本レジュメの最後に付録として Appendix のページを添付している。適宜参照されたい。

1　グラフィカルモデルにおける推論

本セクションではノードの一部分を観測値として固定した際にそのほかのノードに関する事後分布を計算していく[*1]。

まず簡単な例を取り扱う。ここではベイズの定理をグラフ上で解釈することで潜在変数についての事後分布を推論することを考える。図 8.37 のような 2 変数ノードの有向グラフを用意する。親ノード変数を x、子ノード変数を y とおけば、いま同時分布 $p(x,y)$ は $p(x)p(y \mid x)$ のように因数分解できる。ここでやりたいことは y が観測されたときに潜在変数 x に関する事後分布を求めることである。

求めたい事後分布はベイズの定理を用いると

$$p(x \mid y) = \frac{p(y \mid x)p(x)}{p(y)} \tag{1}$$

のように表せる。これにより同時分布は $p(x,y) = p(x \mid y)p(y)$ という形式でも表すことができる。これはグラフの観点で言うと図 8.37(c) のように有向グラフの矢印が反転した形として解釈することができる[*2]。

1.1　連鎖における推論

ここではもう少し複雑な例として図 8.32 のようなノード連鎖のグラフを取り扱う。有向連鎖と無向連鎖は全く同じ条件付き独立性を表現する[*3]。数式の操作を機械的に行えるように無向グラフ形式で議論を行う。

このグラフの同時分布は次の形で書ける。

$$p(\boldsymbol{x}) = \frac{1}{Z}\psi_{1,2}(x_1,x_2)\psi_{2,3}(x_2,x_3)\dots\psi_{N-1,N}(x_{N-1},x_N) \tag{2}$$

各ノード変数は K 状態取るものとする。ポテンシャル関数 $\psi_{n-1,n}(x_{n-1},x_n)$ が両変数について K 個ずつ状態を取ること、この関数の個数が $N-1$ 個であることから、同時分布はトータルで $(N-1)K^2$ 個のパラメタを持つことになる。

連鎖の途中のノード x_n の周辺分布 $p(x_n)$ を求めたい。原理的には x_n 以外の全ての変数について周辺化（総和を取る）すれば周辺分布は

$$p(x_n) = \sum_{x_1}\cdots\sum_{x_{n-1}}\sum_{x_{n+1}}\cdots\sum_{x_N} p(\boldsymbol{x}) \tag{3}$$

のように計算することができる。今 K 状態変数が N 個あることから $\boldsymbol{x}$ のとりうる状態の数は K^N である。このことから同時分布を計算して保存を行い、さらに周辺化して $p(x_n)$ を求めるには連鎖の長さ N に関して指数的な計算負荷が必要となる。

ここではグラフィカルモデルの条件付き独立性を利用することで上の計算を効率よく行えることを学ぶ。そのために式 2 を式 3 へ代入し和と積の順番を入れ替えてやる必要がある。この計算の詳細については App. 1 を参照されたい。

[*1] 本格的に一般論を取り扱うのは次回以降の模様。

[*2] $p(y)$ は観測された時点で確率分布でなくなると思うのだが、観測値つまり定数で書けるようになるということ？

[*3] 頁 104 の議論を参照。

App. 1 の計算によって求めたい周辺分布は

$$p(x_n) = \frac{1}{Z}\mu_\alpha(x_n)\mu_\beta(x_n) \tag{4}$$

という表式に変形できる。ただし $\mu_\alpha(x_n)$ は x_1 から x_{n-1} までに関しての和を含み、$\mu_\beta(x_n)$ は x_{n+1} から x_N までに関しての和を含むものとする。

今回の場合について計算負荷は次の通り。各ポテンシャル関数について片方の変数を周辺化する*4のに必要な計算量が K^2 である。そしてこれが連鎖の数 $N-1$ 個ぶんだけ足し算する必要がある。よってトータルの計算量は $O(NK^2)$ になる。で、合ってる？

このように和と積を入れて変えてやることで計算のオーダーが連鎖の長さについて線形となる*5。これによりもともと式 3 で想定していた指数関数的なオーダーと異なり計算効率の良いアルゴリズムを手に入れることができた。*6

式 4 を見ると周辺分布が 2 つの因子に規格化定数をかけたものになっていることがわかる。ここからはそれぞれの因子について「メッセージ」という概念を定義しあてはめることを考える。細かい定義については読み合わせ。

ここでメッセージの再帰性について補足を加える。例えば $\mu_\alpha(x_n)$ については x_1 から x_{n-1} までの和があり、そのうち x_{n-2} までを取り出してそれを新しく $\mu_\alpha(x_{n-1})$ と約束してやれば簡単に式（8.55）を導くことができる。だからもし $\mu_\alpha(x_n)$ を得たい場合は

・はじめに式（8.56）を計算

・これに対して入力側ノード変数と出力側ノード変数を引数に持つポテンシャル関数を掛け算

・さらに入力側ノード変数について総和を取る

というステップを $n-1$ 回繰り返せば良い。

連鎖の両側からノード x_n に向かってメッセージを送り込んでいくことを「メッセージパッシング」と呼ぶ。メッセージパッシングのグラフ構造は図 8.38 の通り*7。細かい情報は読み合わせ。

ここで規格化定数の計算オーダーを考える。そのために式（8.54）についての規格化条件に注目する。いま規格化定数はこれは x_n に関して K 状態ぶんの総和を取ることで求められる。したがって規格化定数の計算オーダーは $O(K)$ になることがわかる。

次に連鎖上の全てのノード n に対してそれぞれの周辺分布 $p(x_n)$ を求める*8。1 つあたりの $p(x_n)$ を計算するのにかかる負荷は $O(NK^2)$ であることは先ほど確かめた。したがって全 N 個の周辺分布をメッセージパッシングによる手続きで計算を行うと負荷は $O(NK^2)$ となる。実はこのやり方は伝播の中で重複した計算が多くなってしまい無駄が生じるという問題がある。

伝播の始点を変更することで計算負荷を落とすことができるらしい。が、ここはわからなかった。「中間的なメッセージを保存する」の意味が掴めなかった。あと結局この新しいやり方での計算負荷は $2NK^2$ のオーダーになるってことで OK？

ノードを観測することが関数 $I(x_n, \hat{x_n})$ を同時分布に掛け算することに等しいことを確認する。わけわからん。何が言いたい。

*4 1 文字消去とも捉えられるかもしれない。

*5 積和アルゴリズムの詳細については次回範囲となる。

*6 ただしグラフが全結合の場合は条件付き独立が全く成立しないので全体の同時分布を真面目に計算するしかない。条件付き独立性があるからこそ計算アルゴリズムを軽量化することができたのである。

*7 このような形をしたグラフをマルコフ連鎖と呼ぶ。

*8 はっきり言うと $p(x_1), p(x_2), \ldots, p(x_N)$ の全てを計算するということである

突然話は変わるが、連鎖上の 2 つの隣接ノード変数に関する同時分布 $p(x_{n-1}, x_n)$ を求めるフレームワークを確認する（演習 9.15)。これは式（8.52）を起点に考えればよい。非常に簡単ではあるが念のためやってみよう（See App. 2)。問題はそのあと。教科書本文の「周辺分布を〜用いて得られる。」のところがさっぱりでした。

1.2 木

ここでは木構造のグラフについて、無向木、有向木、有向多重木とは何か理解する。今回は演習 8.18 を通じて無向木と有向木の関係性を抑えそれらの持つ分布が等価であることを示す（See App. 3)[*9]。基本は読み合わせ[*10]。

前サブセクションまではノードの連鎖からなるグラフに対してノード数に線形な時間で厳密推論が行えることを確かめた。そこではメッセージパッシングというアルゴリズムを用いた。次回以降は上で紹介した「木」という一般的な構造[*11]に対してもこのアルゴリズムを適用し推論を行っていく。

[*9] まあよくわかりませんでしたが。
[*10] 経路の唯一性やモラル化については手を動かしてやれば納得できる。
[*11] 連鎖は木の中の一部、という位置付け。

App. 1 和と積の入れ替えによる周辺分布の計算

式(2)を(3)に代入する

$$p(x_n) = \sum_{x_1}\cdots\sum_{x_{n-1}}\sum_{x_{n+1}}\cdots\sum_{x_N}\frac{1}{Z}\psi_{1,2}(x_1,x_2)\psi_{2,3}(x_2,x_3)\cdots\psi_{N-1,N}(x_{N-1},x_N)$$

$$= \frac{1}{Z}\sum_{x_1}\cdots\sum_{x_{N-2}}\psi_{1,2}(x_1,x_2)\cdots\underbrace{\left(\sum_{x_{N-1}}\psi_{N-2,N-1}(x_{N-2},x_{N-1})\sum_{x_N}\psi_{N-1,N}(x_{N-1},x_N)\right)}_{\sim x_{N-2}}$$

$$= \frac{1}{Z}\sum_{x_1}\cdots\sum_{x_{N-3}}\psi_{1,2}(x_1,x_2)\cdots\underbrace{\left(\sum_{x_{N-2}}\psi_{N-3,N-2}(x_{N-3},x_{N-2})\sum_{x_{N-1}}\psi_{N-2,N-1}(x_{N-2},x_{N-1})\sum_{x_N}\psi_{N-1,N}(x_{N-1},x_N)\right)}_{\sim x_{N-3}}$$

・ポテンシャル関数を(大きい方から)取り込む

・サブスクリプトの大きいxについて総和をとる

この2つを $\psi_{n,n+1}$ を取り込むまで繰返す

$$= \frac{1}{Z}\sum_{x_1}\cdots\sum_{x_{n-1}}\psi_{1,2}(x_1,x_2)\cdots\psi_{n-1,n}(x_{n-1},x_n)\underbrace{\left[\sum_{x_{n+1}}\psi_{n,n+1}(x_n,x_{n+1})\cdots\sum_{x_N}\psi_{N-1,N}(x_{N-1},x_N)\right]}_{(=\mu_\beta(x_n))}$$

$$= \frac{1}{Z}\underbrace{\left[\sum_{x_{n-1}}\psi_{n-1,n}(x_{n-1},x_n)\cdots\sum_{x_1}\psi_{1,2}(x_1,x_2)\right]}_{(=\mu_\alpha(x_n))}\mu_\beta(x_n)$$

$$= \frac{1}{Z}\mu_\alpha(x_n)\mu_\beta(x_n)$$

教科書を見るとここに「…」がいるんだけどおかしいよね？
いらなくね？

App. 2 $\mu(x_n)$ を用いた同時分布 $p(x_{n-1}, x_n)$ の計算

式(3)について $\sum_{x_{n-1}}$ を外すだけで良い.

$$p(x_{n-1}, x_n) = \frac{1}{Z} \psi_{n-1,n}(x_{n-1}, x_n) \Big[\underbrace{\sum_{x_{n-2}} \psi_{n-2,n-1}(x_{n-2}, x_{n-1}) \cdots \sum_{x_1} \psi_{1,2}(x_1, x_2)}_{(\because (8.55))} \Big] \mu_\beta(x_n)$$

$$= \frac{1}{Z} \psi_{n-1,n}(x_{n-1}, x_n) \boxed{\mu_\alpha(x_{n-1})} \mu_\beta(x_n)$$

$$= \frac{1}{Z} \mu_\alpha(x_{n-1}) \psi_{n-1,n}(x_{n-1}, x_n) \mu_\beta(x_n)$$

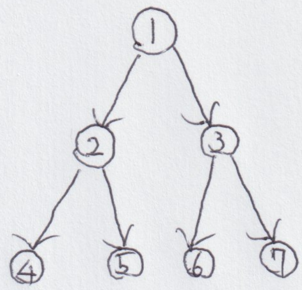
piece of cake!

App.3　有向木と無向木における分布の等価性

8.3.4で論じたようなポテンシャル関数と条件付き分布の対応付けができることを示せばよいと理解した。それ以上はわからん！

一般の場合のギロンもしていない。図8.39(b)を例に取り上げて説明を書いていく。

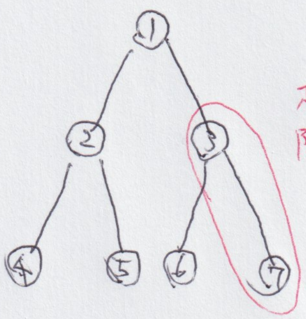
有向グラフ



ノード内の番号が変数のサブスクリプトを表す。同時分布 $p(x)$ は次のようになる。

$$p(x) = p(x_1)p(x_2|x_1)p(x_3|x_1)p(x_4|x_2)p(x_5|x_2)p(x_6|x_3)p(x_7|x_3)$$

無向グラフ（モラル化による）



極大クリークは隣接ノード対

無向グラフ表現では

$$p(x) = \frac{1}{Z}\psi_{1,2}(x_1,x_2)\psi_{1,3}(x_1,x_3)\psi_{2,4}(x_2,x_4)\psi_{2,5}(x_2,x_5)\psi_{3,6}(x_3,x_6)\psi_{3,7}(x_3,x_7)$$

したがって $Z=1$ とすれば以下のような対応関係を構築できる。

$$\begin{cases}
\psi_{1,2}(x_1,x_2) = p(x_1)p(x_2|x_1) \\
\psi_{1,3}(x_1,x_3) = p(x_3|x_1) \\
\psi_{2,4}(x_2,x_4) = p(x_4|x_2) \\
\psi_{2,5}(x_2,x_5) = p(x_5|x_2) \\
\psi_{3,6}(x_3,x_6) = p(x_6|x_3) \\
\psi_{3,7}(x_3,x_7) = p(x_7|x_3)
\end{cases}$$