実践 Python による
ベイズ分析とトピックモデル

—— 先進的なデータ分析へのアプローチ ——

博士（工学） 藤野 巖【著】

コロナ社

【本書ご利用にあたって】

本書で解説している内容を実行・利用したことによる直接あるいは間接的な損害に対して，著作者およびコロナ社は一切の責任を負いかねます。利用についてはすべて読者個人の責任において行ってください。

本書に掲載されている情報は，本書執筆時点のものです。将来にわたって保証されるものではありません。特に，各社が提供しているソフトウェアパッケージは仕様やサービス提供に係る変更が頻繁にあり，Python のライブラリ群等も頻繁にバージョンアップがなされています。これらによっては本書で解説しているアプリケーション等が正常に動作しなくなることもあるので，あらかじめご了承ください。

本書の発行にあたって，読者の皆様に問題なく実践していただけるよう，できる限りの検証をしていますが，以下の環境以外では構築・動作を確認していないので，あらかじめご了承ください。

PC 本体：Windows11 Pro 64 bit（CPU：Intel(R) Core i7，メモリ：32 GB）

開発環境：Anaconda バージョン 2023.09（Python バージョン 3.11）

また，上記環境を整えたいかなる状況においても動作が保証されるものではありません。ネットワークやメモリの使用状況および同一 PC 上にあるほかのソフトウェアの動作状況によって，本書のプログラムが正常に動作できなくなることがあります。併せてご了承ください。

なお，本書に記載している会社名，製品名は，それぞれ各社の商標または登録商標です。

本書の購入者に対する限定サービスとして，本書に掲載しているソースコードは，以下に示すコロナ社の Web ページからダウンロードできます。ぜひご利用ください。

https://www.coronasha.co.jp/np/isbn/9784339029437/

なお，本書に掲載しているソースコードについては，オープンソースソフトウェアの BSD ライセンス下で再利用も再配布も自由です。

ま　え　が　き

世の中にはデータがあふれかえっている。データと向き合うのは，もはや日常茶飯事である。データとは，現実社会の記録である。個々のデータを取り出してみることで，昔のことを具体的に正確に思い出すことができる。さらに，データをたくさん集めておけば，それらを分析することで，世の中の事柄のトレンド，事柄間の関連性のような価値のある情報が浮かび上がる。

データ分析の手法として，以前から平均値，標準偏差，四分位数などの統計量が用いられている。ただしこれらの手法では，収集したデータをそのまま利用しているため，いわゆる潜在的な情報，深層的な情報を発見することができない。

本書は，大量なデータから潜在的で深層的な情報を発掘できるトピックモデルという確率的なデータ分析手法の理論と実践の両方を解説するものである。トピックモデルは，自然言語処理（NLP）の手法の一つとして知られているが，その威力は文書データにとどまらず，画像データや軌跡データの分析にも応用できるような汎用的な技術であり，ディープラーニングと並んで人工知能の基本として認識されるべき技術である。

本書では，簡単な準備学習を経てから，ベイズ分析とトピックモデル，そしてその応用事例の順に，プログラムを作成しながら実践的に学習できるように章立てが用意されている。各章の内容を以下に簡単に紹介する。

1章：本書の学習に必要な確率と確率分布の知識およびそのプログラム実現を説明する。

2章：対比の位置付けとして，従来のデータ分析の基本手法を復習する。

3章：ベイズ分析の基本的な考え方を説明する。併せてベイズ分析のプログラム実現に使われるPyMCライブラリの使い方を紹介する。

4章：対比の位置付けとして，従来の文書データ分析の基本手法を復習する。

5章：ユニグラムモデルを構成して，文書データの分析を行う。また，PyMC ライブラリにより，そのプログラム実現を示す。

6章：トピックの考え方を取り入れて，混合ユニグラムモデルを構成する。また，混合ユニグラムモデルを用いた文書解析プログラム例を示す。

7章：混合ユニグラムモデルをさらに発展させて，トピックモデルを構成する。また，トピックモデルを用いた文書解析プログラム例を示す。

8章：Scikit-learn ライブラリにあるトピックモデルのモジュールの使い方を説明する。それを利用して，20 ニュースグループデータセットの英語文書データからトピックを抽出する。

9章：Gensim というトピックモデルに特化したライブラリの使い方を説明する。それを利用して，Wikipedia の日本語文書データからトピックを抽出する。

10章：トピックモデルを拡張して，著者トピックモデルを構成する。そのうえで，Gensim ライブラリを利用して，Twitter（現，X）から収集した日本語の投稿データからトピックを抽出する。

11章：トピックモデルを画像データセットに応用する。Gensim ライブラリを利用して，Caltech 101 というデータセットから，小さく分割されたセルで表したトピックを抽出する。

12章：トピックモデルを軌跡データセットに応用する。Gensim ライブラリを利用して，船舶の AIS データから，航路（コース）となるようなトピックを抽出する。

読者の皆様におかれましては，本書を読むことで先進的なデータ分析スキルを身に付け，トピックモデルの技術を応用して実務の場で役立てて頂ければ幸いである。

最後に，本書の執筆にあたり，多方面から熱心に支えてくれた家族に感謝の意を表する。また，本書を出版する機会を与えてくださったコロナ社の皆さんに衷心より感謝申し上げる。

2024年2月

藤野　巖

目　　次

1. 確率と確率分布

2. データの統計分析の基本

3. ベイズ分析の基本

4. 文書データ分析の基本

11. 画像データセットからのトピック抽出

12. 船舶の航跡データからのトピック抽出

1

確率と確率分布

本章では，今後のベイズ分析とトピックモデルの学習で必要となる各種確率分布について解説する。まず，確率の基本的な考え方を述べてから，条件確率，同時確率，ベイズの定理の順に基本的な概念を説明する。そのうえで，各種確率分布の計算式とグラフも示す。最後に，各種確率分布のグラフを作成する Python プログラムについて説明する。

1.1 確率と確率分布

1.1.1 確　　率

世の中のことは，確実に起こることもあれば，不確実に起こることもある。コインを高く投げたら，確実に落ちる。けれども，落ちたとき，上を向くのが，表なのか裏なのかは不確実である。不確実に起こることは，一度やってみるだけでは，起こることもあれば，起こらないこともある。けれども，何度もやってみると，だいたいそのうちの何回かが起こることになる。

具体例を使ってもう少し考えてみよう。例えば，コインを 100 回投げたら，そのうち表が出たのは 50 回であったとする。対して，サイコロを 100 回振ったら，そのうち 1 の目が出たのは 16 回であったとする。ここで，コインを投げたりサイコロを振ったりするような行為を**試行**（trial）と言う。また，コインの表やサイコロの 1 の目のような試行の結果を**事象**（event）と言う。つまり，この例では，コインが表という事象 A が 100 回試行のうち 50 回発生した。サイコロの目が 1 という事象 B は 100 回試行のうち 16 回発生した。このような状況を踏まえて，事象 A は事象 B より起こりやすいと考える。このように，不確実なことについては，その起こりやすさに違いがある。ある事象の起こりやすさを使えば，そのような事象の起こる確実性の度合いを表すことができる。

この事象が起こりやすさのことを**確率**（probability）と言う。

1.1.2 確率分布

話をもう一歩進めてみよう。まず，**確率変数**（random variable）c を使ってコインを投げる試行の結果を表すものとする。コインを投げる試行の結果，上に向く面は，表と裏の二つしかないので，確率変数 c の値が表（1）と裏（0）の二つしかない。確率変数 c のすべての値に対する確率が以下のようになる。

$$P(c=1)=\frac{1}{2}, \qquad P(c=0)=\frac{1}{2} \tag{1.1}$$

このような確率変数のすべての値の確率を全体的に表したものを**確率分布**（probability distribution）と言う。

また，確率変数 d を使ってサイコロを振る試行の結果を表すものとする。サイコロの目は，1〜6 まであるので，確率変数 c のすべての値に対する確率は以下のようになる。

$$\begin{aligned} &P(d=1)=\frac{1}{6}, \qquad P(d=2)=\frac{1}{6}, \qquad P(d=3)=\frac{1}{6} \\ &P(d=4)=\frac{1}{6}, \qquad P(d=5)=\frac{1}{6}, \qquad P(d=6)=\frac{1}{6} \end{aligned} \tag{1.2}$$

数式だけでは，全体的なイメージがつかみにくいので，グラフを使ってすべての確率を示すことがより一般的に行われている。また，コインやサイコロのように，確率変数が有限個整数から値を取るとき，**離散型**（discrete）**確率変数**と言い，離散型の確率分布の全体を表した関数を**確率質量関数**（**PMF**：probability mass function）と言う。対して，確率変数が実数から値を取るとき，**連続型**（continuous）**確率変数**と言い，連続型の確率分布の全体を表した関数を**確率密度関数**（**PDF**：probability density function）と言う。

1.2 条件確率と同時確率

簡単な例題を使って説明しよう。**図 1.1** のように，箱の中には，黒いボールが一つと白いボールが二つ入っており，箱からボールを取り出すときの確率を考える。条件確率と同時確率を説明する前に，まずは，普通の確率（無条件の確率）を計算しておく。箱の中からボール b を一つ取り出す試行をしたとき，その結果が黒いボール，または白いボールである確率は以下のようになる。

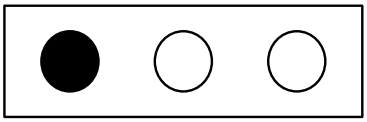

図 1.1 条件確率と同時確率の例題

$$P(b=\text{黒})=\frac{1}{3},\qquad P(b=\text{白})=\frac{2}{3} \tag{1.3}$$

1.2.1 条 件 確 率

これからは，図 1.1 の箱の中からボールを一つ取り出すことを 2 回行うとする。まず，1 回目の試行のボール $b1$ を取り出したあと，2 回目の試行のボール $b2$ が黒または白となる確率を計算する。すべての組合せの計算結果を以下に示す。

$$\begin{aligned} &P(b2=\text{黒}|b1=\text{黒})=0, \qquad P(b2=\text{白}|b1=\text{黒})=1 \\ &P(b2=\text{黒}|b1=\text{白})=\frac{1}{2}, \qquad P(b2=\text{白}|b1=\text{白})=\frac{1}{2} \end{aligned} \tag{1.4}$$

ここでは，1 回目のボールが特定の色となったことを条件として，2 回目のボールが特定の色となる確率を計算している。このような確率のことを**条件確率**（conditional probability）と言う。一般的に言うと，条件確率は，ある事象 A が起こることを仮定したときの別の事象 B が起こる確率で，これを $P(B|A)$ と記す。

1.2.2 同時確率

つぎに，1回目のボールが特定の色となり，かつ2回目のボールが特定の色となる確率を計算する。すべての組合せの計算結果を以下に示す。

$$
\begin{aligned}
P(b1=\text{黒かつ } b2=\text{黒}) &= \frac{1}{3}\times 0 = 0 \\
P(b1=\text{黒かつ } b2=\text{白}) &= \frac{1}{3}\times 1 = \frac{1}{3} \\
P(b1=\text{白かつ } b2=\text{黒}) &= \frac{2}{3}\times \frac{1}{2} = \frac{1}{3} \\
P(b1=\text{白かつ } b2=\text{白}) &= \frac{2}{3}\times \frac{1}{2} = \frac{1}{3}
\end{aligned}
\tag{1.5}
$$

ここでは，1回目のボールが特定の色となるかつ2回目のボールが特定の色となる確率を計算している。このような二つの事象を同時に起きる確率のことを**同時確率**（joint probability）と言う。一般的に言うと，同時確率は，ある事象 A と事象 B が同時に起こる確率で，これを $P(A \cap B)$ †と記す。

1.3 乗法定理とベイズの定理

1.3.1 乗法定理

ここで，1.2節の計算結果をよく観察してみると，同時確率が

$$P(b1=\text{白かつ } b2=\text{黒}) = \frac{2}{3}\times\frac{1}{2} \tag{1.6}$$

となる一方，一回目の試行では，無条件確率が

$$P(b1=\text{白}) = \frac{2}{3} \tag{1.7}$$

となり，2回目の試行では，条件確率が

$$P(b2=\text{黒}\,|\,b1=\text{白}) = \frac{1}{2} \tag{1.8}$$

† 本によっては，同時確率を $P(AB)$ または $P(A,B)$ と記すことがある。

となるので，同時確率が以下のように，無条件確率と条件確率の積で表せることがわかる。

$$P(b1=\text{白かつ}\ b2=\text{黒}) = P(b1=\text{白})P(b2=\text{黒}\,|\,b1=\text{白}) \tag{1.9}$$

これを一般的に言うと，事象 A と事象 B があるとすれば，同時確率

$$P(A \cap B) = P(A|B)P(B) \tag{1.10}$$

と表すことができる。この公式は**乗法定理**（multiplication theorem）と呼ばれる。これを利用すれば次式より条件確率 $P(A|B)$ を計算できるようになる。

$$P(A|B) = \frac{P(A \cap B)}{P(B)} \tag{1.11}$$

さらに，ここで事象 A の確率が事象 B の影響を受けないならば，つまり

$$P(A|B) = P(A)$$

ならば，同時確率 $P(A \cap B)$ は，以下のように，それぞれの事象の無条件確率の積で計算できるようになる。

$$P(A \cap B) = P(A)P(B) \tag{1.12}$$

この場合は，事象 A と事象 B が**独立**（independent）であると言う。

1.3.2　ベイズの定理

乗法定理から，事象 A と事象 B の同時確率は

$$P(A \cap B) = P(A|B)P(B) \tag{1.13}$$

となり，事象 B と事象 A の同時確率は

$$P(B \cap A) = P(B|A)P(A)$$

となる。しかし，この二つの確率が等しいので，$P(A \cap B) = P(B \cap A)$ から以下の**ベイズの定理**（Bayes' theorem）を簡単に導出することができる。

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} \tag{1.14}$$

索　　引

――著 者 略 歴――

1991 年　東海大学大学院工学研究科博士課程修了
　　博士（工学）
1994 年　東海大学短期大学部講師
2001 年　東海大学短期大学部助教授
2004 年　イギリス・サウサンプトン大学客員教授
～05 年
2007 年　東海大学短期大学部教授
2008 年　東海大学情報通信学部教授
　　現在に至る
2016 年　フランス海軍アカデミー招聘研究員
～17 年

実践　Python によるベイズ分析とトピックモデル
―先進的なデータ分析へのアプローチ―

Perfect Practice　Bayesian Analysis and Topic Modeling with Python
―An Advanced Approach to Data Analysis―

2024 年 4 月 25 日　初版第 1 刷発行　★

検印省略

著　者　藤野（ふじの）　巖（いわお）
発行者　株式会社　コロナ社
　　代表者　牛来真也
印刷所　壮光舎印刷株式会社
製本所　株式会社　グリーン

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 978-4-339-02943-7　C3055　Printed in Japan　（松岡）