二つの分散:不偏推定量と最尤推定量のどちらを使うべきか

井手剛 (IBMワトソン研究所)

June 2023

分散の不偏推定量と最尤推定量

スカラーの観測値 $\mathcal{D} = \{x_1, \ldots, x_N\}$ があったとする。**標本平均**はどの本を見ても

$$\bar{x} = \frac{1}{N}\sum_{n=1}^{N} x_n$$

となっている。しかし、**標本分散**に関しては、本により定義が違う。下記のどっちかである。

$$\text{不偏推定量}: \ \hat{s}^2 = \frac{1}{N-1}\sum_{n=1}^{N}(x_n - \bar{x})^2$$

$$\text{最尤推定量}: \ \hat{\sigma}^2 = \frac{1}{N}\sum_{n=1}^{N}(x_n - \bar{x})^2$$

古い本を見ると、最尤推定量 $\hat{\sigma}^2$ は不偏性（unbiasedness）を満たさないので使うべきではないなどと書いてあることがある。本当にそうなのか、という疑問に答えるのが本稿の目的である。実はどちらも理論的な保証がある。

- 不偏推定量は文字通り**不偏性 (unbiasedness)** という性質を持つことを証明できる。ある統計量（標本平均みたいにデータから計算できる何かの量）が不偏であるとは、その期待値が真の値と一致するという性質。
- 最尤推定量は、**一致性 (consistency)** という性質を持つことを証明できる。一致性とは、標本数 N が大きくなれば大きくなるほど、その推定値が真の値に近づいてゆくという性質。

どちらもある意味正しく、かつ N が数百とかそれ以上では実用上は大差はないので、どうでもいいとも言えるが、現代のデータサイエンスでは、最尤推定で統一すべきという考えがだんだん主流になっていると思う。結論は最後にまとめる。

逆に言えば、いまだに、「最尤推定量は間違っている、不偏推定量を使え」、という人がいたとしたら、その人は実世界の問題を解くことにあまり興味がないか、昔授業で教わった知識を自ら更新する力がないか、単に知ったかぶりをしているか、みたいな可能性がある。気を付けた方がいい。

1　分散の最尤推定量の導出

実用上、最尤推定の問題設定は単純である。データは数字として得られたとする。その現実を認めて、（たぶん頻度のグラフを描いたりいろいろ考えた結果として）そのばらつきを正規分布

$$p(x \mid \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

で表すと決心する。そして、当てはまりが一番よくなるように、未知量 μ, σ^2 を決める。当てはまりの度合いとしては、x に観測値の数字、たとえば $x_1 = 1.3$ と $x_2 = 2.0$ を得た時なら、$p(1.3 \mid \mu, \sigma^2)$ と $p(2.0 \mid \mu, \sigma^2)$ を採用することができる。他のデータについても考えて、全体としてその数字が一番大きくなるように μ, σ^2 を決めればよい。そういう基準での当てはまりの良さ表すのが**対数尤度 log-likelihood** という量である。

$$\begin{aligned} L(\mu, \sigma^2) &= \sum_{n=1}^{N} \ln p(x_n \mid \mu, \sigma^2) \\ &= -\frac{N}{2}\ln(2\pi) - \frac{N}{2}\ln(\sigma^2) - \frac{1}{2\sigma^2}\sum_{n=1}^{N}(x_n - \mu)^2 \end{aligned}$$

これは μ, σ^2 についての連続関数だから、微分してゼロとおく、という技で最大点を見出すことができる。まず μ については

$$0 = \frac{\partial L}{\partial \mu} = \frac{1}{\sigma^2}\sum_{n=1}^{N}(x_n - \mu) = \frac{1}{\sigma^2}\sum_{n=1}^{N} x_n - \frac{\mu}{\sigma^2}\sum_{n=1}^{N} 1 = \frac{1}{\sigma^2}\sum_{n=1}^{N} x_n - \frac{\mu N}{\sigma^2}$$

という式が出てくるので、μ の推定値は標本平均

$$\mu = \bar{x} \triangleq \frac{1}{N}\sum_{n=1}^{N} x_n \tag{1}$$

である。一方、σ^2 について、$b \triangleq \sigma^{-2}$ という変数を定義してこれについて最大化することにする。対数尤度は

$$L(\mu, b) = -\frac{N}{2}\ln(2\pi) + \frac{N}{2}\ln b - \frac{b}{2}\sum_{n=1}^{N}(x_n - \mu)^2$$

と書けるから

$$0 = \frac{\partial L}{\partial b} = \frac{N}{2b} - \frac{1}{2}\sum_{n=1}^{N}(x_n - \mu)^2$$

という条件が出る。式(1)で得た $\mu = \bar{x}$ を使って整理すると

$$\frac{1}{b} = \boxed{\sigma^2 = \hat{\sigma}^2 \triangleq \frac{1}{N}\sum_{n=1}^{N}(x_n - \bar{x})^2} \tag{2}$$

という分散の最尤推定量が得られる。

上記の論法には何の無理もない。実測したデータを現実としてありのままに受け入れる。その上で、データに当てはまるかもしれないモデル（分布）を仮定し、それを最尤法でデータに当てはめた。それだけである。最尤法はモデルによらず普遍的に使えるので馬鹿の一つ覚えのように使ってよい。すべての仮定は「表に」出ており、潔い。たしかに、「一致性」という理論的性質を示すためには、裏に真の分布を考え etc. とやる必要があるのだが、データサイエンティストの作業の中にそういう空理空論は入る余地はなく、それに足を取られる危険はない。

2　分散の不偏推定量の導出

不偏分散を導く論法は最尤推定と考え方が違う。上のデータは確かに数値として得られたのだが、それは必ずしも信用できないと考え、「何千組もデータセットを生成しまくった時に、どのくらい推定量がばらつくか」という問題を考える。この考えに基づけば、上の最尤推定量は、データセット $\mathcal{D}$ の関数である。そこでそれを

$$f(\mathcal{D}) = \frac{1}{N}\sum_{n=1}^{N}(x_n - \bar{x})^2, \quad \bar{x} \triangleq \frac{1}{N}\sum_{n=1}^{N} x_n$$

みたいに書いておく。この値は、データセットのばらつきに伴いどのくらいばらつくだろうか。もちろん、「何千組もデータセットを生成しまくった」というのは現実にはやっていないので、ばらつきを計算することは本来できない。そこで、仮に、真の確率密度関数 $p(x)$ の平均と分散が

$$m_0 = \int_{-\infty}^{\infty} \mathrm{d}x\ p(x)x, \qquad \sigma_0^2 = \int_{-\infty}^{\infty} \mathrm{d}x\ p(x)(x - m_0)^2 \tag{3}$$

のようにわかっていると仮定する。この分布の下で $f(\mathcal{D})$ の期待値を計算した場合、それは真の分散 σ_0^2 に一致するだろうか、それとも違うだろうか。それが考えるべき問題である。

この問いに答えるため、まず $\bar{x} \triangleq \frac{1}{N}\sum_{n=1}^{N} x_n$ ゆえ

$$\begin{aligned} f(\mathcal{D}) &= \frac{1}{N}\sum_{n=1}^{N}\left(x_n^2 - 2x_n\bar{x} + \bar{x}^2\right) \\ &= \frac{1}{N}\sum_{n=1}^{N} x_n^2 - \frac{2\bar{x}}{N}\sum_{n=1}^{N} x_n + \frac{\bar{x}^2}{N}\sum_{n=1}^{N} 1 = \frac{1}{N}\sum_{n=1}^{N} x_n^2 - 2\bar{x}\times\bar{x} + \bar{x}^2 \\ &= \frac{1}{N}\sum_{n=1}^{N} x_n^2 - \bar{x}^2 \end{aligned}$$

となることに注意する。$\bar{x}^2 = \frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N} x_i x_j$ だから、x_n^2 と $x_i x_j$ の期待値が計算できればよいことになる。x_n や x_i は、真の分布 $p(x)$ によりばらつく

ことになっているのだから、期待値を計算すると、

$$(x_n^2\text{の期待値}) = \int_{-\infty}^{\infty} \mathrm{d}x_n \; p(x_n) x_n^2 = \sigma_0^2 + m_0^2$$

$$(x_i x_j\text{の期待値}) = \begin{cases} \sigma_0^2 + m_0^2, & i = j \\ m_0^2, & i \neq j \end{cases}$$

のように計算できる。最初の式は $\sigma_0^2 = \int \mathrm{d}x_n \; p(x_n) x_n^2 - m_0^2$ による。これらを使うと

$$\begin{aligned} (f\text{の期待値}) &= \frac{1}{N} \sum_{n=1}^{N} (\sigma_0^2 + m_0^2) - \frac{1}{N^2} [N(\sigma_0^2 + m_0^2) + N(N-1) m_0^2] \\ &= (\sigma_0^2 + m_0^2) - \frac{1}{N} (\sigma_0^2 + N m_0^2) \\ &= \sigma_0^2 \frac{N-1}{N} \end{aligned} \tag{4}$$

となる。残念ながらこれは σ_0^2 とは一致していないことが分かる。しかし、右辺がぴったり σ_0^2 に一致するためには、$f(\mathcal{D})$ の定義を変更して

$$\hat{s}^2 \triangleq f(\mathcal{D}) \times \frac{N}{N-1}$$

のように $\frac{N}{N-1}$ 倍しておけばよい。ゆえ、「期待値が真の分散と一致するべし」という要請を満たすためには、

$$\boxed{\hat{s}^2 = \frac{1}{N-1} \sum_{n=1}^{N} (x_n - \bar{x})^2} \tag{5}$$

を標本分散の定義として採用すればよい、と言うことになる。これがいわゆる不偏分散である。

上でやったことは、まず $f(\mathcal{D})$ の具体的な定義を（天下り式に）与え、「真の分布」についての理論的仮定の下、それの矛盾を指摘したということである。統計量を何か作り、それの数学的性質を論ずる、というのは頻度派 (frequentist) と呼ばれる統計学者が論文を書く時の基本作法である。しかし、実用に使いたい人の観点からすれば、そのロジックは少々わかりにくい。実用的な統計量を導くための議論に、現実に観測したはずのデータをいわば無視して真の分布を考える、みたいな空理空論っぽい仮説が紛れ込んでくるからである。それに、最初の統計量をどうやって思いついたのかさっぱりわからない。分散ならまだいいが、より抽象的な統計量なら困る。

結果として、頻度派的な論法を実務家の観点から見ると、「下々の者どもはわしの言う通りの統計量を（ブラックボックスとして）使ってろ」と言われているかのようである。この辺が（主流派の）統計学者が実務家に尊敬されつつ煙たがれる理由であろう。どういう空理空論を考えようが、その空理空論からどういう結論を導こうが、数学としては自由である。しかし我々はお客様なり上司なりにわかってもらえなければ話にならないのである。

まとめ: データサイエンスとは

ということで、データサイエンスの本質をひと言で答えろと言われたら、「観測データに対してもっとも当てはまりの良いモデルをつくるために、最尤推定を使ってパラメターを決めること」と答えればよいと個人的には思う。

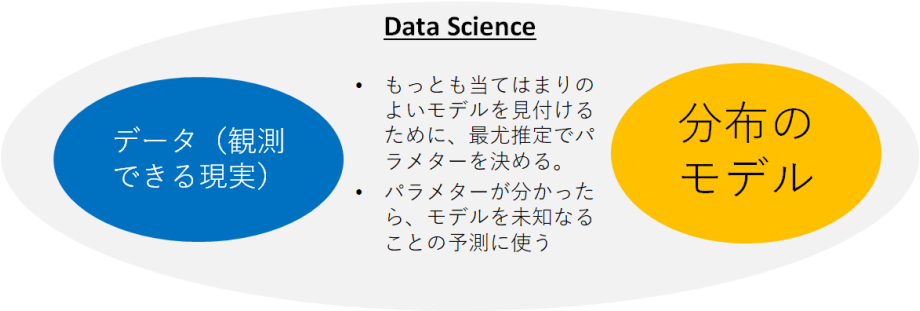

Figure 1: データサイエンス概要

MNISTの数字分類を深層学習でやる場合とか、分布が表向き出てこない場合はよくある。しかし目的関数（損失関数）の式の出所について一度でも考えたことがあれば、どの（まともな）目的関数も、なにかの分布的想定が裏で結び付いているらしい、ということくらいは気づいたかもしれない。例えば、2乗損失は、正規分布の対数尤度から出てくる。では正規分布を別の分布に変えてみたらどうなるだろう。求めるべきパラメターにある程度の曖昧さを許したい時はどうすべきだろう。こういう奥行きに気を配ることが、未知の問題に有効で創造的な解決策を考える力になり、それはただのAPI使いから、データ「サイエンティスト」へ進化するための第一歩なのだと思う。