シリーズ 情報科学における確率モデル 9

Series on Stochastic Models in Informatics and Data Science

ベイズ学習とマルコフ決定過程

中井　達【著】

コロナ社

刊行のことば

われわれを取り巻く環境は，多くの場合，確定的というよりもむしろ不確実性にさらされており，自然科学，人文・社会科学，工学のあらゆる領域において不確実な現象を定量的に取り扱う必然性が生じる。「確率モデル」とは不確実な現象を数理的に記述する手段であり，古くから多くの領域において独自のモデルが考案されてきた経緯がある。情報化社会の成熟期である現在，幅広い裾野をもつ情報科学における多様な分野においてさえも，不確実性下での現象を数理的に記述し，データに基づいた定量的分析を行う必要性が増している。

一言で「確率モデル」といっても，その本質的な意味や粒度は各個別領域ごとに異なっている。統計物理学や数理生物学で現れる確率モデルでは，物理的な現象や実験的観測結果を数理的に記述する過程において不確実性を考慮し，さまざまな現象を説明するための描写をより精緻化することを目指している。一方，統計学やデータサイエンスの文脈で出現する確率モデルは，データ分析技術における数理的な仮定や確率分布関数そのものを表すことが多い。社会科学や工学の領域では，あらかじめモデルの抽象度を規定したうえで，人工物としてのシステムやそれによって派生する複雑な現象をモデルによって表現し，モデルの制御や評価を通じて現実に役立つ知見を導くことが目的となる。

昨今注目を集めている，ビッグデータ解析や人工知能開発の核となる機械学習の分野においても，確率モデルの重要性は十分に認識されていることは周知の通りである。一見して，機械学習技術は，深層学習，強化学習，サポートベクターマシンといったアルゴリズムの違いに基づいた縦串の分類と，自然言語処理，音声・画像認識，ロボット制御などの応用領域の違いによる横串の分類によって特徴づけられる。しかしながら，現実の問題を「モデリング」するためには経験とセンスが必要であるため，既存の手法やアルゴリズムをそのまま

適用するだけでは不十分であることが多い。

本シリーズでは，情報科学分野で必要とされる確率・統計技法に焦点を当て，個別分野ごとに発展してきた確率モデルに関する理論的成果をオムニバス形式で俯瞰することを目指す。各分野固有の理論的な背景を深く理解しながらも，理論展開の主役はあくまでモデリングとアルゴリズムであり，確率論，統計学，最適化理論，学習理論がコア技術に相当する。このように「確率モデル」にスポットライトを当てながら，情報科学の広範な領域を深く概観するシリーズは多く見当たらず，データサイエンス，情報工学，オペレーションズ・リサーチなどの各領域に点在していた成果をモデリングの観点からあらためて整理した内容となっている。

本シリーズを構成する各書目は，おのおのの分野の第一線で活躍する研究者に執筆をお願いしており，初学者を対象とした教科書というよりも，各分野の体系を網羅的に著した専門書の色彩が強い。よって，基本的な数理的技法をマスターしたうえで，各分野における研究の最先端に上り詰めようとする意欲のある研究者や大学院生を読者として想定している。本シリーズの中に，読者の皆さんのアイデアやイマジネーションを掻き立てるような座右の書が含まれていたならば，編者にとっては存外の喜びである。

2018 年 11 月

編集委員長　土肥　正

まえがき

多段決定過程を解決する数学的な方法として，動的計画法（dynamic programming）がベルマン（R. Bellman）により創始され，その初期の結果が 1950 年代末には書物としてまとめられた。この動的計画法は，確率的多段決定過程に応用され，特にマルコフ決定過程についていろいろと研究されている。その中で，状態を直接に知ることができない部分観測可能なマルコフ決定過程（POMDP）については多くの研究が進められ，1982 年にはモナハン（G. Monahan）によるサーベイ論文[33)†]が発表されている。このような部分観測可能なマルコフ決定過程は，確率的多段決定過程を考えるうえで重要な問題であり，理論的な発展と多くの場面での応用が開発されている。この部分観測可能なマルコフ決定過程において，直接に知ることのできない状態に関する情報を解析するための基本的な方法は，ベイズの定理をもとにしたものである。本書では，ベイズの定理に基づく学習と，それをもとにした部分観測可能なマルコフ決定過程の基本的な結果と応用について著すことを目的としたものである。

マルコフ決定過程については，数多くの書籍や研究があるが，本書では動的決定過程の基本的な性質を解説し，あわせて部分観測可能なマルコフ決定過程を解析するために必要な確率的順序関係と，その性質を解説することに主眼をおいた。1 章では，本書で必要となる確率と確率過程の基本的な用語と性質について簡単に述べる。2 章では，部分観測可能なマルコフ決定過程において，ベイズの定理に基づく学習を解析するうえで基本となる確率的順序関係について解説する。このような確率的順序関係は，待ち行列理論や取替問題をはじめ数理ファイナンスなどいろいろな分野で用いられているが，本書では特に必要となる尤度比順序を中心に述べる。3 章では，多段決定過程を解析する手法であ

† 肩付き数字は，巻末の引用・参考文献の番号を表す。

る動的計画法と最適性の原理を解説し，マルコフ決定過程の基本的性質についてまとめる。4 章では，ジョブサーチと確率的逐次割当問題を中心に，マルコフ決定過程の一つである期待値最大化問題を解説する。この確率的逐次割当問題は，1970 年代からいろいろと解析され，最近では新たなアプローチで研究されている。5 章では，部分観測可能なマルコフ連鎖を中心に，ベイズの定理に基づく学習による事前分布と事後分布の関係などをまとめる。6 章では，部分観測可能な 2 状態のマルコフ決定過程として，逐次解析や探索問題をはじめ最適停止問題などについて解説する。7 章では，部分観測可能なマルコフ決定過程としてジョブサーチや確率的逐次割当問題，最適選択問題を取り上げ，その基本的な性質を解析する。

マルコフ決定過程や部分観測可能なマルコフ決定過程については，すでに多くの著書や論文があり，本書を執筆するにあたり，多くの文献を参考にさせていただいた。基本的な予備知識としては，大学 1 年次での線形代数，微分・積分と初歩の確率論と数理統計学を仮定しているが，必要に応じて学習していただければと思う。

本書を執筆するにあたって，執筆する機会を与えてくださり，いろいろなご助言をいただいた本シリーズ編集委員長である広島大学大学院工学研究科教授土肥正氏と，いろいろアドバイスをいただいた愛知大学名誉教授玉置光司氏に，この場を借りて心より感謝申し上げます。さらに，学校法人本山学園理事長室山義正氏にはいろいろお世話になり感謝いたします。また，本シリーズ担当委員には，原稿を丁寧に見ていただき，多くのミスを指摘していただきました。この場を借りて感謝いたします。

2022 年 1 月

中井　達

目　　　次

第1章　確率と確率過程

第2章　確率的順序関係

第3章　マルコフ決定過程

第4章　ジョブサーチと確率的逐次割当問題

第5章　学習と情報

第 6 章　部分観測可能な 2 状態マルコフ決定過程

第 7 章　部分観測可能な逐次割当問題

3 マルコフ決定過程

多段決定過程における動的計画法とマルコフ決定過程について考える。

3.1 動的計画法

3.1.1 多段決定過程

多段決定過程（multistage decision process）は，あらかじめ決められた目的を達するために，段階を追って逐次に決定をとる決定問題である。この多段決定過程を考える段階の数は，有限のときもあれば無限のときもあり，この数を**計画期間**（planning horizon）という。

このような多段決定過程は，四つの要素 $(S, A, r(s,a), \tau(s,a))$ により定式化される。

(1) **状態空間** (state space) S：状態を s で表し，状態全体の集合を S とする。

(2) **行動空間** (action space) A：決定を a で表し，決定全体の集合を A とする。

(3) **利得関数**（reward function）$r(s,a)$：状態が s のときに決定 a をとれば，利得 $r(s,a)$ が得られる。

(4) **推移法則**（transition rule）$\tau(s,a)$：状態が s のときに決定 a をとれば，状態は法則 $\tau(s,a)$ に従い，つぎの状態が定まる。

この多段決定過程は，それぞれの段階で状態 s を知って決定 a をとる ($s \in S, a \in A$)。この状態 s と決定 a により，この段階での利得 $r(s,a)$ が定まり，つぎの段階の状態は推移法則 $\tau(s,a)$ に従って定まる。このような決定と推移を繰り返し，利得の総和を最大にする。また，この多段決定過程の最初の状態を**初**

期状態（initial state）という。ここでは，利得を考えたが，費用あるいは損失 $c(s,a)$ を考えるときは，$-c(s,a)$ を利得と考えればよい。

以下では，離散時間の多段決定過程とし，計画期間を n とし，決定をとる時点を $\{1,2,3,\cdots,n\}$ とする。状態空間は，有限集合あるいは可算集合とし，一般性を失うことなく整数で番号づけられ，$S=\{1,2,3,\cdots\}$ とする[†]。行動空間は，有限集合とする。状態が s のときに，決定 a をとったときの利得 $r(s,a)$ は非負関数とする。状態が s のときに，決定 a をとったときの推移法則 $\tau(s,a)$ は確定的でただ一つ定まる。さらに，状態空間，行動空間，利得関数，推移法則は決定の段階にかかわらず同じとする。これらの条件は，多段決定過程の基本的な性質を説明するためであり，条件が変わっても適当な仮定のもとで，基本的な性質が同じように成り立つ。以降の例 3.1 は，上記の条件から外れたものとなっている。

初期状態が i_1 のとき，この状態から始めて決定 $a_1,\cdots,a_n$ をこの順番にとれば，状態は推移法則に従って

$$i_1, i_2=\tau(i_1,a_1), i_3=\tau(i_2,a_2),\cdots, i_{n+1}=\tau(i_n,a_n)$$

と推移する。この決定の列に対し，それぞれの期の状態と決定により利得が定まり，それぞれ

$$r(i_1,a_1), r(i_2,a_2),\cdots, r(i_n,a_n)$$

となる。この多段決定過程では，n 期目での決定をとったあと，最終の状態が i のときに，各期での利得とは別に，状態 i に応じて利得 $R(i)$ が得られ，この利得を**最終利得**（terminal reward）という。

いま，k 期目での決定 a_k を $(k=1,2,\cdots,n)$，状態 i_k の関数として $a_k=a_k(i_k)$ と表し，この関数を**決定関数**（decision function）という。これらの決定関数の列 $\{a_k(\cdot)\}_{k=1,2,\cdots,n}$ を与えたとき，これを一つの**政策**（policy）という。

† 状態空間が有限集合あるいは可算集合であるときは状態を $i,j,k,\cdots$ で表し，それ以外のときは状態を $s,t,u,\cdots$ で表す。

初期状態が i_1 のとき，$\{a_k(\cdot)\}_{k=1,2,\cdots,n}$ を一つの政策とし，この政策に従って n 期間を終えたあとの総利得 $R_n(i_1|\{a_k(\cdot)\}_{k=1,2,\cdots,n})$ は最終利得を含め

$$R_n(i_1|\{a_k(\cdot)\}_{k=1,2,\cdots,n})=\sum_{k=1}^{n} r(i_k, a_k(i_k))+R(i_{n+1}) \tag{3.1}$$

となる。ただし，$i_{k+1}=\tau(i_k, a_k(i_k))$ とする ($k=1,2,3,\cdots,n$)。式 (3.1) をこの多段決定過程の**目的関数**（objective function）といい，この値を最大にする政策を**最適政策**（optimal policy）という。

3.1.2　最 適 方 程 式

初期状態が i_1 のとき，n 期間にわたり最適政策に従って得られる総利得を $v_n(i_1)$ とおく。$n=1$ のときは

$$v_1(i_1)=\max_{a\in A}(r(i_1,a)+R(\tau(i_1,a))) \tag{3.2}$$

となる。この 1 期間のモデルでの最適政策は，式 (3.2) の右辺を最大にする決定である。さらに，目的関数 (3.1) はそれぞれの期での利得を加えた関数だから，つぎの基本的な関係式が成り立つ。

定理 3.1

$$v_n(i_1)=\max_{a_1(\cdot),\cdots,a_n(\cdot)}\left(\sum_{k=1}^{n} r(i_k, a_k(i_k))+R(i_{n+1})\right) \tag{3.3}$$

のとき

$$v_n(i_1)=\max_{a_1\in A}(r(i_1,a_1)+v_{n-1}(\tau(i_1,a_1))) \tag{3.4}$$

となる ($n=1,2,\cdots$)。ただし，$v_0(i)=R(i)$ である ($i\in S$)。

【証明】　$i_{k+1}=\tau(i_k, a_k(i_k))$ とし ($k=1,2,3,\cdots,n$)，$i_2=\tau(i_1,a_1)$ に注意すれば，式 (3.3) で 1 期目での決定と，2 期目以降の決定の列に分けて

$$v_n(i_1)=\max_{a_1(\cdot),\cdots,a_n(\cdot)}\left(r(i_1,a_1(i_1))+\sum_{k=2}^{n} r(i_k, a_k(i_k))+R(i_{n+1})\right)$$

$$
\begin{aligned}
&= \max_{a_1(\cdot)} \left(r(i_1, a_1(i_1)) + \max_{a_2(\cdot),\cdots,a_n(\cdot)} \left(\sum_{k=2}^{n} r(i_k, a_j(i_k)) + R(i_{n+1}) \right) \right) \\
&= \max_{a_1(\cdot)} \left(r(i_1, a_1(i_1)) + \max_{a_2(\cdot),\cdots,a_n(\cdot)} R_{n-1}(i_2 | \{a_k(\cdot)\}_{k=2,\cdots,n}) \right) \\
&= \max_{a_1 \in A} (r(i_1, a_1) + v_{n-1}(\tau(i_1, a_1))) \qquad (3.5)
\end{aligned}
$$

となり，式 (3.4) が成り立つ。

◇

式 (3.4) をこの多段決定過程の**最適方程式**（optimality equation）という。また, 式 (3.4) を上向きの**再帰方程式**（recursive equation）ともいい, $n=1,2,3,\cdots$ と解いて $v_1(i), v_2(i), v_3(i), \cdots$ が順次求まる。また，n 期間の多段決定過程で，式 (3.4) の右辺を最大にする a_1 の一つを $a_1^*(i_1)$ とおく。これは，初期状態が i_1 のときの n 期間モデルでの 1 期目の最適な決定となる。2 期目以降の最適な決定 a_k^* は $(k=1,2,\cdots,n)$，それぞれ

$$
a_k^*(i_k^*) \quad \text{および} \quad i_k^* = \tau(i_{k-1}^*, a_{k-1}^*(i_{k-1}^*))
$$

となる。ただし, $i_1^*=i_1$ とする。すなわち, 決定関数の列が $a_1^*(\cdot), \cdots, a_{n-1}^*(\cdot), a_n^*(\cdot)$ のとき，$\{i_k^*\}_{k=1,2,\cdots,n}$ は初期状態が $i_1^*=i_1$ のとき最適政策によって推移する状態の列となる。この最適方程式を用いて，最適政策を逐次に求める方法を**動的計画法**（dynamic programming）という。

これまでは，それぞれの期での利得関数は同じとしたが，利得関数が期により異なり，k 期目の状態 i_k で決定 a_k をとったときの利得を $r_k(i_k, a_k)$ と拡張することもできる $(k=1,2,\cdots,n)$。

3.1.3　最適性の原理

初期状態が i の n 期間の多段決定過程の最適政策を $a_1^*(\cdot), a_2^*(\cdot), \cdots, a_n^*(\cdot)$ とすると，式 (3.4) は

$$
v_n(i_1) = r(i_1, a_1^*(i_1)) + v_{n-1}(\tau(i_1, a_1^*(i_1))) \qquad (3.6)
$$

となる。式 (3.6) から，初期状態が i_1 のとき最適政策によって得られる総利得

は，初期状態 i_1 で最適政策により得られる 1 期目での利得と，2 期目での状態 $\tau(i_1, a_1^*(i_1))$ を初期状態とした $n-1$ 期間の多段決定過程で最適政策により得られる総利得の和に等しい。

式 (3.1) と式 (3.4) より，$i_2^* = \tau(i_1, a_1^*(i_1))$ となり

$$v_n(i_1) = r(i_1, a_1^*(i_1)) + R_{n-1}(i_2^* | \{a_k^*(\cdot)\}_{k=2,3,\cdots,n})$$

となる。したがって，式 (3.6) より

$$v_{n-1}(\tau(i_1, a_1^*(i_1))) = R_{n-1}(i_2^* | \{a_k^*(\cdot)\}_{k=2,3,\cdots,n})$$

となり，つぎのことが成り立つ。

定理 3.2（**最適性の原理**（**principle of optimality**））　初期状態が i_1 のとき，n 期間の多段決定過程の最適政策を $a_1^*(\cdot), a_2^*(\cdot), \cdots, a_n^*(\cdot)$ とすれば，$a_2^*(\cdot), a_3^*(\cdot), \cdots, a_n^*(\cdot)$ は，$i_2^* = \tau(i_1, a_1^*(i_1))$ を初期状態とする $n-1$ 期間の多段決定過程の最適政策である。

定理 3.3　最適性の原理（定理 3.2）から定理 3.1 が導ける。

【証明】　式 (3.3) より

$$v_n(i_1) = \max_{a_1(\cdot), \cdots, a_n(\cdot)} R_n(i_1 | \{a_k(\cdot)\}_{k=1,\cdots,n})$$

とする。1 期目に決定 a をとり，つぎに直面する状態 $\tau(i_1, a)$ を初期状態とする $n-1$ 期間の多段決定過程では，その最適政策に従う政策は，この n 期間モデルの政策の一つとなる。よって，最初の期での任意の決定 $a \in A$ に対し

$$v_n(i_1) \geqq r(i_1, a) + v_{n-1}(\tau(i_1, a))$$

となり

$$v_n(i_1) \geqq \max_{a \in A}(r(i_1, a) + v_{n-1}(\tau(i_1, a))) \tag{3.7}$$

索　　　　引

—— 著 者 略 歴 ——

1975年　京都大学理学部卒業
1978年　大阪大学大学院基礎工学研究科博士前期課程修了（数理系専攻）
1981年　大阪大学大学院基礎工学研究科博士後期課程退学（数理系専攻）
1981年　大阪府立大学助手
1985年　工学博士（大阪大学）
1987年　神戸大学助教授
1991年　九州大学助教授
1996年　九州大学教授
2008年　千葉大学教授
2018年　千葉大学名誉教授
2020年　学校法人本山学園岡山医療専門職大学理事
　　　　現在に至る

ベイズ学習とマルコフ決定過程

Markov Decision Process Based on Bayes' Theorem

2022 年 3 月25日　初版第 1 刷発行

検印省略

著　　者　中（なか）　井（い）　　達（とおる）
発 行 者　株式会社　コ ロ ナ 社
　　　　　代 表 者　牛 来 真 也
印 刷 所　三 美 印 刷 株 式 会 社
製 本 所　有限会社　愛 千 製 本 所

112–0011　東京都文京区千石 4–46–10
発 行 所　株式会社　コ ロ ナ 社
CORONA PUBLISHING CO., LTD.
Tokyo Japan
振替 00140–8–14844・電話(03)3941–3131(代)
ホームページ　https://www.coronasha.co.jp

ISBN 978–4–339–02839–3　C3355　Printed in Japan　（新井）