

## 確率的モデルによる情報処理：解説

### 1 例

図1に示すようなマルコフ的信息源を考えよう。

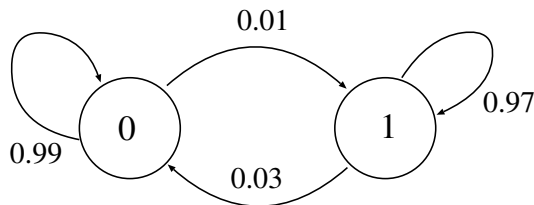


図1: 状態遷移図

図1は、状態が0, 1の2つであるようなマルコフ的信息源を表す遷移図である。状態は丸で囲われており、状態の遷移の確率は状態間に引かれた矢印のついた線の上に書かれている。図では、状態0にあったときに、次もまた状態0のままの確率が0.99、状態1に移る確率が0.01であることを示している。

このような確率的な状態遷移を500回繰り返したときの一例を、図2左側に描いている。おおよそ100回の状態遷移のうち2,3回、異なった状態に遷移する様子が見える。図2左に示し

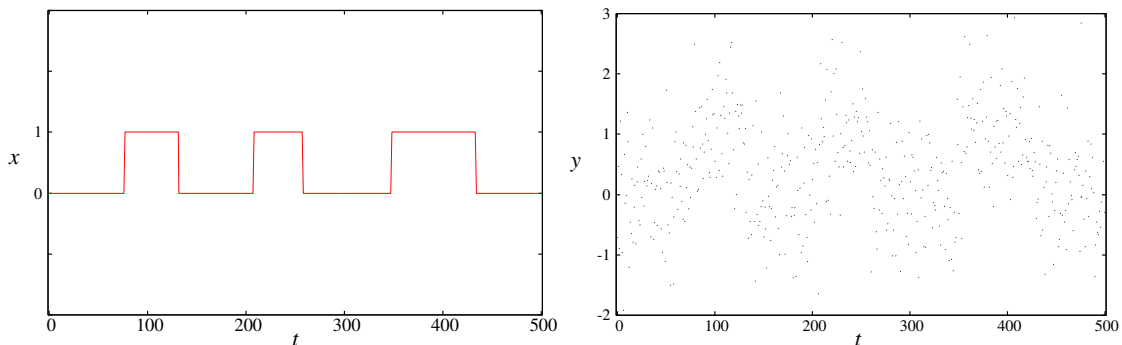


図2: 左: ランダムサンプル. 右: もとのデータにノイズが加わった観測データ

たものは、図1に示したモデルより確率的に生成された結果という意味で、ランダムサンプルと呼ばれている。おなじ事をもう一度おこなうと、100回に数回、状態を遷移するという傾向は同じであるが、図2左に示すものとは異なった、もう一つのランダムサンプルが得られる。出現する可能性のある系列は $2^{500}$ 通りあり、その中には00000000...000というものから0101010101...0101という滅多なことではみることができない系列も含まれる。

このように生成されたデータを遠く離れた地点に送ることを考えてみよう。ケーブルの性質が悪く、実際には、図2左にあるような信号のはずが、はるかかなたのデータを観測するため、もとのデータ(01系列)にノイズが加わり、観測できたのは図2右に示すものだったとする。くやしいが、このデータを観測して、もとのデータがなんであったか。できるだけ正確に推定するほかない。

不幸中の幸いで、今、以下のことはわかっているとしよう。

- もとのデータは図 1 に示すモデルから生成されている。
- ノイズは平均 0，分散  $\sigma^2 = 0.7^2$  の正規分布にしたがっている。
- ノイズは各時間で独立（異なる時刻で観測されるノイズの相関はない）。

解決策の一つは、図 2 右をにらんで、どこで状態遷移がおこったか目で判断することである。それでいい結果が得られるならそれでよい。

図 3 には、これから紹介する動的計画法という手法を使って、もとのデータを推定した結果を示している。推定した結果をもとのデータと重ね合わせると見にくいので推定結果を上方向に 3 だけずらしている。

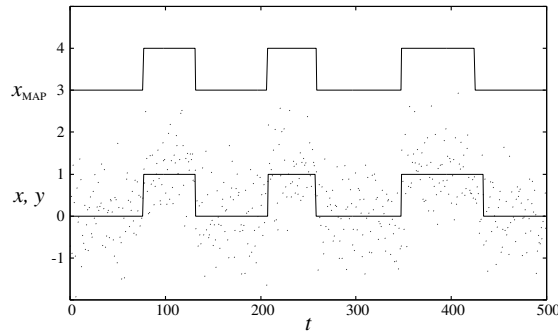


図 3: 事後確率を最大にする  $x$  の推定値  $x_{\text{MAP}}$

状態の変化  $0 \rightarrow 1$  や  $1 \rightarrow 0$  が起った箇所を完全ではないものの、かなり正確に当てていることがわかる。少なくとも人間の目で判断するのと同様以上の推定能力をもっていそうである。

## 2 マルコフチェーン

準備として 0 状態と 1 状態を遷移する単純なマルコフチェーンを考えよう。まずは初期状態であるが、時刻  $t = 0$  において、状態 0 にいる確率を  $p_0 = 0.5$  とする（確率は足し算すると 1 であるから  $p_1 = 0.5$  であることもすぐにわかる）。時刻  $t$  で  $X_t = i$  という状態にいたと仮定したとき、時刻  $t + 1$  で  $X_{t+1} = j$  という状態に遷移している確率を  $p_{ij}$  とする。この  $p_{ij}$  を並べてつくった  $P = \{p_{ij}\}$  という行列を遷移行列といい、図 1 の場合、

$$\begin{bmatrix} p_{00} & p_{01} \\ p_{10} & p_{11} \end{bmatrix} = \begin{bmatrix} 0.99 & 0.01 \\ 0.03 & 0.97 \end{bmatrix} \quad (1)$$

と書ける。

$T$  回遷移を繰り返すとして（上の例では  $T = 500$  ）、このマルコフチェーンから生成される最も尤もらしい 0 と 1 の系列は何だろうか。たとえば 11111...111 とすべて 1 ができる確率は  $0.5 \times (0.97)^T$  である。可能な 01 の系列は  $2^T$  個ある。いろいろ考えるまでもなく、00000...000 とすべて 0 である系列が最も高い確率（ $0.5 \times (0.99)^T$ ）で出現することに気づく。

この生成される最も尤もらしい 0 と 1 の系列を求めることを数学的に記述すると

$$\operatorname{argmax}_{x_0, x_1, \dots, x_T} \operatorname{Prob}(X_0 = x_0, X_1 = x_1, \dots, X_T = x_T) \quad (2)$$

となる。Prob の部分を確率変数  $X_0, X_1, \dots, X_T$  の同時確率分布という。argmax という記号は、その同時確率  $P$  を最大にする  $x_0, x_1, \dots, x_T$  の組み合わせ (0,1 の要素からなる  $T$  次元ベクトル) という意味である。こういう風には書くと記述が長くなるので、これからは

$$\operatorname{argmax}_{x_0, x_1, \dots, x_T} \operatorname{Prob}(x_0, x_1, \dots, x_T) \quad (3)$$

と書くようにする。

現在の状態に依存して、確率的に状態が遷移していく過程をマルコフ連鎖という。つまり過去の状態遷移履歴は、現在の状態にとりこまれているので、上式は

$$\operatorname{argmax}_{x_0, x_1, \dots, x_T} \operatorname{Prob}(x_0) \operatorname{Prob}(x_1|x_0) \operatorname{Prob}(x_2|x_1) \cdots \operatorname{Prob}(x_T|x_{T-1}) \quad (4)$$

と条件付き確率を使って掛け算の形でかける。こう書けるところがミソである。これももっと短く記述したいので、

$$\operatorname{argmax}_{x_0, x_1, \dots, x_T} p_{x_0} p_{x_0 x_1} p_{x_1 x_2} \cdots p_{x_{T-1} x_T} \quad (5)$$

と書く。例えば、 $p_0$  は初期状態で状態 0 にいる確率、 $p_{01}$  は時刻  $t$  で  $X_t = 0$  という状態にいたと仮定した場合、時刻  $t+1$  で  $X_{t+1} = 1$  という状態に遷移している確率である。ここでは簡単のため、この遷移確率は時間的には変動しないことにする。この節では、当たり前のことをくどくど書くか、これはあとあと必要となる表現なので、辛抱していただきたい。

さて、さきの例で  $T = 500$  のとき、上記の確率は  $0.5 \times (0.99)^{500}$  になる。これは非常に小さな正の数である。どのくらい小さな数か実際にコンピュータを使って計算しようとするとき、アンダーフローがおこり計算できない (どうなるかやってみるといい)。そういうときには  $\log$  をとればよいというわけで、 $\log 0.5 + T \log(0.99)$  を計算し、その後  $e$  の片に載せる。つまり  $0.5 \times (0.99)^T = e^{\log 0.5 + T \log(0.99)}$  とすると計算できる。

式 (5) を計算するためには確率を最大にする系列を見つけることが必要である。ここで  $\log$  が単調増加関数であることを思いおこせば、式 (5) の中身の  $\log$  を最大化しても同じであることがわかる、したがって問題は

$$\operatorname{argmax}_{x_0, x_1, \dots, x_T} \left\{ \log p_{x_0} + \sum_{t=1}^T \log p_{x_{t-1} x_t} \right\} \quad (6)$$

と書ける。

ここでいきなりではあるが、

$$C_0(i) = \log(p_i) \quad (7)$$

$$C_n(i) = \max_{x_0, x_1, \dots, x_{n-1}} \left\{ \log p_{x_0} + \sum_{t=1}^{n-1} \log p_{x_{t-1} x_t} + \log p_{x_{n-1} i} \right\} \quad (8)$$

という量を定義しておこう。 $C_n(i)$  は状態遷移を時刻  $n$  まで繰り返し、値  $i \in \{0, 1\}$  で終わっているときの最も尤もらしい系列 (best path と呼ぶ) の発生確率である。たとえば、時刻  $n = 1$  では

$$C_1(i) = \max_{x_0} \{ \log p_{x_0} + \log p_{x_0 i} \} \quad (9)$$

となる．さらに，この best path のうち  $X_{n-1}$  の状態を  $S_n(i)$  と書くことにする．時刻  $n$  での  $C_n(i), i \in \{0, 1\}$  が与えられたとすると，

$$S_{n+1}(j) = \operatorname{argmax}_i \{C_n(i) + \log p_{ij}\} \quad (10)$$

と書け，

$$C_{n+1}(j) = C_n(S_{n+1}(j)) + \log p_{S_{n+1}(j)j} \quad (11)$$

と書ける．表現がややこしく見えるが，この式はおさえておこう． $S_n(j)$  は時刻  $n-1$  の状態なので  $S_{n-1}(n, j)$  と書いた方がわかりやすいかもしれないが，つねに  $n-1, n$  の関係なので，余計なインデックスをつけないほうがいいし，このまま計算機プログラムを書けるといいういみで，このままにしておく．

$S_{n+1}(j)$  の意味するところは，時刻  $n+1$  で状態  $j$  にあると仮定したとき「最も尤もらしい系列の時刻  $n$  での状態」である．例えば，

$$S_1(j) = \operatorname{argmax}_i \{C_0(i) + \log p_{ij}\} \quad (12)$$

は，時刻 1 で状態  $j$  にいる ( $x_1 = j$ ) と仮定したときの  $x_0$  の最適な推定値となる．

今はまだ  $x_1 = 0$  なのか  $x_1 = 1$  なのか分からないので，4通りの系列のうち，時刻 1 でのとりうる 2つの状態に対してそれぞれ最適な  $x_0$  の値  $S_1(0), S_1(1)$  を覚えておく必要がある．この  $S_1(0)$  と  $S_1(1)$  の値は具体的に計算できる．

$x_1 = j$  と仮定した場合，最適な  $x_0$  が  $S_1(j)$  である．これを式で書くと

$$C_1(j) = C_0(S_1(j)) + \log p_{S_1(j)j} \quad (13)$$

となる．これも具体的に計算できる．

このようなことを一般的に書いたものが最初の式である．時刻  $n$  での最適な状態がわかると，時刻  $n$  までの系列の発生確率に，それぞれの状態に遷移する確率をかければ，時刻  $n+1$  で状態  $j$  でいる最も尤もらしい系列の発生確率を計算できる．

1)  $x_n = 0, 1$  それぞれの場合について最適な経路とその発生確率が既に分かっている．2) それをもとにして  $x_{n+1} = j$  の場合の最適な  $x_n$  と全体の系列の発生確率を計算する．これを続ける．つまり，4通りの系列のうち，時刻  $n$  でのとりうる 2つの状態に対して最適な経路を常に 2つをキープしておくことになる．

時刻  $T$  まで観測し，そのときの最も尤もらしい系列は

$$\hat{x}_T = \operatorname{argmax}_i C_T(i) \quad (14)$$

$$\hat{x}_{T-1} = S_T(\hat{x}_T) \quad (15)$$

...

$$\hat{x}_{n-1} = S_n(\hat{x}_n)$$

...

$$\hat{x}_1 = S_2(\hat{x}_2)$$

$$\hat{x}_0 = S_1(\hat{x}_1)$$

であると，順々に計算できる．

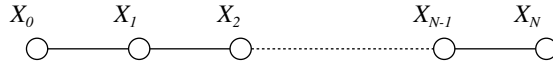


図 4: 確率変数の依存性を描いた図

確率変数の依存性をはっきりさせるため、今後、確率変数間の依存性を図 4 に示すように表示する。線でつながっていないのは、依存性がないことを示している。たとえば  $X_5$  のとりうる値は  $X_4$  と  $X_6$  にのみ依存している、といった具合である。このような確率変数間の依存関係は図 1 のような状態遷移図をみていただけではわからない。図 4 のようなグラフを依存性グラフ、マルコフ確率場、状態空間モデルなどと呼ぶことがある。

以上の計算は、マルコフチェーンという、現在の状態は、古い過去には依存せず、直前にどの状態にいたかにだけで確率的に決まるという性質があるからこういうふうを書いた。普通はこんなにうまくはいかない。

### 3 隠れマルコフモデル

前節では、あたりまえのことをなぜ面倒な式を使って説明するのかも思っただけかもしれない。本節では図 2 右のような状況を考える。観測できるのは  $x_0, x_1, \dots$  ではなく、 $x_0, x_1, \dots$  にノ

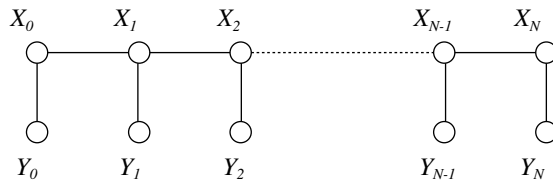


図 5: 確率変数の依存性：1次元の隠れマルコフモデル

イズが加わった  $y_0, y_1, \dots$  である。

さて  $x_0, x_1, \dots$  は前節で考えたマルコフ性をもつ系列である。観測する  $y_0, y_1, \dots$  は図 5 に示すように、 $Y_i = f(X_i) = X_i + Z_i$  という関係を満たしているものとする。ここで  $Z_i$  は各時刻  $i$  に独立に、確率に平均 0、標準偏差 0.7 の正規分布にしたがっているとすると、このような状況を

$$Y_n = X_n + Z_n, Z_n \sim \mathcal{N}(0, (0.7)^2) \quad (16)$$

とかく。

観測データ  $y_0, y_1, \dots$  は  $x_0, x_1, \dots$  に依存して決まる。データを生成する源の  $\{X_i\}$  はマルコフ性をもつが、 $\{X_i\}$  の関数である  $\{Y_i\}$  はマルコフ性をもたない。このようなモデルを隠れマルコフモデルという ( $x$  と  $y$  を関係付ける関数が 1 対 1 の非確率的な関数であれば  $\{Y_i\}$  もマルコフ性をもつが、それは特殊な場合)。

目的は  $y_0, y_1, \dots$  を観測し、もとの  $x_0, x_1, \dots$  を推定することである。つまり  $y_0, y_1, \dots$  を観測し、最も尤もらしい  $\hat{x}_0, \hat{x}_1, \dots$  を求めることである。これを式で書くと

$$\operatorname{argmax}_{x_0, x_1, \dots, x_T} \operatorname{Prob}(x_0, x_1, \dots, x_T \mid y_0, y_1, \dots, y_T) \quad (17)$$

となる .

$$\text{Prob}(x_0, \dots, x_T | y_0, \dots, y_T) = \frac{\text{Prob}(y_0, \dots, y_T | x_0, \dots, x_T) \text{Prob}(x_0, \dots, x_T)}{\text{Prob}(y_0, y_1, \dots, y_T)} \quad (18)$$

$$\text{Prob}(x_0, \dots, x_T | y_0, \dots, y_T) = \frac{\text{Prob}(x_0, \dots, x_T, y_0, \dots, y_T)}{\text{Prob}(y_0, y_1, \dots, y_T)} \quad (19)$$

であり (ベイズの公式), 確率変数間の依存性は図 5 に示すように,  $X_i$  はマルコフ性を持ち,  $Y_i$  は  $X_i$  に依存することがわかっているため, 求めたい  $x_0, x_1, \dots$  は

$$\text{argmax}_{x_0, x_1, \dots, x_T} \text{Prob}(x_0, x_1, \dots, x_T, y_0, y_1, \dots, y_T) \quad (20)$$

$$\text{argmax}_{x_0, x_1, \dots, x_T} \prod_{n=0}^T \text{Prob}(y_n | x_n) \text{Prob}(x_0) \prod_{n=1}^T \text{Prob}(x_n | x_{n-1}) \quad (21)$$

となる . ここで式 (19) の分母  $\text{Prob}(y_0, y_1, \dots, y_T)$  は  $\text{Prob}(x_0, \dots, x_T | y_0, \dots, y_T)$  の具体的な値をもとめたい場合には必要であるが, この確率を最大にする  $x_0, \dots, x_T$  を求めることについてだけであれば, 前述のとおり関係がない . 具体的にコンピュータを使って計算する場合は, 前節と同じように式の中身の  $\log$  をとり

$$\text{argmax}_{x_0, x_1, \dots, x_T} \left\{ \log p_{x_0} + \sum_{t=1}^T \log p_{x_{t-1}x_t} + \sum_{t=0}^T \log q_{x_t y_t} \right\} \quad (22)$$

を求める . このような手法を事後確率最大化といい, 推定した結果  $\hat{x}_0, \hat{x}_1, \dots$  を事後確率最大推定値 (maximum a posteriori estimator) という .

前節と同様に

$$C_0(i) = \log(p_i) + \log(q_{iy_0}) \quad (23)$$

$$S_{n+1}(j) = \text{argmax}_i \{ C_n(i) + \log p_{ij} + \log q_{jy_{n+1}} \} \quad (24)$$

$$C_{n+1}(j) = C_n(S_{n+1}(j)) + \log p_{S_{n+1}(j)j} + \log q_{jy_{n+1}} \quad (25)$$

$$n = 0, 1, \dots, T-1$$

と書く . そうすると, 時刻  $T$  での  $x_T$  の推定値を

$$\hat{x}_T = \text{argmax}_i C_T(i) \quad (26)$$

と求め,

$$\hat{x}_n = S_{n+1}(\hat{x}_{n+1}) \quad (27)$$

$$n = T-1, \dots, 0$$

と順々に計算できる .

## 4 パラメータ推定

モデルのパラメータ  $p_{00} = 0.99, p_0 = 0.5$  などはわかっているものとして話を進めてきた . 実際は, パラメータの値がわかっていない場合でも, パラメータの値を観測データから推定し, うまくもとのデータを復元できる方法がある . これは EM (Expectation-Maximization) アルゴリズムと呼ばれている .

1. まずパラメータの初期値  $\{p_{ij}\}$  を乱数に設定する .
2. 前節のとおり MAP 解を求める .
3. MAP 解  $\hat{x}_0, \hat{x}_1, \dots, \hat{x}_T$  をもとにパラメータを推定する . その仕方はラグランジェの未定係数法を使って導き出せる . ここでは , パラメータの値をどう更新すればよいか , その結果だけを示す .

$X$  が状態  $i$  にあるとした場合 , 次時刻に状態  $j$  に遷移することは何回あったか , データ  $\hat{x}_0, \hat{x}_1, \dots, \hat{x}_T$  から経験的に計算する . これを  $n_{ij}$  と書く .

$$n_{ij} = \text{\#times make a transition from } i \text{ to } j \quad (28)$$

具体的には  $n_{00}, n_{01}, n_{10}, n_{11}$  を計算する . パラメータ  $p_{ij}$  の値を以下の  $\tilde{p}_{ij}$  の値に更新する .

$$\tilde{p}_{ij} = \frac{n_{ij}}{\sum_{j=0}^1 n_{ij}} = \frac{n_{ij}}{n_{i0} + n_{i1}} \quad (29)$$

4. 2 に戻り , パラメータの値が収束するまで , この作業を繰り返す .

## 5 周辺分布 $Y$ を利用したデータのモデル化

$X_0, X_1, \dots, X_T$  はマルコフという単純なものであるが , 図 5 に示したモデル (線型グラフモデル) を使って , 確率変数間  $Y_0, Y_1, \dots, Y_T$  にどんな依存性があっても , その構造をいくらかでも精度良くモデル化できることが知られている . 具体的には , 先の例では  $Y_0, Y_1, \dots, Y_T$  は単にノイズが含まれたデータであったが , 音声データや画像の確率的な構造をモデル化できる . つまり , このタイプのモデルは任意の  $\text{Prob}(y_0, y_1, \dots, y_T)$  を表現できる能力がある .

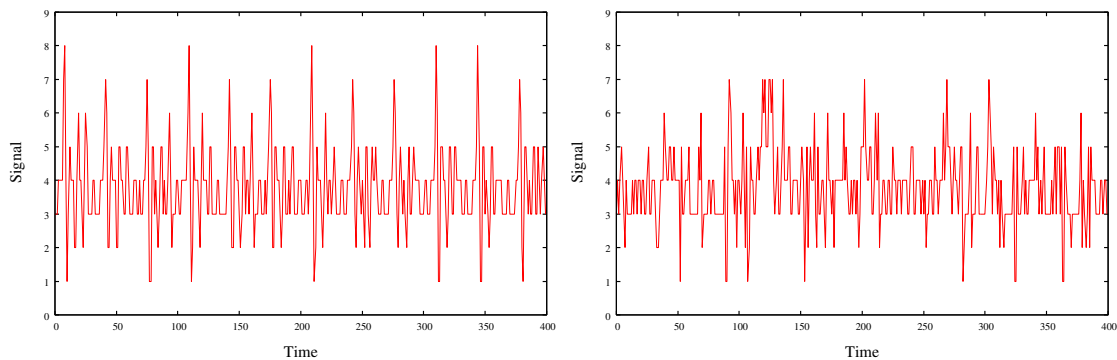


図 6: 左 : 粗くサンプリングした音声データ . 右 : 状態数 20 でモデル化し , そのモデルからのランダムサンプル

もちろん  $x_i$  のとりうる値の数や  $x_i \rightarrow y_i$  の関数をどのように設計するか考える必要がある . 図 6 では , 音声データをモデル化する具体例を示している . 具体的には  $X_i \in \{1, \dots, 20\}$  として  $Y_i = 1 + X_i \bmod 8$  としている . 学習したパラメータは  $X$  の状態遷移の確率  $\{p_{ij}\}$  だけである . その学習方法は上に示してあるものと同じ方法を用いた . 図 6 の左側には「ア」という音声をサンプリング周波数 8kHz , 量子化ビット数 8 (256 値) で記録したものを , 400 個

(100msec) の 8 値データにまで粗くサンプリングしなおしたものである。図 6 の右側には、このデータだけから、状態遷移確率を上記の EM アルゴリズムによりもとめ、その後、その学習済のモデルからランダムサンプルをおこなった結果を示している。

## 6 画像のモデル化

画像をモデル化する場合、2次元格子モデルがよく使われる。この場合、線型グラフの際に用いた動的計画法が使えなくなる。ただし、並列処理で、事後確率を最大にする MAP 推定値を近似的に計算する Gibbs Sampler という方法が使える。なお、ツリー型のモデルを使えば、簡単な画像であれば、事後確率を最大にする推定値を動的計画法を使って求めることができる。

## 7 正確な事後確率の計算

事後確率を最大にする MAP 推定値  $\hat{x}_0, \hat{x}_1, \dots, \hat{x}_T$  を計算することはできたが、その正確な事後確率

$$\operatorname{argmax}_{x_0, x_1, \dots, x_T} \operatorname{Prob}(x_0, x_1, \dots, x_T \mid y_0, y_1, \dots, y_T) \quad (30)$$

の値は非常に小さく、通常は計算できない。実は、線型グラフの場合は、これを計算機で正確に効率よく計算する方法がある。

【ヒント】分子と分母を入れ替えて 3 節の計算と似た作業をおこなう。

## 8 マルコフ確率場とギブス分布

確率変数の依存関係が一般的にはグラフの構造で書ける。「マルコフ確率場とギブス分布の同値性」について説明できなかった。申し訳ないので、互いの定義だけ以下に書いておく。以下では、 $X$  を確率変数のベクトル、 $x$  を具体的な値のベクトル、 ${}_i x$  を  $x$  から  $x_i$  を除いたベクトル、 $\text{NB}(i)$  を  $i$  番目のノードと近傍関係にあるノードの集合とする。

定義：マルコフ確率場 (MRF)

確率分布が

$$P(X_i = x_i \mid \mathbf{X} = \mathbf{x}) = P(X_i = x_i \mid \mathbf{X}_{\text{NB}(i)} = \mathbf{x}_{\text{NB}(i)}) \quad (31)$$

を満たすとき、その分布はグラフ  $G$  に関するマルコフ確率場であるという。上の式は長たらしい。通常は  $P(x_i \mid \mathbf{x}) = P(x_i \mid \mathbf{x}_{\text{NB}(i)})$  と略して書く。

定義：Gibbs 分布

どの  $x$  に対しても  $P(x) > 0$  であり、 $C$  をクリークの集合としたとき

$$P(x) = \prod_{c \in C} f_c(x_c) \text{ と書けるならば、その確率分布は Gibbs 分布である。} \quad (32)$$

ほかにも面白い性質がたくさんある...

## 参考文献

- [1] H. Künsch, S. Geman, and A. Kehagias, “Hidden Markov random fields,” *Annals of Applied Probability*, vol.5, no.3, pp.557–602, 1995.
- [2] S. Geman and K. Kochanek. “Dynamic programming and the graphical representation of error-correcting codes,” *IEEE Trans. Information Theory*, vol.47, no.2, pp.549–567, Feb. 2001.