目的

我々は現象や観測を説明するモデルを作る．モデルからデータを生成するのがシミュレーションである．モデルは我々が勝手に使ったもの，いわば妄想に過ぎない．しかし将来を予測したり，思わぬことを教えてくれたりと，ときに有用である．それでは，現象や観測のデータが与えられたとき，モデルに含まれるパラメータなどをどのように推論すればよいのだろうか．他のモデルとの比較をどのようにすればよいだろうか．これらの推論を行うための統一的な枠組みを与えてくれるのが，**ベイズ推論** *(Bayesian inference)* である．簡単な例を使って，これをマスターしよう．

有名な警句[1]

> . . . all models are approximations. Essentially, all models are wrong, but some are useful. However, the approximate nature of the model must always be borne in mind . . .

私のブログから

私は講義やゼミの中で、最初は条件付確率にまったく不案内だった学生や社会人がこの「社長と秘書」の話を聞きながら、自分で具体的に確率を計算して考えることで、条件付確率の意味を正しく把握できるようになることを経験してきた。さらに、ベイズ推定や状態空間モデルを含む多くの応用、ひいてはモデルとシミュレーションの本質に話をつなげていくことができる。統計の教科書でこのあたりのことが強調されていないのはとても残念である。半日程度のセミナーでお話しすることができるので、興味を持たれた読者はぜひお声がけいただきたい。

自然現象であれ、社会現象であれ、現象のモデルは人がでっち上げる妄想であり、しかし時に極めて有用である。ただし、モデルには未知のパラメータや量が含まれるから、データからそれらを推論する必要がある。またモデルはただ一つではない。「秘書」がデータならば「社長」はモデルである。一方、モデルからシミュレーションを行ってデータを生成することはいつでも可能だ。したがって、秘書から社長の機嫌を推論することは「逆シミュレーション」であり、データ同化など別の名前でもよばれている。

[1] Box, G. E. P. and Draper, N. R., *Empirical Model-Building and Response Surfaces*, (John Wiley & Sons, 1987), p. 424.

藤原 義久 / yoshi.fujiwara@gmail.com

興味ある現象に関する観測や実験の $\mathcal{D}$ とする．我々はそれを説明する仮説すなわちモデルを作る．それを $\mathcal{M}$ と書こう．モデルは我々が作ったものであるから，それに含まれるパラメータや初期条件・境界条件など (まとめてパラメータとよぼう) を決めれば，モデルからデータを生成する，すなわち「シミュレーション」$\mathcal{M} \to \mathcal{D}$ を実行することは (少なくとも原理的には) 簡単であるはずだ．しかし，実際に必要なことはデータからモデルのパラメータについて推論すること，いわば「逆シミュレーション」$\mathcal{D} \to \mathcal{M}$ である．

条件付確率の練習問題で見たように,「逆シミュレーション」は直観的には理解しにくい．そこで以下のようにベイズの定理をを用いることを考えよう．

$$P(\mathcal{M}|\mathcal{D}) = \frac{\mathcal{P}(\mathcal{D}|\mathcal{M})\,\mathcal{P}(\mathcal{M})}{\mathcal{P}(\mathcal{D})} \tag{1}$$

(1) 式の右辺の各因子は以下の意味をもつ．

- 分子の 1 番目の因子 $P(\mathcal{D}|\mathcal{M})$ は「シミュレーション」$\mathcal{M} \to \mathcal{D}$ から計算できるはずだ．この因子は**尤度** *(likelihood)* とよばれる[2]．
- 分子の 2 番目の因子 $P(\mathcal{M})$ はモデルのパラメータについて，我々が事前に知っている情報を組み込むことができるものである．この因子は**事前確率** *(prior probability)* とよばれる．
- 分母の因子 $P(\mathcal{D})$ は規格化 (normalization) 因子，すなわち (1) の両辺を $\mathcal{M}$ に含まれるパラメータの可能なすべての値について積分すると全確率 (total probability) 1 になることを保証するためのものである．したがって

  $$P(\mathcal{D}) = \sum_{\mathcal{M}} \mathcal{P}(\mathcal{D}|\mathcal{M})\,\mathcal{P}(\mathcal{M}) \tag{2}$$

  $P(\mathcal{D})$ は周辺確率であるから，$\mathcal{M}$ を含んでいないので，モデルのパラメータを推定する話には関係がない[3]

(1) 式の右辺を計算した結果として得られる左辺を**事後確率** *(posterior probability)* とよぶ．

以上の枠組みを**ベイズ推論** *(Bayesian inference* とよぶ．最初の応用例として次の「スリッパ投げ」をあげる．

[2]「尤」は「もっともらしい」の意味 (犬ではない) で，尤度は「ゆうど」と読む．

[3]複数のモデルを比較してどのモデルがもっともらしいかを決める，モデル選択 (model selection) ではこの因子 (evidence とよばれる) が活躍する．

藤原 義久 / yoshi.fujiwara@gmail.com

例: スリッパ投げ (tossing a slipper)

スリッパを一つ床に投げて，表か裏かどちらかが出るとしよう．どちらの面が出るのかは実験のたびに変わるので，確率変数として扱う．何度も実験をして表裏の観測データを収集して，スリッパ投げをコイン投げと見なすようなモデルを考えたとしよう．モデルにはただ一つのパラメータ，すなわち面が出る確率がある．100 回実験を行った結果，表が 34 回出たとすると

$$\text{面が出る確率} = \frac{34}{100}$$

と考えるかもしれない．ではなぜそのように考えるのだろうか．そもそも面が出る確率というパラメータを含むモデルを考えるのは，将来同様の実験を行ったときに得られるデータを説明したいためであるだろう．実験を行うたびごとに結果は変わってくるから，上の値からはずれるに違いない．ではそのずれをどのように理解すればよいだろうか．

データ $\mathcal{D}$ ：N 回の独立なスリッパ投げの観測結果

$$\{x_1, x_2, x_3, x_4, \cdots, x_N\} = \{\text{H,H,T,H,}\cdots\text{,T}\} \tag{3}$$

ここで，表と裏をそれぞれ head=H, tail=T と書いた．以下，このデータを $x_{1:N}$ と表すことにする．表が出た回数を n と表す．裏が出た回数は $N - n$ である．

モデル $\mathcal{M}$ ：表 (head=H) の出る確率 θ $(0 \le \theta \le 1)$．したがって裏 (tail=T) の出る確率は $1 - \theta$

モデルが与えられたとき，それによって実際にデータが生成される確率は

$$P(x_{1:N}|\theta) = \theta^n (1-\theta)^{N-n} \tag{4}$$

これを尤度 (likelihood) とよぶ．

ベイズ推論 $\mathcal{D} \to \mathcal{M}$ ：データが得られる前のモデル θ の値についての信念の度合 (degree of belief) を $P(\theta)$ とする．信念の度合を確率として表現する枠組みがベイズ推論である．これを事前確率 (prior probability) とよぶ．

逆にデータが与えられたとき，そのモデルに対する確率 (信念の度合) は，ベイズの定理により

$$P(\theta|x_{1:N}) = \frac{P(x_{1:N}|\theta)\, P(\theta)}{P(x_{1:N})} \tag{5}$$

これを事後確率 (posterior probability) とよぶ．

ここでは簡単のため，モデルについて事前には何の知識もないとして，$P(\theta)$ は一様な確率であるとする．

$$P(\theta) = 1 \qquad (0 \le \theta \le 1) \tag{6}$$

(5) の分母は

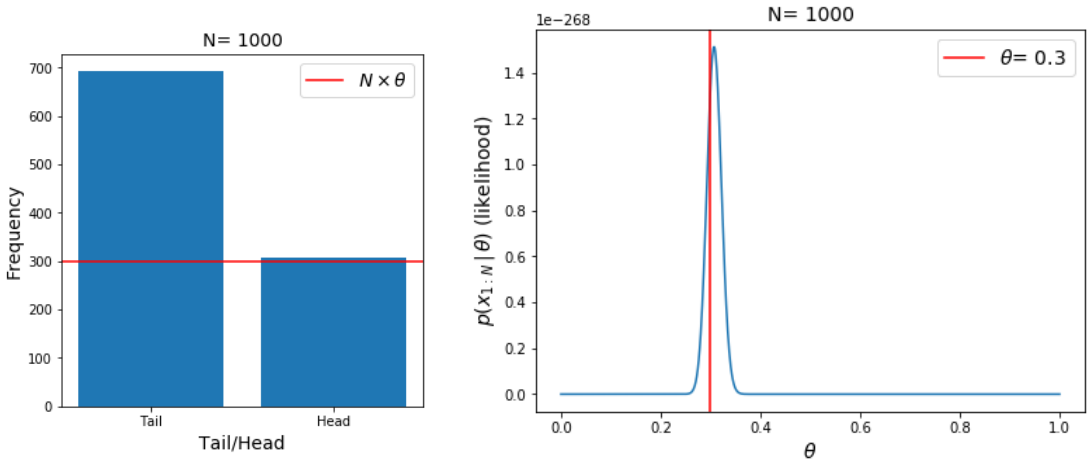
$$\int_0^1 P(\theta|x_{1:N})\, d\theta = 1 \tag{7}$$

の規格化のための因子であり，特にパラメータ θ に依存しない．

したがって，事前知識がないとしたいまの場合には，データが与えられたときにそのモデル (パラメータ) に対する確率 (信念の度合) は (5) の分子にある尤度 (4) で決まる．

藤原 義久 / yoshi.fujiwara@gmail.com

やってみよう (`prog_bayes_infer.ipynb`)
モデルを知っている場合にシミュレーションを行ってデータを生成して，正しくモデル (パラメータ) を推定できるかどうかやってみよう．



(a) ヒストグラム (b) 事後確率 (5)

図 1: スリッパ投げのシミュレーションとベイズ推定の結果．(a) の水平線は表の出る回数の期待値．(b) の縦の線は真のパラメータの値．

事前には何ら知識がなく，θ の値は (6) とまったく不確かであったが，データが得られたおかげで事後には (5) として図 1 (b) のように推論できた．事後確率 (5) の最大値を計算すると

$$\theta = \frac{n}{N} \tag{8}$$

であることが分かる[4]．事後確率 (5) は，事前確率が (6) の場合，(5) は尤度 (4) に比例するので，その最大値は尤度 (4) の最大値に等しい．このようにモデルのパラメータを決める方法は**最尤法** *(maximum likelihood)* とよばれる．

しかし，事後確率 (5) は図 1 (b) にあるように最大値以外の情報として (8) からのずれを教えてくれる．有限の回数 N の実験結果一セットから推論したパラメータの値がどの程度不確かなのか，その推定結果も教えてくれるのである．

[4]読者自らが計算して確かめよ．(5) を θ について微分したものがゼロとなる必要条件を解けばよい．対数をとってから微分すると見通しがよい．

後記

事前確率を変えると事後確率はもちろん変わってくる．これを不思議に思ったり，あるいは得られた事後確率は正しいのか不安に感じたりする人がいるかもしれない．先験的，絶対的に正しい確率の値がどこかにあって，確率論とはそれを求めるものであるという迷信を信じている人が少なくないからであろう．そのような誤った考えは捨てるべきである．コインを投げて表が出る確率はいくらか．ほとんどの人が 1/2 であると答える．なぜそのように考えるのか，と聞かれたら戸惑うだろう．1/2 という確率は先験的や絶対的に正しいものではなくて，コイン投げを理解するためのモデル，すなわち我々の妄想かつ便利な道具であって，こう考えるとどうやら手元に得られた実験を理解できるようだと考えるべきものである．

円にでたらめに弦を引くとき，その長さが円の内接正三角形の一辺よりも長くなる確率はどれだけか．これには無限の解答が存在する (Bertand のパラドックス)．また，物理学に馴染みのある読者は統計力学における Maxwell-Boltzmann 統計・Bose-Einstein 統計・Fermi-Dirac 統計の起源を思い出そう．

したがって，事前確率を変えると事後確率が変わっていくというのは，我々がモデルをいろいろ変形したとき，手元にあるデータと整合的に何が帰結できるかを導くための大変便利な枠組みなのである．

文献 [1] の最初の章をぜひ読まれたい．

参考文献

[1] 小針 睨宏「確率・統計入門」(岩波書店，1973).
名著．第 1 章「確率モデル」に確率とは何かについて分かりやすく含蓄に富む説明がある．

藤原 義久 / yoshi.fujiwara@gmail.com