# 最尤法とモデル選択

2018 年講義資料 （作成者：大塚 淳）

## 1 尤度と最尤法

我々は今まで、統計学における仮説の取り扱いを見てきた。仮説とは、つまるところ、関心ある確率変数の確率分布についての命題である。例えばコインが公平であるという仮説は、$P(X=1)=P(X=0)=0.5$ という命題として表せる。仮説は、与えられたデータを予測・説明する[*1]。例えば、コインを 10 回投げてちょうど 5 回表が出たとしよう。このデータは、上の仮説によって良く説明される。しかしたった 1 回しか表が出なかったとしたら、この結果は当該仮説ではあまり良く説明されないだろう。では、あるデータが与えられたとして、これを最も良く説明する仮説は何か。これを問うのが、最尤法 (method of maximum likelihood) である。

最尤法のアイデアを一言で言えば「与えられたデータのもとで尤度が最も高い仮説が、そのデータを最も良く説明する」というものだ。既に見たように尤度とは、仮説のもとでの証拠・データの確率 $P(E|H)$ であった。これが高いほど、仮説がデータを良く予測・説明できている。データとして 10 回のコイン投げで表がでる回数 $X$、仮説として「コインの表がでる確率は $\theta$ である」という命題を考えよう。$0 \leq \theta \leq 1$ なので、0 から 1 まで無限個の仮説が考えられる。今、表が 6 回出たとしよう $(X=6)$。さて、この無限個の仮説のうち、この結果・データを最も良く説明する仮説、つまり尤度 $P(X=6|\theta)$ を最大化する $\theta$ は何だろうか？

二項分布の確率分布より、尤度 $P(X=6|\theta)$ は以下によって表される

$$ {}_{10}C_6\ \theta^6(1-\theta)^4 \tag{1}$$

これは $\theta$ の関数であり、$\theta$ を横軸、尤度を縦軸にとってプロットすると図 1 のようになる。この図によれば、$\theta=0.6$ のとき最も尤度が高い。つまり 10 回中 6 回表が出たという結果を最もありそうにする・説明するのは、そのコインが 6/10 の確率で表になるという仮説である。

ここでの確率モデルは二項分布であり、その母数は $\theta$ 一つだけであった。あるデータが与えられたとき、そのモデルの尤度を最大にするような母数を求めることができる。これを母数の**最尤推定量** (maximum likelihood estimator; MLE) という。また、そのときの尤度をモデル $M$ の最大尤度といい、$L(M)$ で表す[*2]。最尤推定量は、尤度を母数に対してプロットしたグラフの「頂上」である。

---

[*1] 哲学的には、予測と説明は異なった概念である。例えば、我々は旗竿から伸びる影からその高さを予測できるが、それがなぜその高さなのかを説明することはできない。しかしここではそうした差異にはこだわらないことにしよう。

[*2] つまり母数 $\boldsymbol{\theta}$ によって特徴づけられるモデル $M(\boldsymbol{\theta})$ および観察されたデータ $\mathbf{x}$ とすると、MLE は $\hat{\boldsymbol{\theta}} =$

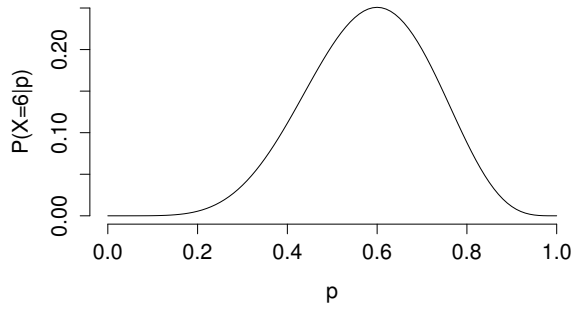


図 1　10 回コインを投げて 6 回表が出たときの、仮説 $\theta$ の尤度

最尤推定量への「山登り」には、微分を用いる。微分は、地点地点での関数・グラフの傾きを与えることを思い出そう。頂上では傾きがゼロとなるので、上述のコイン投げのケースでの最尤推定量は、尤度関数 (式 1) を $\theta$ で微分して得られる式をゼロについて解けば求まる。せっかくなので具体的にやってみよう。今、最初に表が 6 回、次に裏が 4 回続けて出たとし、このデータセットを $\mathbf{x}$ と呼ぼう。このもとでの尤度は

$$P(\mathbf{x}|\theta) = \theta^6(1-\theta)^4 \tag{2}$$

これを最大化するような $\theta$ を求めるために、上式を $\theta$ で微分して

$$\frac{dP(D|\theta)}{d\theta} = \theta^5(1-\theta)^3(6-10\theta) \tag{3}$$

これが上図のグラフの傾きを与える。この式をゼロにするような $\theta$ を求めると、括弧の中より $\theta = 0.6$ となり、最尤推定量が求まる。

この例のように母数が一つだけの場合、頂上へは一次元 ($\theta$ 軸) の「山登り」となる。しかし一般に $n$ 個 の母数を持つモデルでは、頂上・最尤推定量へと到達するためには $n$ 次元空間での山登りをしなければならない。また今回のケースでは山頂が一つのみの「富士山型」であり、解析的に答えを導き出せたが、より複雑な尤度関数の場合は実際に一歩一歩「山を登って」求めなければならないことがしばしばである。その場合、まず出発点 $\theta_0$ を決めて、その場所で尤度関数を微分してその場所での「傾斜」を求める。そしてその傾斜がプラスの方向に一歩進み、それを $\theta_1$ とする。そして次はその場所でまた微分して傾きを求め・・・というように、その都度尤度関数の勾配を確かめつつ上っていかねばならない。またその場合でも、山頂に到達できず、局所最適点に陥ってしまうという問題が生じる。このことについてはまた後、深層学習のところで見ていく。

---

$\arg\max_{\boldsymbol{\theta}} P(\mathbf{x}|M(\boldsymbol{\theta}))$ であり、最大尤度 $L(M) = P(\mathbf{x}|M(\hat{\boldsymbol{\theta}}))$ である。

## 2　モデル選択

“All models are wrong. Some are useful.” — G. Box

“Pluralitas non est ponenda sine neccesitate.”— William of Ockham

モデル選択理論の目的は、二つ以上のモデルが与えられたときに、手持ちのデータをもとにそれらの予測性能の良さを比較することにある。例えば、年齢 $X$ から身長 $Y$ を予測したいとしよう。このとき次のようなモデル（回帰モデル）を考えることができる。

$$y = \alpha + \beta x + \epsilon \qquad (M_1)$$
$$y = \alpha + \beta x + \gamma x^2 + \epsilon \qquad (M_2)$$

ただし $\epsilon$ は一定の分布（ここでは正規分布としよう）に従う誤差項だとする。両モデルとも身長は「年齢の関数 + 確率的ノイズ」によって決まると主張するが、その関数の次数が異なる。二次関数は一次関数よりも複雑なので、$M_2$ は $M_1$ よりも複雑なモデルだと言える。

このとき、二つのモデルの尤度について、次が成り立つ：すなわちどのようなデータ $(\mathbf{x}, \mathbf{y})$ が与えられても、必ず $L(M_2) \geq L(M_1)$ となる。つまり手持ちのデータに限って言えば、複雑モデルの方が必ず説明力が高くなる。しかし問題は、手持ちのデータではなく、むしろ将来得られるはずのデータを説明・予測する力である。つまり我々が知りたいのは、モデルの平均予測性能、つまり「将来にわたって同じようなサイズのデータを用いて $X$ から $Y$ の予測を行うとしたら、平均して、$M_1$ と $M_2$ どちらがより現実に近い結果をもたらしてくれるだろうか」ということだ。赤池は、このモデルの平均予測性能が、以下によって推定されることを示した：

$$\mathrm{Score}(M) = \log L(M) - k(M)$$

ただし $k(M)$ はモデルが持つ母数の数である。ちなみにここでの $\mathrm{Score}(M)$ に $-2$ をかけたものを、赤池情報量基準 (AIC) という[*3]。

赤池の結果によれば、モデル $M$ の平均予測性能 $\mathrm{Score}(M)$ は、二つの要素から決まる。

1. モデルの対数[*4]最大尤度 $\log L(M)$：モデルが手持ちのデータをどれだけ説明できるか
2. 母数の数 $k(M)$：モデルの複雑さを表し、予測性能に対してペナルティとして働く

例えば、上であげた $M_1, M_2$ をとると、先ほど述べた理由から $L(M_1) \leq L(M_2)$ となり、最初の要素では $M_2$ が勝る。しかし $k(M_1) = 1 < k(M_2) = 2$ より、二つ目の要素では $M_1$ が勝る。最終的に $\mathrm{Score}(M_2)$ が $\mathrm{Score}(M_1)$ に勝るのは、後者の面でのペナルティを補えるくらいの尤度の上昇が見込めるときである。このように、赤池の理論ではモデルの説明力と複雑さを天秤にかけることで、その

[*3] つまり AIC$= -2\log L(M) + 2k(M)$. AIC は真なる分布とモデルの間のズレ (Kullback-Leibler divergence) の平均の推定なので、小さいほど良いモデルということになる。

[*4] 対数 log は増加関数なので、$L(M_i) > L(M_j)$ であれば常に $\log L(M_i) > \log L(M_j)$ となる、つまり尤度における優劣の順序を変えない。

平均予測能力を評価する。両者が常にトレードオフにある限り、複雑なモデルが常に良い予測をもたらすとは限らないのである。