# 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
情報理工学研究科　情報数理プログラム

## スケジュール（予定）

- 4月11日　授業の概要とガイダンス
- 4月18日　ベイズの定理
- 4月25日　ベイズの定理はどのように誕生したのか？
- 5月2日　ベイズはコンピュータ、人工知能の父である！！
- 5月9日　アランチューリングとベイズ
- 5月16日　ベイズから機械学習へ
- 5月23日　確率の基礎の復習
- 5月30日　ビリーフとベイズ
- 6月6日　尤度と最尤推定
- 6月13日　数値計算法による推定
- 6月20日　ベイズ推定と事前分布
- 6月27日　マルコフチェーンモンテカルロ(MCMC)法
- 7月4日　ベイジアンネットワーク
- 7月11日　ベイジアンネットワークと機械学習
- 7月25日　テストと総括

## 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

## ビック・データ時代

- 90年代ー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ーテンシー
- 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

## 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

## 確率推論では

- より一般化した表現
- 同時確率分布

## 例：性別、髪の長さ、背の高さの同時確率分布

データより性別、髪の長さ、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する。

P(男、髪短い、背高い)=0.2
P(男、髪長い、背高い)=0.125
P(男、髪長い、背低い)=0.05
P(男、髪短い、背低い)=0.125
P(女、髪短い、背高い)=0.05
P(女、髪長い、背高い)=0.125
P(女、髪長い、背低い)=0.2
P(女、髪短い、背低い)=0.125

P(男)=0.5, P(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(男、髪短い、背高い)=0.2

P(男、髪短い、背低い)=0.125
P(女、髪短い、背高い)=0.05

P(女、髪短い、背低い)=0.125

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(男、髪短い、背高い)=0.2

P(女、髪短い、背高い)=0.05

男の確率=0.2/0.25=0.8

## 確率推論の数学的定式化

データ$\mathrm{x}_d$が得られたときの$\mathrm{x}_i$の確率は

$$p(\mathrm{x}_i|\mathrm{x}_d) = \sum_{j \neq i} p(\mathrm{x}_1, \mathrm{x}_2, \cdots \mathrm{x}_N|\mathrm{x}_d)$$

## 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！！

## 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。

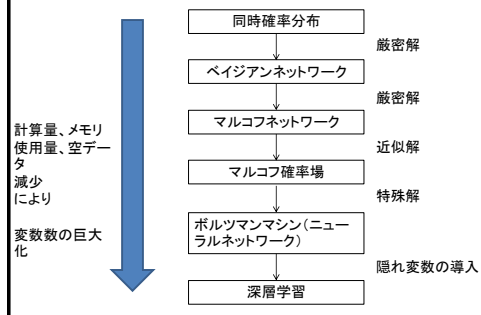
## 二つの問題

- 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）

## 解決のための数理モデル




## 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！！

## 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

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

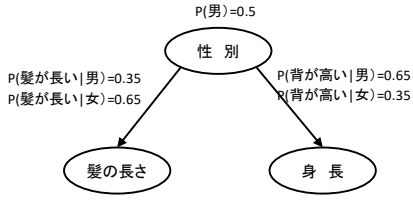
ここで、

$\alpha_{ijk} = 1/K$

$K$: 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

## ベイジアン・ネットワーク




$p$(性別、髪の長さ、身長$|G$) =
$p$(性別)$p$(髪の長さ|性別)$p$(身長|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の問題

- 欠点として計算量の多さ
- 現在　厳密学習では、2000ノードのネットワーク(Natori, Uto, Ueno 2017))
- 厳密推論では200ノードのネットワーク
(Li and Ueno 2017)
- 将来的にはこれを克服すれば最強ツールになる！！

## マルコフネットワーク

- 無向グラフ構造
- ベイジアンネットワークの互換モデル

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から変換できるがその逆は不可
- 構造の学習ができない
- パラメータ数が多い

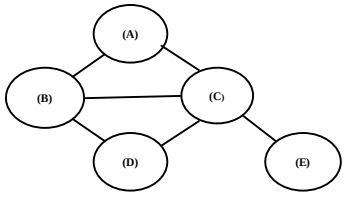
## マルコフネットワークの同時確率分布 クリーク分解定理

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

$C$はクリーク集合

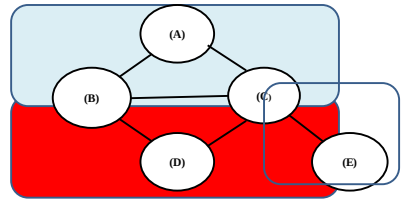
## クリーク

- 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



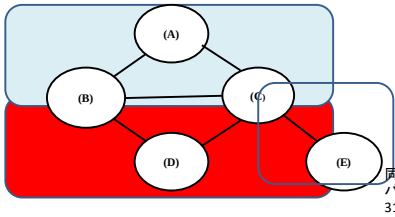

## クリーク

- 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



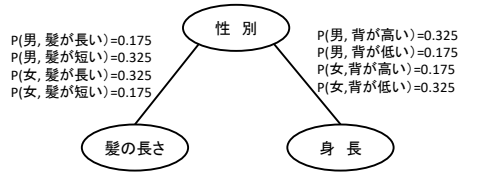

## 計算例　2値の場合

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




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

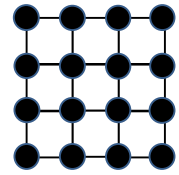
- パラメータ数はクリークの大きさに対して指数的に増え計算量が増えてしまうのでベイジアンネットワークより計算量が大きい。
- 数学的厳密性を緩和して計算が簡易なモデルの必要性

## 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。

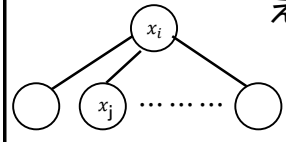
すなわち

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

## 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




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

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

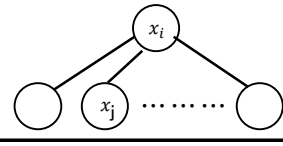
$$= \frac{1}{1 + \exp\left(\sum_i E\ (x_i = 1) - \sum_i E\ (x_i = 0) + \sum_{(i,j)} E(x_i = 1, x_j) - \sum_{(i,j)} E(x_i = 0, x_j)\right)}$$

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

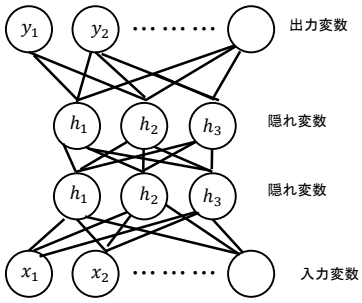
## ボルツマンマシン（ニューラルネットワーク）

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

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



## 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。
- 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.

## やはり脳モデルはすごい！！

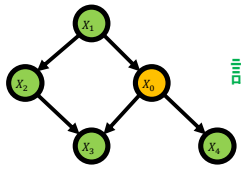
- ビッグデータにおける同時確率分布の問題は変数の値のパターンがコンピュータや人間のメモリに入らないこと、計算速度が遅すぎること、パターンが多すぎて空データが増えてしまうことである！！
- 脳モデルは　メモリに乗らないほどの変数パターンは計算せず、すべて独立変数のように扱い、隠れ変数が仲介する階層モデルにより、結果として変数間の依存性を補完する.
- 計算速度、メモリ使用量、欠損データ、近似精度のトレードオフをすべて解決する！！

## 隠れ変数の数と構造の最適化が今後のビッグチャレンジ

- 問題：周辺尤度や従来の情報量規準は隠れ変数の数と構造を決めることはできない。ただ、モデルが正則性を満たさないということだけではない。

- 隠れ変数の数と構造を最適にする規準（スコア）は何か？
- ベイジアンネットワークで得られる構造のパラメータ数とどのような関係にあるのか？
- 数学的に解明できるのか？
- その構造を学習できるのか？

## ベイジアンネットワーク分類器 (Bayesian Network Classifier: BNC)




目的変数：$X_0$

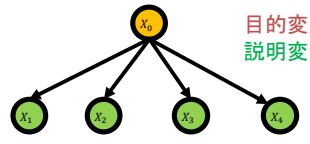
説明変数：$X_1, \cdots, X_4$

$P(X_0, X_1, \cdots, X_n \mid G)$から$P(X_0 \mid X_1, \cdots, X_n)$を計算して，分類を行う.

## Friedmanら(1997)による批判

**周辺尤度で学習した構造の分類精度が，単純な構造をとるNaïve 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.

## Friedmanら(1997)による批判

BDeu：同時確率分布$P(X_0, X_1, \cdots, X_n \mid G)$を表現する
生成モデルをモデル化

分類が目的ならば，条件付確率分布$P(X_0 \mid X_1, \cdots, X_n, G)$を表現する識別モデルを学習するべき

**条件付き対数尤度スコア(Conditional Log Likelihood: CLL)の提案**

$$\sum_{d=1}^{N} \log P(x_0^d \mid x_1^d, \ldots, x_n^d, G, \Theta)$$

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

## 条件付き対数尤度の問題点

**条件付き対数尤度**スコアは周辺尤度とは異なり閉形式で表せないため，構造探索に効率的なアルゴリズムが適用できず，厳密学習に膨大な時間がかかる．

## 条件付き対数尤度の近似手法

- 構造探索に対してHill-Climbing法の適用した近似学習(Grossman et al.,2004)
- 構造探索に効率的なアルゴリズムを用いることができるように修正した近似条件付き対数尤度スコアapproximate CLL(aCLL)(Carvalho et al.,2013)

**上記の近似学習手法の方が，周辺尤度を用いた学習法より分類精度の高い分類器が得られることが経験的に示されている．**

A. M. Carvalho, P. Ado, and P. Mateus. Efficient Approximation of the Conditional Relative Entropy with Applications to Discriminative Learning of Bayesian Network Classifiers. *Entropy*, 15(7): 2716-2735,2013.

## 本当に条件付き対数尤度は良いのか？

- **周辺尤度最大化より条件付き対数尤度最大化の方がなぜ良いのかという理由については未だ明らかになっていない．**
- **周辺尤度を最大化する構造を厳密に学習できるにもかかわらず，既存研究では近似学習を行っている．**

**このため，探索精度の悪さが結果に影響したのかもしれない．**

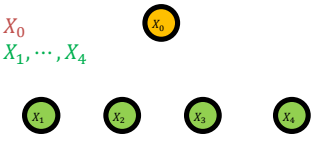
## 周辺尤度による厳密学習と条件付き対数尤度による近似学習の比較

- **Sugahara, Uto, and Ueno(2018)は，サンプルサイズが大きいときは周辺尤度を用いた厳密学習の方が条件付き対数尤度の近似学習よりも高い分類精度を示すことを経験的に示した．**
- **一方で，サンプルサイズの小さい場合，周辺尤度は目的変数の親変数が多い構造を学習する傾向があり，パラメータ数が指数的に増えてしまい分類精度が著しく悪くなることもわかった．**

Sugahara, S.; Uto, M.; and Ueno, M. 2018. Exact learning augmented naive Bayes classifier. In Proceedings of the 9th International Conference on Probabilistic Graphical Models, volume 72 of Proceedings of Machine Learning Research, 439-450. PMLR.

## 提案手法

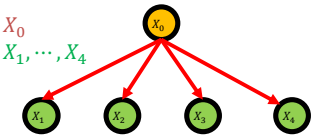
目的変数：$X_0$
説明変数：$X_1, \cdots, X_4$



## 提案手法

まず強制的に目的変数から全説明変数へエッジを引く.

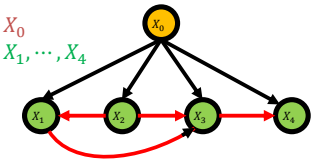
目的変数：$X_0$
説明変数：$X_1, \cdots, X_4$




## 提案手法

まず強制的に目的変数から全説明変数へエッジを引く.

目的変数：$X_0$
説明変数：$X_1, \cdots, X_4$

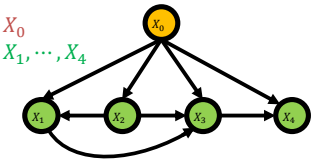



説明変数間の構造を学習.

## 提案手法

このように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) 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を回避できる.

## ANB厳密学習の漸近的性質

- 定理 (Sugahara and Ueno, 2021)
  厳密学習されたANBは真の構造と分類確率が等価な構造に漸近的に一致する.

Sugahara, S.; Ueno, M. Exact Learning Augmented Naive Bayes Classifier. *Entropy* **2021**, *23*, 1703. https://doi.org/10.3390/e23121703

## Deep Leaningとの比較

- Deep Learning では分類確率までは一致性がないため，この性質は確率推論を含む意思決定問題などで大きな意味を持つ.
- Deep Learningでは説明性・解釈性が乏しいことが問題となっているが 因果をグラフィカルモデルで表現できるベイジアンネットワークは利点が多い

## Deep Learning のDropout 同様にモデル平均を用いる

*1, Shouta Sugahara, Itsuki Aomi, Maomi Ueno Proceedings of the 10th International Conference on Probabilistic Graphical Models*, PMLR 138:461-472, 2020

2. Sugahara, S.; Aomi, I.; Ueno, M. Bayesian Network Model Averaging Classifiers by Subbagging. *Entropy* **2022**, *24*, 743. https://doi.org/10.3390/e24050743

ベイジアンネットワーク分類機>> Deep Learning