6. 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

今後のスケジュール（予定）

- 4月10日　授業の概要とガイダンス
- 4月17日 ベイズの定理
- 4月24日　ベイズはどのように誕生したか？
- 5月1日　ベイズはコンピュータ、人工知能の父である！！
- 5月8日　アランチューリングとベイズ
- 5月15日　ベイズから機械学習へ
- 5月22日　確率の基礎の復習
- 5月29日　ビリーフとベイズ
- 6月12日　尤度と最尤推定
- 6月19日　数値計算法による推定
- 6月26日　ベイズ推定と事前分布
- 7月3日　マルコフチェーンモンテカルロ(MCMC)法
- 7月10日　ベイジアンネットワーク
- 7月17日　ベイジアンネットワークと機械学習
- 7月24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x_dが得られたときの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。

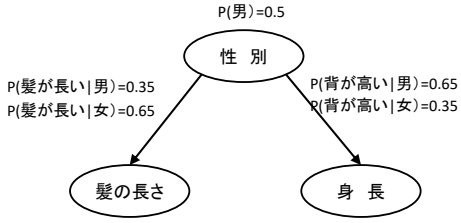
二つの問題

- 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）

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

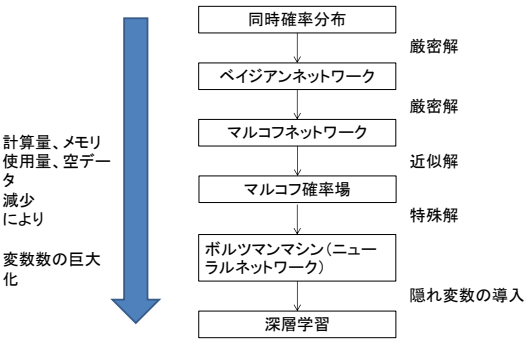
ベイジアンネットワークの問題

- 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！！

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モデル

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$\{p(x_i|\Pi_i, \mathrm{G})\}$，$(i = 1, \dots, N)$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.

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を直接計算

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

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

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

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

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

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

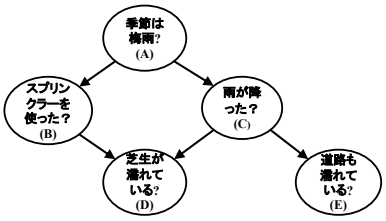
ここで、

$\alpha_{ijk} = 1/K$

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

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)]

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

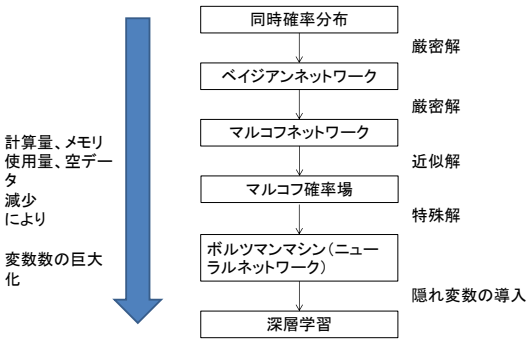
周辺確率の計算効率化

例えば，図3.12について周辺確率$p(E=1|\mathrm{G})$を求める。

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

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

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
- 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

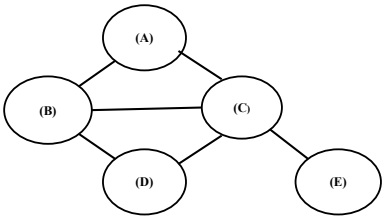
マルコフネットワークの同時確率分布 クリーク分解定理

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

Cはクリーク集合

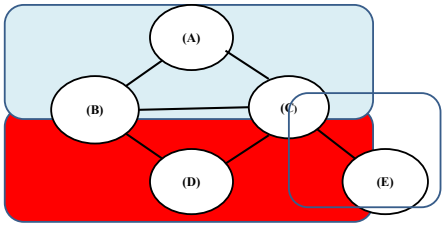
クリーク

- 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



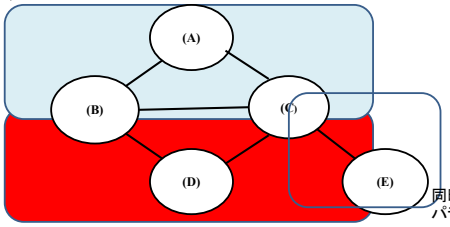
クリーク

- 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



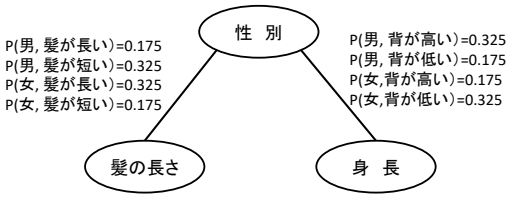
計算例　2値の場合

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



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

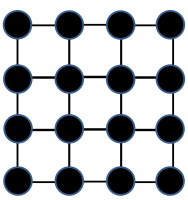
マルコフネットワークの問題

- 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。

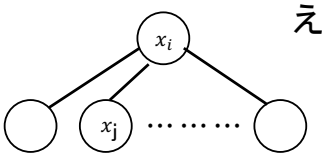
すなわち

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

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



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

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

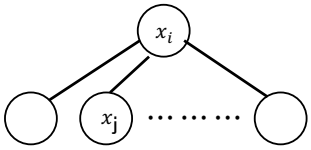
$$= \frac{1}{1 + \exp\left(\sum_i E\ (x_i = 1) - \sum_i E\ (x_i = 0) + \sum_{(i,j)} E(x_i = 1, x_j) - \sum_{(i,j)} E(x_i = 0, x_j)\right)}$$

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

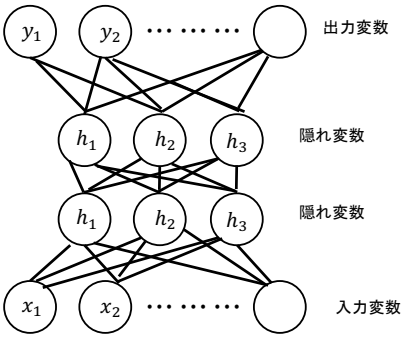
ボルツマンマシン（ニューラルネットワーク）

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

分類器

分類器はディープラーニングアプローチと確率アプローチに大別される

- ディープラーニングアプローチ

分類精度が高い
解釈性が低い
真の分類確率を推定できない

例）深層ニューラルネットワーク

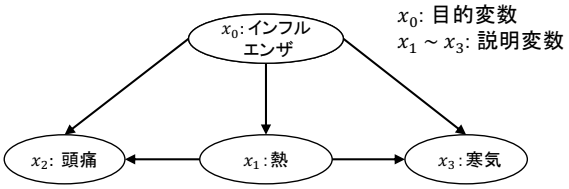
- 確率アプローチ

解釈性が高い
真の分類確率を推定できる
分類精度が低い

例）ベイジアンネットワーク分類器

ベイジアンネットワーク分類器

ベイジアンネットワーク分類器:
ベイジアンネットワークにおける一つのノードを目的変数とし，その他のノードを説明変数とした分類器



x_0: 目的変数
$x_1 \sim x_3$: 説明変数

目的変数パラメータ数最小化によるベイジアンネットワーク分類器学習

最も分類精度の高いベイジアンネットワーク分類器として，目的変数パラメータ数(NCP)を最小にして真の分類確率に漸近収束する構造を学習する手法が提案されている

$$NCP(G) = \sum_{i=0}^{N} NCP_i(\Pi_i),$$

$$NCP_i(\Pi_i) = \begin{cases} (r_i - 1)q_i & (i = 0 \vee x_0 \in \Pi_i) \\ 0 & (\text{その他の場合}) \end{cases}$$

Π_iは変数x_iの親変数集合，r_iはx_iの状態数，
q_iはΠ_iのとりうるパターン数を意味する

評価実験

- 比較手法:
  1. ランダムフォレスト (RF)
  2. サポートベクターマシーン (SVM)
  3. 深層ニューラルネットワーク (MLP)
  4. ベイジアンネットワーク分類器 (BNC)
- 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，分類精度とした

各手法の分類精度

No.	Dataset	Variables	RF	SVM	MLP	BNC
1	Lenses	5	0.7833	0.7833	0.7833	0.8750
2	Hayes-Roth	5	0.5742	0.6566	0.7198	0.8333
3	Iris	5	0.7533	0.8267	0.8267	0.8200
4	Balance Scale	5	0.8272	0.9088	0.9792	0.9152
5	Banknote	5	0.8433	0.8819	0.8790	0.8819
6	LED7	8	0.7191	0.7356	0.7347	0.7325
7	HTRU2	9	0.9084	0.9141	0.9112	0.9140
8	Breast Cancer	10	0.7324	0.7284	0.7108	0.7401
9	BCW	10	0.9677	0.9722	0.9693	0.9751
10	Solar Flare	11	0.8431	0.8431	0.8402	0.8431
11	Wine	14	0.9379	0.9157	0.9382	0.9494
12	Heart	14	0.8407	0.8407	0.8185	0.8407
13	Australian	15	0.8232	0.8435	0.8522	0.8464
14	EEG	15	0.5943	0.7273	0.7288	0.7135
15	Zoo	17	0.7927	0.8909	0.9118	0.9406
16	Congressional	17	0.9522	0.9565	0.9482	0.9698
17	Pendigits	17	0.6887	0.9446	0.9430	0.9344
18	Letter	17	0.3724	0.6606	0.6602	0.6297
19	Lymphography	19	0.7486	0.8043	0.8390	0.7905
20	Image Segmentation	19	0.6675	0.8255	0.8338	0.8277
21	Hepatitis	20	0.8250	0.8625	0.8375	0.8000
	average		0.7712	0.8344	0.8412	0.8463

おわり

半年間　ご清聴ありがとうございました！！