予測と推定の計算理論的基礎

石井 信 (奈良先端科学技術大学院大学)
佐藤 雅昭 (国際電気通信基礎技術研究所)

1 はじめに

我々がある状況において行動する過程について考えてみる。時刻 t における状況を $x(t)$、その時の行動を $y(t)$ とすると、行動の選択過程は $y(t) = F(x(t))$ と表現することができる。脳の行なっていることの多くを高度に抽象化すれば、この F を決めていることに他ならない。例えば人間の運動の場合、$x(t)$ は環境および身体の状態全てを含むものであり、$y(t)$ は筋肉を制御する信号である。

何らかの経験によって F を決定することを「学習」と総称する。各 $x(t)$ について模範となるべき $y(t)$ が経験を通じて教えられる時、教師あり学習という。あるいは、関数 F を決めるという意味で、関数近似とも呼ばれる。本稿では、この関数近似の問題について述べる。

脳において関数近似ができたとして、その重要な機能は予測である。我々は日常の多くの場面で予測を行なっている。例えば自動車の動きが予測できなければ道路も横断できない。話相手の反応を予測できなければスムーズなコミュニケーションもできないであろう。こういう予測は時系列の予測として抽象化できる。

予測のもう一つの重要な問題は内部状態の予測である。我々が大きな自由度を持つシステムの挙動の予測を行なうことができるのは、その内部状態に関する予測を行なっているからである。例えば新聞に目を通して、政治や経済の情勢から株式投資銘柄を決めることができるのは、政治経済システムの状態をおぼろげながらも予測しているからに他ならない。コミュニケーションにおいても、相手との会話を通じて相手の内部状態の予測を行ない、予測に基づき良い行動の選択を行なっていると考えられる。

脳の計算理論を考える上では、関数近似と予測は重要な問題である。そこで、本稿では、その理論的な基礎を説明する。

2 最小二乗法と関数近似

入力変数 x と出力変数 y との間の関数関係を観測データから推定するのが関数近似の問題である。その際にはいかなる関数 (族) を用いて近似するのかをアプリオリに与えることがしばしば行なわれる。この場合、関数族をモデル、関数族から一つの関数を選び出すことをパラメータ推定と呼ぶ。

最も簡単な場合として、T 個の観測データからなるデータセット $\{(x(t), y(t))|t = 1, ..., T\}$ に対して、線形モデル $y = ax$ を仮定する。x および y は 1 次元 (スカラー) とする。スカラー a を決めるのがパラメータ推定である。その推定の規範として、二乗誤差

$$E = \sum_{t=1}^{T} (y(t) - ax(t))^2 \tag{1}$$

を最小にすることを考える。二乗誤差関数 E はデータセットが与えられた際のパラメータ a についての関数である。最小点では停留条件が成り立つので、

$$\frac{\partial E}{\partial a} = -2\sum_{t=1}^{T}(y(t) - ax(t))x(t) = 2\sum_{t=1}^{T}(ax^2(t) - x(t)y(t)) = 0 \tag{2}$$

となる。すなわち必要条件 (十分条件でもある) は、

$$a = \frac{\sum_{t=1}^{T} x(t)y(t)}{\sum_{t=1}^{T} x^2(t)} \tag{3}$$

である。(3) 式の分子は入出力データの相関、分母は入力データの分散という形をしている。これを最小二乗法という。ここでの議論は 1 次元で行なったが、スカラー同士の積を内積に替えることにより、ベクトル (多次元) の場合も全く同様になる。

しかし $y = ax$ は原点を通る直線であり、それによって近似できる関数は極めて限られる。そこで M 個の関数のセット $\{k_i(x)|i = 1, ..., M\}$ を用いて、その線形結合により関数を近似することを考える。

$$y = \sum_{j=1}^{M} a_j k_j(x) \tag{4}$$

ここで各 $k_i(x)$ はカーネル関数と呼ばれる。カーネル関数は一般に非線形なものが用いられるが、パラメータセット $\{a_i|i = 1, ..., M\}$ から見ると線形モデルであるため、パラメータ線形モデルと呼ばれる。パラメータ線形モデルの例をいくつか挙げておく。

- 多項式近似

$$y = a_0 + a_1 x + a_2 x^2 + ... + a_{M-1}x^{M-1} = \sum_{j=0}^{M-1} a_j x^j \tag{5}$$

- 動径基底関数 (Radial basis functions; RBF)

$$y = \sum_{j=1}^{M} a_j \exp\left(-\frac{1}{2\sigma^2}|x - \mu_j|^2\right) \tag{6}$$

ここで $|\cdot|$ はユークリッドノルムである。

近年盛んに研究されているサポートベクタマシン (Support vector machine; SVM) もパラメータ線形モデルの一つであるが、最小二乗法と若干異なる規範でパラメータ推定が行なわれる。

データセット $\{(x(t), y(t))|t = 1, ..., T\}$ に対する二乗誤差

$$E = \sum_{t=1}^{T}\left|y(t) - \sum_{j=1}^{M} a_j k_j(x(t))\right|^2 \tag{7}$$

による最小二乗法を用いてパラメータを決める。上と同様に、

$$\begin{aligned}\frac{\partial E}{\partial a_i} &= -2\sum_{t=1}^{T}\left(y(t) - \sum_{j=1}^{M} a_j k_j(x(t))\right)k_i(x(t)) \\ &= 2\sum_{t=1}^{T} k_i(x(t))\sum_{j=1}^{M} a_j k_j(x(t)) - 2\sum_{t=1}^{T} k_i(x(t))y(t) = 0\end{aligned} \tag{8}$$

となる。

$$z_i \equiv \frac{1}{T}\sum_{t=1}^{T} k_i(x(t))y(t) \tag{9a}$$

$$K_{i,j} \equiv \frac{1}{T}\sum_{t=1}^{T} k_i(x(t))k_j(x(t)) \tag{9b}$$

を定義することにより、必要条件 (8) は

$$\sum_{j=1}^{M} K_{i,j}a_j = z_i \qquad (i = 1, ..., M) \tag{10}$$

と M 次の連立方程式、あるいは、ベクトル表記で

$$Ka = z \tag{11}$$

と書くことができる。カーネル共分散行列 K は対称 $K_{i,j} = K_{j,i}$ であり、かつ任意のベクトル v に対して、$v'Kv = \frac{1}{T}\sum_{t=1}^{T}\left(\sum_{i=1}^{M} k_i(x(t))v_i\right)^2 \geq 0$ であるから、非負行列である。ここでプライム ($'$) は転置を表す。また特別な場合を除いて K は正定値行列であるので、逆行列が存在して、

$$a = K^{-1}z \tag{12}$$

となる。これがパラメータ線形モデルに対する最小二乗法である。また K の正定値性は、最小二乗解の十分性を意味している。

パラメータに関して非線形なモデルも用いられる。最も有名なモデルが階層型パーセプトロン (Multi-layered perceptron; MLP) である。今までの記述との対応を考慮した例として、以下の 3 層の MLP を考える。

$$y = \sum_{i=1}^{M} a_i \mathrm{sig}(h_i) \tag{13a}$$

$$h_i = \sum_{j=1}^{N} b_{i,j}x_j \tag{13b}$$

N は入力次元、M はカーネル (中間層ユニット) の数である。関数 sig は非線形で有界な関数なら何でも良いが、一般にはシグモイド関数

$$\mathrm{sig}(h) \equiv \frac{1}{1+\exp(-h)} \tag{14}$$

が用いられる。MLP では 2 種のパラメータがある。中間層から出力への重み a_i については、上記の最小二乗法で行なうこともできるが、一般には $a_i, b_{i,j}$ の学習ともに、勾配法

$$\Delta\theta_k = -\eta\frac{\partial E}{\partial \theta_k} \tag{15}$$

が用いられる。ここで E は二乗誤差関数であり、パラメータである $a_i, b_{i,j}$ についての関数である。θ_k は a_i と $b_{i,j}$ のいずれかを表現している。また η は正数であり、学習係数と呼ばれる。

勾配法がうまく行く原理を簡単に述べる。パラメータ θ_k が $\Delta\theta_k$ だけ変更される際の二乗誤差関数 E の変化 ΔE について考える。

$$\Delta E \approx \sum_k \Delta\theta_k \frac{\partial E}{\partial \theta_k} = -\eta \sum_k \left(\frac{\partial E}{\partial \theta_k}\right)^2 \leq 0 \tag{16}$$

となるので、二乗誤差関数は (ほぼ) 減少することが分かる。二乗誤差関数には明らかな下界 (0) が存在するので、少なくともそれ以上の値をとる極小点で収束する。すなわち最小二乗推定が実現できる。ただし、パラメータ線形の場合と異なり、大域的な最適性は保証できず、局所的な最適性である。なお、一般に MLP モデルの学習法は誤差逆伝搬法と呼ばれるが、勾配法による (局所) 最小二乗法に他ならない。実際の繁雑なアルゴリズムは本質的でなく、その原理は (15) 式で十分である。

(演習 1)　データセットおよび基底の中心位置が与えられた場合の動径基底関数のパラメータ推定を最小二乗法を用いて行なう。

$$y = \sum_{j=1}^{M} a_j \exp\left(-\frac{1}{2\sigma^2}|x-\mu_j|^2\right) \tag{17}$$

基底の中心位置は、しばしば最尤推定法 (後述) で決定される。

3　ダイナミクスの学習

観測変数 x の時系列 $\{x(t)|t=1,...,T\}$ から、その時系列を発生したダイナミクスを推定し、時系列の将来を予測することは重要な問題である。時刻 $t, t-1, ..., t-M+1$ における観測値の線形結合により、次の時刻での観測変数の推定値 $\hat{x}(t+1)$ を推定するモデルを自己回帰 (Auto regressive; AR) モデルと呼ぶ。

$$\hat{x}(t+1) = a_1 x(t) + ... + a_M x(t-M+1) \tag{18}$$

M は AR の次数と呼ばれる。

時刻 t について、入力 $X(t) \equiv (x(t), x(t-1), ..., x(t-M+1))$、出力 $y(t) \equiv \hat{x}(t+1)$ の関数近似問題とみなすと、パラメータ線形モデル

$$y(t) = \sum_{i=1}^{M} a_i X_i \tag{19}$$

による推定問題である。したがって二乗誤差関数

$$E = \sum_{t=M}^{T-1} (x(t+1) - a' X(t))^2 \tag{20}$$

を定義することにより、前章と同じように最小二乗法によりパラメータベクトル $\boldsymbol{a}$ を決めることができる。最小二乗法のための線形方程式 (11) を AR では特に Yule-Walker 方程式と呼ぶこともある。

さて (19) 式の右辺は状態 $X(t)$ の関数であり、

$$y(t) \equiv \hat{x}(t+1) = F(X(t)) \tag{21}$$

の形をしている。これを以後、モデルの時間発展方程式と呼ぶことにする。AR では F として特に線形関数を考えた。一方で、F として例えば MLP などの非線形関数を考えることもできる。学習、すなわちパラメータの推定の目的は、モデルの時間発展がシステムのダイナミクスに近くなるようにすることである。

データセットを用いてパラメータを決定した後の時間発展方程式、すなわち模倣されたダイナミクスを用いて、時系列の推定を行なうことができる。AR では (18) 式である。これを 1 ステップ予測と呼ぶ。推定された $\hat{x}(t+1)$ を用いると、$\hat{X}(t+1) = (\hat{x}(t+1), x(t), ..., x(t-M+2))$ が得られる。これと関数 F から $\hat{x}(t+2)$ が得られ、それから $\hat{X}(t+2)$ が得られる。こうして次々と時系列の推定ができる。こうした逐次的な推定をマルチステップ予測と呼ぶ。時系列を生成しているシステムがカオスなどの非線形性の強いシステムである場合 AR モデルは適当ではないことが多い。その場合、MLP や RBF などの非線形モデルを用いて推定を行なう必要がある。

システムに外部入力 $u(t)$ がある場合は、(18) 式を拡張して、

$$\hat{x}(t+1) = a_1 x(t) + ... + a_M x(t-M+1) + b_1 u(t) + ... + b_K u(t-K-1) \tag{22}$$

で推定する手法がある。これを ARMA(Auto regressive moving average) と呼ぶ。K は MA の次数と呼ばれる。AR の場合と同様にパラメータ a, b は $x(t)$ と $u(t)$ の観測データを用いて最小二乗法で決めることができる。

さてこれまでは、暗黙に、システムのダイナミクスを規定する状態変数が観測変数と一致するという状況を仮定していた。したがってモデルの時間発展方程式である (21) 式はシステムの状態変数を用いて記述されていることになる。ここでより一般的な場合を考える。システムの状態変数を $Z(t) \equiv (z_1(t), ..., z_L(t))$ として、システムのダイナミクスを

$$Z(t+1) = F^*(Z(t)) \tag{23}$$

とする。L をシステム次元と呼ぶ。簡単のためノイズ (システムノイズ) は考えない。観測変数 $X(t)$ は $X(t) \equiv (x_1, ..., x_N) = CZ(t)$ で与えられるものとする。N を観測次元と呼ぶ。行列 C は $N \times L$ の観測行列である。

ここで特に考える状況は、観測次元がシステム次元よりも小さい、すなわち $N < L$ の場合である。これを部分観測という。例えば $X(t) = z_1(t)$ とすれば、状態変数の一つだけが観測できることになる。こうした場合でも、システムのダイナミクスである $F^*(Z)$ を、線形モデル、あるいは MLP や RBF などの非線形モデルによって近似することができる。学習には例えば最小二乗法を用いることができる。この時、システムの状態変数 $Z(t)$ は観測できない変数ということで、隠れ変数、あるいは内部状態変数と呼ばれる。システムが隠れ変数を持つ場合の学習法の多くでは、パラメータの推定と同時に、隠れ変数の推定をも行なう必要がある。決定論的な時間発展方程式によりモデル化する場合は、$Z(t)$ の初期状態が与えられれば、モデルを用いて $Z(t)$ の推定時系列が計算できるので、誤差関数のパラメータによる微分は時間逆向きの誤差逆伝搬法などを用いて計算できる。一方、4 章で述べる確率モデルでは隠れ変数 $Z(t)$ がしたがう確率分布を観測時系列 $\{X(t)|t = 1, ..., T\}$ から推定する必要がある。

しかし、隠れ変数の推定を含むダイナミクスの推定は、システムの状態変数が全て観測できる場合に比べて複雑で学習にも多大な計算時間を要する。一方で、埋め込み法を用いることにより、観測時系列から、隠れ変数を推定することなしに、システムのダイナミクスをモデル化することができる。例として状態変数 $Z(t)$ の一つの成分 $z_1(t)$ のみが観測できる場合を考える。時系列 $\{z_1(t)|t = 1, ...\}$ から遅れ座標 $Y(t)$ を以下のように定義する。

$$Y(t) \equiv (z_1(t), z_1(t-\tau), ..., z_1(t-\tau(M-1))) \qquad (t = \tau(M-1)+1, ...) \tag{24}$$

τ は遅れ時間、M は埋め込み次元と呼ばれる。$M \geq 2L+1$ が成り立つ時、状態変数 $Z(t)$ で記述されたシステムのダイナミクスを、遅れ座標 $Y(t)$ を用いたダイナミクスに変換できることが示されている。そこで遅れ座標を用いた時間発展則

$$z_1(t+1) = F(Y(t)) \tag{25}$$

でシステムのダイナミクスをモデル化することができる。(18) 式あるいは (21) 式は、(24)(25) 式において $\tau = 1$ とした場合になっている。

(演習 2)　時系列 $\{x(t)|t = 1, ..., T\}$ に対して AR 法を適用せよ。この時系列は実際に AR モデルによって生成されたものである。次数を色々と変えて実験してみよ。マルチステップ予測を行ない、実際の系列と比較せよ。

4　最尤推定法と EM アルゴリズム

確率変数 x の二つの確率分布 $a(x)$,$b(x)$ について、Kullback-Leibler (KL) ダイバージェンス $D(a\|b)$ を以下で定義する。

$$D(a\|b) \equiv \int dx a(x) \log\left(\frac{a(x)}{b(x)}\right) \tag{26}$$

KL ダイバージェンスは非負であり、0 となるのは (ほとんど全ての x について) $a(x) = b(x)$ である時に限る。これは以下で証明できる。$z > 0$ に対して、

$$z - \log z \geq 1 \qquad (\text{等号は } z = 1 \text{ の時}) \tag{27}$$

であるので、

$$\frac{b(x)}{a(x)} - \log\left(\frac{b(x)}{a(x)}\right) \geq 1 \tag{28}$$

が成り立つ。したがって、

$$\int dx a(x) \left(\frac{b(x)}{a(x)} - \log\left(\frac{b(x)}{a(x)}\right)\right) \geq \int dx a(x) = 1 \tag{29}$$

$a(x)$ が x の確率分布であるという性質を用いた。(29) 式の左辺は $\int dx b(x) + D(a\|b) = D(a\|b) + 1$ であるから

$$D(a\|b) \geq 0 \qquad (\text{等号は、ほとんど全ての } x \text{ について } a(x) = b(x) \text{ の時}) \tag{30}$$

となる。以上の性質により KL ダイバージェンスは二つの確率分布間の距離 (差) を計るのにしばしば用いられる。ただし $D(a\|b) \neq D(b\|a)$ と距離公理のうち対称性を満たしていない。そのため偽距離と呼ばれることもある。

確率変数 x に関する未知のデータ分布 $\rho(x)$ を θ によるパラメータ族 $P(x|\theta)$ で近似することを考える。KL ダイバージェンスは

$$\begin{aligned} D(\rho\|P) &= \int dx \rho(x) \log\left(\frac{\rho(x)}{P(x|\theta)}\right) \\ &= \int dx \rho(x) \log \rho(x) - \int dx \rho(x) \log P(x|\theta) \end{aligned} \tag{31}$$

で与えられる。KL ダイバージェンスを最小にすることが近似の目標であり、その時のパラメータ θ が最適と考える。(31) 式右辺第 1 項はパラメータ θ によらない定数であるから、KL ダイバージェンスの最小化は

右辺第 2 項の最大化と等価である。ここで右辺第 2 項は、$\log P(x|\theta)$ の未知分布 $\rho(x)$ についての期待値である点に注意する。さて、実際のデータ分布が分からないので、右辺第 2 項の期待値の計算は厳密にはできない。しかし、データ数がある程度あれば期待値はデータ平均で近似できる。すなわち、未知の分布 $\rho(x)$ からデータセット $\{x(t)|t=1,...,T\}$ が発生した時、次式が成り立つ。

$$\int dx\rho(x)\log P(x|\theta) \approx \frac{1}{T}\sum_{t=1}^{T}\log P(x(t)|\theta) \tag{32}$$

右辺は左辺の不偏推定量である。以上のことから、

$$L(\theta) \equiv \sum_{t=1}^{T}\log P(x(t)|\theta) \tag{33}$$

の最大化を θ について行えば良い。これが最尤推定法である。パラメータ θ についての関数 (33) を対数尤度関数、最尤推定法によって求められたパラメータを最尤推定量という。

さて、入力変数 x が与えられた時に、出力変数 $y(D$ 次元とする) が $y=ax+u$ で与えられると仮定して、データセット $\{(x(t),y(t))|t=1,...,T\}$ から、パラメータ a を求める問題を考える。ただし u は分散 σ^2 のガウスノイズである。このとき、y は中心位置 ax、分散 σ^2 の正規分布 $\mathcal{N}(y|ax,\sigma^2)$ にしたがう。

$$P(y|x,a) = \mathcal{N}(y|ax,\sigma^2) \equiv (2\pi\sigma^2)^{-\frac{D}{2}}\exp\left(-\frac{1}{2\sigma^2}(y-ax)^2\right) \tag{34}$$

ここで $\mathcal{N}(z|\mu,\sigma^2)$ は中心 μ、分散 σ^2 の確率変数 z についての正規分布である。中心位置が入力 x に依存しているため、(34) 式で与えられる y の分布は x の条件付きである点に注意する。(34) 式より、

$$\log P(y|x,a) = -\frac{1}{2\sigma^2}(y-ax)^2 - \frac{D}{2}\log(2\pi\sigma^2) \tag{35}$$

であるので、対数尤度関数は

$$L(a) = -\frac{1}{2\sigma^2}\sum_{t=1}^{T}(y(t)-ax(t))^2 + (a\text{ に依存しない項}) \tag{36}$$

となる。$L(a)$ を a について最大化することと、(1) 式の E を a について最小化することは等価である。したがってこの場合の最尤推定法は最小二乗法と等価である。これは最も簡単な場合について述べたが、2 章で述べたような一般的な関数近似モデルに関する最小二乗法も最尤推定法から導くことができる。モデルにガウスノイズが付加された形の確率モデルを考えると、その確率分布は $\exp(-\frac{E}{2\sigma^2})$ の形をしている。ここで E は二乗誤差関数である。すなわち、この場合にも最尤推定法は最小二乗法と等価になる。

最尤推定法を用いるためにはデータの分布に関する確率分布族をアプリオリに定める必要がある。最小二乗法の場合と同様にこの分布族をモデルと呼ぶが、特に確率分布のモデルであるので確率モデルと呼ばれる。また、データの出現の仕方をモデル化しているので (データ) 生成モデルと呼ばれることもある。

データの分布を近似することの応用の一つがクラスタリングである。今、M 個のクラスタの各々が等方等分散の正規分布をなしていて、そこからデータ $\{x(t)|t=1,...,T\}$ が観測されたと仮定する。以下の生成モデルを仮定する。

$$P(i|\theta) = \nu_i \tag{37a}$$

$$P(x|i,\theta) = \mathcal{N}(x|\mu_i,\sigma_i^2) \tag{37b}$$

パラメータは $\theta \equiv \{\nu_i, \mu_i, \sigma_i^2 | i = 1, ..., M\}$ である。ν_i は i 番めのクラスタからデータが出てきた確率を表し、混合比と呼ばれる。$P(i|\theta)$ が確率であるために、$\sum_{i=1}^{M} \nu_i = 1$ を満たす必要がある。この式から、

$$P(x|\theta) = \sum_{i=1}^{M} P(x, i|\theta) = \sum_{i=1}^{M} P(x|i, \theta) P(i|\theta) = \sum_{i=1}^{M} \nu_i \mathcal{N}(x|\mu_i, \sigma_i^2) \tag{38}$$

となる。データが複数の正規分布の混合から生成されたことを仮定しているので、これを混合正規分布と呼ぶ。

注意すべき点は、混合正規分布 (38) において、クラスタの指標 i は隠れ変数になっていることである。すなわち、データを観測した点では、(37) 式に対応する生成過程は観測できないため、データがどのクラスタから出てきたかは厳密には分からない。できるのは推測することだけである。

一般に、隠れ変数 z を持つ確率モデル $P(x|\theta) = \sum_z P(x, z|\theta)$ の最尤推定法について考える。説明の都合上 z を離散変数とするが、連続の場合でも同様である。データセット $\{x(t)|t = 1, ..., T\}$ に対する対数尤度

$$L(\theta) \equiv \sum_{t=1}^{T} \log P(x(t)|\theta) = \sum_{t=1}^{T} \log \left(\sum_z P(x(t), z|\theta) \right) \tag{39}$$

を最大化する θ を求めたい。必要条件である停留条件 $\partial L / \partial \theta = 0$ は以下のように計算できる。

$$\begin{aligned} \frac{\partial L(\theta)}{\partial \theta} &= \sum_{t=1}^{T} \sum_z \frac{\partial P(x(t), z|\theta)/\partial \theta}{\sum_{z'} P(x(t), z'|\theta)} \\ &= \sum_{t=1}^{T} \sum_z P(z|x(t), \theta) \frac{\partial}{\partial \theta} \log P(x(t), z|\theta) = 0 \end{aligned} \tag{40}$$

ここで、

$$P(z|x(t), \theta) \equiv \frac{P(x(t), z|\theta)}{\sum_{z'} P(x(t), z'|\theta)} \tag{41}$$

はデータ $x(t)$ を観測した際の、隠れ変数 z の事後確率 (posterior) と呼ばれる。

非線形方程式である (40) 式をモデルパラメータ θ について解くことは、例えば勾配法を用いれば可能であるが、ここでは (40) 式の形に注目して、以下のような繰り返しアルゴリズムを考える。

1. E (Expectation) ステップ

   (a) 各データ $x(t)$ に対して、現在のパラメータの推定値 $\bar{\theta}$ を用いて隠れ変数 z の事後確率 $P(z|x(t), \bar{\theta})$ を (41) 式により計算する。

   (b) 隠れ変数を含む (完全) データセット $\{(x(t), z(t))|t = 1, ..., T\}$ に対する対数尤度 $\sum_{t=1}^{T} \log P(x(t), z(t)|\theta)$ の、隠れ変数の予測事後確率についての期待値

   $$Q(\theta|\bar{\theta}) = \sum_{t=1}^{T} \sum_{z(t)} P(z(t)|x(t), \bar{\theta}) \log P(x(t), z(t)|\theta) \tag{42}$$

   を計算する。

2. M (Maximization) ステップ

   期待対数尤度 $Q(\theta|\bar{\theta})$ をパラメータ θ について最大化する。すなわち

   $$\frac{\partial Q(\theta|\bar{\theta})}{\partial \theta} = \sum_{t=1}^{T} \sum_z P(z|x(t), \bar{\theta}) \frac{\partial}{\partial \theta} \log P(x(t), z(t)|\theta) = 0 \tag{43}$$

   の解を求める。

3. 求められたパラメータを $\bar{\theta}$ としてステップ 1 に戻る。ステップ 1 と 2 の繰り返しが収束すれば終る。

このアルゴリズムを EM アルゴリズムと呼ぶ。(40) 式と (43) 式の違いに注意する。(43) 式は (40) 式中の一方の θ を $\bar{\theta}$ に置き換えたものである。一般に (40) 式が解析的に解けない場合でも (43) 式は解ける場合があり、それが EM アルゴリズムのメリットである。

E ステップと M ステップを交互に繰り返すことにより、対数尤度が増大することが証明できるため、漸近的にパラメータ θ の最尤推定量を求めることができる。仮に EM アルゴリズムが収束したとすると、$\theta = \bar{\theta}$ となるが、この時、(43) 式と (40) 式は同じになるので、最尤推定量が求められていることはすぐに分かる。なお、ここで得られる推定量は一般的に局所的最尤解である。

最も簡単な場合の EM アルゴリズムを導出するために、$\nu_i = 1/M, \sigma_i^2 = \sigma^2 (i = 1, ..., M)$ の場合の混合正規分布を考える。入力 x の次元は N とする。ここで σ^2 は固定パラメータであり推定はしないものとする。

$$P(x, i|\theta) = \frac{1}{M}(2\pi\sigma^2)^{-\frac{N}{2}} \exp\left(-\frac{1}{2\sigma^2}|x - \mu_i|^2\right) \tag{44a}$$

$$P(x|\theta) = \sum_{i=1}^{M} P(x, i|\theta) \tag{44b}$$

パラメータは $\theta \equiv \{\mu_i | i = 1, ..., M\}$ である。

- E ステップ

$$P(i|x(t), \bar{\theta}) = \frac{P(x(t), i|\bar{\theta})}{\sum_{j=1}^{M} P(x(t), j|\bar{\theta})} = \frac{\exp\left(-\frac{1}{2\sigma^2}|x(t) - \bar{\mu}_i|^2\right)}{\sum_{j=1}^{M} \exp\left(-\frac{1}{2\sigma^2}|x(t) - \bar{\mu}_j|^2\right)} \tag{45}$$

- M ステップ

$$\log P(x, i|\theta) = -\frac{1}{2\sigma^2}|x - \mu_i|^2 - \log M - \frac{N}{2}\log(2\pi\sigma^2) \tag{46}$$

であるので、

$$\begin{aligned} Q(\theta|\bar{\theta}) &= \sum_{t=1}^{T}\sum_{i=1}^{M} P(i|x(t), \bar{\theta}) \log P(x(t), i|\theta) \\ &= -\frac{1}{2\sigma^2}\sum_{t=1}^{T}\sum_{i=1}^{M} P(i|x(t), \bar{\theta})|x - \mu_i|^2 - T\left(\log M + \frac{N}{2}\log(2\pi\sigma^2)\right) \qquad (47) \\ &= -\frac{T}{2\sigma^2}\sum_{i=1}^{M}\left(\langle|x|^2\rangle_i - 2\langle x\rangle_i\mu_i + \langle 1\rangle_i|\mu_i|^2\right) - T\left(\log M + \frac{N}{2}\log(2\pi\sigma^2)\right) \qquad (48) \end{aligned}$$

ここで、

$$\langle 1\rangle_i \equiv \frac{1}{T}\sum_{t=1}^{T} P(i|x(t), \bar{\theta}) \tag{49a}$$

$$\langle x\rangle_i \equiv \frac{1}{T}\sum_{t=1}^{T} x(t) P(i|x(t), \bar{\theta}) \tag{49b}$$

$$\langle|x|^2\rangle_i \equiv \frac{1}{T}\sum_{t=1}^{T} |x(t)|^2 P(i|x(t), \bar{\theta}) \tag{49c}$$

を用いた。Q をパラメータ θ(この場合は μ_i) について最大化することは、

$$\frac{\partial Q}{\partial \mu_i} = -\frac{T}{2\sigma^2}\left(2\langle 1\rangle_i\mu_i - 2\langle x\rangle_i\right) = 0 \tag{50}$$

を解くことによって得られる。すなわち

$$\mu_i = \langle x \rangle_i / \langle 1 \rangle_i \tag{51}$$

となる。

結局、E ステップは (45) 式、M ステップは (51) 式で与えられる。この結果についての解釈を与える。(45) 式は Gaussian soft-max と呼ばれる関数である。理解の都合上、分散 σ^2 が小さいとする。(45) 式は、各データ $x(t)$ について、現在の中心位置 $\bar{\mu}_i$ が $x(t)$ にユークリッド距離の意味で最も近いものを選び、そのクラスタへの所属確率をほぼ 1 としている。すなわち現在の中心位置を用いた最小距離規範 (Nearest neighbor 法; NN 法) によるクラスタリングを行なっている。(51) 式では、NN 法によってクラスタリングを行なった後の、各クラスタ内のデータの平均値によってクラスタ中心である μ_i を変更している。(51) 式の分子は各クラスタの構成要素についてデータの和、分子は要素数に対応している。すなわちこの EM アルゴリズムはクラスタリングでしばしば用いられる K 平均法そのものである。違う言い方をすれば、K 平均法は EM アルゴリズムを極めて簡素化したものである。

(演習 3-1)　上記の最も簡単な場合の EM アルゴリズムを用いて実際の 2 次元散布データセット $\{x(t)|t = 1, ..., T\}$ のクラスタリングを行なう。クラスタの数 M を色々と変えて実験をしてみよ。
(演習 3-2)　(45) 式において、分散 σ^2 を可変パラメータとした時の EM アルゴリズムを導出せよ。

5　カルマンフィルター

3 章で述べた隠れ変数 (内部状態変数) を含むダイナミクスのモデル化の手法の一つとしてカルマンフィルターについて述べる。カルマンフィルタは観測時系列からシステムの内部状態変数を推定する手法として広く用いられている。

対象とするシステムの状態変数が $z(t) \equiv (z_1(t), ..., z_L(t))$ であり、システムのダイナミクスが以下のような確率的線形方程式で表わされるものとする。

$$z(t+1) = Az(t) + u(t) + \xi(t) \tag{52}$$

A は $L \times L$ 行列で、$u(t)$ は時刻 t における外部入力である。$\xi(t)$ は白色ガウスノイズであり、システムノイズと呼ばれる。すなわち $\xi(t)$ は平均 0、共分散行列 U をもつ正規分布にしたがい、$t \neq s$ の時、$\xi(t)$ と $\xi(s)$ は無相関である。以上より、

$$P(z(t+1)|z(t)) \propto \exp\left[-\frac{1}{2}(z(t+1) - Az(t) - u(t))' U^{-1} (z(t+1) - Az(t) - u(t))\right] \tag{53}$$

が成り立つ。なお本章では正規分布の正則化係数は重要な役割を果たさないので、簡単のために省略する。条件付き確率 (53) は、時刻 t での状態変数が $z(t)$ であった時に、次の時刻 $t+1$ での状態変数が $z(t+1)$ となる状態遷移確率を表わしている。このように、次の時刻の状態がシステムの履歴によらずに現在の状態のみで決まるような確率モデルを、マルコフ過程と呼ぶ。

観測変数 $x(t) \equiv (x_1(t), ..., x_N(t))$ は状態変数 $z(t)$ から

$$x(t) = Cz(t) + w(t) \tag{54}$$

という関係式で変換されているとする。C は $N \times L$ の観測行列である。$w(t)$ は平均 0、共分散行列 V をもつ白色ガウスノイズであり、観測ノイズと呼ばれる。これから

$$P(x(t)|z(t)) \propto \exp\left[-\frac{1}{2}(x(t)-Cz(t))'V^{-1}(x(t)-Cz(t))\right] \tag{55}$$

となる。

以後、行列 A, C, V, U が時刻によらずかつ既知であり、外部入力が無い $(u(t)=0)$ 場合に、観測時系列 $\{x(t)|t=1,...\}$ から内部状態変数 $z(t)$ の推定を行なう問題を考える。時刻 t までの観測時系列 $X\{t\} = \{x(s)|s=1,...,t\}$ をもとに時刻 t での状態変数 $z(t)$ が、平均 $\hat{z}(t)$、共分散行列 $Q(t)$ を持つ正規分布にしたがうことがわかっているものとする。すなわち、

$$P(z(t)|X\{t\}) \propto \exp\left[-\frac{1}{2}(z(t)-\hat{z}(t))'Q^{-1}(t)(z(t)-\hat{z}(t))\right] \tag{56}$$

である。これをもとに、次の時刻 $t+1$ での状態変数の分布を推定することができる。

$$P(z(t+1)|X\{t\}) = \int dz(t) P(z(t+1)|z(t))P(z(t)|X\{t\}) \tag{57}$$

(53) 式と (56) 式から、$P(z(t+1)|z(t))P(z(t)|X\{t\})$ は次式のようになる。

$$P(z(t+1)|z(t))P(z(t)|X\{t\}) \propto \exp\left[-\frac{1}{2}E_S(z(t+1),z(t))\right] \tag{58a}$$

$$E_S(z(t+1),z(t)) = (z(t+1)-Az(t))'U^{-1}(z(t+1)-Az(t)) + (z(t)-\hat{z}(t))'Q^{-1}(t)(z(t)-\hat{z}(t)) \tag{58b}$$

また E_S は $z(t+1)$ と $z(t)$ に関して 2 次形式になっており、以下のように計算することができる。

$$\begin{aligned} E_S(y,z) &= (y-Az)'U^{-1}(y-Az) + (z-\hat{z})Q^{-1}(z-\hat{z}) \\ &= (z-\bar{z})R^{-1}(z-\bar{z}) + (y-A\hat{z})\tilde{Q}^{-1}(y-A\hat{z}) \end{aligned} \tag{59}$$

ここで、

$$\bar{z} = \hat{z} + RA'U^{-1}(y-A\hat{z}) \tag{60a}$$

$$R = (Q^{-1}+A'U^{-1}A)^{-1} \tag{60b}$$

$$\tilde{Q} = U + AQA' \tag{60c}$$

である。(58b)(59) 式より (57) 式中の $z(t)$ に関する積分はガウス積分になる。積分を行なうと、(59) 式右辺第 1 項は定数になるので、右辺第 2 項のみが残る。すなわち次式が成り立つ。

$$P(z(t+1)|X\{t\}) \propto \exp\left[-\frac{1}{2}(z(t+1)-A\hat{z}(t))'\tilde{Q}^{-1}(t+1)(z(t+1)-A\hat{z}(t))\right] \tag{61a}$$

$$\tilde{Q}(t+1) = U + AQ(t)A' \tag{61b}$$

$\tilde{Q}(t+1)$ は、時刻 t における状態分布の知識 (56) とシステムダイナミクスの知識 (53) をもとに、次時刻 $t+1$ の状態分布推定を行なった時の状態分布の共分散行列である。(61b) 式からこの共分散が、システムのダイナミクスで定まる値 $AQ(t)A'$ よりもシステムノイズ分の U だけ増大していることがわかる。一方 (61a) 式より、状態変数の期待値は次式で与えられる。

$$\tilde{z}(t+1) = A\hat{z}(t) \tag{62}$$

以上は、時刻 t までの観測時系列 $X\{t\}$ をもとに時刻 $t+1$ での状態分布の推定を行なう問題を考えてきた。このように新たな観測を行なわずに、ダイナミクスの知識のみを用いて状態分布推定を行なった場合、ノイズのために毎時刻分散が増大してゆき、推定精度が落ちてゆく。また (61) 式より、状態変数の初期分布が正規分布で与えられる場合、以後の状態分布はずっと正規分布になることが分かる。

次に時刻 $t+1$ で新たな観測データ $x(t+1)$ が得られた時に、状態分布の推定がどのように変更されるかを調べる。ベイズの定理より次式が成り立つ。

$$P(z(t+1)|X\{t+1\}) = P(z(t+1)|x(t+1), X\{t\}) = \frac{P(z(t+1), x(t+1)|X\{t\})}{P(x(t+1)|X\{t\})} \tag{63}$$

ここで $P(x(t+1)|X\{t+1\})$ は未知だが、$z(t+1)$ に依存しないので、$z(t+1)$ の分布を求める際には単なる定数として扱える。これから

$$P(z(t+1)|X\{t+1\}) \propto P(z(t+1), x(t+1)|X\{t\}) = P(x(t+1)|z(t+1))P(z(t+1)|X\{t\}) \tag{64}$$

が成り立つ。(64) 式の右辺は (55)(61)(62) 式から以下で与えられる。

$$P(x(t+1)|z(t+1))P(z(t+1)|X\{t\}) \propto \exp\left[-\frac{1}{2}E_O(x(t+1), z(t+1))\right] \tag{65a}$$

$$\begin{aligned} E_O(x(t+1), z(t+1)) &= (x(t+1) - Cz(t+1))'V^{-1}(x(t+1) - Cz(t+1)) \\ &\quad +(z(t+1) - \tilde{z}(t+1))'\tilde{Q}^{-1}(t+1)(z(t+1) - \tilde{z}(t+1)) \end{aligned} \tag{65b}$$

E_O に対して、(59) 式の E_S と同様の計算をすることができる。

$$\begin{aligned} E_O(x, z) &= (x - Cz)'V^{-1}(x - Cz) + (z - \tilde{z})'\tilde{Q}^{-1}(z - \tilde{z}) \\ &= (z - \bar{z})\hat{Q}^{-1}(z - \bar{z}) + (x - C\tilde{z})W^{-1}(x - C\tilde{z}) \end{aligned} \tag{66}$$

ここで、

$$\bar{z} = \tilde{z} + \hat{Q}C'V^{-1}(x - C\tilde{z}) \tag{67a}$$

$$\hat{Q} = (\tilde{Q}^{-1} + C'V^{-1}C)^{-1} \tag{67b}$$

$$W = V + C\tilde{Q}C' \tag{67c}$$

である。(66) 式の右辺第 2 項は z に依存しないので定数として扱える。結局次式が成り立つ。

$$P(z(t+1)|X\{t+1\}) \propto \exp\left[-\frac{1}{2}(z(t+1) - \hat{z}(t+1))'Q^{-1}(t+1)(z(t+1) - \hat{z}(t+1))\right] \tag{68}$$

ここで $\bar{z}$ を $\hat{z}$ に書き直した。$x(t+1)$ を観測した後での $z(t+1)$ の期待値 $\hat{z}(t+1)$ と共分散 $Q(t+1)$ は (67) 式を用いて下のように与えられる。

$$\hat{z}(t+1) = \tilde{z}(t+1) + K(t+1)(x(t+1) - C\tilde{z}(t+1)) \tag{69a}$$

$$Q(t+1) = (\tilde{Q}^{-1}(t+1) + C'V^{-1}C)^{-1} \tag{69b}$$

$$K(t+1) = Q(t+1)C'V^{-1} \tag{69c}$$

(61b)(62)(69) 式がカルマンフィルターである。以下で式の意味を簡単に説明する。時刻 t での状態の期待値 $\hat{z}(t)$ がわかっている時、システムダイナミクスを用いた予測により $\tilde{z}(t+1) = A\hat{z}(t)$ が得られる ((62)

式)。この予測は観測データ $x(t+1)$ を用いて修正される。すなわち、観測データと内部状態の予測値 $\tilde{z}(t+1)$ から予測される観測値との差 $(x(t+1) - C\tilde{z}(t+1))$ を用いて $\tilde{z}(t+1)$ が修正される ((69a) 式)。この修正のための比例係数行列 $K(t+1)$ はカルマンゲインと呼ばれる。

状態分布の共分散の大きさがどのように変化するかを見るために、特に 1 次元の場合を考える。すなわち、システムノイズの分散 $U = \sigma_S^2$、観測ノイズの分散 $V = \sigma_O^2$、$z(t)$ の分散を $Q(t) = \sigma(t)^2$ とする。(61b) 式と (69b) 式より

$$\sigma^2(t+1) = \left(\frac{1}{\sigma_S^2 + A^2\sigma^2(t)} + \frac{C^2}{\sigma_O^2} \right)^{-1} \tag{70}$$

となる。この式から観測を行なった後での分散 $\sigma^2(t+1)$ は観測を行う前の分散 $\tilde{\sigma}^2(t+1) = A^2\sigma^2(t) + \sigma_S^2$ よりも小さくなっていることがわかる。また観測ノイズの分散 σ_O^2 が非常に小さい時 $\sigma^2(t+1) \approx \sigma_O^2/C^2$ となり状態の分散も小さくなり推定精度が上がることがわかる。

これまで A, C, V, U は既知だとしてきたが、未知の場合でも EM アルゴリズムを用いて状態推定とパラメータ推定を同時に行なうことができる。

(演習 4)　システムノイズと観測ノイズがある状況での単振子の状態推定問題に対してカルマンフィルターを適用する。振子の鉛直方向からの角度を θ とすると、θ があまり大きくない場合の振子の運動方程式は

$$\frac{d^2\theta}{dt^2} = -\omega^2\theta, \qquad\qquad \omega^2 = \frac{g}{r}$$

で与えられる。ここで g は重力加速度、r は振子の長さである。角速度を $v \equiv d\theta/dt$ で定義すると運動方程式は、

$$\frac{d\theta}{dt} = v, \qquad\qquad \frac{dv}{dt} = -\omega^2\theta$$

となる。時間間隔 Δt でサンプリングして、オイラー差分により離散時間方程式に変換すると、

$$\begin{aligned} \theta(n+1) &= \theta(n) + \Delta t \cdot v(n) \\ v(n+1) &= v(n) - \Delta t \cdot \omega^2\theta(n) \end{aligned}$$

となる。分散 σ_S^2 をもつシステムノイズと分散 σ_O^2 をもつ観測ノイズとが付加されたシステムから、(ノイズの付加された) 角度の時系列 $\{\theta(n)|n = 1, ...\}$ が観測できる状況で、状態変数である $\theta(n)$ と $v(n)$ を推定せよ。