6. マルコフグラフ（マルコフ確率場、深層学習）

電気通信大学大学院
情報理工学研究科
植野　真臣

本日の目標

- ベイジアンネットワークの互換モデルであるマルコフネットワークについて学ぶ
- マルコフネットワークは、マルコフ確率場、条件付確率場、深層学習のボルツマン分布の上位モデルである

マルコフネットワーク

- BNでは確率構造がDAGのときのみ、

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

という分解が可能になる。

もし、確率構造がDAGにならない変数集合は分解できないので同時確率分布のまま扱わねばならない。分解できない変数集合の同時確率分布を一つのノードとしてグラフィカルモデルとして表現したものをマルコフネットワークと呼ぶ。

問

- $I(A, C|B, D)_G$ and $I(B, D|A, C)_G$

もしくは　$\mathrm{A} \perp C|B, D$ and $\mathrm{B} \perp D|A, C$

を満たすベイジアンネットワークは存在するか？

無向グラフでの表現

- $I(A, C|B, D)_G$ and $I(B, D|A, C)_G$

もしくは$\mathrm{A} \perp C|B, D$ and $\mathrm{B} \perp D|A, C$

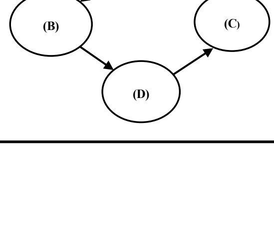
はベイジアンネットワークでは表現できない。

循環構造

- $I(A, C|B, D)_G$ and $I(B, D|A, C)_G$

もしくは$\mathrm{A} \perp C|B, D$ and $\mathrm{B} \perp D|A, C$

はベイジアンネットワークでは表現できない。

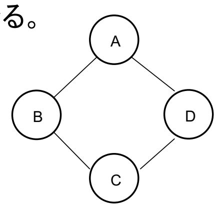


無向グラフでの表現

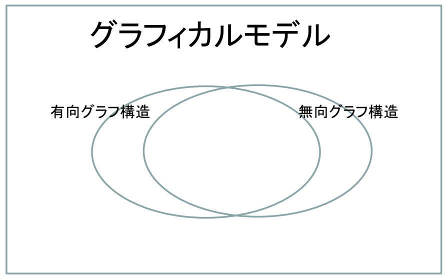
- $I(A,C|B,D)_G$ and $I(B,D|A,C)_G$

もしくは$A \perp C|B,D$ and $B \perp D|A,C$

は有向グラフでは表現できないが無向グラフでは表現できる。



有向グラフと無向グラフの確率分布の表現力



有向グラフから無向グラフへの変換

- 問
- 有向グラフの方向を取り除けば無向グラフになるのか？

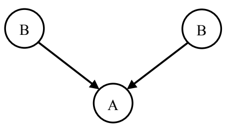
有向グラフから無向グラフへの変換と条件付き独立性

- 問
- 有向グラフの方向を取り除けば同じ条件付き独立性を持つ無向グラフになるのか？

- NO

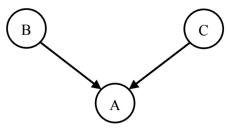
V構造

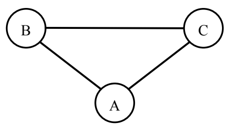
- 無向グラフに変換せよ。



V構造

- 無向グラフに変換せよ。



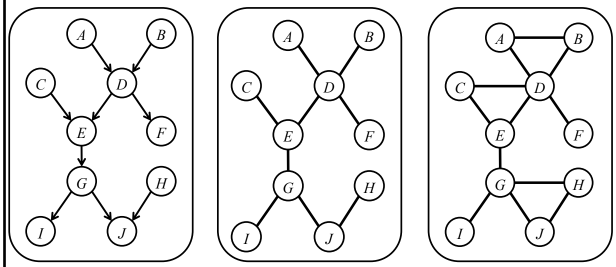


モラルグラフ（復習）

定義

有向グラフにおいて共通の子ノードをもつすべての親ノードの対にエッジを張り，方向性を取り除いて構築された「有向グラフに対応した無向グラフ」を**モラルグラフ**（moral graph）と呼ぶ．エッジを張ることをモラル化と呼ぶ。

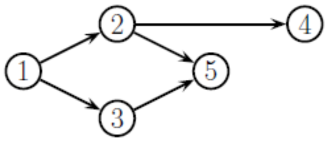
モラルグラフの例



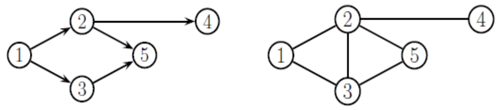
(a) 有向グラフ　(b) 有向グラフに対応した無向グラフ　(c) 対応したモラルグラフ

モラルグラフの例

問：以下の有向グラフを無向グラフに変換せよ。



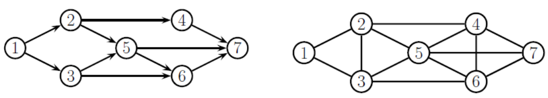
問：以下の有向グラフを無向グラフに変換せよ。



(a)　(b)

モラル化は完全ではない

- 4 ⊥ 5|2を確認せよ モラル化されたグラフでは？



(a)　(b)

4,6にもモラル化してエッジを引いてしまったため

有向グラフをコーダルグラフに変換

- コーダル グラフ(Chordal Graph)によりモデル化される。
- DAGをコーダル化してクリークごとをノードに置き換えると木が作成される。DAGは必ずコーダル化できる。
- DAGが最初から木の場合、それはコーダル化しなくてもコーダルグラフである。

コーダルグラフ：弦

定義
ループの**弦**(chord)とは，ループ中の2 ノードに張られたエッジを示し，そのループを分断し二つのループに分解するようなものをいう.
例
図**2.19** では，エッジ$E - G$ はループ$E - F - G - D - E$ の弦である.

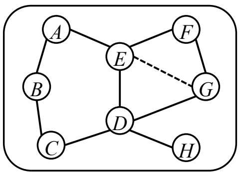


図2.19 弦を持つループの例

コーダルグラフ

4 以上の長さをもつすべてのループが少なくとも一つの弦をもつとき，**コーダルグラフ**(chordal graph)と呼ぶ.

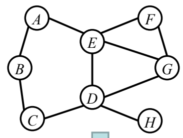
コーダルグラフの作成法

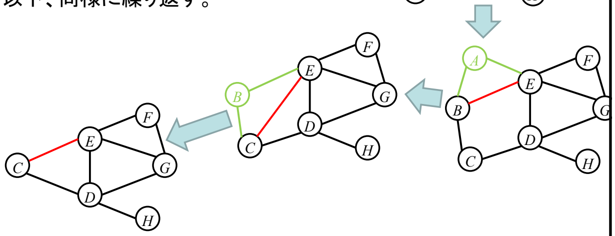
ベイジアンネットワークの変数消去に対応

1. 変数消去順序i=1,…Nを決定
2. ベイジアンネットワークをモラル化し、無向化する
3. ノードiの隣接ノード同士でエッジがないノードにエッジ(弦)を張り、iを消去。
4. 変数消去順序i=1,…Nについて2を実行し、張られた新しいすべてのエッジ(弦)を元の無向グラフに張ればコーダルグラフになる。

コーダルグラフの作成法

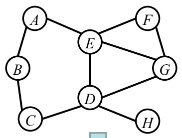
変数消去順　ABCDEFGH
Aの隣接ノードBとEにはエッジがないので、エッジ(弦)を張り、Aを消去。
Bの隣接ノードCとEにはエッジがないので、エッジ(弦)を張り、Bを消去。
以下、同様に繰り返す。

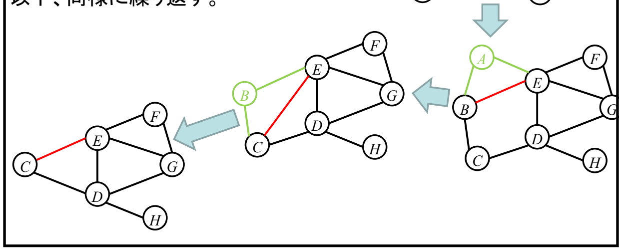




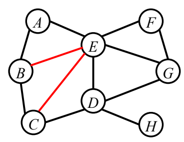
コーダルグラフの作成法

変数消去順　ABCDEFGH
Aの隣接ノードBとEにはエッジがないので、エッジ(弦)を張り、Aを消去。
Bの隣接ノードCとEにはエッジがないので、エッジ(弦)を張り、Bを消去。
以下、同様に繰り返す。





得られたコーダルグラフ



参考

- 確率推論は、実際にはコーダルグラフを作成し、すべてのクリークを列挙する。（クリークとは、グラフのノード部分集合に属するすべての二つのノードをつなぐ辺がある部分グラフのことをいう。）各クリークをノードとする、ジョインツリーと呼ばれる木を作成し、その木を用いて確率を伝搬させる手法が用いられる。
  本授業で述べた変数消去法は，計算量は $O(N^2\exp(w))$であった.
  ここでwはファクター中の最大ウィズである. コーダルグラフを用いたジョインツリーアルゴリズムは，$O(\mathrm{N}\exp(w))$ まで減じることができる.

可能な変換は一方通行

- 有向グラフ $\Rightarrow$ 無向グラフ
- 有向グラフ $\not\Leftarrow$ 無向グラフ

同時確率分布

- ベイジアンネットワークなど有向グラフではチェーンルールが適用でき、同時確率分布が表現できた。
- 無向グラフ構造でも同時確率分布が表現できるのか？

ファクター

- 定義
- 確率変数xについて、その値から実数空間への非負値への関数ϕをファクター（またはクリークポテンシャル、ポテンシャル）と呼ぶ。

Hammersley Cliffordの定理

- 無向グラフGのすべての極大クリーク集合Cについて　ファクターϕ_cの積として
- $P(x_1, x_2, \cdots, x_N|G) = \frac{1}{Z(\theta)} \prod_{c \in C} \phi_c(\,x_c|\theta_c)$
- ここで$Z(\theta)$はパーティション関数(partition function)で$x_1, x_2, \cdots, x_N$のすべての取りえる値のパターンについての和$\sum_{x_1,x_2,\cdots,x_N}$を用いて以下のように示せる。

$$Z(\theta) \equiv \sum_{x_1,x_2,\cdots,x_N} \prod_{c \in C} \phi_c(\,x_c|\theta_c)$$

マルコフネットワーク

- *このとき*
- $P(x_1, x_2, \cdots, x_N|G) = \frac{1}{Z(\theta)} \prod_{c \in C} \phi_c(\,x_c|\theta_c)$
- をギブス分布（Gibbs Distribution)と呼ぶ。
- ファクターϕ_cのことを「クリークポテンシャル」、もしくは単に「ポテンシャル」ということもある。
- 無向確率ネットワークをマルコフネットワークと呼ぶ。

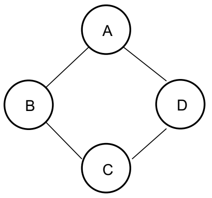
例題

A B	$\phi_1(A,B)$
真 偽	1
真 真	10
偽 真	5
偽 偽	30

B C	$\phi_2(B,C)$
真 偽	1
真 真	100
偽 真	1
偽 偽	100

C D	$\phi_3(C,D)$
真 偽	100
真 真	1
偽 真	100
偽 偽	1

D A	$\phi_4(D,A)$
真 偽	1
真 真	100
偽 真	1
偽 偽	100



$A \perp C | B, D \ and \ B \perp D | A, C$

例題
アリス(A)、ボブ(B)、チャールズ(C)、デビ(D)の４人は、話し合い授業でペアを組んで一緒に学んでいる。ある課題ごとに賛成（真）と反対（偽）の評決をとり、これまでのそれぞれの回数は上の表のとおりになった。ボブとチャールズはすごく意見が合うし、チャールズとデビはまったく意見が合わないようである。
上のファクターとHammersley Cliffordの定理を用いてA, B,C,Dの同時確率分布表を作成せよ。

計算の仕方

$$P(A,B,C,D|G) = \frac{1}{Z}\phi_1(A,B)\cdot\phi_2(B,C)\cdot\phi_3(C,D)\cdot\phi_4(D,A)$$

ここで

$$Z = \sum_{A,B,C,D}\phi_1(A,B)\cdot\phi_2(B,C)\cdot\phi_3(C,D)\cdot\phi_4(D,A)$$

計算アルゴリズム

1. すべての*A,B,C,D*の取りえるパターンについて
   $\phi_1(A,B)\cdot\phi_2(B,C)\cdot\phi_3(C,D)\cdot\phi_4(D,A)$
を計算する。
2.すべての*A,B,C,D*の取りえるパターンについて1.で求めた結果をZとして求め、それぞれのパターンについて
$\frac{1}{Z}\phi_1(A,B)\cdot\phi_2(B,C)\cdot\phi_3(C,D)\cdot\phi_4(D,A)$
を求める。

$\phi_1(A=0,B=0)\cdot\phi_2(B=0,C=0)\cdot\phi_3(C=0,D=0)\cdot\phi_4(D=0,A=0)$

を求めよ。

A B	$\phi_1(A,B)$
真 偽	1
真 真	10
偽 真	5
偽 偽	30

B C	$\phi_2(B,C)$
真 偽	1
真 真	100
偽 真	1
偽 偽	100

C D	$\phi_3(C,D)$
真 偽	100
真 真	1
偽 真	100
偽 偽	1

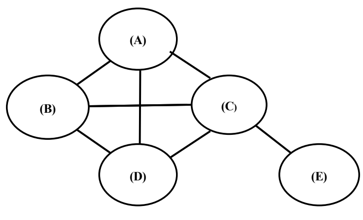
D A	$\phi_4(D,A)$
真 偽	1
真 真	100
偽 真	1
偽 偽	100

$\phi_1(A=0,B=0)\cdot\phi_2(B=0,C=0)\cdot\phi_3(C=0,D=0)\cdot\phi_4(D=0,A=0)$
=30 ·100 ·1 ·100=300, 000

同時確率分布表

				非正規化	正規化
A	B	C	D	1 (,)· 2 (,)· 3 (,)· 4 (,)	P(A,B,C,D)
0	0	0	0	300,000	0.04
0	0	0	1	300,000	0.04
0	0	1	0	300,000	0.04
0	0	1	1	30	$4.1\cdot10^{-6}$
0	1	0	0	500	$6.9\cdot10^{-5}$
0	1	0	1	500	$6.9\cdot10^{-5}$
0	1	1	0	5,000,000	$6.9\cdot10^{-1}$
0	1	1	1	500	$6.9\cdot10^{-5}$
1	0	0	0	100	$1.4\cdot10^{-5}$
1	0	0	1	1,000,000	$1.4\cdot10^{-1}$
1	0	1	0	100	$1.4\cdot10^{-5}$
1	0	1	1	100	$1.4\cdot10^{-5}$
1	1	0	0	10	$1.4\cdot10^{-6}$
1	1	0	1	100,000	$1.4\cdot10^{-2}$
1	1	1	0	100,000	$1.4\cdot10^{-2}$
1	1	1	1	100,000	$1.4\cdot10^{-2}$
合計				7,201,840	1

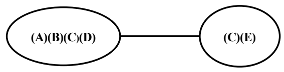
マルコフネットワーク



p(A,B,C,D)

A	B	C	D	p(A,B,C,D)
1	1	1	1	0.06384
1	1	1	1	0.02736
1	1	1	0	0.00336
1	1	1	0	0.00144
1	1	0	1	0.02160
1	1	0	0	0.00240
1	0	1	1	0.3072
1	0	1	0	0.0768
1	0	0	1	0.0
1	0	0	0	0.09600
0	1	1	1	0.01995
0	1	1	1	0.00855
0	1	1	0	0.00105
0	1	1	0	0.00045
0	1	0	1	0.24300
0	1	0	0	0.02700
0	0	1	1	0.00800
0	0	1	0	0.00200
0	0	0	1	0.0
0	0	0	0	0.09000

マルコフネットワーク



p(C,E)

C	E	p(C,D,E)
1	1	0.3634
1	0	0.1560
0	1	0.0
0	0	0.4800

マルコフネットワークの問題

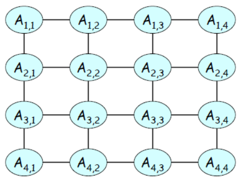
- 工夫すればベイジアンネットワークとほぼ等価に表現できる。
- しかし、クリークが大きい場合はパラメータ数は指数的に増え計算量　大
- より　計算が簡易なモデルの開発
- ⇓
- マルコフ確率場

マルコフ確率場（Pairwise Markov Random Field）

各ノードのファクター集合($\phi(X_i), i = 1, \cdots, N$)と各エッジのファクター($\phi(X_i, X_j), (X_i, X_j) \in E$)によってファクターが定義されるモデル。クリークが2ノードのみによって構成されるという制約。

右図は完全グラフのマルコフ確率場。

$p(A_{1,1}, \cdots, A_{4,4}|G) =$
$\frac{1}{Z}\phi_{12}(A_{1,1}, A_{1,2}) \cdot \phi_{21}(A_{1,1}, A_{2,1})$
$\cdots \phi_{34}(A_{3,4}, A_{4,4})$

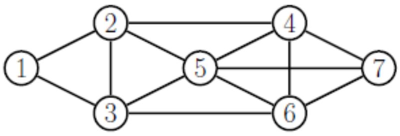


マルコフ確率場

- 確率構造に無向グラフ構造を持つグラフィカルモデルをマルコフ確率場(Markov Random Field) もしくはマルコフネットと呼ぶ。
- $I(A, B|C)_G$:グラフGでCを除くとAとBを結ぶ路（Path) がないとき、Cを所与としてAとBは条件付き独立である。

例

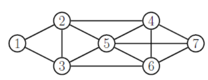
- $I(1,7|2,3)_G$　もしくは　$1 \perp 7|2,3$
- $I(1,7|4,5,6)_G$もしくは　$1 \perp 7|4,5,6$
- $I(2,6|3,4,5)_G$もしくは　$2 \perp 6|3,4,5$



マルコフ　ブランケット

- 定義
- ノード集合Xのマルコフ　ブランケットは、Xの要素のすべての隣接ノード集合
- 例
- 5のマルコフブランケットは

{2,3,4,6,7}



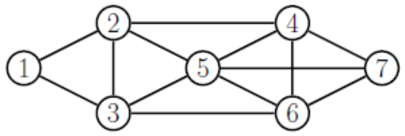
同時確率分布計算例

- マルコフネットワーク

$$p(x_1,\cdots,x_7|G)$$
$$=\frac{1}{Z}\phi_{123}(x_1,x_2,x_3)\phi_{235}(x_2,x_3,x_5)\phi_{245}(x_2,x_4,x_5)\phi_{356}(x_3,x_5,x_6)\phi_{4567}(x_4,x_5,x_6,x_7)$$

- マルコフ確率場

$$p(x_1,\cdots,x_7|G)$$
$$=\frac{1}{Z}\phi_{12}(x_1,x_2)\phi_{13}(x_1,x_3)\phi_{23}(x_2,x_3)\phi_{24}(x_2,x_4)\phi_{25}(x_2,x_5)\phi_{35}(x_3,x_5)\phi_{36}(x_3,x_6)$$
$$\phi_{45}(x_4,x_5)\phi_{46}(x_4,x_6)\phi_{47}(x_4,x_7)\phi_{56}(x_5,x_6)\phi_{57}(x_5,x_7)\phi_{67}(x_6,x_7)$$

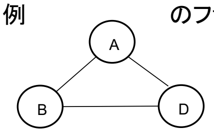


マルコフ確率場の利点

1. 変数間で対象であるので、方向を持たないような変数（位置データや関係性データ）を扱う場合に自然である
2. 自然言語処理における自動ラベリング問題では、隠れマルコフなどの有向グラフよりも条件付き確率場(Conditional Random Field)のほうがラベルバイアス問題がなくなり良い

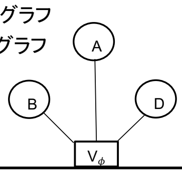
Factor Graph

- ○で示される変数ノード（variable nodes）と□で示されるファクターノード（factor nodes）によって表現されるグラフで構造を可視化する。
- 例　　　　　　のファクターフラフを考える。

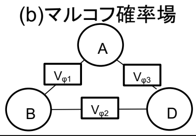


- ファクターグラフ

(a) マルコフグラフ



(b)マルコフ確率場



Factor Graphの利点

- Factor Graphはファクター構造を明示することができる。例えば、完全グラフであるN変数マルコフランダムフィールドでは、$\binom{N}{2}$個のファクターノードが明示される。

ファクターのParameterization

- ベイジアンネットワークでは、条件付き確率表がパラメータであった。ファクター$\phi_c(x_c|\theta_c)$はどのようにパラメータ化するのであろうか？

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

各ノードのファクター集合($\phi(X_i), i = 1, \cdots, N$)と各エッジのファクター($\phi(X_i, X_j), (X_i, X_j) \in E$)によってファクターが定義される。

$$E = \sum_{i} \phi(X_i) + \sum_{(i,j)} \phi(X_i, X_j)$$

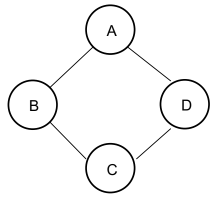
例題

A B	$\phi_1(A,B)$
真 偽	1
真 真	10
偽 真	5
偽 偽	30

B C	$\phi_2(B,C)$
真 偽	1
真 真	100
偽 真	1
偽 偽	100

C D	$\phi_3(C,D)$
真 偽	100
真 真	1
偽 真	100
偽 偽	1

D A	$\phi_4(D,A)$
真 偽	1
真 真	100
偽 真	1
偽 偽	100



$A \perp C | B, D \ and \ B \perp D | A, C$

例題
アリス(A)、ボブ(B)、チャールズ(C)、デビ(D)の４人は、話し合い授業でペアを組んで一緒に学んでいる。ある課題ごとに賛成（真）と反対（偽）の評決をとり、これまでのそれぞれの回数は上の表のとおりになった。ボブとチャールズはすごく意見が合うし、チャールズとデビはまったく意見が合わないようである。
上のファクターからエネルギー関数を求めよ。

例題　次のファクターから　エネルギー関数を求めよ。

ファクター

A B	$\phi_1(A,B)$
真 偽	1
真 真	10
偽 真	5
偽 偽	30

B C	$\phi_2(B,C)$
真 偽	1
真 真	100
偽 真	1
偽 偽	100

C D	$\phi_3(C,D)$
真 偽	100
真 真	1
偽 真	100
偽 偽	1

D A	$\phi_4(D,A)$
真 偽	1
真 真	100
偽 真	1
偽 偽	100

エネルギー関数

A B	$\phi_1(A,B)$
真 偽	0
真 真	-2.3
偽 真	-1.61
偽 偽	-3.4

B C	$\phi_2(B,C)$
真 偽	0
真 真	-4.61
偽 真	0
偽 偽	-4.61

C D	$\phi_3(C,D)$
真 偽	-4.61
真 真	0
偽 真	-4.61
偽 偽	0

D A	$\phi_4(D,A)$
真 偽	0
真 真	-4.61
偽 真	0
偽 偽	-4.61

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

エネルギーが0は、まったく同時確率分布に寄与しないことを意味する

Log-Linear マルコフネットワーク

一般に

- 特徴集合(a set of features) $F = \{f_1(C_1), \cdots, f_k(C_k)\}$
- C_i: i 番目のクリーク
- 重み集合 $w_1, \cdots, w_k$

としたとき、分布Pが

$$P(x_1, x_2, \cdots, x_N | G) = \frac{1}{Z(\theta)} \exp\left(-\sum_{i=1}^{k} w_i f_i(C_i)\right)$$

で表現できるとき、Log-linear マルコフネットワークと呼ぶ。
マルコフ確率場の定義を満たすものをLog-linearマルコフ確率場と呼ぶ。

イジングモデル（Ising Model）

最初に提案されたLog-linear マルコフ確率場の一つ。原子の動作を表現するための統計物理モデル。それぞれ{+1,-1}をとる変数集合(x_i)がある。各二変数間(x_i, x_j)のエッジのエネルギー関数は重みw_{ij}として

$$E(X_i, X_j) = w_{ij} \, x_i x_j$$

で示される。すなわち、$x_i = x_j$（二つが同じ回転）のとき、$E(X_i, X_j) = w_{ij}$となり、それ以外のとき$E(X_i, X_j) = -w_{ij}$となる。

Ising Model

一般的に以下で表現される。

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

$w_{ij} > 0$のとき、モデルは$x_i = x_j$（二つが同じ回転）を好み、強磁性（ferromagnetic）と呼ばれる。$w_{ij} < 0$のとき、モデルは$x_i \neq x_j$（二つが逆の回転）を好み、反強磁性（antiferromagnetic）と呼ばれる。
$w_{ij} = 0$のとき、二つの原子にはまったく相互作用がなく、無相互作用(non-interaction)と呼ばれる。

ボルツマン分布

- 熱平衡状態にある温度 T の系がエネルギー E を取る確率は
- $P(E) = \frac{1}{Z}\exp(-\frac{E}{T})$
- で与えられるとき、この確率分布をBolzmann分布と呼ぶ。エネルギーがパラメータの線形回帰であれば指数型分布族となる。
- 当然、ギブス分布の一つである。

ボルツマンマシン

{0,1}をとる変数集合(x_i)(ニューロンが発火すると1、それ以外0)がある。各二変数間(x_i, x_j)のエッジのエネルギー関数は重みw_{ij}としてx_jの入力を所与として、以下のように発火確率が求められる。

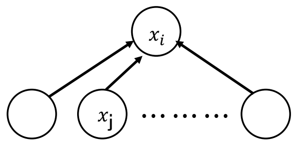
$$P(x_1, x_2, \cdots, x_N|\theta) = \frac{1}{Z(\theta)}\exp\left(-\sum_{(i,j)} w_{ij}\, x_i x_j - \sum_i b_i\, x_i\right)$$

ボルツマンマシンはイジングモデルの{0,1}を取るのみに修正した変形モデルである。また、隠れ変数も許す。(x_i, x_j)がともに発火した1の場合のみに$\sum_{(i,j)} w_{ij}\, x_i x_j = 1$となる。

ボルツマンマシン

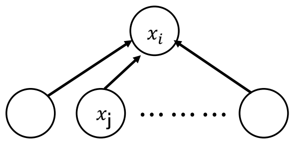
各二変数間(x_i, x_j)のエッジのエネルギー関数は重みw_{ij}としてx_jの入力を所与として、以下のように発火確率が求められる。

$$P\left(x_i = 1 \middle| x_j \neq x_i\right) = \frac{1}{1 + \exp\left(-\sum_j w_{ij} x_j - b_i\right)}$$

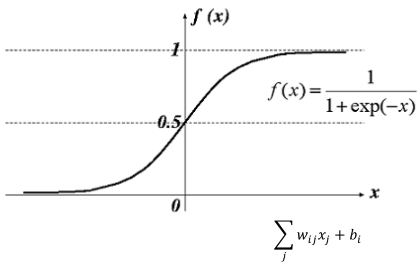


ボルツマンマシン

$$P\left(x_i = 1 \middle| x_j \neq x_i\right) = \frac{1}{1 + \exp\left(-\sum_j w_{ij} x_j - b_i\right)}$$

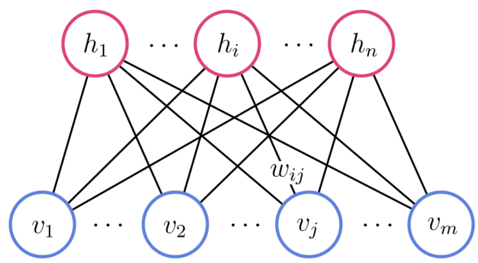


シグモイド関数



制限ボルツマンマシン

隠れ変数



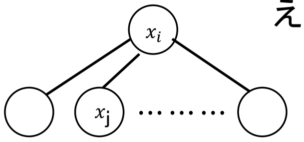
観測変数

まとめると

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



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

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

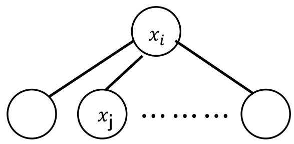
$$= \frac{1}{1 + \exp\left(-\sum_i E\ (x_i = 0) - \sum_{(i,j)} E(x_i = 0, x_j)\right)}$$

$E\ (x_i = 0)$=b_i, $E(x_i = 0, x_j)$=$w_{ij} x_j$とおくと

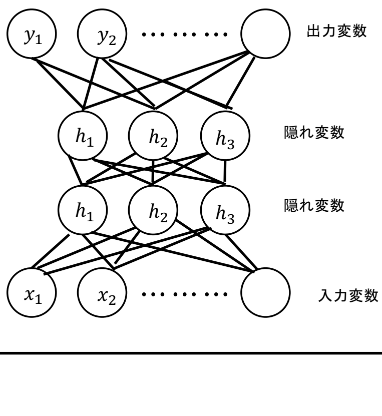
ボルツマンマシン（ニューラルネットワーク）

$E\ (x_i = 0)$=b_i, $E(x_i = 0, x_j)$=$w_{ij} x_j$とおくと

$$P\left(x_i = 1 | x_j \neq x_i\right) = \frac{1}{1 + \exp\left(-\sum_j w_{ij} x_j - b_i\right)}$$



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.

隠れ変数の有効性

定理 Ueno (2011, UAI)

ベイジアンネットワーク学習において、疎なデータ(パラメータ数に対してデータ数が少ない)の場合、周辺尤度最大化で得られる構造よりもより冗長な構造が真の構造である。

⇛

パターン数が多いグラフィカルモデルはほとんどがデータ数が足りない。隠れ変数を階層的に導入するとパラメータ推定が可能な範囲で爆発的にパラメータ数を増やせる。

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。

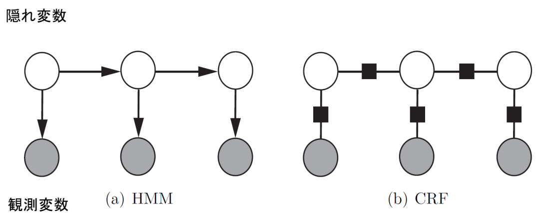
- 隠れ変数の数と構造を最適にする規準（スコア）は何か？
- ベイジアンネットワークで得られる構造のパラメータ数とどのような関係にあるのか？
- 数学的に解明できるのか？
- その構造を学習できるのか？

条件付き確率場(CRF; Conditional Random Field)

- MRFによる条件付確率の予測で分類器の一つ。

- $P(Y|X) = P(y_1, y_2, \cdots, y_N | x_1, x_2, \cdots, x_N)$
- $= \frac{1}{\sum_Y P(Y,X)} P(Y,X) = \frac{1}{\sum_Y P(Y,X)} \prod_i \phi_i(Y,X)$

隠れマルコフモデルHMMとの違い



HMM

- $P(X,Y) = \prod_{t=1}^{T} P(y_t|y_{t-1}) P(x_t|y_t)$

このモデルは同時確率分布を得ることはでき、データ発生モデルではあるが分類器（識別機）ではない。

HEMM(McCallum et al 2000)

- HMMの識別モデル
- $P(Y|X) = \prod_{t=1}^{T} P(y_t|y_{t-1}, X)$

HMMのエッジの向きを逆にしたモデル

HMM分類器とCRF分類器

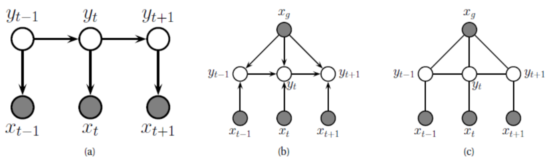
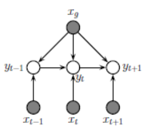


Figure 19.14 Various models for sequential data. (a) A generative directed HMM. (b) A discriminative directed MEMM. (c) A discriminative undirected CRF.

CRF は系列ラベリングに有効

- HMMなどの有向グラフでは、後の変数データは前の変数の推論に影響しない。
- HMM分類器では x_t Biasとy_{t-1}はd分離されている(Label Bias, Lafferty et al 2001)。
-  HMMは文脈などが反映されない。

例

- 自動翻訳 (bank)
- 意味 銀行、堤防、丘

そのあとに fishという単語が出た場合、この単語がCRFでは「堤防」の意味であることがわかるが、HMMでは反映されない。

グラフィカルモデル

- ベイジアンネットワーク

Naïve Bayes, Augmented Naïve Bayes, TAN, LDA, Markov model, HMM, HEMM

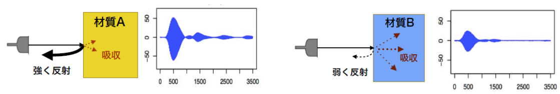
- マルコフネットワーク

マルコフ確率場、(深層)ボルツマンマシン、条件付き確率場

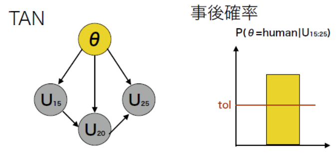
応用研究

超音波データ解析に基づく ベイズ的人体検出ロボットの開発

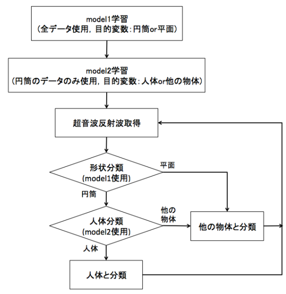
超音波の反射波の形状の違いをもとに人体を検出

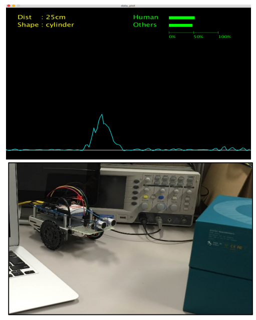


反射波のデータから学習した
ベイジアンネットワーク分類器により判定

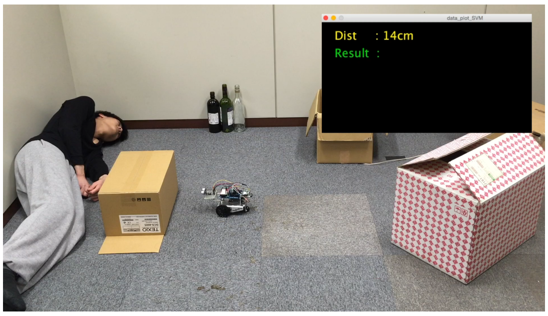


超音波データ解析に基づく ベイズ的人体検出ロボットの開発



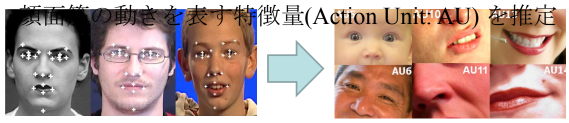


災害時などの視界が悪い環境下で有効

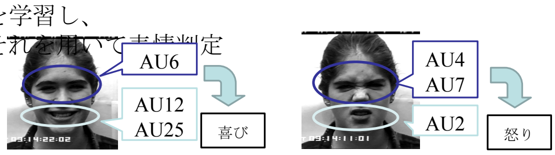


顔表情認識技術の開発

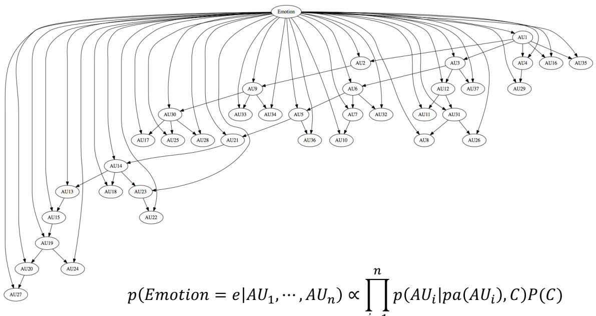
カメラから得られる目・鼻・口などの特徴点のデータから、
顔面筋の動きを表す特徴量(Action Unit: AU) を推定



特徴量を変数とするベイジアンネットワーク分類器を学習し、
それを用いて表情判定



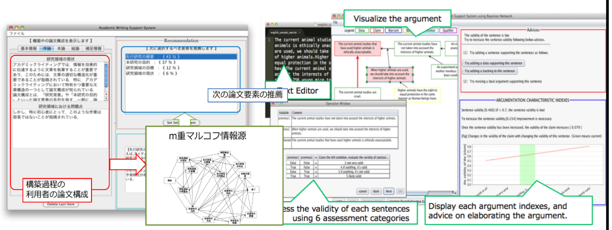
学習したベイジアンネットワーク分類器



$$p(Emotion = e|AU_1, \cdots, AU_n) \propto \prod_{i=1}^{n} p(AU_i|pa(AU_i), C)P(C)$$

ネットワーク構造を所与とした
目的変数の事後確率から表情を推定
８９％の正確度

自然言語処理



自然言語処理、線形計画法とベイジアンネットワークにより文章の論理構造を自動理解するシステムの開発
論述式試験の自動評価などに適用可能

ビッグデータ時代のデータサイエンス

- データ数の増加により十分なデータが与えられていることを意味しない。
- 精緻な現象予測のために、変数の組み合わせ爆発が起こり、計算量の爆発、データ不足を如何に解決するかの学問。

- ご清聴ありがとうございました！！