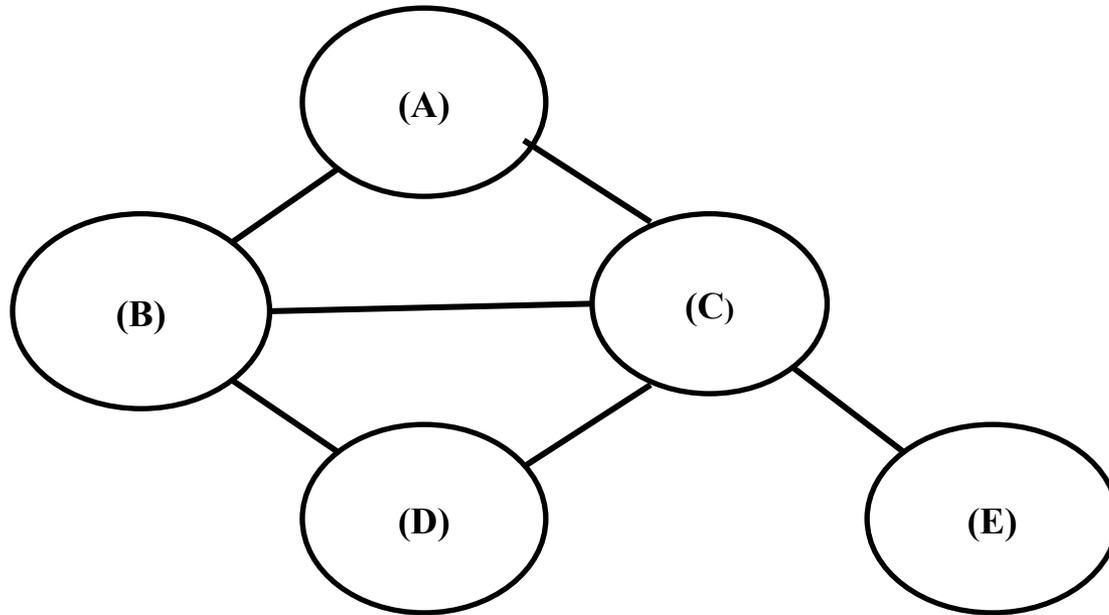
24. マルコフネットワークの同時確率分布のクリーク分解定理

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

C はクリーク集合

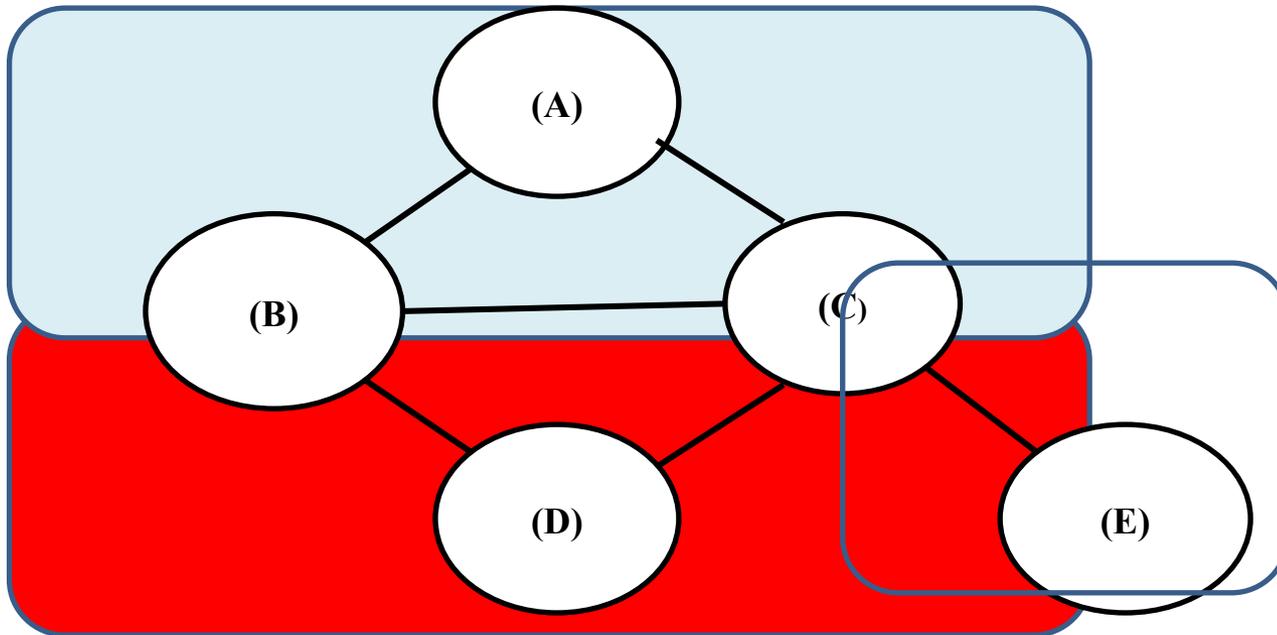
クリーク

- 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



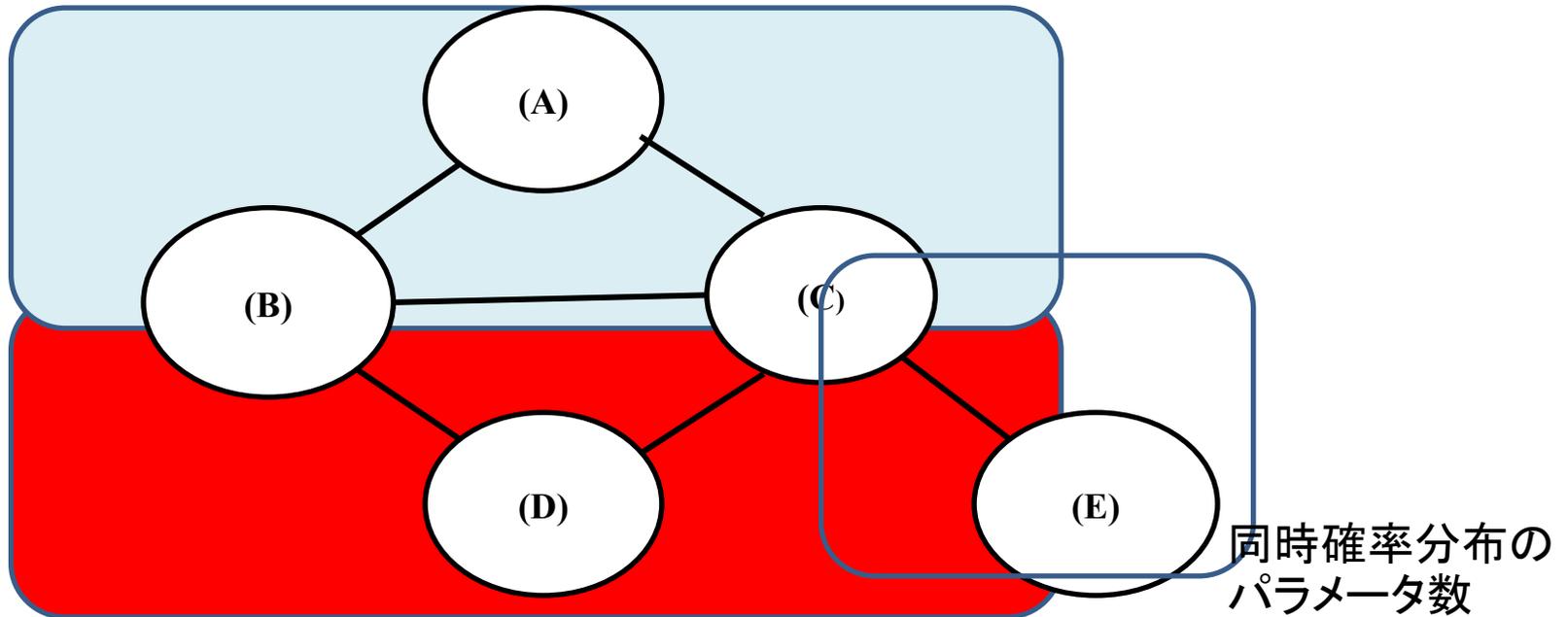
クリーク

- 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

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

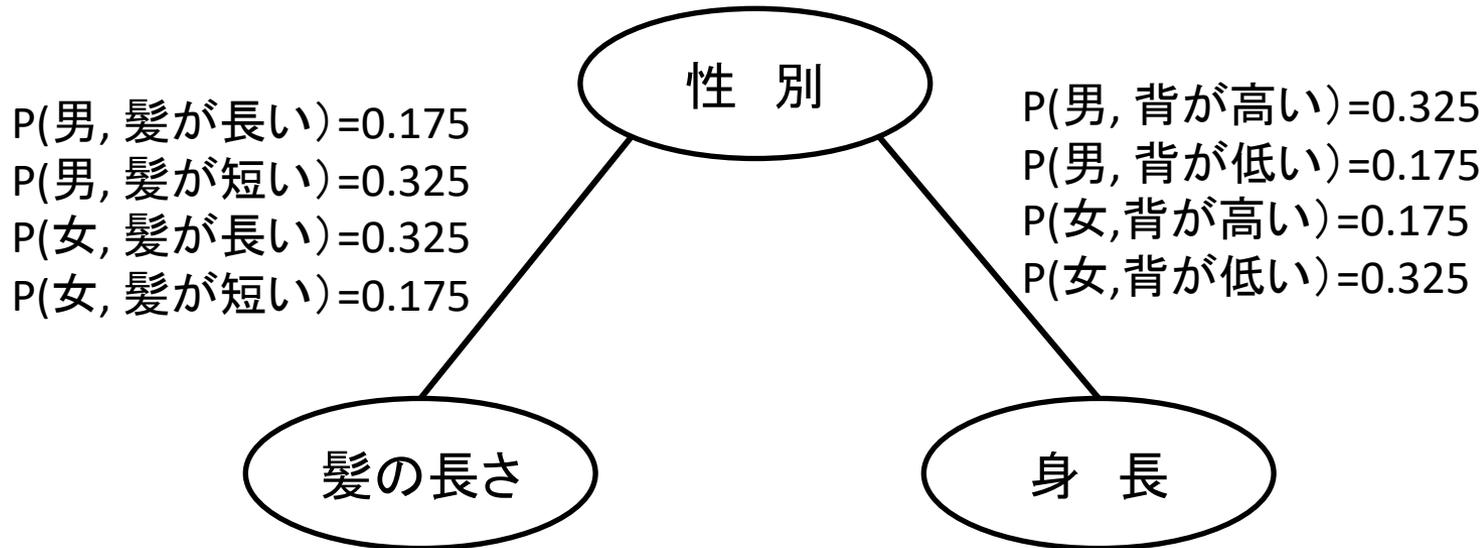


31

⇒

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

25. マルコフネットワークの問題

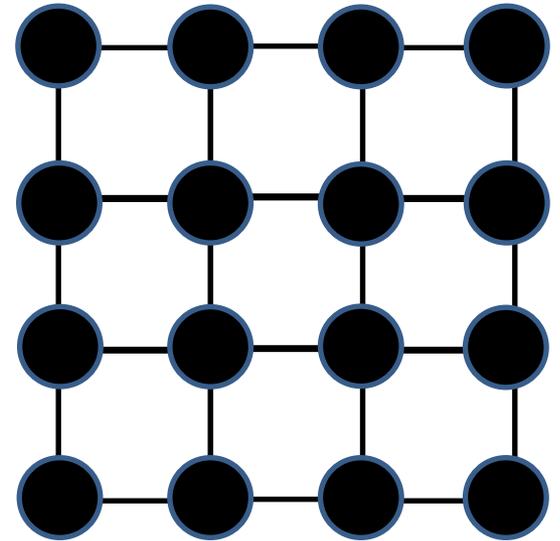
- パラメータ数はクリークの数に対して指数的に増え計算量が増えてしまうのでベイジアンネットワークより計算量が大きい。
- 数学的厳密性を緩和して計算が簡易なモデルの必要性

26. マルコフ確率場

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

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

- 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。

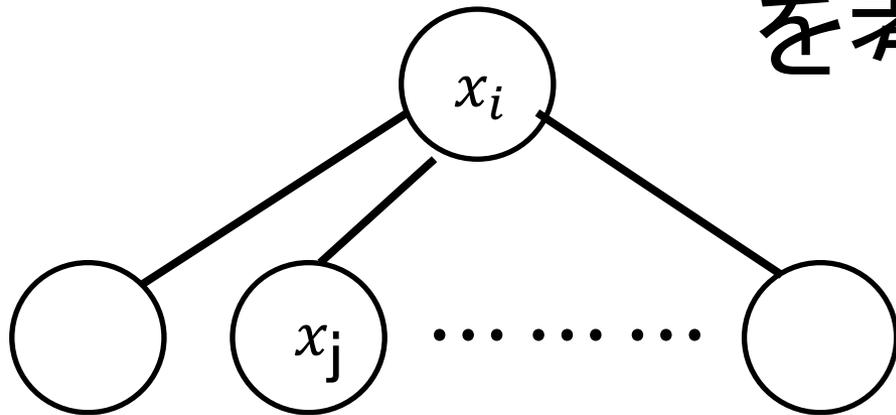
すなわち

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

マルコフ確率場のLog-Linear モデル表現

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

27. x は二値しかとらずに以下の構造を考える



$$\begin{aligned}
 P(x_i = 1 | x_1, x_2, \dots, x_N, G) &= \frac{1}{Z(\theta)} \exp\left(-\sum_i E(x_i = 1) - \sum_{(i,j)} E(x_i = 1, x_j)\right) \\
 &= \frac{\exp(-\sum_i E(x_i = 1) - \sum_{(i,j)} E(x_i = 1, x_j))}{\exp(-\sum_i E(x_i = 1) - \sum_{(i,j)} E(x_i = 1, x_j)) + \exp(-\sum_i E(x_i = 0) - \sum_{(i,j)} E(x_i = 0, x_j))} \\
 &= \frac{1}{1 + \exp(\sum_i E(x_i = 1) - \sum_i E(x_i = 0) + \sum_{(i,j)} E(x_i = 1, x_j) - \sum_{(i,j)} E(x_i = 0, x_j))}
 \end{aligned}$$

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

ボルツマンマシン

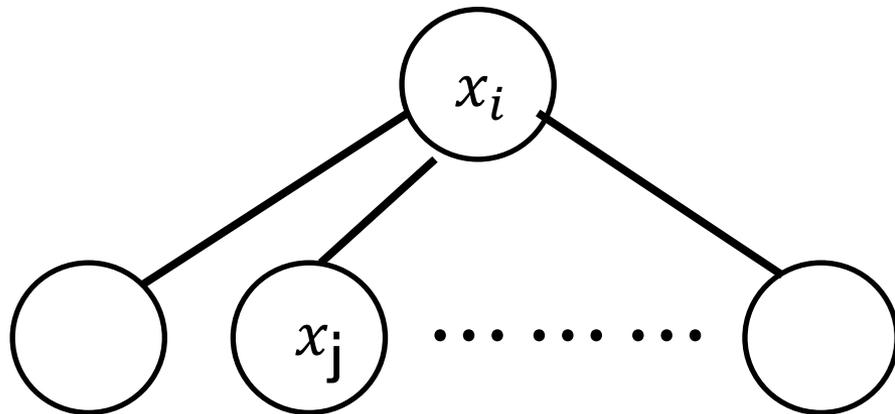
(ニューラルネットワーク)

$$-\sum_i E(x_i = 1) + \sum_i E(x_i = 0) = b_i,$$

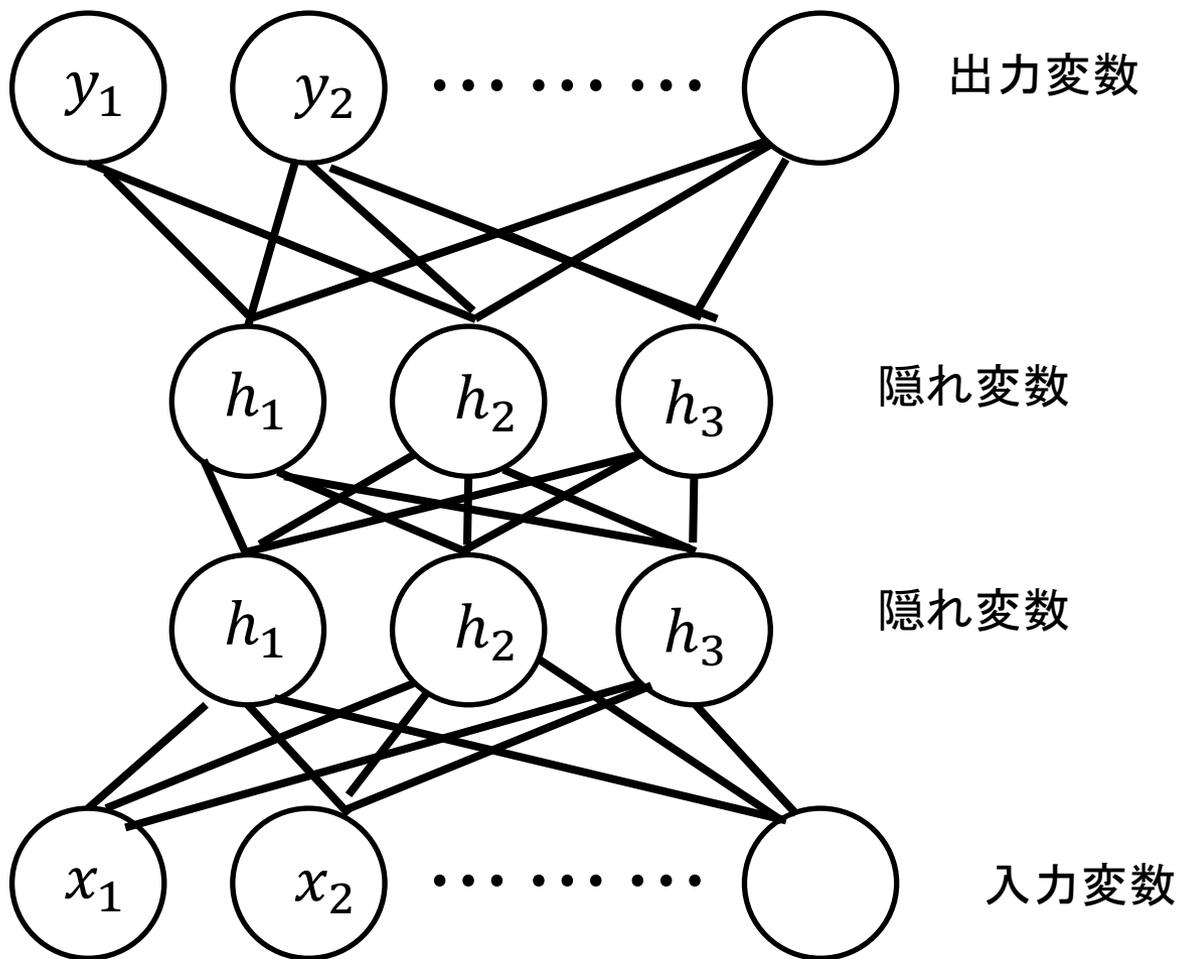
$$\sum_{(i,j)} E(x_i = 1, x_j) - \sum_{(i,j)} E(x_i = 0, x_j) = w_{ij}x_j$$

とおくと

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



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。
- 完全グラフ構造において、各辺の重みを最適化することにより、マルコフグラフの構造も同時に推定できる。
- 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。
- 真の確率構造が複雑な場合、隠れ変数層を増やさなければならないはず。
- ベイジアンネットワークで学習されるエッジ数が隠れ変数の数に関係している可能性が高い。

深層学習はすごい！！

- ビッグデータにおける同時確率分布の問題は変数の値のパターンがコンピュータや人間のメモリに入らないこと、計算速度が遅すぎること、パターンが多すぎて空データが増えてしまうことである！！
- 脳モデルはメモリに乗らないほどの変数パターンは計算せず、すべて独立変数のように扱い、隠れ変数が仲介する階層モデルにより、結果として変数間の依存性を補完する。
- 計算速度、メモリ使用量、欠損データ、近似精度のトレードオフをすべて解決する！！

28. まとめ

- データサイエンスのための因果推論を行うためには確率構造の推定が必要
- 条件付き独立性による確率構造の表現をグラフィカルモデルという
- 同時確率分布の厳密解にはベイジアンネットワークとマルコフネットワークがある。
- 深層学習はベイジアンネットワーク。マルコフネットワークの近似である。