隠れマルコフモデルの計算ノート

@teramonagi

2012 年 12 月 22 日

1　イントロダクション

このノートは『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』(以下、PRML) の第 13 章「系列データ」に記載されているトピックである「隠れマルコフモデル (以下、HMM)」の数式変形、特に

- HMM の最尤推定（Baum-Welch アルゴリズム）(本稿 5 章)
- フォワード・バックワードアルゴリズム（本稿 6 章）

をかっちりと数式を使って導出する事を目的とした個人的な記録を兼ねたノートである。PRML[1] の第 13 章「系列データ」は 8 章のグラフィカルモデリングとの結びつきが強く、PRML[1] でもよく参照されているが、本稿では出来るだけこのドキュメント内で閉じた話とするためグラフィカルモデリングとの兼ね合いについては一切触れておらず、またそのため若干構成が異なる点に注意されたい。対象読者のレベルとしては条件付き確率の定義やベイズの定理を知っており、すらすら式変形ができる程度が望ましい。

本稿の構成は次の通りである。2 章で隠れマルコフモデルを説明するための礎となるマルコフモデルに対する簡単な説明を与え、3 章でその発展形としての隠れマルコフモデルについての説明を行う。続く 4 章では、後の計算で必要になる数学的な道具の定義やその説明を記し、5、6 章にて本稿の主題である HMM の最尤推定とフォワード・バックワードアルゴリズムの解説を行う。

2　（一次）マルコフモデル

まず、隠れマルコフモデルを定義する前にマルコフモデルを説明する。系列データ[*1]を扱う際に最も簡略化した仮定を置こうとすると**それぞれのデータは独立・同一な確率分布に従う**と仮定するのが最も簡単であろう。しかし、この方法では一切の**データ間の相互依存関係**を記述することができない。そこでデータ間の相互依存性を導入する上で最も簡単な方法の 1 つが次に述べる**マルコフモデル**を考えることである。

まず、ある観測系列データ $X := \{x_n, n = 1, \ldots, N\}$ の同時分布関数 $p(x_1, \ldots, x_N)$ を、条件付き確率の性質

$$P(A, B) = P(A|B)P(B), (^{\forall}A, ^{\forall} B \subset \Omega) \tag{1}$$

[*1] ここでは単に時系列データとでも考えておけばよい。適当に順序付けられたデータの事である。

を繰り返し適用することで一般性を失わずに

$$\begin{aligned} p(x_1,\ldots,x_N) &= p(x_N|x_1,\ldots,x_{N-1})p(x_1,\ldots,x_{N-1}) \\ &= p(x_N|x_1,\ldots,x_{N-1})p(x_{N-1}|x_1,\ldots,x_{N-2})p(x_1,\ldots,x_{N-2}) \\ &= p(x_N|x_1,\ldots,x_{N-1})p(x_{N-1}|x_1,\ldots,x_{N-2})p(x_{N-2}|x_1,\ldots,x_{N-3})p(x_1,\ldots,x_{N-3}) \\ &\cdots \\ &= p(x_N|x_1,\ldots,x_{N-1})p(x_{N-1}|x_1,\ldots,x_{N-2})p(x_{N-2}|x_1,\ldots,x_{N-3})\cdots p(x_3|x_1,x_2)p(x_2|x_1)p(x_1) \\ &= p(x_1)\prod_{n=2}^{N} p(x_n|x_1,\ldots,x_{n-1}) \end{aligned} \tag{2}$$

と展開することができる[*2]。ここで条件付き分布の各々が最も直近の観測値にだけに依存し、それよりも過去のデータに依存しないという条件を課したい。すなわち数式で書くと

$$p(x_n|x_1,\ldots,x_{n-1}) = p(x_n|x_{n-1}) \tag{3}$$

が成立すると仮定すると、(2) 式は

$$p(x_1,\ldots,x_N) = p(x_1)\prod_{n=2}^{N} p(x_n|x_{n-1}) \tag{4}$$

と変形することができ、**一次マルコフ連鎖**と呼ばれるモデルの分布関数を得る [*3]。このモデルは**各時系列データの生成確率がその一時点前のデータにのみ依存する**という特徴的な構造になっている。

3 隠れマルコフモデルとは

次にこの一次マルコフ連鎖の特色を活かしつつ、離散的な値を取る**潜在変数系列** $Z := \{z_n, n = 1,\ldots,N\}$ を導入する事によってよりモデルの表現力をより高めることを検討しよう。考え方としては

- 観測変数 x_n は潜在変数 z_n に影響を受けて生成される
- 潜在変数 z_n は一次マルコフ連鎖に従い z_{n-1} の影響を受けて生成される
- 観測変数 x_n 同士は独立である

という方法を採用する。この条件を満たすように構築したモデルの 1 つが**隠れマルコフモデル**であり、先ほど導入した潜在変数系列 Z と観測変数系列 X の同時確率分布を以下のように表現するものである。

$$p(\boldsymbol{x}_1,\ldots,\boldsymbol{x}_N,\boldsymbol{z}_1,\ldots,\boldsymbol{z}_N) := p(\boldsymbol{z}_1)\left[\prod_{k=2}^{N} p(\boldsymbol{z}_k|\boldsymbol{z}_{k-1})\right]\prod_{k=1}^{N} p(\boldsymbol{x}_k|\boldsymbol{z}_k)$$

潜在変数系列 Z が一次マルコフ連鎖 (4) の構造を取っている点に加えて、既述の一次マルコフ連鎖では x_n が x_{n-1} により条件付けられて生成されていたが、隠れマルコフモデルでは潜在変数 z_n により条件付けられることになるが特徴である。

このモデルにおいて、(特にほかならぬ私が) 興味があり、本稿で扱うトピックは

1. モデルを現実の系に当てはめた際のパラメーター推計の方法 (本稿 5 章)
2. 観測変数系列 X を取得した際の潜在状態の推定 (本稿 6 章)

である。本題に入る前に次章では数学的な準備・定式化を行っておく。

[*2] PRML[1](13.1) 式
[*3] PRML[1](13.2) 式

4　数学的な準備

具体的な計算に入る前に数学的な定式化をここに記しておく。以下の議論は適当な確率測度 P が定義された空間上で行うものとする[*4]。

既述ではあるが、本稿では系列データとして扱うデータとして以下の二種類を用意する。

- 観測変数系列 $X := \{x_n, n = 1, \ldots, N\}$:実際に観測・取得することが出来るデータ（離散・連続変数どちらでも良い）
- 潜在変数系列 $Z := \{z_n, n = 1, \ldots, N\}$:実際に観測・取得する事が出来ないデータ（離散変数）

x_n の次元については今回直接に言及しないので特段明記することはないが例えば $x_n \in \mathbb{R}^d (d$ 次元計量線形ベクトル空間の元等）とでも考えておけばよい。一方、潜在変数系列 Z については以下のように仮定を置く。問題として取り扱いたい・考えているシステム（系）には K 個の潜在変数の取り得る値（状態）があるとし、各時点 n での状態を表す潜在変数 $z_n, n \in \{1, \ldots, N\}$ を K 次元ベクトル空間 $\mathbb{K}$ 上の元とし、以下のように書くこととする。

$$
\begin{aligned}
z_n &= \begin{pmatrix} z_{n1} \\ z_{n2} \\ \vdots \\ z_{nK} \end{pmatrix} \\
z_{nk} &\in \{0, 1\}, {}^{\forall} n \in \{1, \ldots, N\}, {}^{\forall} k \in \{1, \ldots, K\} \\
\sum_{k=1}^{K} z_{nk} &= 1, {}^{\forall} n \in \{1, \ldots, N\}
\end{aligned}
\tag{5}
$$

こう書く事によるメリットは K 個ある（と仮定している）潜在変数の取り得る値（状態）を K 次元ベクトル z_n の要素 z_{nk} が 1 となる箇所に対応させて判断することができるということである[*5]。具体的に書くと、潜在変数 z は以下の K 個のベクトル値のいずれかを取るということである。

$$
z = \left.
\begin{pmatrix} 1 \\ 0 \\ \vdots \\ 0 \\ 0 \end{pmatrix} \text{or} \begin{pmatrix} 0 \\ 1 \\ \vdots \\ 0 \\ 0 \end{pmatrix} \text{or} \cdots \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 1 \\ 0 \end{pmatrix} \text{or} \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ 1 \end{pmatrix}
\right\} K \text{ 個}
\tag{6}
$$

このように書くことで見通し良く数式変形していくことが可能になる。

さらに隠れマルコフモデルにおいて潜在状態間の遷移を表す条件付き分布 $p(z_n|z_{n-1})$ を特徴づけるパラメーター、$K \times K$ 行列 A を導入しよう。この行列 A の各成分は

$$
A_{jk} := p(z_{nk} = 1 | z_{n-1,j} = 1)
\tag{7}
$$

として定義するものであり、解釈としては**潜在変数の状態が状態 j から状態 k へと遷移する確率**である。遷移**確率**であることから行列 A は規格化されており、“系はいずれかの状態に遷移する” という意味で行き先の状態 k に対して和を取ると 1 になる。すなわち

$$
\sum_{k=1}^{K} A_{jk} = 1, {}^{\forall} k \in \{1, 2, \ldots, K\}
\tag{8}
$$

[*4] 真面目に書くならフィルター付き確率空間 $(\Omega, \mathcal{F}(= \mathcal{F}_N), \mathbb{F}, P)$ を定義し、n で添え字付けられた x_n や z_n と言った確率変数に関してすべて $\mathcal{F}_n$-可測であるとするという感じか

[*5] PRML[1] でいう 1 対 K 符号化法

が成立する。$z_{nk} \in \{0,1\}, {}^\forall n \in \{1,\ldots,N\}, {}^\forall k \in \{1,\ldots,K\}$ であることを思い出し、行列 A を使うと条件付き分布 $p(z_n|z_{n-1})$ は明示的にかけて

$$p(z_n|z_{n-1}) := p(z_n|z_{n-1},A) = \prod_{k=1}^{K}\prod_{j=1}^{K}(A_{jk})^{z_{n-1,j}\times z_{nk}} \tag{9}$$

とすることが出来る[*6]。例えば潜在変数 z_n, z_{n-1} がそれぞれ $z_{nk'}=1, z_{n-1,j'}=1$ に対応する状態 $\boldsymbol{k'}, \boldsymbol{j'} \in \mathbb{K}$ を取っていたとすると、(9) 式の右辺の積 $\prod$ は $k=k', j=j'$ の箇所のみが有効な値 $(A_{j'k'})^{1\times 1} = A_{j'k'}$ を取り、残りの項は $(A_{jk})^0 = 1$ となるので

$$p(z_n = \boldsymbol{k'}|z_{n-1} = \boldsymbol{j'}) = p(z_{nk'}=1|z_{n-1,j'}=1,A) = \prod_{k=1}^{K}\prod_{j=1}^{K}(A_{jk})^{z_{n-1,j}\times z_{nk}} = A_{j'k'} \tag{10}$$

というように、ある特定の状態間（この場合 $\boldsymbol{k'}$ と $\boldsymbol{j'}$ 間）の遷移を行列の要素に紐付けてみる事が出来る。

潜在変数の初期状態 z_1 も同様に K 次元ベクトルのパラメーター $\pi \in \mathbb{K}$ を

$$\begin{gathered}\pi = \begin{pmatrix}\pi_1\\ \pi_2\\ \vdots\\ \pi_K\end{pmatrix}\\ \sum_{k=1}^{K}\pi_k = 1\end{gathered} \tag{11}$$

と導入することで特徴づけてやることにより

$$p(z_1) := p(z_1|\pi) = \prod_{k=1}^{K}\pi_k^{z_{1k}} \tag{12}$$

とすることができる[*7]。

最後に本稿で具体的な導出は行わないが、隠れマルコフモデル (5) の説明していない最後の項 $p(x_k|z_k)$ も分布を特徴づけるパラメータベクトル ϕ を[*8]

$$\phi = \begin{pmatrix}\phi_1\\ \phi_2\\ \vdots\\ \phi_K\end{pmatrix} \tag{13}$$

と導入することで

$$p(x_k|z_k) := p(x_k|z_k,\phi) = \prod_{k=1}^{K}p(x_n|\phi_k)^{z_{nk}} \tag{14}$$

とすることができる。各々の ϕ_k は例えば平均 μ・分散 σ^2 の正規分布の場合は、その平均と分散のセット $\phi_k = (\mu,\sigma^2) \in \mathbb{R}\times\mathbb{R}_+$ となる。以降では、本章で導入したパラメーター A,π,ϕ をまとめて $\theta := \{\pi, A, \phi\}$ と書く事にしよう。

[*6] PRML[1](13.7) 式。元のテキストだと $z_{n-1,j}z_{nk}$ がべき乗のように見えるのであえて括弧で括って書いている。

[*7] PRML[1](13.8)

[*8] この ϕ の属する空間はモデル・分布の設定によるので明記しない

5　HMM の最尤推定

本章では本稿の目的の 1 つである観測変数系列 $\boldsymbol{X} = \{\boldsymbol{x}_1, \cdots, \boldsymbol{x}_N\}$ を既知、すなわち既にデータが手に入った物として、モデルのパラメーター $\theta = \{\pi, A, \phi\}$ を最尤推定により求めることを考える。

まず観測変数系列 X と潜在変数系列 Z の同時分布関数 $p(X, Z|\theta)$ を Z について和を取ることで周辺化し、尤度関数を導入する[*9][*10]。

$$L(\theta|X) = p(X|\theta) = \sum_Z p(X, Z|\theta) \tag{15}$$

この尤度関数 $L(\theta|X)$ を直に最大化し最尤推定を実行しようとすると $\sum_Z$ の計算をする事になるが、N 個ある潜在変数系列に対して各々 K 個の状態を取るので結局 K^N 項の和を計算する事になるので、取り扱おうとするデータサイズが大きいとこれは実質不可能である。

ここではこの困難を克服するために EM アルゴリズムを用いる。EM アルゴリズムは 2 つのステップ

- E ステップ：現在推定されている潜在変数の分布に基づいて、モデルの尤度の期待値を計算
- M ステップ：E ステップで求まった尤度の期待値を最大化するパラメータを求める

で構成されているアルゴリズムであり、このそれぞれを交互に繰り返すことによりパラメータを推計するアルゴリズムである。

5.1　E ステップ

本稿の対象である HMM において、E ステップではまず最初にパラメータをある適当な値 θ^{old} と設定し、潜在変数の事後分布 $p(Z|X, \theta^{old})$ を求め、完全データに対する尤度関数[*11] の対数の期待値 $Q(\theta, \theta^{old})$ を

$$\begin{aligned} Q(\theta, \theta^{old}) &= \mathbb{E}_Z \left[\ln p(X, Z|\theta)\right] \\ &= \sum_Z p(Z|X, \theta^{old}) \ln p(X, Z|\theta) \end{aligned} \tag{16}$$

として求める[*12]。

ここで今後の式変形を見通し良くするために潜在変数 z_n の周辺事後分布 $\gamma(z_n)$、2 つの連続した潜在変数に対する同時分布 $\xi(z_{n-1}, z_n)$ を導入する[*13]。

$$\gamma(z_n) = p(z_n|X, \theta^{old}), \qquad \gamma : \mathbb{K} \to [0, 1] \tag{17}$$

$$\xi(z_{n-1}, z_n) = p(z_{n-1}, z_n|X, \theta^{old}), \qquad \xi : \mathbb{K} \times \mathbb{K} \to [0, 1] \tag{18}$$

この $\gamma(z_n), \xi(z_{n-1}, z_n)$ について潜在変数 z と同じく、各要素での表現

$$\begin{aligned} \gamma(z_{nk}) &:= \mathbb{E}\left[z_{nk}\right] = \sum_{z_n} \gamma(z_n) z_{nk} = p(z_{nk} = 1|X, \theta^{old}) \\ \xi(z_{n-1,j}, z_{nk}) &:= \mathbb{E}\left[z_{n-1,j} z_{nk}\right] = \sum_{z_{n-1}, z_n} z_{n-1,j} z_{nk} = p(z_{n-1,j} = 1, z_{nk} = 1|X, \theta^{old}) \end{aligned} \tag{19}$$

[*9] パラメーターへの依存性を明確にするため θ で条件付けている

[*10] PRML[1](13.11) 式

[*11] X, Z 共に解っている状況での尤度関数。すなわち上述の $L(\theta|X)$ ではなく、強いて書くなら $L'(\theta|X, Z)$

[*12] PRML[1](13.12) 式

[*13] PRML[1](13.13)(13.14) 式

を導入しておく。

(5) 式を (16) に代入し、γ, ξ の定義 (19) を用いて式変形してくと $Q(\theta, \theta^{old})$ は

$$
\begin{aligned}
Q(\theta, \theta^{old}) &= \sum_Z p(Z|X, \theta^{old}) \ln p(X, Z|\theta) \\
&= \sum_Z p(Z|X, \theta^{old}) \left\{ \ln p(z_1|\pi) + \sum_{k=2}^{N} \ln p(z_k|z_{k-1}, A) + \sum_{k=1}^{N} \ln p(z_k|x_k, \phi_k) \right\} \\
&= \sum_Z p(Z|X, \theta^{old}) \ln p(z_1|\pi) + \sum_Z p(Z|X, \theta^{old}) \sum_{k=2}^{N} \ln p(z_k|z_{k-1}, A) + \sum_Z p(Z|X, \theta^{old}) \sum_{k=1}^{N} \ln p(z_k|x_k, \phi_k) \\
&= \sum_{z_1} p(z_1|X, \theta^{old}) \ln p(z_1|\pi) + \sum_{k=2}^{N} \sum_Z p(Z|X, \theta^{old}) \ln p(z_k|z_{k-1}, A) + \sum_{k=1}^{N} \sum_Z p(Z|X, \theta^{old}) \ln p(z_k|x_k, \phi_k) \\
&= \sum_{z_1} p(z_1|X, \theta^{old}) \ln p(z_1|\pi) + \sum_{k=2}^{N} \sum_{z_k, z_{k-1}} p(z_k, z_{k-1}|X, \theta^{old}) \ln p(z_k|z_{k-1}, A) + \sum_{k=1}^{N} \sum_{z_k} p(z_k|X, \theta^{old}) \ln p(z_k|x_k, \phi_k) \\
&= \sum_{z_1} \sum_{k=1}^{K} \gamma(z_1) z_{1k} \ln \pi_k + \sum_{k=2}^{N} \sum_{z_k, z_{k-1}} \sum_{j=1}^{K} \sum_{k=1}^{K} \xi(z_{k-1}, z_k) z_{k-1,j} zk, k \ln A_{jk} + \sum_{k=1}^{N} \sum_{k=1}^{k} \gamma(z_k) z_{kk} \ln p(x_k|\phi_k) \\
&= \sum_{k=1}^{K} \gamma(z_{1,k}) \pi_k + \sum_{n=2}^{N} \sum_{j=1}^{K} \sum_{k=1}^{K} \xi(z_{n-1,j}, z_{n,k}) \ln A_{jk} + \sum_{n=1}^{N} \sum_{k=1}^{K} \gamma(z_{n,k}) \ln p(x_k|\phi_k)
\end{aligned}
\tag{20}
$$

と変形する事が出来る[*14]。(20) 式の最後の行を見ると理解できるようにあとは γ, ξ を計算できれば $Q(\theta, \theta^{old})$ を計算する事が出来る。この話題については [**?**] 章「フォワード・バックワードアルゴリズム」で行うことにして、EM アルゴリズムの M ステップを先に考える。

5.2 M ステップ

EM アルゴリズムの M ステップでは、パラメーター θ を適当に選択することで $Q(\theta, \theta^{old})$ の最大化を行う事を考える。

ここで遷移行列 A と初期分布 π には条件

$$
\begin{aligned}
\sum_k A_{jk} &= 1, {}^{\forall} k \in \{1, \ldots, K\} \\
\sum_{k=1}^{K} \pi_k &= 1
\end{aligned}
$$

が付いていることを思い出そう。従って、ラグランジュの未定乗数 $(\lambda_1, \ldots, \lambda_{K+1})$ を導入し、以下のような拘束条件付きの最大化問題として $Q(\theta, \theta^{old})$ の最大化問題を解く事にする。

$$
\begin{aligned}
\theta_{max} &= \arg\max_{\theta=(\pi,A,\phi)} F(\pi, A, \phi) \\
F(\pi, A, \phi) &:= Q(\theta, \theta^{old}) - \lambda_1 \left(\sum_{k=1}^{K} \pi_k - 1 \right) + \sum_{j=1}^{K} \lambda_{j+1} \left(\sum_{k=1}^{K} A_{jk} - 1 \right)
\end{aligned}
\tag{21}
$$

[*14] PRML[1](13.17) 式

この関数 $F(\pi, A, \phi)$ は素直に π, A の各ベクトル・行列要素で微分することで

$$\begin{aligned}
\frac{\partial F}{\partial \pi_k} &= \frac{\gamma(z_{1k})}{\pi_k} - \lambda_1 = 0, \therefore \pi_k = \frac{\gamma(z_{1k})}{\lambda_1} \\
\frac{\partial F}{\partial A_{jk}} &= \frac{\sum_{n=2}^{N} \xi(z_{n-1,j}, z_{nk})}{A_{jk}} - \lambda_{j+1}, \therefore A_{jk} = \frac{\sum_{n=2}^{N} \xi(z_{n-1,j}, z_{nk})}{\lambda_{j+1}}
\end{aligned} \tag{22}$$

とすることが出来る。残りのパラメーター ϕ についての微分は観測変数 x がどのような分布に従うと設定するかに応じて結果が変わるものであるので、ここでは示さない。ここで算出した (22) を拘束条件であるに代入して $(\lambda_1, \ldots, \lambda_{K+1})$ を求める。

$$\begin{aligned}
\sum_{k=1}^{K} \pi_k &= \sum_{k=1}^{K} \frac{\gamma(z_{1k})}{\lambda_1} = 1, \therefore \lambda_1 = \sum_{j=1}^{K} \gamma(z_{1j}) \\
\sum_{k=1}^{K} A_{jk} &= \sum_{k=1}^{K} \frac{\sum_{n=2}^{N} \xi(z_{n-1,j}, z_{nk})}{\lambda_{j+1}} = 1, \therefore \lambda_{j+1} = \sum_{l=1}^{K} \sum_{n=2}^{N} \xi(z_{n-1,j}, z_{nl})
\end{aligned} \tag{23}$$

ただしここで和記号 $\sum$ の変数を変更している点に注意されたい。。上記の (23) の結果を再び (22) に戻すことで

$$\begin{aligned}
\pi_k &= \frac{\gamma(z_{1k})}{\lambda_1} = \frac{\gamma(z_{1k})}{\sum_{j=1}^{K} \gamma(z_{1j})} \\
A_{jk} &= \frac{\sum_{n=2}^{N} \xi(z_{n-1,j}, z_{nk})}{\lambda_{j+1}} = \frac{\sum_{n=2}^{N} \xi(z_{n-1,j}, z_{nk})}{\sum_{l=1}^{K} \sum_{n=2}^{N} \xi(z_{n-1,j}, z_{nl})}
\end{aligned} \tag{24}$$

とすることが出来る[*15]。

このように E ステップ・M ステップそれぞれを構築することで EM アルゴリズムでの定式化は完了である。残された問題は $Q(\theta, \theta^{old})$ の最尤推定を行うための $\gamma(z_{nk}), \xi(z_{n-1,j}, z_{nk})$ の効率的な算出であり、これが次のフォワード・バックワードアルゴリズムへとつながる。

6 フォワード・バックワードアルゴリズム

前述のように $\gamma(z_{nk}), \xi(z_{n-1,j}, z_{nk})$ の効率的な算出方法を検討したい。この章では PRML[1] にならいパラメーター θ への依存性を明記しないこととする。例えば

$$p(X, z_n) = p(X|z_n, \theta)$$

という略記法を用いる。

ここで

$$p(X, z_n) = p(x_1, \ldots, x_n|z_n)p(x_{n+1}, \ldots, x_N|z_n)p(z_n) \tag{25}$$

という関係が成り立つ[*16]ので、$\gamma(z_n)$ は以下のように変形される。

$$\begin{aligned}
\gamma(z_n) = p(z_n|X) &= \frac{p(X|z_n)p(z_n)}{p(X)} \\
&= \frac{p(x_1, \ldots, x_n|z_n)p(x_{n+1}, \ldots, x_N|z_n)p(z_n)}{p(X)} \\
&= \frac{p(x_1, \ldots, x_n, z_n)p(x_{n+1}, \ldots, x_N|z_n)}{p(X)} \\
&= \frac{\alpha(z_n)\beta(z_n)}{p(X)}
\end{aligned} \tag{26}$$

[*15] PRML[1] の (13.18),(13.19) 式

[*16] PRML[1](13.24)。証明は余力があるときに Appendix に記載される

ただしここで

$$\alpha(z_n) := p(x_1, \ldots, x_n, z_n) \tag{27}$$
$$\beta(z_n) := p(x_{n+1}, \ldots, x_N | z_n) \tag{28}$$

と定義している[*17]。

$\alpha(z_n)$ と $\beta(z_n)$ を効率的に求めるための計算方法を構築する。そのために以下のように $\alpha(z_n)$ を変形する。

$$\begin{aligned}
\alpha(z_n) &= p(x_1, \ldots, x_n, z_n) \\
&= p(x_1, \ldots, x_n | z_n) p(z_n) \\
&= p(x_n | z_n, x_1, \ldots, x_{n-1}) p(x_1, \ldots, x_{n-1} | z_n) p(z_n) \\
&= p(x_n | z_n) p(x_1, \ldots, x_{n-1} | z_n) p(z_n) \\
&= p(x_n | z_n) p(x_1, \ldots, x_{n-1}, z_n) \\
&= p(x_n | z_n) \sum_{z_{n-1}} p(x_1, \ldots, x_{n-1}, z_{n-1}, z_n) \\
&= p(x_n | z_n) \sum_{z_{n-1}} p(x_1, \ldots, x_{n-1}, z_n | z_{n-1}) p(z_{n-1}) \\
&= p(x_n | z_n) \sum_{z_{n-1}} p(x_1, \ldots, x_{n-1} | z_n, z_{n-1}) p(z_n | z_{n-1}) p(z_{n-1}) \\
&= p(x_n | z_n) \sum_{z_{n-1}} p(x_1, \ldots, x_{n-1} | z_{n-1}) p(z_n | z_{n-1}) p(z_{n-1}) \\
&= p(x_n | z_n) \sum_{z_{n-1}} p(x_1, \ldots, x_{n-1}, z_{n-1}) p(z_n | z_{n-1}) \\
&= p(x_n | z_n) \sum_{z_{n-1}} \alpha(z_{n-1}) p(z_n | z_{n-1})
\end{aligned} \tag{29}$$

この式は $\alpha(z_{n-1})$ を算出してから $\alpha(z_n)$ を算出するという意味で、n に関して、言うなれば時間の向きに関して**前向きな（フォワード）再帰**計算となっているので、フォワード α 再帰と呼ぶ事にする。このフォワード再帰計算を行うための初期条件 $\alpha(z_1)$ は

$$\begin{aligned}
\alpha(z_1) &= p(x_1, z_1) = p(x_1 | z_1) p(z_1) \\
&= \left(\prod_{k=1}^{K} \pi_k^{z_{1k}} \right) \left(\prod_{k=1}^{K} p(x_k | \phi_k)^{z_{1k}} \right) \\
&= \prod_{k=1}^{K} \{ \pi_k p(x_1 | \phi_k) \}^{z_{1k}}
\end{aligned} \tag{30}$$

[*17] PRML[1](13.34)(13.35)

として求める事が出来る。同様に $\beta(z_n)$ に関する再帰式を導出する。

$$\begin{aligned}
\beta(z_n) &= p(x_{n+1},\ldots,x_N|z_n) \\
&= \sum_{z_{n+1}} p(x_{n+1},\ldots,x_N,z_{n+1}|z_n) \\
&= \sum_{z_{n+1}} p(x_{n+1},\ldots,x_N|z_{n+1},z_n)p(z_{n+1}|z_n) \\
&= \sum_{z_{n+1}} p(x_{n+1},\ldots,x_N|z_{n+1})p(z_{n+1}|z_n) \\
&= \sum_{z_{n+1}} p(x_{n+2},\ldots,x_N|z_{n+1},x_{n+1})p(x_{n+1}|z_{n+1})p(z_{n+1}|z_n) \\
&= \sum_{z_{n+1}} p(x_{n+2},\ldots,x_N|z_{n+1})p(x_{n+1}|z_{n+1})p(z_{n+1}|z_n) \\
&= \sum_{z_{n+1}} \beta(z_{n+1})p(x_{n+1}|z_{n+1})p(z_{n+1}|z_n)
\end{aligned} \tag{31}$$

これは α の再帰式とは逆に $\beta(z_{n+1}$ を算出した後 $\beta(z_n)$ を算出するという n の向きに対して**後向き（バックワード）再帰**計算となっているので、バックワード β 再帰と呼ぶ事にする。これで PRML[1] の (13.38) 式を導出することが出来た。

β の再帰計算の初期値となる $\beta(z_N)$ は $\gamma(z_N) = p(z_N|X)$,$\alpha(z_N) = p(x_1,\ldots,x_N,z_N)$ である一方、

$$\gamma(z_N) = \frac{\alpha(z_N)\beta(z_N)}{p(X)} \tag{32}$$

となるので

$$\beta(z_N) = 1 \tag{33}$$

となる。

$$\sum_{z_n} \gamma(z_n) = \sum_{z_n} p(z_n|X) = 1 \tag{34}$$

$$1 = \sum_{z_n} \gamma(z_n) = \sum_{z_n} \frac{p(\alpha(z_n)\beta(z_n)}{p(X)} = \frac{1}{p(X)} \sum_{z_n} p(\alpha(z_n)\beta(z_n) \tag{35}$$

$$\therefore p(X) = \sum_{z_n} p(\alpha(z_n)\beta(z_n) \tag{36}$$

として尤度関数 $p(X)$ を計算することが出来る。上式は任意の $n \in \{1,\cdots,N\}$ に対して成立する事、またで見たように $\beta(z_N) = 1$ である事を勘案すると

$$p(X) = \sum_{z_N} p(\alpha(z_N)) \tag{37}$$

と書くことが出来る。これで PRML[1] の (13.41)(13.42) を導出する事が出来た。

$\xi(z_{n-1}, z_n)$ を求めるには以下のように計算する。

$$\begin{aligned}
\xi(z_{n-1}, z_n) &= p(z_{n-1}, z_n|X) \\
&= \frac{p(X|z_{n-1}, z_n)p(z_{n-1}, z_n)}{p(X)} \\
&= \frac{p(x_{1:n-1}|z_{n-1})p(x_n|z_n)p(x_{n+1:N}|z_n)p(x_{n+1:N}|z_n)p(z_n|z_{n-1})p(z_{n-1}}{p(X)} \\
&= \frac{p(x_{1:n-1}|z_{n-1})p(z_{n-1})p(x_n|z_n)p(x_{n+1:N}|z_n)p(x_{n+1:N}|z_n)p(z_n|z_{n-1}}{p(X)} \\
&= \frac{\alpha(z_{n-1})p(x_n|z_n)p(x_{n+1:N}|z_n)p(x_{n+1:N}|z_n)\beta(z_n)}{p(X)}
\end{aligned} \tag{38}$$

フォワード (α) 再帰とバックワード (β) 再帰の計算結果を用いて $\xi(z_{n-1}, z_n)$ を計算する事が出来る。

ここまで算出してきた道具で EM アルゴリズムを完成させることが出来るのでその手順をまとめておこう。

1. パラメータ $\theta^{old} = (\pi, A, \phi)$ を適当に定める
2. フォワード α 再帰、バックワード β 再帰を実行し、$\gamma(z_n), \xi(z_{n-1}, z_n)$ を求める（E ステップ完了）
3. 式を用いて更新されたパラメータ θ^{new} を求める（M ステップ）
4. 2,3 をパラメータ θ が収束するまで交互に繰り返す

参考文献

[1] C.M. ビショップ, 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, 丸善出版, 2008.