確率伝播アルゴリズムとは

— 脳の認識機構と関係が深い「確率伝播アルゴリズム」について、極力分かりやすく解説する —

2010-02-26

産業技術総合研究所 脳神経情報研究部門

一杉 裕志

内容

- 確率論の基礎知識の復習
- ベイジアンネットとは
- 確率伝播アルゴリズム

背景

- ベイジアンネットを用いて大脳皮質の機能を実現する研究が大きく進展しつつある。しかし関連論文はベイジアンネットの基礎知識がないと重要性を理解しずらい。
- ベイジアンネットの基本事項は数学的には実は難しくない。四則演算しか使わない。
- しかし確率論の独特の記法のため独学には敷居が高い。本資料ではできるかぎり親切に説明することを試みる。

確率論の基礎知識の復習

確率論の基礎知識の復習

- 確率変数 X が x という値を取る確率：$P(X = x)$
  - 省略した記法：$P(x)$
- 確率変数 X の確率分布：　$P(X)$
- 同時確率：　$P(X, Y)$
- 周辺化：　$P(X) = \sum_{Y} P(X, Y)$
- 条件付確率：　$P(X, Y) = P(X \mid Y) P(Y)$
- ベイズの定理：$P(Y \mid X) = P(X \mid Y) P(Y) / P(X)$

P ってなに？

- 中に書かれる文字列と文脈によって異なる意味を持つ。
- 仮に「確率変数 X が値 x をとる確率」を表す関数を p_X(x) とすると:
  - 確率 P(X=x) （略記 P(x)）は値 p_X(x) のこと。
  - 確率分布 P(X) は関数 p_X のこと。
- 例えば $P(X)=\sum_{Y} P(X,Y)$ は、こういう意味:

$$\mathrm{p\_X} = g \quad where \quad g(x) = \sum_{y \in Domain(Y)} \mathrm{p\_X\_Y}(x, y)$$

- 参考:「確率の記法 - カーネル法Wiki」

http://www.neurosci.aist.go.jp/~akaho/kernel/index.php?%B3%CE%CE%A8%A4%CE%B5%AD%CB%A1

Ｐ以外も要注意

- Ｐ単独で出てきて関数 p_X を表すこともある。小文字のｐやｑも同様。
- Ｐと同様に、Ｘごとに定義される関数が $\lambda_X(x)$ と書かれずに $\lambda(x)$ と書かれたり、ＸとＹの組み合わせごとに定義される関数が $\lambda_{YX}(x)$ と書かれずに $\lambda_Y(x)$ と書かれることもある。
  - （本資料ではこういう省略は避けている。）
  - プログラミング言語的に解釈すると、すべての確率変数は型が違うから関数名の多重定義が可能ということかもしれない。

Σの文法

- 総和の記号Σは足し算よりは優先度が高いが掛け算よりは低い。

$$A\sum_B CD + E\sum_F GH = A\left(\sum_B (CD)\right) + E\left(\sum_F (GH)\right)$$

- 右から左に結合。

$$\sum_A B\sum_C D\sum_E F = \sum_A\left(B\left(\sum_C\left(D\left(\sum_E F\right)\right)\right)\right)$$

- 総積の記号ΠもΣと同様。

$$\sum_A B\prod_C D = \sum_A\left(B\prod_C D\right)$$

少なくとも[Pearl 1988]と「ビショップ本」はこういう文法のようである。
ただしたまにΠが掛け算よりも優先度が高い（ΠaΠb=(Πa)(Πb)）こともあるので要注意。

周辺化(marginalization)について

- 周辺化とは、いくつかの確率変数を「無視」(marginalize)した同時確率を求めること。
  - marginalize :社会的に無視[過小評価]する, 主流から追いやる
    三省堂 EXCEED英和辞典 http://dictionary.goo.ne.jp/leaf/ej2/43872/m0u/marginalize/

$$P(X)=\sum_Y P(X,Y)$$

例:

P(X=0,Y=0) =0.2	P(X=1,Y=0) =0.1	P(Y=0)=0.3
P(X=0,Y=1) =0.4	P(X=1,Y=1) =0.3	P(Y=1)=0.7
P(X=0)=0.6	P(X=1)=0.4	

- 変数がたくさんあっても同様。

$$P(C,E)=\sum_A\sum_B\sum_D\sum_F P(A,B,C,D,E,F)$$

- 条件付確率でも同様。

$$P(X\mid Z)=\sum_Y P(X,Y\mid Z)$$

- これはだめ:

$$P(X\mid Z)=\sum_Y P(X\mid Y,Z)???$$

ベイズの定理について

対称性から $P(X,Y) = P(Y,X)$

条件付確率の定義から $P(X,Y) = P(X \mid Y)P(Y)$

$$P(Y,X) = P(Y \mid X)P(X)$$

よって $P(X \mid Y)P(Y) = P(Y \mid X)P(X)$

P(X)>0 ならば両辺を P(X) で割って、ベイズの定理が得られる。

$$P(Y \mid X) = P(X \mid Y)P(Y) / P(X)$$

なお、 $P(X) = \sum_{Y} P(X,Y) = \sum_{Y} P(X \mid Y)P(Y)$ なので、

こうも書く: $$P(Y \mid X) = \frac{P(X \mid Y)P(Y)}{\sum_{Y} P(X \mid Y)P(Y)}$$

X の値が与えられているなら、分子はΣP(Y|X)=1にするための正規化定数と考えて、こうも書く。

$$P(Y \mid X) = \frac{1}{\alpha} P(X \mid Y)P(Y)$$

ベイジアンネットとは

同時確率が分かっていればいろいろなことが計算できる

- 例 : 芝生が濡れているとき、雨が降った確率を求めよ。ただし P(S,W,R,C) は与えられているものとする。

S: スプリンクラーが動いた
W: 芝生が濡れている
R: 雨が降った
C: 雲が出ている

$$P(R = yes \mid W = yes)$$

$$= \frac{P(W = yes, R = yes)}{P(W = yes)}$$ 条件付確率の定義より

$$= \frac{\sum_S \sum_C P(S, W = yes, R = yes, C)}{\sum_S \sum_R \sum_C P(S, W = yes, R, C)}$$ 周辺化

同時確率の素朴な記憶の仕方

- 全ての値をテーブルで持つ。
- 変数が n 個あると $O(2^n)$ のサイズが必要。
  → n が大きくなると**実用的でない**！

S	W	R	C	P(S,W,R,C)
no	no	no	no	0.68
no	no	no	yes	0.08
…	…	…	…	…
yes	yes	yes	yes	0.02

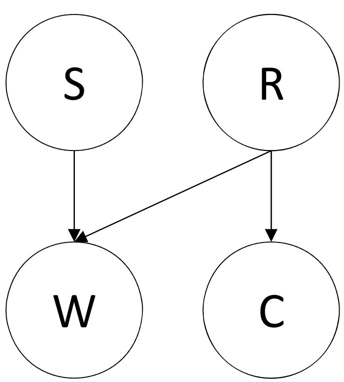
サイズ $2^4=16$

ベイジアンネット

- 同時確率の表を小さなサイズの条件付確率表の積に分解して**効率的**に（近似）表現できる**場合がある**。
→ そうやって表現したものがベイジアンネット

例：$P(S,W,R,C) = P(W \mid S,R)P(C \mid R)P(S)P(R)$

上の式の内容を図で書くとこう：



必要な条件付確率表：

P(S=yes)
0.2

P(R=yes)
0.02

S	R	P(W=yes|S,R)
no	no	0.12
no	yes	0.8
yes	no	0.9
yes	yes	0.98

R	P(C=yes|R)
no	0.3
yes	0.995

サイズ 4+2+1+1=8

矢印について

- 矢印は必ずしも「因果関係」である必要はない。
- 矢印は推論時の情報の流れではない。あとで見るように、推論時には逆向きにも情報が流れる。
- n 個の確率変数の同時確率は実は n(n-1)/2 本の矢印を使って必ず分解できるが、その場合は効率は良くならない。少ない矢印で表現できてはじめて効率がよくなる。

確率伝播アルゴリズム

確率伝播アルゴリズム(belief propagation algorithm)とは

- ベイジアンネットにおいて、確率変数の事後確率を効率的に計算するアルゴリズム。
  - 「伝播」は正しくは「でんぱ」と読むが「でんぱん」と誤読されることも多い。「伝搬」は誤読に起因する誤字。
  - 信念伝播アルゴリズムとも訳される。

原理

- 掛け算の足し算に対する分配法則:

  a(b+c) = ab + ac

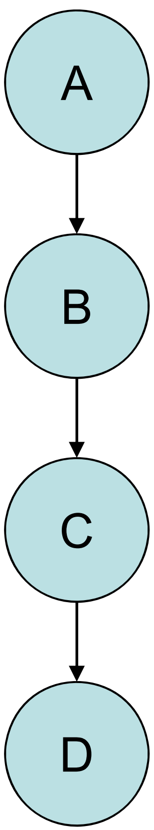
  – 左辺は演算2回だが右辺は演算3回

  $$a\sum_{i=1}^{n} b_i = \sum_{i=1}^{n} ab_i$$

  – 左辺は演算 n 回だが右辺は 2n-1 回。
- 因子のくくりだしを積極的に適用すれば、事後確率計算に必要な演算数を劇的に減らせる。
- さらに、演算の中間結果の重複計算を避けることで、効率を上げている。

簡単なベイジアンネットの例

- 簡単なネットワークならば、それ専用の確率伝播アルゴリズムの導出は難しくない。



  - このネットワークを例に、D の観測値が与えられた時の事後確率 P(A|D), P(B|D), P(C|D) を計算するアルゴリズムを導出してみる。

$$P(A,B,C,D)$$
$$= P(D|C)P(C|B)P(B|A)P(A)$$

事後確率の計算式

$$P(C \mid D) = \frac{P(C,D)}{P(D)} = \frac{1}{\alpha}\sum_{B}\sum_{A} P(A,B,C,D)$$

ただし正規化定数
$\alpha = P(D)$ とする

$$= \frac{1}{\alpha}\sum_{B}\sum_{A}\big(P(D \mid C)P(C \mid B)P(B \mid A)P(A)\big)$$

$$= \frac{1}{\alpha}\underline{P(D \mid C)}\left(\sum_{B}\underline{P(C \mid B)}\left(\sum_{A} P(B \mid A)P(A)\right)\right)$$

P(D|C)はB, A に依存せず、
P(C|B)はAに依存しないので
Σの外にくくりだせる。

同様に

$$P(B \mid D) = \frac{1}{\beta}\left(\sum_{C} P(D \mid C)P(C \mid B)\right)\left(\sum_{A} P(B \mid A)P(A)\right)$$

$$P(A \mid D) = \frac{1}{\gamma}\left(\sum_{B}\left(\sum_{C} P(D \mid C)P(C \mid B)\right)P(B \mid A)\right)P(A)$$

注：因子のくくり
だし方は一通り
とは限らないが、
ここではこうくく
りだす。

メッセージ

- 関数 $m_{AB}, m_{BC}, m_{DC}, m_{CB}, m_{BA}$ を下記のように定義。m_{XY} を「XからYへのメッセージ」と呼ぶ。

$$m_{AB}(a) = P(A = a)$$

$$m_{BC}(b) = \sum_{A} P(B = b \mid A)P(A) = \sum_{A} P(B = b \mid A) m_{AB}(A)$$

$$m_{DC}(c) = P(D = d \mid C = c)$$

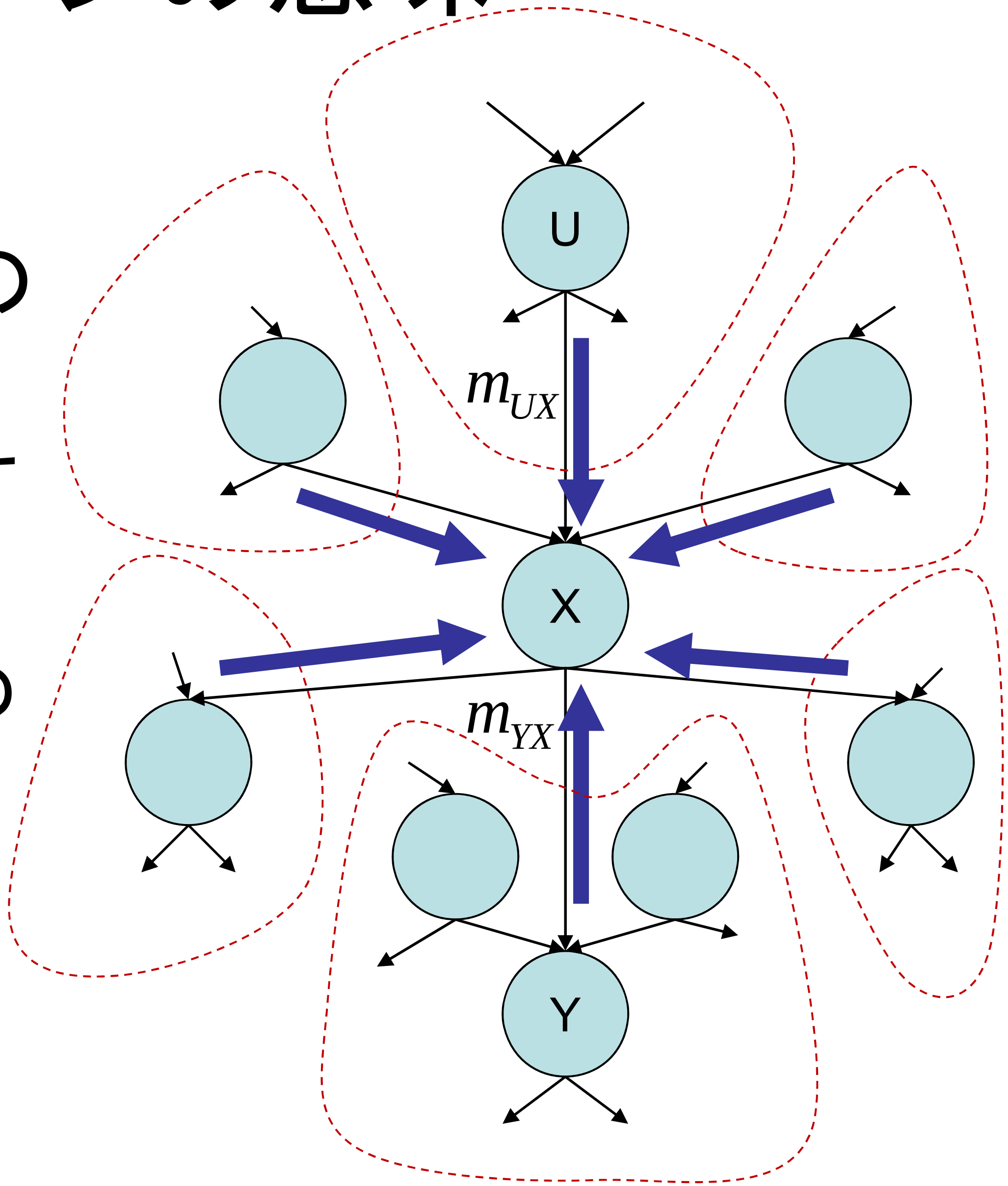
$$m_{CB}(b) = \sum_{C} P(D = d \mid C)P(C \mid B = b) = \sum_{C} m_{DC}(C)P(C \mid B = b)$$

$$m_{BA}(a) = \sum_{B} \left(\sum_{C} P(D = d \mid C)P(C \mid B) \right) P(B \mid A = a)$$

$$= \sum_{B} m_{CB}(B)P(B \mid A = a)$$

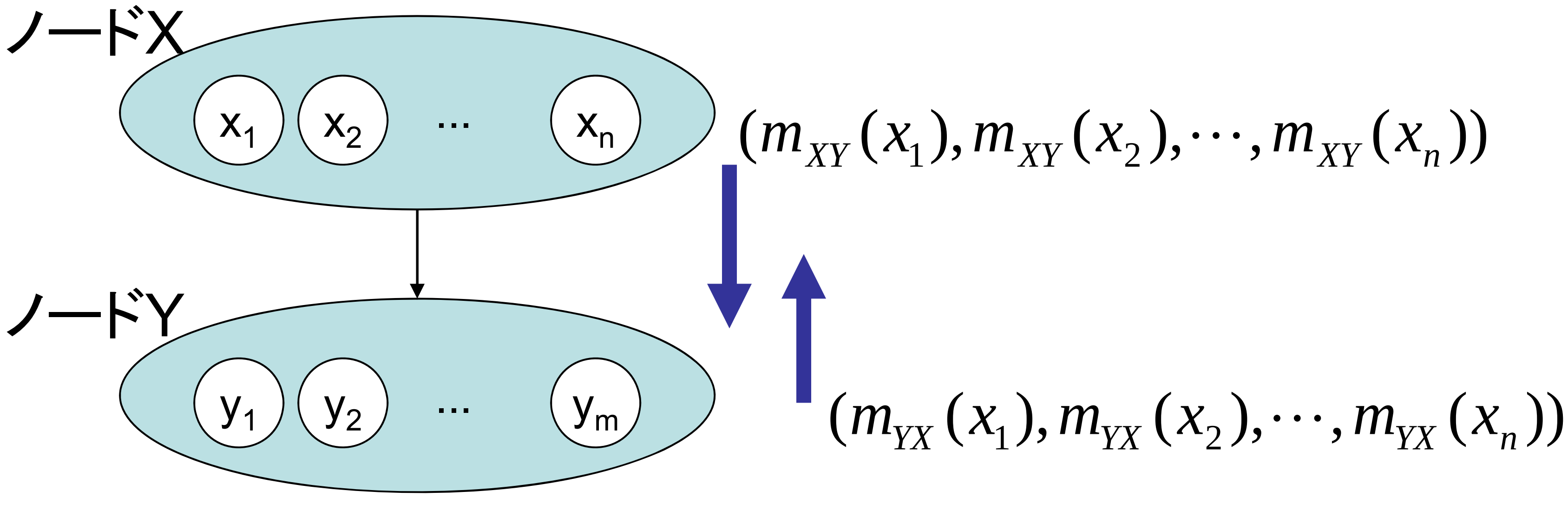
メッセージの意味

- メッセージ m_{ZX}は、事後確率 P(X|D) の計算に必要な情報のうち、X に隣接するノード Z より先にあるものを全て集めたもの。



メッセージの中身

- ノード X が n 個の値を取り得るとすると、$m_{XY}(x)$ 、$m_{YX}(x)$ として送られるものは実質的には n 次元数値ベクトル。



確率伝播アルゴリズム

$$m_{AB}(a) = P(a)$$

$$m_{BC}(b) = \sum_a P(b \mid a) m_{AB}(a)$$

$$m_{DC}(c) = P(d \mid c)$$

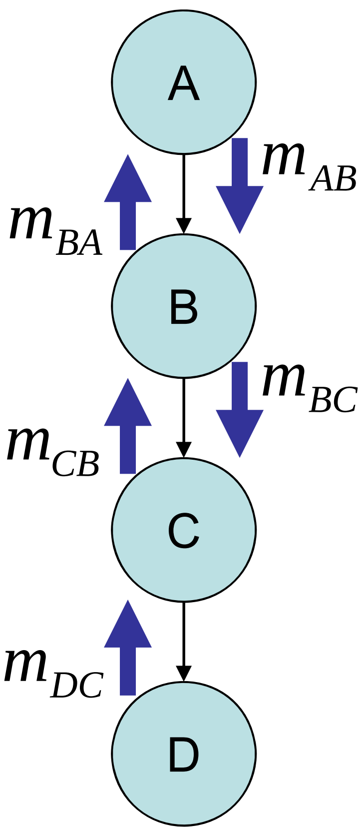
$$m_{CB}(b) = \sum_c m_{DC}(c) P(c \mid b)$$

$$m_{BA}(a) = \sum_b m_{CB}(b) P(b \mid a)$$

$$P(c \mid d) = \frac{1}{\alpha} m_{DC}(c) \sum_b P(c \mid b) m_{BC}(b)$$

$$P(b \mid d) = \frac{1}{\beta} m_{CB}(b) m_{BC}(b)$$

$$P(a \mid d) = \frac{1}{\gamma} m_{BA}(a) m_{AB}(a)$$



事後確率の計算式を、メッセージを使った式に書きかえると、確率伝播アルゴリズムが得られる。

注：引数を小文字（値）に統一したが、大文字（確率変数）にしても意味するところは変わらない。

確率伝播アルゴリズムの性質

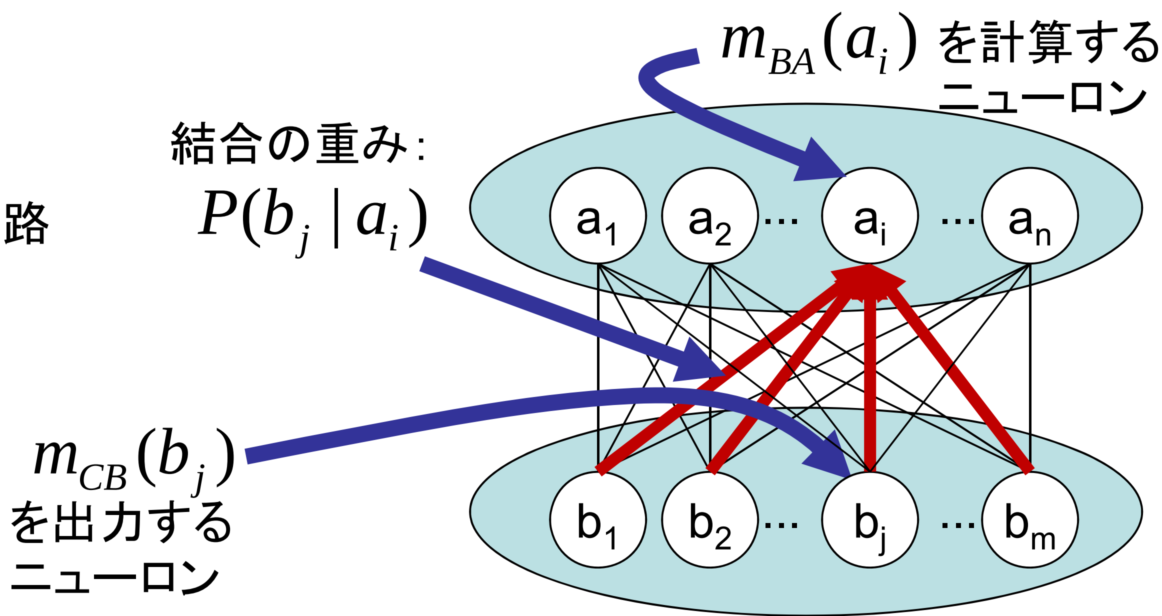
- 一般に、X が送るメッセージおよび X の事後確率は、X に送られてくるメッセージと X の条件付確率表だけから計算できる！（ただしメッセージを送る順序には注意。）
  - 局所的な情報だけを使ってアルゴリズムが実行されると言う意味で、神経回路と似ている。
- ここでは観測変数を D としたが、どれが観測変数であっても、観測変数以外のノードがやる計算は変わらない！
  - １つの固定した神経回路を使って様々な推論が行えることを示している。

行われている計算

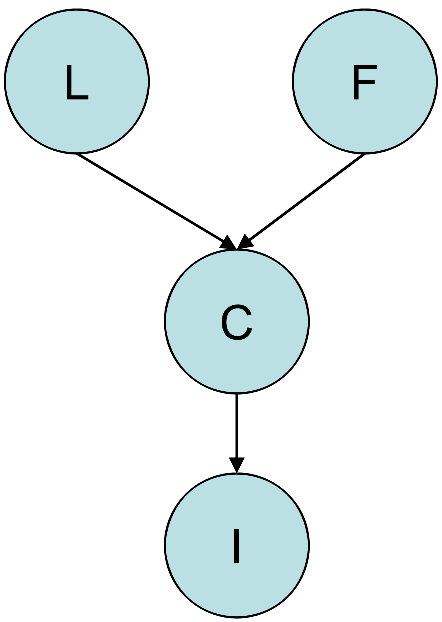
- $\sum_a P(b|a) m_{AB}(a)$ とか $\sum_b m_{CB}(b) P(b|a)$ という形の計算がたくさん出てくるところに注目。これは2つのベクトルの内積計算。
  - 簡単な神経回路で実現可能。
    - （ただし複雑なベイジアンネットではこうならないので注意。）

例：右図は下記の計算をする神経回路

$$m_{BA}(a) = \sum_b m_{CB}(b) P(b|a)$$



もう1つ簡単な例:
[Rao 2005] のネットワーク



$$P(I, C, L, F)$$
$$= P(I \mid C) P(C \mid L, F) P(L) P(F)$$

同時確率の計算式

$$P(C\,|\,I) = \frac{P(I,C)}{P(I)} = \frac{1}{\alpha}\sum_{L}\sum_{F} P(I,C,L,F)$$

ただし正規化定数 $\alpha = P(I)$

$$= \frac{1}{\alpha}\sum_{L}\sum_{F} P(I\,|\,C)P(C\,|\,L,F)P(L)P(F)$$

$$= \frac{1}{\alpha}\underline{P(I\,|\,C)}\sum_{L}\sum_{F} P(C\,|\,L,F)P(L)P(F)$$

L と F に依存しない
因子 P(I|C) をΣの外にくくりだす。

同様に

$$P(L\,|\,I) = \frac{1}{\beta}\left(\sum_{C} P(I\,|\,C)\left(\sum_{F} P(C\,|\,L,F)P(F)\right)\right)P(L)$$

$$P(F\,|\,I) = \frac{1}{\gamma}\left(\sum_{C} P(I\,|\,C)\left(\sum_{L} P(C\,|\,L,F)P(L)\right)\right)P(F)$$

メッセージ

- 関数 $m_{LC}, m_{FC}, m_{IC}, m_{CL}, m_{CF}$ を下記のように定義。

$$m_{LC}(L) = P(L)$$

$$m_{FC}(F) = P(F)$$

$$m_{IC}(C) = P(I = I' | C)$$

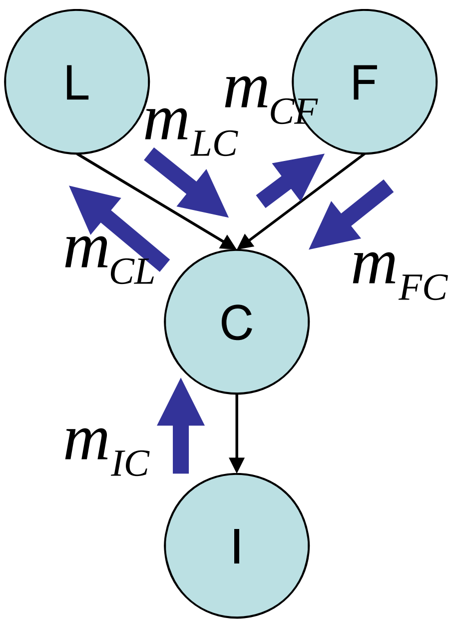
$$m_{CL}(L) = \sum_C P(I = I' | C) \left(\sum_F P(C | L, F) P(F) \right)$$

$$= \sum_C m_{IC}(C) \sum_F P(C | L, F) m_{FC}(F)$$

$$m_{CF}(F) = \sum_C P(I = I' | C) \left(\sum_L P(C | L, F) P(L) \right)$$

$$= \sum_C m_{IC}(C) \sum_L P(C | L, F) m_{LC}(L)$$

確率伝播アルゴリズム



$$m_{LC}(l) = P(l)$$

$$m_{FC}(f) = P(f)$$

$$m_{IC}(c) = P(i \mid c)$$

$$m_{CL}(l) = \sum_c m_{IC}(c) \sum_f P(c \mid l, f) m_{FC}(f)$$

$$m_{CF}(f) = \sum_c m_{IC}(c) \sum_l P(c \mid l, f) m_{LC}(l)$$

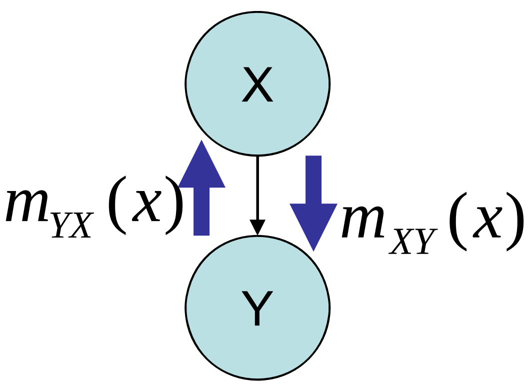
$$P(c \mid i) = \frac{1}{\alpha} m_{IC}(c) \sum_f \sum_l P(c \mid l, f) m_{LC}(l) m_{FC}(f)$$

$$P(l \mid i) = \frac{1}{\beta} m_{CL}(l) m_{LC}(l)$$

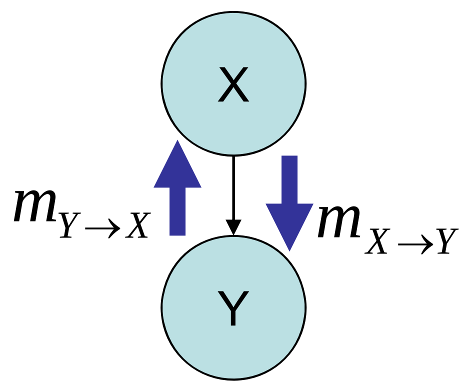
$$P(f \mid i) = \frac{1}{\gamma} m_{CF}(f) m_{FC}(f)$$

メッセージの様々な記法

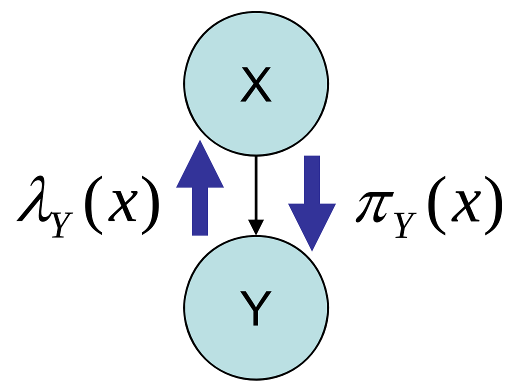
- どれも意味は同じ。



本資料



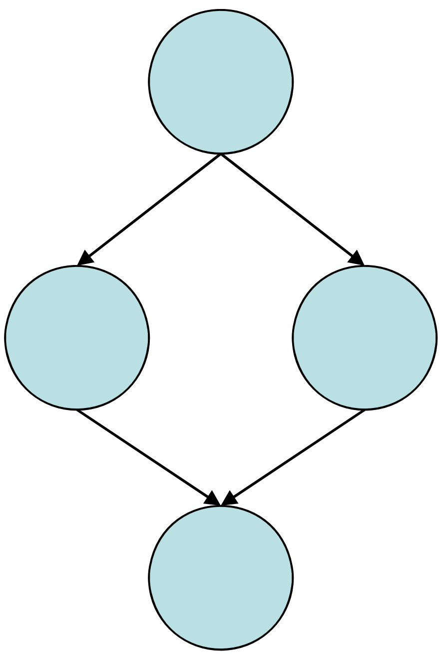
[Rao 2005]
[Chikkerur, Serre and Poggio 2009]



[Pearl 1988]
[George and Hawkins 2005]
[Ichisugi 2007]

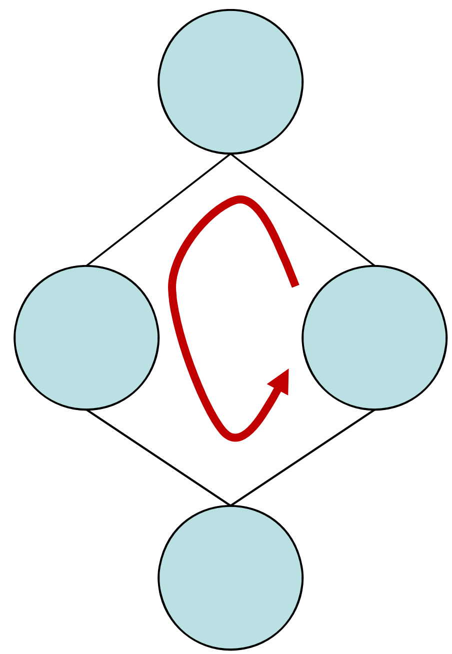
「loop のある」ベイジアンネット

- 矢印の向きを無視するとサイクル（循環構造）があるベイジアンネットを「loop がある」と言う。



loop のあるベイジアンネットの例

矢印の向きを無視したときサイクルがある。



loopy belief propagation

- ベイジアンネットに loop があると、確率伝播アルゴリズムは使えない。
  - メッセージの計算式が相互依存して計算できない。
- それでもあえて確率伝播アルゴリズムを適用し、適当な初期値からはじめてメッセージの再計算・伝播を繰り返すと、多くの場合メッセージの値が収束する。
  - これを loopy belief propagation と呼び、近似解法としてよく使われている。
  - 伝播の順序は重要でない。→ 非同期神経回路！

マルコフ確率場 (Markov random field)

- ベイジアンネットはノード間の結合に向きがある（非対称な結合）。向きのない（対称な）結合を使ったモデルはマルコフ確率場と呼ばれる。
- マルコフ確率場に対しても確率伝播アルゴリズムが定義される。考え方は同じ。

belief revision algorithm

- 各変数の周辺事後確率ではなく、MPE（尤度最大の値の組）を求めるためのアルゴリズム。
- 考え方は belief propagation と同じ。和演算の代わりに max を使う。

- 脳の認識アルゴリズムは loopy belief revision の近似かも。
[Litvak and Ullman 2009]
http://www.mitpressjournals.org/doi/abs/10.1162/neco.2009.05-08-783

大脳皮質のモデルの対立軸

- 下記の対立軸があるが、現在特に論争になっているというほどではない。
  - belief propagation vs. belief revision
  - 尤度 vs. 対数尤度
  - ベイジアンネット vs. マルコフ確率場 vs. ヘルムホルツマシン
  - 発火頻度コーディング vs. 膜電位コーディング

    - （私の予想は、belief revision, 尤度、ベイジアンネット、発火頻度コーディング。）

参考URL

- 「脳とベイジアンネットFAQ」
http://staff.aist.go.jp/y-ichisugi/besom/brain-bayesnet-faq.html
- 「脳を理解するための情報源メモ」＃ベイジアンネット
http://staff.aist.go.jp/y-ichisugi/besom/brain-memo.html#Bayesian-network

参考文献

- [Pearl 1988]
  Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference
  Judea Pearl: Morgan Kaufmann Pub ; ISBN: 1558604790
- 「ビショップ本」下 第８章 グラフィカルモデル
  ・パターン認識と機械学習 上 - ベイズ理論による統計的予測
  C. M. ビショップ (著), 元田 浩 (翻訳), 栗田 多喜夫 (翻訳), 樋口 知之 (翻訳), 松本 裕治 (翻訳), 村田 昇 (翻訳) 出版社: シュプリンガー・ジャパン株式会社 (2007/12/10) ISBN-13: 978-4431100133
  ・パターン認識と機械学習 下 - ベイズ理論による統計的予測
  C. M. ビショップ (著), 元田 浩 (翻訳), 栗田 多喜夫 (翻訳), 樋口 知之 (翻訳), 松本 裕治 (翻訳), 村田 昇 (翻訳) 出版社: シュプリンガー・ジャパン株式会社 (2008/7/11) ISBN-13: 978-4431100317

最後に

- このくらいの基礎知識があれば、下記の論文が読めるはず。（？）

[Lee 2003]
http://www.cnbc.cmu.edu/~tai/papers/lee_mumford_josa.pdf
[George and Hawkins 2005]
http://www.cnbc.cmu.edu/cns/papers/GeorgeHawkinsIJCNN05.pdf
[Rao 2005]
http://www.cs.washington.edu/homes/rao/nreport_bayes_atten05.pdf
[Ichisugi 2007]
http://staff.aist.go.jp/y-ichisugi/besom/20070509ijcnn-paper.pdf
[Chikkerur, Serre and Poggio 2009]
http://dspace.mit.edu/bitstream/handle/1721.1/49416/MIT-CSAIL-TR-2009-047.pdf?sequence=1
[Litvak and Ullman 2009]
http://www.mitpressjournals.org/doi/abs/10.1162/neco.2009.05-08-783

- 質問大歓迎。

大脳皮質とベイジアンネット 関連文献

- [Lee 2003] Lee, T.S., Mumford, D. (2003) Hierarchical Bayesian inference in the visual cortex. Journal of Optical Society of America, A. . 20(7): 1434-1448. http://www.cnbc.cmu.edu/~tai/papers/lee_mumford_josa.pdf
- [George and Hawkins 2005] George, D. Hawkins, J., A hierarchical Bayesian model of invariant pattern recognition in the visual cortex, In proc. of IJCNN 2005, vol. 3, pp.1812-1817, 2005. http://www.cnbc.cmu.edu/cns/papers/GeorgeHawkinsIJCNN05.pdf
- [Rao 2005] R. Rao., Bayesian inference and attention in the visual cortex. Neuroreport 16(16), 1843-1848, 2005. http://www.cs.washington.edu/homes/rao/nreport_bayes_atten05.pdf

大脳皮質とベイジアンネット 関連文献

- [Ichisugi 2007] Yuuji ICHISUGI,The cerebral cortex model that self-organizes conditional probability tables and executes belief propagation,In Proc. of International Joint Conference on Neural Networks (IJCNN2007), pp.1065--1070, Aug 2007. http://staff.aist.go.jp/y-ichisugi/besom/20070509ijcnn-paper.pdf
- [Chikkerur, Serre and Poggio 2009] Sharat Chikkerur, Thomas Serre and Tomaso Poggio: A Bayesian inference theory of attention: neuroscience and algorithms, Computer Science and Artificial Intelligence Laboratory Technical Report, MIT-CSAIL-TR-2009-047 CBCL-280, October 3, 2009 http://dspace.mit.edu/bitstream/handle/1721.1/49416/MIT-CSAIL-TR-2009-047.pdf?sequence=1
- [Litvak and Ullman 2009] Shai Litvak, Shimon Ullman: Cortical Circuitry Implementing Graphical Models, Neural Computation 21, 3010.3056 (2009) http://www.mitpressjournals.org/doi/abs/10.1162/neco.2009.05-08-783