統計的推測の基礎 (ver. 0.1)

かしゃるふぁ

2021 年 6 月 5 日

概要

このノートは渡辺澄夫先生のベイズ統計の教科書[1]の勉強ノートです。

1　統計的推測の枠組み

1.1　統計的推測の外観

まず、統計的推測を最尤推定やベイズ推定を含む一般的な枠組みで議論しよう。統計的推測では、ある未知の確率分布関数$q(x)$が存在して、これが我々が根本的には求めたい確率分布関数であると仮定する。この分布$q(x)$を真の分布と呼ぶことにする。例えば、$q(x)$は我々が得ているデータを生成している未知の分布であり、それを今持っているデータから推測したいという状況を考えている。このとき一般には、真の分布は「何らかのパラメトリックな確率モデル$p(x|\theta)$と真のパラメータ$\theta_\star$が存在して、$q(x) = p(x|\theta_\star)$となっている」とは限らないし、また「パラメータがθが何らかの確率分布$\varphi(\theta)$で生成されていて、$q(x) = \int p(x|\theta)\varphi(\theta)d\theta$のようになっている」とも限らない。このノートでは$q(x)$について上記のような仮定は一切置かない。

次に統計的推測を行うために使うサンプルについて説明をする。このサンプルとはすでに得られているデータに相当している。統計的推測では、サンプルはある確率変数の実現値であると考える。そこでサンプルの個数をnとして、それらのサンプルを表す確率変数を$X^n = (X_1, X_2, \ldots, X_n)$とする。一般にはこれらは独立であるとは限らないし、確率変数X_iが真の分布$q(x)$に従っているとは限らない。しかし、このサンプルX^nについて何らかの仮定を置かなければ、統計的推測について議論をすることは不可能である。そこで多くの教科書で行われているように、このノートではサンプルX^nは真の分布$q(X)$に従う独立な確率変数であるという強い仮定を置く。

以上の設定のもとで議論を進めいてく。未知の分布$q(x)$について、我々がサンプルX^nをもとにして推定した分布のことを**予測分布**と呼び、$p_{\rm pred}(x|X^n)$で表す。この時点では、$p_{\rm pred}(x|X^n)$は一般の確率分布であるとする。あとの節で、統計的推測を行う上で便利な予測分布の形について言及する。

1.2 予測分布と汎化損失

予測分布$p_{\mathrm{pred}}(x|X^n)$は何らかの意味で、$q(x)$を近似していなければならない。そこで近似度合いを評価する指標について考えていく。 2つの分布$p_{\mathrm{pred}}(x|X^n)$と$q(x)$の相違度、つまり2つの確率分布の相違度を定量化する指標は複数あるが、よく使われるのはKullback–Leibler divergenceである。一般に分布qとpのKullback–Leibler divergenceは

$$D(q\|p) \coloneqq \int q(x) \log \frac{q(x)}{p(x)} dx \tag{1}$$

で定義される。 Kullback–Leibler divergenceを変形すると、

$$D(q\|p) = \int q(x) \log q(x) dx - \int q(x) \log p(x) dx \tag{2}$$

$$= -S(q) - \int q(x) \log p(x) dx \tag{3}$$

となる。ここで右辺の第一項目はqだけに依存するから、相対的な差を考える上では第二項目だけを考えれば良い。そこで予測分布$p_{\mathrm{pred}}(x|X^n)$による$q(x)$の推測の精度を議論するために

$$G_n \coloneqq -\int q(x) \log p_{\mathrm{pred}}(x|X^n) dx \tag{4}$$

を考える。これを**汎化損失**と呼ぶ。汎化損失の最小値は$S(q)$であり、これは$p_{\mathrm{pred}}(x|X^n)$が$q(x)$に一致する場合に実現される。汎化損失を統計的推測の精度指標として用いるのであれば、統計的推測とは汎化損失が小さくなるような予測分布$p_{\mathrm{pred}}(x|X^n)$を求めることである。そこで以下のような疑問が生まれる。

疑問 1　与えれれたqに対して$\lim_{n\to\infty} G_n = S(q)$となるような予測分布$p_{\mathrm{pred}}(x|X^n)$の（必要）十分条件は何か？

この答えは???である。

統計的推測では、G_nが小さくなるように予測分布$p_{\mathrm{pred}}(x|X^n)$を試行錯誤して決めていけば良いが、ここには問題が生じる。それは$q(x)$は未知の分布であるため、G_nを直接的に計算することはできないことである。そこで何らかの指標を使って、汎化損失G_nを推測する。例えば$q(x)$をサンプルX^nによる経験分布で近似した**経験損失**

$$T_n \coloneqq -\frac{1}{n} \sum_{i=1}^{n} \log p_{\mathrm{pred}}(X_i|X^n) \tag{5}$$

のような指標が考えられる。では経験損失は、汎化損失に対する良い推定量になっているのだろうか。残念ながら経験損失は汎化損失の不偏推定量ではない。つまり、一般に$\mathbb{E}[T_n] \neq \mathbb{E}[G_n]$である。ただし、いくつかの仮定の下で、$\mathbb{E}[T_n] = \mathbb{E}[G_n] + O(n^{-1})$が成り立つ[1]。つまりサンプル数が多い漸近的な領域であれば、経験損失の期待値は汎化損失の期待値とほぼ等しいとみなせる。

汎化損失のよりよい推定量は、WAIC（widely applicable information criterion）である。ここではWAICの定義は述べないが、WAICをW_nで表すと、$\mathbb{E}[W_n] = \mathbb{E}[G_n] + o(n^{-1})$となる。

1.3 確率モデルと事後分布

さて、今までの議論では一般の形の予測分布を考えていたが、予測分布が$q(x)$に近づくように最適化する上では、一般の分布を考えるよりも特定のクラスの確率分布を考えるほうが都合が良い。そこで、予測分布を調整するためのパラメータ$w \in W$を導入して、次の形の分布を考える。

$$p_{\mathrm{pred}}(x|X^n) = \int_W p_{\mathrm{model}}(x|w)p_{\mathrm{post}}(w|X^n)dw \tag{6}$$

ここで条件付き確率$p_{\mathrm{model}}(x|w)$を（パラメトリックな）**確率モデル**と呼び、$p_{\mathrm{post}}(w|X^n)$をパラメータwの**事後分布**（posterior distribution）と呼ぶ。ここで確率モデル$p_{\mathrm{model}}(x|w)$が予測分布を探索するためのクラスを定め、パラメータwをサンプルX^nからどのように求めるのかを決めているのが事後分布$p_{\mathrm{post}}(w|X^n)$である。

例えば最尤推定法の場合は、パラメータwのサンプルX^nによる最尤推定量を$w_{\mathrm{ML}}(X^n)$として、$p_{\mathrm{post}}(w|X^n) = \delta(w - w_{\mathrm{ML}})$となる。ここで$\delta(w)$はデルタ分布である。この枠組みでは、最尤推定もベイズ推定も同じようにより扱うことができるし、パラメータwはサンプルX^nに依存した確率変数となる。もちろん、最初に設定したようにサンプル、つまりデータも確率変数である。ここで登場した確率モデルとの事後分布2つを調整するのが統計モデリングである。

渡辺ベイズ本[1]では事後分布の形として、

$$p_{\mathrm{post}}(w|X^n) = \frac{1}{Z_n(\beta)}\varphi(w)\prod_i \{p_{\mathrm{model}}(X_i|w)\}^\beta \tag{7}$$

というのを考えている。この形の事後分布を考えると多くの有益な性質を示すことができる。このノートでは後の節で、なぜこの形が理論的な解析を行う上で便利なのかについて説明する。

1.4 確率モデルの評価

前の節で予測分布を確率モデルと事後分布という2つの部分に分けた。この節では、ひとまず事後分布についてはおいておき、確率モデルに焦点を当て、真の分布$q(x)$と確率モデル$p_{\mathrm{model}}(x|w)$の関係を評価しよう。この節ではパラメータは確率変数ではなく、ただの変数である。

さて**平均対数損失関数**を

$$L(w) \coloneqq -\int q(x)\log p_{\mathrm{model}}(x|w)dx \tag{8}$$

で定義する。平均対数損失関数は、汎化損失と同じやり方で真の分布と確率モデルの違いを定量化し、それを確率モデルのパラメータ$w \in W$の関数として見たものである。汎化損失と同様に、$q(x)$は未知の分布であるから、我々は平均対数損失関数を直接的に計算することはできない。そこ

で**経験対数損失関数**を

$$L_n(w) \coloneqq -\frac{1}{n}\sum_{i=1}^{n}\log p_{\mathrm{model}}(X_i|w) \tag{9}$$

で定義する。経験対数損失関数は、平均対数損失関数の定義に現れる真の分布を経験分布で置き換えたものである。このとき定義から明らかなように、$\mathbb{E}[L_n(w)] = L(w)$が成立する。つまり、経験対数損失関数は平均対数損失関数の不偏推定量となっている。これは汎化損失と経験損失の関係とは、異なっている。

確率モデル$p_{\mathrm{model}}(x|w)$の中で、最も上手く真の分布$q(x)$を近似していると言えるのは、平均対数損失関数が最小値を取るうようなパラメータを使ったモデルである。そこで、パラメータの集合W_0を

$$W_0 \coloneqq \arg\min_{w\in W} L(w) \tag{10}$$

で定義し、**真の分布に対して最適なパラメータの集合**と呼ぶ。ここで任意の最適なパラメータ$w_0 \in W$について、確率モデル$p_{\mathrm{model}}(x|w_0)$がユニークな確率分布を表すとき、真の分布に対して最適な確率分布は**実質的にユニーク**であるという。最適な確率分布が実質的にユニークのとき、そのユニークな確率分布を$p_0(x) \coloneqq p_{\mathrm{model}}(x|w_0)$と書くことにする。 $p_0(x)$の定義から、平均対数損失関数について

$$L(w) \geq -\int q(x)\log p_0(x)dx =: L_0 \tag{11}$$

が成り立つ。ここで等号が成立するのは、wが真の分布に対して最適なパラメータの場合のみである。平均対数損失関数を最小値が0になるように定数分調整した量を

$$K(w) \coloneqq L(w) - L_0 \tag{12}$$

として、これを**平均誤差関数**と呼ぶ。平均誤差関数は変形すると

$$K(w) = -\int q(x)\log\frac{p_{\mathrm{model}}(x|w)}{p_0(x)}dx \tag{13}$$

と表すこともできる。ここでも真の分布を経験分布で置き換えた量

$$K_n(w) = -\frac{1}{n}\sum_{i=1}^{n}\log\frac{p_{\mathrm{model}}(X_i|w)}{p_0(X_i)} \tag{14}$$

を考え、これを**経験誤差関数**と呼ぶ。定義から明らかなように、経験誤差関数は平均誤差関数の不偏推定量である。

1.5 事後確率の考慮

さてパラメータwが事後分布$p_{\mathrm{post}}(w|X^n)$に従う確率変数だと考えよう。このときの経験誤差関数$K_n(w)$の振る舞いが知りたい。ここでは$K_n(w)$も確率変数である。そこで$K_n(w)$のモーメント

母関数

$$\mathcal{K}_n(\alpha) \coloneqq \int_W e^{-\alpha K_n(w)} p_{\rm post}(w|X^n) dw \tag{15}$$

を考える。

ここで事後分布$p_{\rm post}(w|X^n)$が一般的な形の場合だと議論が難しいので、$\mathcal{K}_n(\alpha)$の計算が簡単になるような事後分布$p_{\rm post}(w|X^n)$を考えたい。そこで、次のような分布を考える。

$$p_{\rm post}(w|X^n) = \frac{1}{Z_n(\beta)} \exp(-\beta K_n(w))\varphi(w) \tag{16}$$

このとき$\varphi(w)$はパラメータwの事前分布と呼ばれる。ここで規格化定数$Z_n(\beta)$は

$$Z_n(\beta) \coloneqq \int_W \exp(-\beta K_n(w))\varphi(w) dw \tag{17}$$

である。この特殊形の場合、$K_n(w)$のモーメント母関数は

$$\mathcal{K}_n(\alpha) \coloneqq \frac{1}{Z_n(\beta)} \int_W e^{-(\alpha+\beta)K_n(w)} \varphi(w) dw = \frac{Z_n(\alpha+\beta)}{Z_n(\beta)} \tag{18}$$

となり、規格化定数$Z_n(\beta)$だけの関数となる。つまり経験誤差関数$K_n(w)$の振る舞いを調べることは、規格化定数$Z_n(\beta)$の振る舞いを調べることに帰着される。経験誤差関数の定義を$Z_n(\beta)$に代入すると、

$$Z_n(\beta) = \int_W \prod_{i=1}^n \left\{ \frac{p_{\rm model}(X_i|w)}{p_0(X_i)} \right\}^\beta \varphi(w) dw \tag{19}$$

であることが分かる。同様に事後分布を計算すると、

$$p_{\rm post}(w|X^n) \propto \varphi(w) \prod_{i=1}^n \{p_{\rm model}(X_i|w)\}^\beta \tag{20}$$

となる。これは渡辺ベイズで議論している事後分布の形に他ならない。

参考文献

[1] 渡辺澄夫, 『ベイズ統計の理論と方法』