最尤推定法

確率を用いた推定理論

14T4035F　白井俊祐

最小二乗法で「すべての点を正確に通る関数」を見つけても「未来予測」には役立たない・・・

↓

データに本質的に何らかの誤差を含んでいる。

↓

「どの程度の誤差を持つデータなのか」を含めて分析する必要

最尤推定法

「あるデータが得られる確率」を設定し、そこから最良のパラメータを決定していくアプローチ。

仮定：M次多項式の関係があり、さらに標準偏差σの誤差が含まれているとする。

標準偏差σ

およそ$\pm\sigma$の範囲で観測データが変動する。

多項式の関係仮定は最小二乗法と同じだが、誤差についての仮定の追加が新しい部分。

仮定を数式で表現する

まず特徴変数xと目的変数tの間のM次多項式

$$f(x) = w_0 + w_1 x + \cdots + w_M x^M$$
$$= \sum_{m=0}^{M} w_m x^m \qquad (1)$$

標準偏差σ

$$N(t_0 | f(x_n), \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(t - f(x_n))^2} \qquad (2)$$

このモデルに含まれるパラメータは？

→係数$\{w_m\}_{m=0}^{M}$と標準偏差σ

このパラメータ値を評価する基準を設定して、最良の値を決定することが目的。

パラメータ評価の基準設定

(1)、(2)式を用いて「トレーニングセットに含まれるデータ$\{(x_n, t_n)\}_{n=1}^{N}$が得られる確率を計算してみる。

まずある特定の観測点x_nについて考えて、そこでt_nが得られる確率は(2)に$t = t_n$を代入して表す。

$$N(t_n|f(x_n), \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(t_n - f(x_n))^2} \qquad (3)$$

(3)式について、すべての観測点について併せて考えると、トレーニングセットのデータが得られる確率Pはそれぞれの積になる。

$$P = N(t_1|f(x_1), \sigma^2) \times \cdots \times N(t_N|f(x_N), \sigma^2)$$

$$= \prod_{n=1}^{N} N(t_n|f(x_n), \sigma^2)$$

(4)

(4)式は、パラメータによって値が変わるので、パラメータの関数と考えることができる。

↓

（4）式を**尤度関数**と呼ぶ。

ここで次の仮説を立てる

仮説：観測されたデータ（トレーニングセット）は、最も発生確率が高い確率に違いない。

この仮説の元、（4）式で計算される確率Pが最大になるようにパラメータを決定する手法を最尤推定法をいう。

パラメータの計算

(4)式のPを最大化するパラメータを求める。
まず(3)式を(4)式に代入して整理。

$$P = \prod_{n=1}^{N} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(t_n - f(x_n))^2}$$

$$= \left(\frac{1}{2\pi\sigma^2}\right)^{\frac{N}{2}} \exp\left[-\frac{1}{2\sigma^2}\sum_{n=1}^{N}\{t_n - f(x_n)\}^2\right] \quad (5)$$

ここで、(5)式を見ると、最小二乗法で使用した二乗誤差E_Dが含まれている。

$$E_D = \frac{1}{2}\sum_{n=1}^{N}\{f(x_n) - t_n\}^2 \tag{6}$$

よって、先ほどの(5),(6)式より尤度関数は以下のようにあらわすことができる。

$$P = \left(\frac{1}{2\pi\sigma^2}\right)^{\frac{N}{2}} e^{-\frac{1}{\sigma^2}E_D} \tag{7}$$

ここで、パラメータに対する依存度を確認する。

$\beta = \frac{1}{\sigma^2}$とする。

E_Dは多項式の係数wに依存しているので

$$P(\beta, w) = (\frac{\beta}{2\pi})^{\frac{N}{2}} e^{-\beta E_D(w)}$$

となり、これを最大にする(β, w)を求めればよい。

更に計算を簡単にするためにPの対数$\ln P$を最大化

$$\ln P(\beta, w) = \frac{N}{2}\ln\beta - \frac{N}{2}\ln 2\pi - \beta E_D(w) \tag{8}$$

対数は単調増加関数なので$\ln P$が最大＝Pが最大です。

これを対数尤度関数と呼ぶ。

対数尤度関数を最大にする値は次の条件で決まる。

$$\frac{\partial(\ln P)}{\partial w_m} = 0 \quad \text{(m=0...,M)} \qquad (9)$$

$$\frac{\partial(\ln P)}{\partial \beta} = 0 \qquad (10)$$

(8)を（9）式に代入すると

$$\frac{\partial E_D}{\partial w_m} = 0$$

が得られる。

これは、二乗誤差を最小にする条件と全く同じであり、誤差関数を最小にする条件をたどると、最小二乗法と同じ結論を得ることができる。

つまり、多項式の係数は最小二乗法と同じ値に決まる。

一方(8)式を(10)式に代入すると

$$\frac{1}{\beta} = \frac{2E_D}{N}$$

が求められ、$\beta = \frac{1}{\sigma^2}$なので

$$\sigma = \sqrt{\frac{1}{\beta}} = \sqrt{\frac{2E_D}{N}} = E_{RMS}$$

E_{RMS}は、前章で定義した平方根平均二乗誤差である。

つまり、標準偏差値σの推定値は、

トレーニングセットに含まれるデータの「多項式で推定される値$f(x_n)$に対する平均的な誤差」

である。

以上より

$$w = (\varphi^T \varphi)^{-1} \varphi^T t$$

$$\sigma = E_{RMS} = \sqrt{\frac{1}{N} \sum_{n=1}^{N} \left(\sum_{m=0}^{M} w_m x_n^m - t_n \right)^2}$$

となる。

なぜ最小二乗法と同じ結果が得られたか

尤度関数の中に二乗誤差E_Dがあることが原因である。

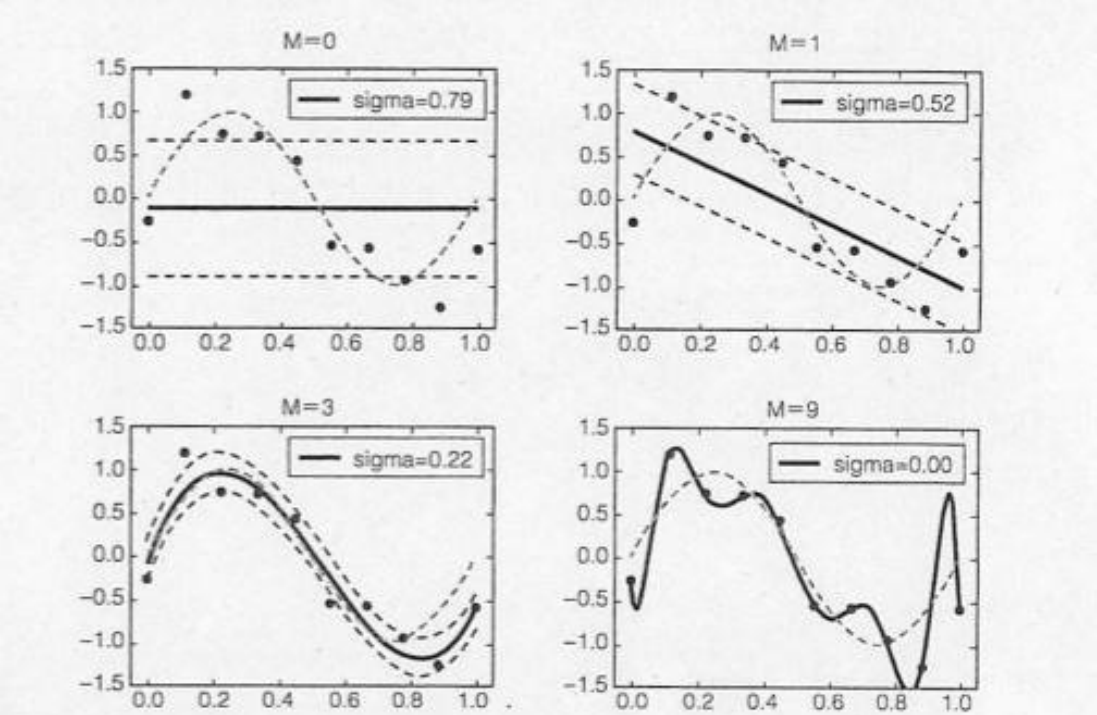
標準偏差σを仮定する際に正規分布を利用しているため二乗誤差が出てきた。

このことから、最小二乗法は最尤推定法の中 でも、正規分布の誤差を仮定した特別な場合に対応するとみなすことができる。

サンプルコードによる確認

グラフの形は最小二乗法で得たものと同じである。

標準偏差の幅を示す破線がグラフに追加されている。



▲図3.5 最尤推定法による標準偏差を含めた推定結果

観測したデータを見ると、多項式で予測される値とトレーニングセットに含まれる観測データの「ズレ」が、標準偏差としてうまく表現されていることが分かる。

・・・しかし、これはあくまでもトレーニングセットの話。

例えば、Ｍ＝9の場合、グラフはすべての点を通り
標準偏差は0となっている。

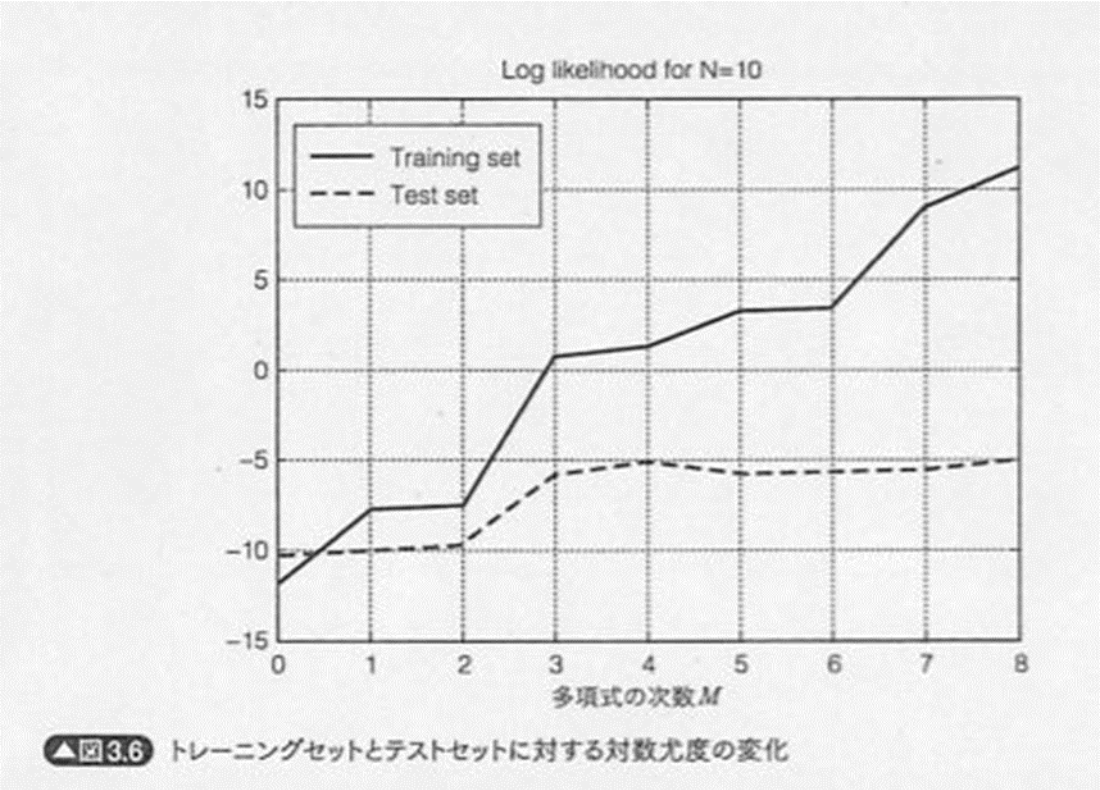
↓

オーバーフィッティングが発生。

最尤推定法では、尤度関数の値の変化を見てオーバーフィッティングの検出を行う。

（5）式にトレーニングセット、テストセットのデータを代入して計算する。

多項式の次数を変化による対数尤度関数の変化を表したグラフ



▲図3.6 トレーニングセットとテストセットに対する対数尤度の変化

Ｍ＝3あたりでオーバーフィッティングが発生
それ以降の尤度関数はほぼ変化がなくなる。

観測点を一つにして考えた場合

ある観測点から繰り返し観測値tを取得すると、

ある値を中心に散らばったデータ群$\{t_n\}_{n=1}^{N}$が得られる。

このデータを正規分布に従って散らばるものと仮定する。

正規分布のパラメトリックモデル

先ほどの計算（最尤推定）を正規分布のパラメトリックモデルで行う。

仮定：平均μ、標準偏差σの正規分布

ある特定のデータ$t = t_n$が求められる確率

$$N(t_n|f(x_n), \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(t_n - f(x_n))^2}$$

また、すべての観測値にまとめて考えて一連のデータ群が得られる確率Pはそれぞれの確率の積となる。

$$P = \prod_{n=1}^{N} N(t_n|\mu, \sigma^2)$$

細かい計算はほぼ同様なので省略

得られる結果は

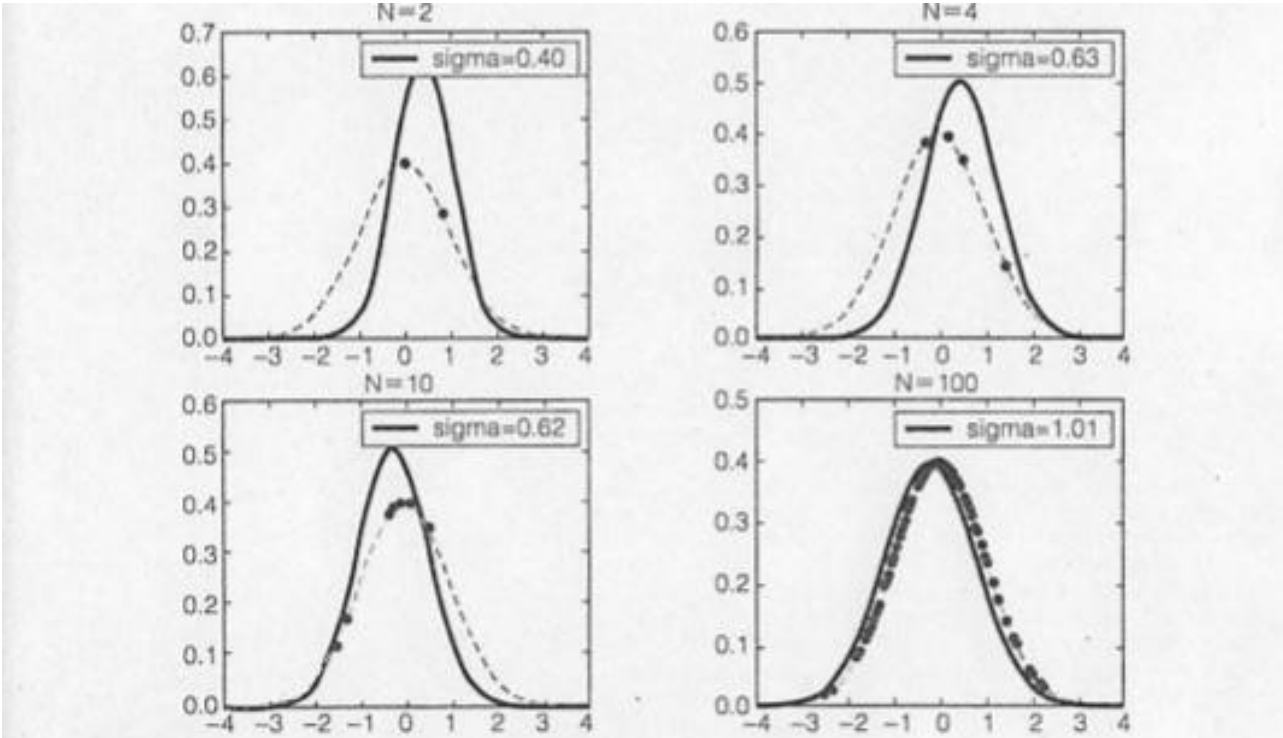
$$\mu = \frac{1}{N}\sum_{n=1}^{N} t_n$$

$$\sigma^2 = \frac{1}{N}\sum_{n=1}^{N} (t_n - \mu)^2$$

となる。

サンプルコードによる確認

サンプルコードを用いて、実際に作成される正規分布のグラフと、推定された正規分布のグラフがどれほど一致するか確認する。



▲図3.11 最尤推定法による正規分布の推定結果

結果を見ると与えられるデータが多ければ多いほど推定結果が正しいものに近くなりやすい

逆に、データが少ないとごく一部しか正しい値が得られないので、与えられたデータから正しい結果の全体像を推定することが難しくなる。

これは、データ数が少ない場合に標準偏差が小さく推定されてしまうことが原因となっている。

推定量の評価方法（一致性と不偏性）

先ほどの結果から最尤推定はかならずしも正解を与えるものではない。

↓

機械学習での結果を無条件に信じず、テストセット検証、クロスバリデーションにより、モデルの繁華能力を評価することが重要。

しかし・・・

先ほどの結果から標準偏差の推定値は実際のものより小さくなる傾向がある　ということがわかる。

推定値を少し大きくした値を分散の推定値として採用することが可能！

↓

どのくらい大きくすればいいのかわからない・・・

・推定量

何らかの理論に基づいて推定値を計算する方法が得られたときの方法。

一致性、不偏性を持つものは性質がよい。

先ほどの例でいうとデータ数を大きくすることで真の値に近づいていく

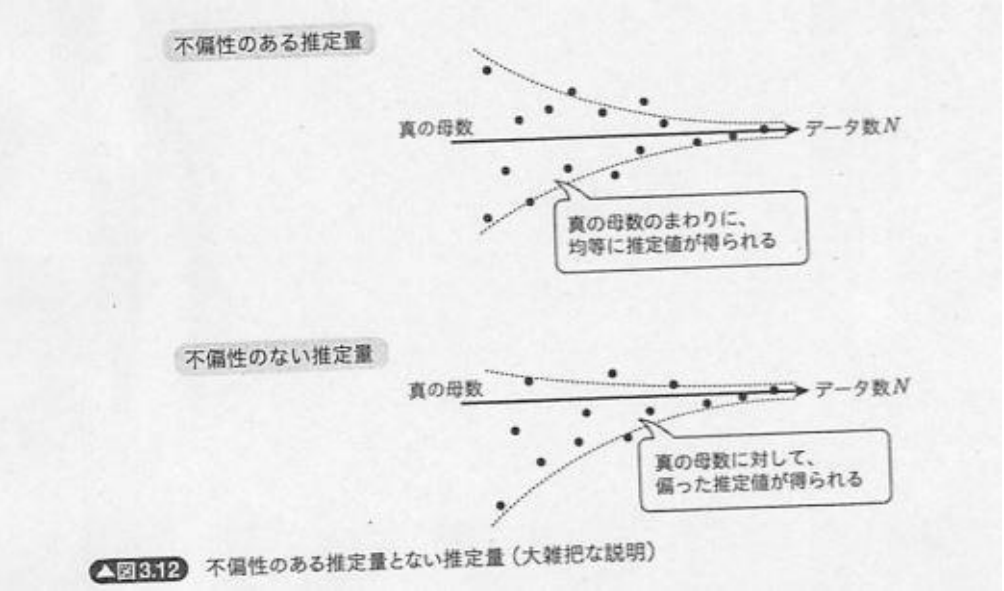
→一致性

一致性を持つ推定量は一致推定量と呼ばれる。

不偏性

推定値を多数計算し、得られた結果の平均を撮ったときに正解に近づいていく性質

不偏性を持つ推定量を不偏推定量と呼ぶ。



▲図3.12 不偏性のある推定量とない推定量（大雑把な説明）