神経回路網特論
2010 年 6 月 25 日

確率モデルによる情報処理

目 次

1　例

図 1 に示すようなマルコフ的情報源を考えよう．

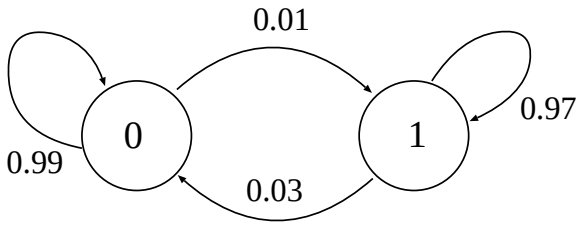

図 1: 状態遷移図

この情報源は 0, 1 の 2 つの状態をもち，状態間を時々刻々遷移する．状態は丸で囲われており，状態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 1 では，状態 0 にあったときに，次もまた状態 0 のままでいる確率が 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何度も繰り返したとしよう．状態の時間的変化の様子を図示することができる（図 2 左）．おおよそ 100 回の状態遷移のうち 2,3 回，現在とは別の状態に遷移する様子がわかる．この 0, 1 の数字の系列を ランダムサンプル と呼ぶ．図 1 に示したモデルより確率的に生成された例という意味である．同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

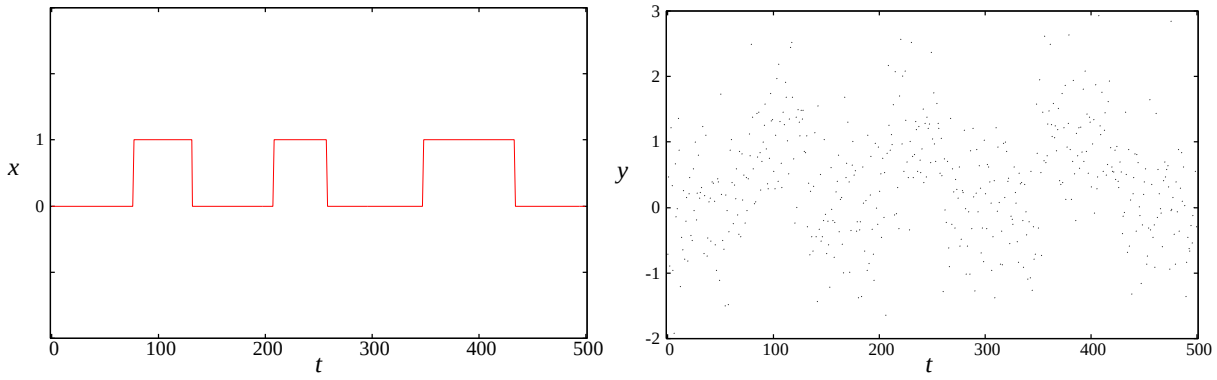

図 2: 左：ランダムサンプル．右：もとのデータにノイズが加わった観測データ

傾向は同じであるが，図 2 左とは異なった，もう一つのランダムサンプルが得られる．500 個の 0, 1 からなる系列を考えると，出現する可能性のある系列は 2^{500} 通りある．その中には $00000000\cdots000$ というもののから $0101010101\cdots0101$ という，出現する確率が 0 ではないが，ほぼ 0 に近い系列も含まれる．

図 1 のモデルから生成されたデータを遠く離れた地点に送ることを考えよう．ケーブルの性質が悪く，実際に届けられたデータ（図 2 右）は，もとのデータ（図 2 左）にひどくノイズが加わったものになっていた．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したい．これが問題である．今，次のことはわかっている．

- 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）．
- 状態遷移の確率は時間的に変動しない．

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，これから紹介する事後確率最大化と呼ばれる手法により，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）．状態の変化 $0 \to 1$ や $1 \to 0$ が起った時間を完全ではないもののの，かなり正確

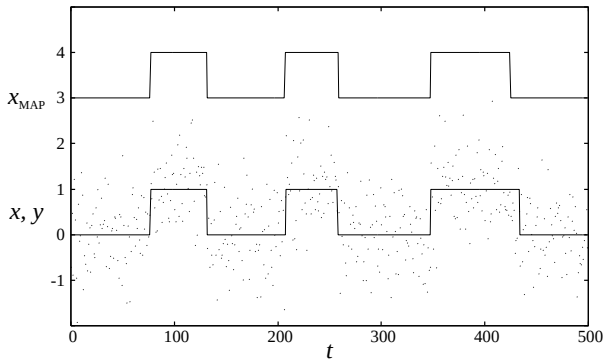

図 3: 事後確率を最大にする x の推定値 x_{MAP}

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等以上の推定能力をもっていそうである．この推論手法を学ぶのが今回の目的である．

2 マルコフ的情報源

2.1 マルコフ連鎖

準備として状態 0 と状態 1 を遷移する単純なマルコフ連鎖（Markov chain）を考えよう．まずは初期状態として，時間 $t = 0$ において，状態 0 にいる確率を $p_0 = 0.5$，状態 1 にいる確率を $p_1 = 0.5$ とする（確率は足し算すると 1.）．時間 t で $X_t = i$ という状態にいた場合に，時間 $t + 1$ で $X_{t+1} = j$ という状態に遷移する確率 $\mathrm{Prob}\{X_{t+1} = j | X_t = i\}$ を p_{ij} と書く．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，

$$P = \begin{bmatrix} p_{00} & p_{01} \\ p_{10} & p_{11} \end{bmatrix} = \begin{bmatrix} 0.99 & 0.01 \\ 0.03 & 0.97 \end{bmatrix} \tag{1}$$

と書ける．このように，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．

2.2 事前確率分布: $p(\boldsymbol{x})$

このモデルから $T(=500)$ 個の $0,1$ 系列が生成されるとしよう（$\boldsymbol{x} = x_0x_1x_2\cdots x_{T-1}$）．このうち，最も生成される確率の高い $\{0,1\}$ の系列は何だろうか．たとえば $11111\cdots111$ と，すべてが 1 である系列が出現する確率は $0.5\times(0.97)^{T-1}$ である．可能な 01 の系列は 2^T 個ある．$01010101\cdots$ が出現する確率がほぼ 0 であると先に述べたが，その出現確率は $0.5\times(0.01)^{249}\times 0.03^{250}$ である．

さて，このマルコフチェーンから生成される最も確率の大きい系列は，$00000\cdots000$ とすべて 0 である系列であり，その確率は $0.5\times(0.99)^{T-1}$ である．この最も生成されやすい T 個の 0 と 1 の系列（T 次元ベクトル）を記号を使い

$$\hat{\boldsymbol{x}} = \underset{x_0,x_1,\cdots,x_{T-1}}{\operatorname{argmax}} \operatorname{Prob}(X_0=x_0, X_1=x_1,\cdots,X_{T-1}=x_{T-1}) \tag{2}$$

と表現する．$\operatorname{Prob}(X_0=x_0,X_1=x_1,\cdots,X_{T-1}=x_{T-1})$ は確率変数 $X_0,X_1,\cdots,X_{T-1}$ の同時確率である．argmax という記号には慣れていないかもしれない．この式には，ありとあらゆる 2^T 個ある x_i の組み合わせのうち，この同時確率を最大にする組み合わせを取ってくるという意味がある．今後もこの記号はよくでてくるので，やさしい例で確認しておこう．$f(x,y)=-(x-1)^2-5(y-3)^2$ とすると

$$\underset{x,y\in\mathcal{R}}{\operatorname{argmax}} f(x,y) \quad = \quad (\hat{x},\hat{y}) = (1,3) \tag{3}$$

である．式 (2) は正式な記述であるが，簡単に，

$$\hat{\boldsymbol{x}} = \underset{\boldsymbol{x}}{\operatorname{argmax}} \operatorname{Prob}(\boldsymbol{x}) \tag{4}$$

と書くことが多い．今の場合，$\hat{\boldsymbol{x}} = 000000\cdots0$ である．

出現可能な $\boldsymbol{x}$ の系列は 2^{500} 通り存在する．各信号系列は，どれも同じ確率で出現するわけではなく，出現のしやすさに偏りがある．2^{500} 通りある系列，どの系列も出現確率は小さいと言っても 0 ではなく，それぞれの出現確率を足すと 1 になっている．この確率分布を事前確率分布という．いま考えているマルコフ情報源の場合，初期状態の確率 p_0,p_1 と状態遷移行列 P を指定することで，2^{500} とおりある各系列の出現確率が決まる．言い換えれば，3 つのパラメータ p_0,p_{00},p_{11} で決まる．(他のパラメータの値は $p_1=1-p_0, p_{01}=1-p_{00}$ などと，自動的に決まる)．受信者は，この各信号の出現のしやすさの事前情報をもとに，観測したデータ $\boldsymbol{y}$ を解釈することになる．

情報源がマルコフ連鎖である場合，過去の状態遷移履歴が現在の状態にとりこまれているので，$\boldsymbol{x}$ の同時確率は

$$\operatorname{Prob}(x_0,x_1,\cdots,x_{T-1}) \quad = \operatorname{Prob}(x_0)\operatorname{Prob}(x_1|x_0)\operatorname{Prob}(x_2|x_1)\cdots\operatorname{Prob}(x_{T-1}|x_{T-2}) \tag{5}$$

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鍵である．次節以降，これが効いてくる．

2.3 コンピュータにより計算する際の注意点

さて，$\hat{\boldsymbol{x}}$ は出現可能性のある 2^T 個のうち最も出現確率の高い系列である．ただし，その出現確率は $0.5\times(0.99)^{499}$ である．これは非常に小さい．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，アンダーフローがおこり計算できない（実際に試してみるとよい）．このような時には対数 log をとった値で評価すればいい．

2.4 確率変数の依存性グラフ

確率変数間の依存性を示すグラフを依存性グラフ（dependency graph）と呼ぶ．Prob($\boldsymbol{x}$) の依存性グラフを図 4 に示す．たとえば確率変数 X_2 の値が固定されると，X_0 と X_1 は，X_3 以降の変数とは独立であることを示している．現在の状態は，古い過去には依存せず，直前の状態だけで確率的に決

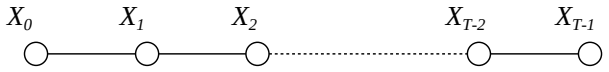

図 4: 確率変数間の依存性を描いたグラフ．Prob($\boldsymbol{x}$)

まるマルコフチェーンの性質から単純な線型のグラフになっている．一般には，すべてのノード (確率変数) が互いにつながった完全グラフになる．確率変数をノード，確率変数間の依存性を線で示すモデルをマルコフ確率場，状態空間モデルなどとも呼ぶことがある．

3 隠れマルコフモデル

3.1 データモデル: $p(\boldsymbol{y}|\boldsymbol{x})$

目的は $\boldsymbol{y} = y_0 y_1 y_2 \cdots$ を観測し，もとの $\boldsymbol{x} = x_0 x_1 x_2 \cdots$ を推定することである．前節で定義した確率分布 Prob($\boldsymbol{x}$) を使いノイズの含まれたデータ（$\boldsymbol{y}$）を解釈する．このことから，確率分布 Prob($\boldsymbol{x}$) を 解釈モデル と呼ぶ．これに対し，本節では データモデルProb($\boldsymbol{y}|\boldsymbol{x}$) を定義する．また，それらを合わせた同時分布

$$\mathrm{Prob}(\boldsymbol{x}, \boldsymbol{y}) = \mathrm{Prob}(\boldsymbol{y}|\boldsymbol{x})\mathrm{Prob}(\boldsymbol{x}) \tag{6}$$

を 生成モデル と呼ぶ.

具体的に Prob($\boldsymbol{y}|\boldsymbol{x}$) を考えよう．観測データ $\boldsymbol{y}$ は

$$y_t = f(x_t) = x_t + n_t \tag{7}$$

という関係を満たしているものとする．ここで n_t は各時間 t で独立で，平均 0，標準偏差 $\sigma = 0.7$ の正規分布にしたがっているとする．このような状況を

$$n_t \quad \sim \quad \mathcal{N}(0, \sigma^2) \tag{8}$$

と記述する．正確に記述すると観測データ y_t は x_t にのみ依存し

$$\mathrm{Prob}(Y_t < y_t | x_t) = \frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^{y_t} \exp\left\{ -\frac{(y - x_t)^2}{2\sigma^2} \right\} \mathrm{d}y \tag{9}$$

のように確率的に決まる．これで Prob($\boldsymbol{y}|\boldsymbol{x}$) が定まった.

データを生成する源の $\{X_t\}$ はマルコフ性をもつが，f という確率的な関数で変換された後の $\{Y_t\}$ 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（関数 f が 1 対 1 であれば $\{Y_t\}$ もマルコフ性をもつ）.

3.2 事後確率最大化: $\mathrm{argmax}_{\boldsymbol{x}}\, p(\boldsymbol{x}|\boldsymbol{y})$

$\boldsymbol{y}$ を観測し，もとの $\boldsymbol{x}$ を推定することが目的である．最も尤もらしい $\hat{\boldsymbol{x}}$ を求めよう．これを式で書くと

$$\hat{\boldsymbol{x}} = \operatorname*{argmax}_{\boldsymbol{x}} \mathrm{Prob}(\boldsymbol{x}|\boldsymbol{y}) \tag{10}$$

となる．ベイズの公式を使えば，これは

$$\mathrm{Prob}(\boldsymbol{x}|\boldsymbol{y}) \;=\; \frac{\mathrm{Prob}(\boldsymbol{y}|\boldsymbol{x})\mathrm{Prob}(\boldsymbol{x})}{\mathrm{Prob}(\boldsymbol{y})} = \frac{\mathrm{Prob}(\boldsymbol{x},\boldsymbol{y})}{\mathrm{Prob}(\boldsymbol{y})} \tag{11}$$

と表現できる．式 (11) の分母 $\mathrm{Prob}(\boldsymbol{y})$ は 事後確率 $\mathrm{Prob}(\boldsymbol{x}|\boldsymbol{y})$ の具体的な値を求める場合には必要である．しかし $\mathrm{Prob}(\boldsymbol{y})$ は正の定数であるため，事後確率を最大にする $\boldsymbol{x}$ を求めることには関係しない．観測データ $\boldsymbol{y}$ を受信する前の時点では，各 $\boldsymbol{x}$ のおこりやすさは $\mathrm{Prob}(\boldsymbol{x})$（事前確率）であったが，実際にデータ $\boldsymbol{y}$ を見た後は，各 $\boldsymbol{x}$ のおこりやすさが $\mathrm{Prob}(\boldsymbol{x}|\boldsymbol{y})$（事後確率）と変わる．

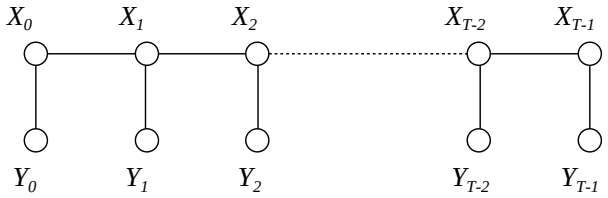

図 5: 確率変数の依存性：隠れマルコフモデル．$\mathrm{Prob}(\boldsymbol{x},\boldsymbol{y})$

X_t はマルコフ性をもち， Y_t は X_t にのみ依存することがわかっているので，求めたい $x_0, x_1, \cdots$ は

$$\hat{\boldsymbol{x}} \;=\; \operatorname*{argmax}_{x_0,x_1,\cdots,x_{T-1}} \mathrm{Prob}(x_0, x_1, \cdots, x_{T-1},\; y_0, y_1, \cdots, y_{T-1}) \tag{12}$$

$$\;=\; \operatorname*{argmax}_{x_0,x_1,\cdots,x_{T-1}} \mathrm{Prob}(x_0) \prod_{t=1}^{T-1} \mathrm{Prob}(x_t|x_{t-1}) \prod_{t=0}^{T-1} \mathrm{Prob}(y_t|x_t) \tag{13}$$

と書ける．このモデルの依存性グラフを描くと（図 5），各条件付確率が，一本の線に対応していることに気づく．ここで log が単調増加関数であることを思いおこせば，式の中身の log を最大化しても同じであることがわかる，したがって具体的にコンピュータで計算する場合は式 (13) の対数をとり，

$$\hat{\boldsymbol{x}} \;=\; \operatorname*{argmax}_{x_0,x_1,\cdots,x_{T-1}} \{\log F_0(x_0) + \sum_{t=1}^{T-1} \log F(x_{t-1}, x_t) + \sum_{t=0}^{T-1} \log G(x_t, y_t)\} \tag{14}$$

を求める．ここで

$$F_0(x_0) \;=\; \mathrm{Prob}(x_0) \tag{15}$$

$$F(x_{t-1}, x_t) \;=\; \mathrm{Prob}(x_t|x_{t-1}) \tag{16}$$

$$G(x_t, y_t) \;=\; \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{(y_t - x_t)^2}{2\sigma^2}\right\} \tag{17}$$

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

3.3 動的計画法

データ $\boldsymbol{y}$ はすでに与えられている．いま，ある $\boldsymbol{x}$ を仮定すれば，式 (14) の中身

$$\log F_0(x_0) + \sum_{t=1}^{T-1} \log F(x_{t-1}, x_t) + \sum_{t=0}^{T-1} \log G(x_t, y_t)$$

の値が求まる．2^{500} 通りある，ありとあらゆる $\boldsymbol{x}$ について試し，これが最大になる $\boldsymbol{x}$ を求めれば良い．これは現実的には不可能である．($2^{500} \approx 10^{150}$．100 年は $100 \times 365 \times 24 \times 60 \times 60 \approx 3.2 \times 10^9$ 秒．大きく見積もって 100 年は 10^{10} 秒．1 秒間に 100 億 $= 10^{10}$ 通りの結果を計算できるとして，100 年で計算できるのは $10^{10} \times 10^{10} = 10^{20}$ 通り．1 秒間に 1 無量大数（$= 10^{68}$）計算できるとしても，100 年で計算できるのは 10^{78} 通りしか計算できない.)

グラフの各ノードを結ぶ線（リンク）は一つ一つの条件付確率に対応している．この依存性グラフの構造（図 5）に着目しよう．これに気づけば，計算量を大幅に減らすことができる．以下で，この「からくり」を説明する．日本語で説明するよりは，数式を使った方がすっきりと理解できる．以下の記述が理解できなければ，グラフと数式を睨み，何回でも読み返してほしい.

まず各 $x_0 \in \{0, 1\}$ に対し

$$C_0(x_0) \quad = \quad \log F_0(x_0) + \log G(x_0, y_0) \tag{18}$$

を計算し，後で使うため記憶しておく．次に各 $x_1 \in \{0, 1\}$ に対し

$$S_0(x_1) \quad = \quad \underset{x \in \{0,1\}}{\operatorname{argmax}} \{C_0(x) + \log F(x, x_1) + \log G(x_1, y_1)\} \tag{19}$$

を計算する．$S_0(x_1)$ に $\hat{\boldsymbol{x}}$ を構成する X_1 が値 x_1 をとる場合の x_0 の値を格納する．式 (19) を計算する際，同時に，

$$C_1(x_1) \quad = \quad C_0(S_0(x_1)) + \log F(S_0(x_1), x_1) + \log G(x_1, y_1) \tag{20}$$

を記憶しておく．同様の作業を繰り返し，各 $x_{t+1} \in \{0, 1\}$ に対し

$$S_t(x_{t+1}) \quad = \quad \underset{x}{\operatorname{argmax}} \{C_t(x) + \log F(x, x_{t+1}) + \log G(x_{t+1}, y_{t+1})\} \tag{21}$$

$$C_{t+1}(x_{t+1}) \quad = \quad C_t(S_t(x_{t+1})) + \log F(S_t(x_{t+1}), x_{t+1}) + \log G(x_{t+1}, y_{t+1}) \tag{22}$$

$t = 0, 1, \cdots, 498$，を計算する．最後に $C_{499}(0), C_{499}(1)$ が得られる．この結果を利用し，

$$\hat{x}_{499} \quad = \quad \underset{x}{\operatorname{argmax}} \{C_{499}(x)\} \tag{23}$$

$$\hat{x}_{498} \quad = \quad S_{498}(\hat{x}_{499}) \tag{24}$$

$$\vdots$$

$$\hat{x}_0 \quad = \quad S_0(\hat{x}_1) \tag{25}$$

と順々に求まる．このようにして最も尤もらしい系列 （best path と呼ぶ）$\hat{\boldsymbol{x}}$ が求まる．記号がややこしく見えるかもしれないが，コンピュータでプログラムを書く場合には，データ構造として $S_t(x), C_t(x), t = 0, 1, \cdots, T-1, x = 0, 1$ という 2 次元の配列を採用すればよい.

4 事後確率分布: $p(\boldsymbol{x}|\boldsymbol{y})$

4.1 正確な事後確率の計算

事後確率を最大にする MAP 推定値 $\hat{\boldsymbol{x}} = \hat{x}_0\hat{x}_1\cdots\hat{x}_{T-1}$ は求まった．しかし，その正確な事後確率

$$\mathrm{Prob}(\hat{\boldsymbol{x}}|\boldsymbol{y}) \tag{26}$$

の値は非常に小さく，通常は計算できない．実は，変数間の依存性が単純な線型のグラフで表現できる場合は，これを計算機で正確に効率よく計算する方法がある．以下では事後確率 $\mathrm{Prob}(\hat{\boldsymbol{x}}|\boldsymbol{y})$ を計算する方法を説明する．事後確率の逆数を計算することがここでの鍵である．

$$\frac{1}{p(\boldsymbol{x}|\boldsymbol{y})} = \frac{p(\boldsymbol{y})}{p(\boldsymbol{x},\boldsymbol{y})} = \frac{\sum_{\tilde{\boldsymbol{x}}} p(\tilde{\boldsymbol{x}},\boldsymbol{y})}{p(\boldsymbol{x},\boldsymbol{y})} \tag{27}$$

$$= \frac{\sum_{\tilde{x}_0,\tilde{x}_1,\tilde{x}_2,\cdots} F_0(\tilde{x_0})F(\tilde{x_0},\tilde{x}_1)F(\tilde{x}_1,\tilde{x}_2)G(\tilde{x}_0,y_0)G(\tilde{x}_1,y_1)G(\tilde{x}_2,y_2)\cdots}{F_0(x_0)F(x_0,x_1)F(x_1,x_2)G(x_0,y_0)G(x_1,y_1)G(x_2,y_2)\cdots} \tag{28}$$

$$= \cdots \frac{\sum_{\tilde{x}_2} F(\tilde{x}_2,\tilde{x}_3)G(\tilde{x}_2,y_2)}{F(x_2,x_3)G(x_2,y_2)} \frac{\sum_{\tilde{x}_1} F(\tilde{x}_1,\tilde{x}_2)G(\tilde{x}_1,y_1)}{F(x_1,x_2)G(x_1,y_1)} \frac{\sum_{\tilde{x}_0} F(\tilde{x}_0,\tilde{x}_1)G(\tilde{x}_0,y_0)F_0(\tilde{x_0})}{F(x_0,x_1)G(x_0,y_0)F_0(x_0)} \tag{29}$$

こう書きつづけるのは大変であるので，

$$F'(x_t,x_{t+1},y_t) = F(x_t,x_{t+1})G(x_t,y_t) \tag{30}$$

と書き直すと，スペースの節約にもなるし意味がわかりやすい．

$$\frac{1}{p(\boldsymbol{x}|\boldsymbol{y})} = \cdots \frac{\sum_{\tilde{x}_2} F'(\tilde{x}_2,\tilde{x}_3,y_2)}{F'(x_2,x_3,y_2)} \frac{\sum_{\tilde{x}_1} F'(\tilde{x}_1,\tilde{x}_2,y_1)}{F'(x_1,x_2,y_1)} \frac{\sum_{\tilde{x}_0} F'(\tilde{x}_0,\tilde{x}_1,y_0)F_0(\tilde{x_0})}{F'(x_0,x_1,y_0)F_0(x_0)} \tag{31}$$

$$= \cdots \frac{\sum_{\tilde{x}_2} F'(\tilde{x}_2,\tilde{x}_3,y_2)}{F'(x_2,x_3,y_2)} \frac{\sum_{\tilde{x}_1} F'(\tilde{x}_1,\tilde{x}_2,y_1)H_1(\tilde{x}_1)}{F'(x_1,x_2,y_1)} \tag{32}$$

$$= \cdots \frac{\sum_{\tilde{x}_3} F'(\tilde{x}_3,\tilde{x}_4,y_3)}{F'(x_3,x_4,y_3)} \frac{\sum_{\tilde{x}_2} F'(\tilde{x}_2,\tilde{x}_3,y_2)H_2(\tilde{x}_2)}{F'(x_2,x_3,y_2)} \tag{33}$$

$$= \cdots \frac{\sum_{\tilde{x}_3} F'(\tilde{x}_3,\tilde{x}_4,y_3)H_3(\tilde{x}_3)}{F'(x_3,x_4,y_3)} \tag{34}$$

$$= \frac{\sum_{\tilde{x}_{499}} G(\tilde{x}_{499},y_{499})}{G(x_{499},y_{499})} \frac{\sum_{\tilde{x}_{498}} F'(\tilde{x}_{498},\tilde{x}_{499},y_{498})H_{498}(\tilde{x}_{498})}{F'(x_{498},x_{499},y_{498})} \tag{35}$$

$$= \frac{\sum_{\tilde{x}_{499}} G(\tilde{x}_{499},y_{499})H_{499}(\tilde{x}_{499})}{G(x_{499},y_{499})} \tag{36}$$

ここで H_t は上式のとおり，それまでの計算をまとめたものである．この値の逆数をとると $p(\boldsymbol{x}|\boldsymbol{y})$ が求まる．コンピュータで計算する際，アンダーフローが起こらないよう分子と分母を入れ替えた点と，グラフの構造を利用した効率的な計算順序を導入した点が鍵である．

4.2 事後確率分布からの正確なサンプリング

事後確率分布 $\Pr(\boldsymbol{x}|\boldsymbol{y})$ にしたがうサンプル $\boldsymbol{x}$ は以下のようにして得られる．まず計算の準備として，

$$\begin{aligned} T_0(x_1) &= \sum_{x_0} F_0(x_0)F(x_0,x_1)G(x_0,y_0) \\ &= \sum_{x_0} F_0(x_0)F'(x_0,x_1,y_0) \end{aligned}$$

と関数 T_0 を定義する．ここで

$$\begin{aligned} F_0(x_0) &= \Pr(x_0) \\ F(x_{t-1},x_t) &= \Pr(x_t|x_{t-1}) \\ G(x_t,y_t) &= \frac{1}{\sqrt{2\pi}\sigma}\exp\left\{-\frac{(y_t-x_t)^2}{2\sigma^2}\right\} \end{aligned}$$

であったことを確認しておく．同様に，

$$\begin{aligned} T_1(x_2) &= \sum_{x_1} T_0(x_1)F'(x_1,x_2,y_1) \\ T_2(x_3) &= \sum_{x_2} T_1(x_2)F'(x_2,x_3,y_2) \\ &\vdots \\ T_{198}(x_{199}) &= \sum_{x_{198}} T_{197}(x_{198})F'(x_{198},x_{199},y_{198}) \\ T_{199} &= \sum_{x_{199}} T_{198}(x_{199})G(x_{199},y_{199}) = Z \end{aligned}$$

とする（準備終了）．まず始めに，確率 $\Pr(X_{199}=x_{199}|\boldsymbol{y})$ を求め，X_{199} の標本値 x_{199} を得ることを考えよう．

$$\begin{aligned} \Pr(x_{199}|\boldsymbol{y}) = \frac{\Pr(x_{199},\boldsymbol{y})}{\Pr(\boldsymbol{y})} &= \frac{T_{198}(x_{199})G(x_{199},y_{199})}{T_{199}} \\ &= \frac{T_{198}(x_{199})G(x_{199},y_{199})}{\sum_{\tilde{x}_{199}} T_{198}(\tilde{x}_{199})G(\tilde{x}_{199},y_{199})} \end{aligned}$$

計算機でアンダーフローが起こらないよう計算するためには，この逆数の計算を試みればよい．

$$\begin{aligned} \frac{1}{\Pr(x_{199}|\boldsymbol{y})} &= \frac{\sum_{\tilde{x}_{199}} T_{198}(\tilde{x}_{199})G(\tilde{x}_{199},y_{199})}{T_{198}(x_{199})G(x_{199},y_{199})} \\ &= \frac{\sum_{\tilde{x}_{199}}\sum_{\tilde{x}_{198}} T_{197}(\tilde{x}_{198})F'(\tilde{x}_{198},\tilde{x}_{199},y_{198})G(\tilde{x}_{199},y_{199})}{\sum_{\tilde{x}_{198}} T_{197}(\tilde{x}_{198})F'(\tilde{x}_{198},x_{199},y_{198})G(x_{199},y_{199})} \\ &= \sum_{\tilde{x}_{199}}\sum_{\tilde{x}_{198}} \frac{T_{197}(\tilde{x}_{198})F'(\tilde{x}_{198},\tilde{x}_{199},y_{198})G(\tilde{x}_{199},y_{199})}{T_{197}(\tilde{x}_{198})F'(\tilde{x}_{198},x_{199},y_{198})G(x_{199},y_{199})} \\ &= \sum_{\tilde{x}_{199}}\sum_{\tilde{x}_{198}} \frac{F'(\tilde{x}_{198},\tilde{x}_{199},y_{198})G(\tilde{x}_{199},y_{199})}{F'(\tilde{x}_{198},x_{199},y_{198})G(x_{199},y_{199})} \\ &= \sum_{\tilde{x}_{199}} \frac{G(\tilde{x}_{199},y_{199})}{G(x_{199},y_{199})} \sum_{\tilde{x}_{198}} \frac{F(\tilde{x}_{198},\tilde{x}_{199})}{F(\tilde{x}_{198},x_{199})} \end{aligned}$$

$x_{199}=0,1$ それぞれの場合について計算できる．このようにして事後確率分布 $\Pr(\boldsymbol{x}|\boldsymbol{y})$ から X_{199} の標本値が得られる．ここで，X_{199} の標本値を得るために，$\boldsymbol{y}$ のうち y_{199} の値しか使っていないことに注意したい．次に，確率分布 $\Pr(x_{198}|x_{199},\boldsymbol{y})$ から X_{198} の標本値を得ることを考えよう．

$$\begin{aligned}\mathrm{Prob}(x_{198}|x_{199},\boldsymbol{y}) &= \frac{\mathrm{Prob}(x_{198},x_{199},\boldsymbol{y})}{\mathrm{Prob}(x_{199},\boldsymbol{y})}\\ &= \frac{T_{197}(x_{198})F'(x_{198},x_{199},y_{198})G(x_{199},y_{199})}{T_{198}(x_{199})G(x_{199},y_{199})}\\ &= \frac{T_{197}(x_{198})F'(x_{198},x_{199},y_{198})}{T_{198}(x_{199})}\end{aligned}$$

例によって，この逆数を計算しよう．

$$\begin{aligned}\frac{1}{\Pr(x_{198}|x_{199},\boldsymbol{y})} &= \frac{T_{198}(x_{199})}{T_{197}(x_{198})F'(x_{198},x_{199},y_{198})}\\ &= \frac{\sum_{\tilde{x}_{198}}T_{197}(\tilde{x}_{198})F'(\tilde{x}_{198},x_{199},y_{198})}{T_{197}(x_{198})F'(x_{198},x_{199},y_{198})}\\ &= \frac{\sum_{\tilde{x}_{198}}\sum_{\tilde{x}_{197}}T_{496}(\tilde{x}_{197})F'(\tilde{x}_{197},\tilde{x}_{198},y_{197})F'(\tilde{x}_{198},x_{199},y_{198})}{\sum_{\tilde{x}_{197}}T_{496}(\tilde{x}_{197})F'(\tilde{x}_{197},x_{198},y_{197})F'(x_{198},x_{199},y_{198})}\\ &= \sum_{\tilde{x}_{198}}\frac{F'(\tilde{x}_{198},x_{199},y_{198})}{F'(x_{198},x_{199},y_{198})}\sum_{\tilde{x}_{197}}\frac{F'(\tilde{x}_{197},\tilde{x}_{198},y_{197})}{F'(\tilde{x}_{197},x_{198},y_{197})}\\ &= \sum_{\tilde{x}_{198}}\frac{F'(\tilde{x}_{198},x_{199},y_{198})}{F'(x_{198},x_{199},y_{198})}\sum_{\tilde{x}_{197}}\frac{F(\tilde{x}_{197},\tilde{x}_{198})}{F(\tilde{x}_{197},x_{198})}\end{aligned}$$

これを $x_{198}=0,1$ それぞれの場合について，計算すればよい．結局，X_{198} の標本値を得るためには，先に得られた x_{199} とデータ y_{198} の 2 つの値しか使っていない．これを続けていこう．

$$\begin{aligned}\frac{1}{\Pr(x_1|x_2,\cdots,x_{199},\boldsymbol{y})} &= \frac{\Pr(x_2,\cdots,x_{199},\boldsymbol{y})}{\Pr(x_1,x_2,\cdots,x_{199},\boldsymbol{y})}\\ &= \frac{T_1(x_2)F'(x_2,x_3,y_2)\cdots}{T_0(x_1)F'(x_1,x_2,y_1)F'(x_2,x_3,y_2)\cdots}\\ &= \frac{\sum_{\tilde{x}_1}T_0(\tilde{x}_1)F'(\tilde{x}_1,x_2,y_1)}{T_0(x_1)F'(x_1,x_2,y_1)}\\ &= \sum_{\tilde{x}_1}\frac{F'(\tilde{x}_1,x_2,y_1)}{F'(x_1,x_2,y_1)}\sum_{\tilde{x}_0}\frac{F_0(\tilde{x}_0)F'(\tilde{x}_0,\tilde{x}_1,y_0)}{F_0(\tilde{x}_0)F'(\tilde{x}_0,x_1,y_0)}\\ &= \sum_{\tilde{x}_1}\frac{F'(\tilde{x}_1,x_2,y_1)}{F'(x_1,x_2,y_1)}\sum_{\tilde{x}_0}\frac{F(\tilde{x}_0,\tilde{x}_1)}{F(\tilde{x}_0,x_1)}\end{aligned}$$

これを $x_1=0,1$ それぞれの場合について，計算すればよい．先ほどと同様，X_1 の標本値を得るために，先に得られた x_2 と y_1 の 2 つの値しか使っていない．次が最後である．少しだけ気をつけよう．

$$\begin{aligned}\frac{1}{\Pr(x_0|x_1,\cdots,x_{199},\boldsymbol{y})} &= \frac{\Pr(x_1,\cdots,x_{199},\boldsymbol{y})}{\Pr(x_0,x_1,\cdots,x_{199},\boldsymbol{y})}\\ &= \frac{T_0(x_1)F'(x_1,x_2,y_1)\cdots}{F_0(x_0)F'(x_0,x_1,y_0)F'(x_1,x_2,y_1)\cdots}\\ &= \sum_{\tilde{x}_0}\frac{F_0(\tilde{x}_0)F'(\tilde{x}_0,x_1,y_0)}{F_0(x_0)F'(x_0,x_1,y_0)}\end{aligned}$$

5　確率モデル

5.1　線型の依存性グラフを持つ隠れマルコフモデルの利点

確率モデルは，現在，音声認識や画像理解，自然言語処理など，さまざまな領域で，ますます重要な役割を果たすようになっている．これまで扱ってきた単純なモデルが，現実の問題と，どう関係があるのか，どういう意味があるのか，結びつかないかもしれない．なかなか理解しにくい部分ではあるが，ここではその説明を試みる．

マルコフ連鎖から出力された信号 $X_0, X_1 \cdots$ にノイズが加わったデータ $Y_0, Y_1, \cdots$ を観測し，そこから，もとのデータを推定する問題を扱ってきた．マルコフ連鎖の状態遷移確率や，ノイズの性質はあらかじめ分かっているものとして，この推定をおこなった．現実の問題では，情報が生成されるメカニズムは，通常，不明である．ここでは，パラメータの値を既知とすることで，問題を易しくしていた．しかしながら,「対象の規則性を正しくモデル化できていれば確率モデルは機能する」ということは強調しておきたい．このことは感じとれる例題である．ここで「対象の規則性を正しくモデル化した確率モデル」とは認識対象を生成する能力のある生成モデルでもある．

これからは，観測データ Y として実際の音声や画像データを考える．このとき X の各変数にどんな意味のある情報を埋め込むか，X の各変数間にどのような規則性を埋め込むかが重要になる．依存性グラフはこれまでと同じ線型のモデルを用いる．この利点は 2 つある．以下にそれを整理しておく．

1. 事後確率分布 $p(\boldsymbol{x}|\boldsymbol{y})$ の変数依存性グラフは単純な線型のグラフ（図 6）になる．このおかげで事後確率分布について確率変数の様々な量，例えば周辺事後確率 $\mathrm{Prob}(x_2, x_3|\boldsymbol{y})$，期待値 $E[X_3|\boldsymbol{y}]$ などが簡単に正確に計算できる．特に，これはパラメータ推定（次節）の際に有効となる．

2. 観測データ Y の周辺確率分布 $\mathrm{Prob}(\boldsymbol{y})$ の変数依存性グラフは完全結合型（図 7）になる．これは，直観的には，どんな複雑な確率分布も表現できることを意味している．実際，任意の同時確率分布 $\mathrm{Prob}(Y)$ をいくらでも精度良く近似しモデル化できることが知られている．

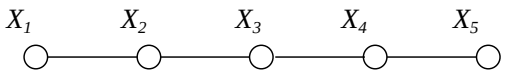

図 6: 確率変数の依存性： $\mathrm{Prob}(\boldsymbol{x}|\boldsymbol{y})$

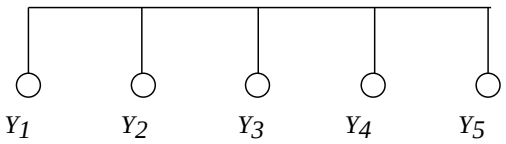

図 7: 確率変数の依存性：$\mathrm{Prob}(\boldsymbol{y}) = \sum_{\boldsymbol{x}} \mathrm{Prob}(\boldsymbol{x}, \boldsymbol{y})$

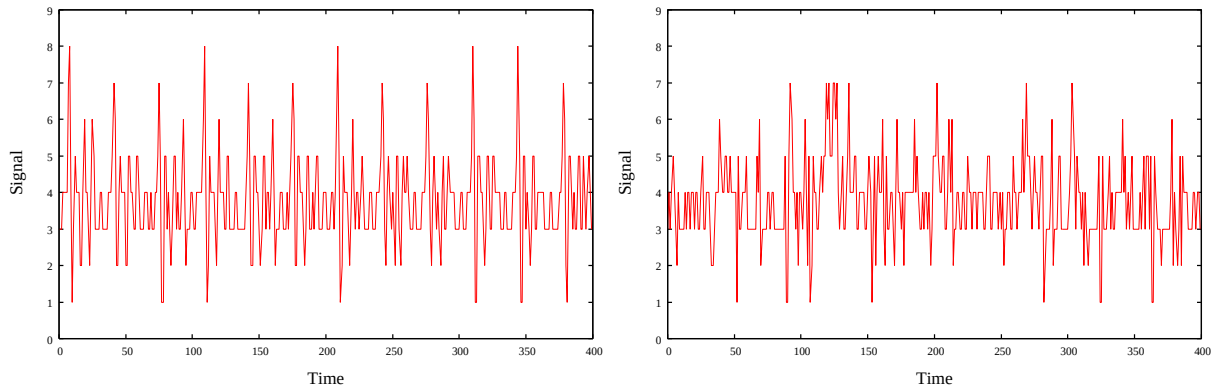

図 8: 左：粗くサンプリングした音声データ．右：状態数 20 でモデル化し，そのモデルからのランダムサンプル

5.2 周辺分布 Y を利用したデータのモデル化

線型の依存性グラフをもつモデルを用い，音声データをモデル化する例を示そう（図 8）．こでまでとの違いは，各 x_i が取りうる値が $0, 1$ の 2 通りではなく，多値（例えば 20 値，50 値）になる点である．y_i として，具体的には $x_i \in \{1, \cdots, 20\}$ として

$$y_i \quad = \quad 1 + x_i \bmod 8 \tag{37}$$

を考えよう．モデルのパラメータは X の状態遷移確率 $\{p_{ij}\}$ だけである．y_i は x_i から決定論的に決まる（ただし 1 対 1 ではない）．モデルは，$20 \times 21/2 = 210$ 個あるパラメータを指定すると定まる．これらのパラメータを調整し，対象をモデル化する．パラメータの調整を学習と呼ぶ．それについては次節で扱う．

先に実験結果を示そう．「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結果を図 8 の左側に示している．これがデータ $\boldsymbol{y}$ である．このデータ $\boldsymbol{y}$ をもとに，世の中はこういう規則性があるということをモデルに学習させよう．この $y_0, y_1, \cdots, y_i \in 1, \cdots, 8$ という 400 個のデータから，モデルの状態遷移確率（380 個のパラメータ）を学習により求める．これでモデルは完成する．もし $x_i \in \{1, \cdots, 50\}$ とすると調節するパラメータの数は $50 \times 51/2 = 1275$ 個ある．データの個数より，パラメータの数が多くなる点に注意したい．完成したモデルからデータを一つ生成してみれば，どのくらい精度よく対象をモデル化できたかが分かる．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図 8 の右側に示している．

6 パラメータ推定

7 画像のモデル化

8 マルコフ確率場とギブス分布

9 Gibbs Sampler