1　Bayes 理論の目的

Bayes 理論は、パラメータの集合 $W \subset \mathbb{R}^d$、真の分布 $q(x)$、確率モデル $p(x|w)$、事前分布 $\varphi(w)$ が事前に与えられて構築される理論である。理論の目的は以下のようにまとめられる：

Bayes 理論の目的

真の分布 $q(x)$、確率モデル $p(x|w)$、事前分布 $\varphi(w)$ が与えられたとする。自然数 $n = 1, 2, 3, \ldots$ に対して、我々は何らかの処方箋に従って予測分布と呼ばれる確率分布の系列 $\{p_n^*(x)\}_{n=1}^{\infty}$ を構成する。予測分布の構成手法は Bayes 推測、事後確率最大化法、最尤推定法、平均プラグイン法などが存在する。たとえば Bayes 推測を用いた場合では、予測分布を

$$p_n^*(x) \coloneqq \int p(x|w)p(w|X^n)\ \mathrm{d}w \tag{1}$$

と定める。ここで、$p(w|X^n)$ は事後分布と呼ばれる確率分布であり、真の分布 $q(x)$ から得られる n 個の独立なサンプル X^n（データセット）を用いて、

$$p(w|X^n) \propto \varphi(w) \prod_{i=1}^{n} p(X_i|w)^{\beta} \tag{2}$$

と定義される。Bayes 理論の主な目的は、予測分布の構成手法として Bayes 推測を用いた場合、n の増大に対して

$$D_{\mathrm{KL}}[q(x) \parallel p_n^*(x)] = \int q(x) \ln \frac{q(x)}{p_n^*(x)}\ \mathrm{d}x \tag{3}$$

がどのような振る舞いをするか、すなわちどのように予測分布 $p_n^*(x)$ が真の分布 $q(x)$ に近づいていくかを調べることが目的である。また、複数のモデル・予測分布の構成手法に対して、それらの良し悪し（どのような意味での良し悪しかは後で）を定量的に評価できる情報量規準の導入も行う。

Bayes 理論の目的にあるように、式 (3) の n に対する振る舞いを調べていく。そのために必要な概念の定義を以下で行う。

1.1　いくつかの概念の定義と本書の仮定について

真の分布に対して最適なパラメータの集合

パラメータの集合 W に対し、真の分布 $q(x)$ とモデル $p(x|w)$ の間の KL divergence を最小にするパラメータの集合を W_0 と定義する：

$$W_0 \coloneqq \{w \in W \mid w = \arg\min_{w} D_{\mathrm{KL}}[q(x) \parallel p(x|w)]\} \tag{4}$$

この集合のことを、真の分布に対して最適なパラメータの集合と呼ぶ。

実質的にユニーク

任意の $w_0 \in W_0$ について、$p(x|w_0)$ がユニークな確率分布 $p_0(x)$ 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。

以下では実質的にユニークが常に実現されていることを仮定する。さらにこの本では、より強い仮定として、「相対的に有限な分散を持つ」という状況が達成されていることを仮定する。[*1]

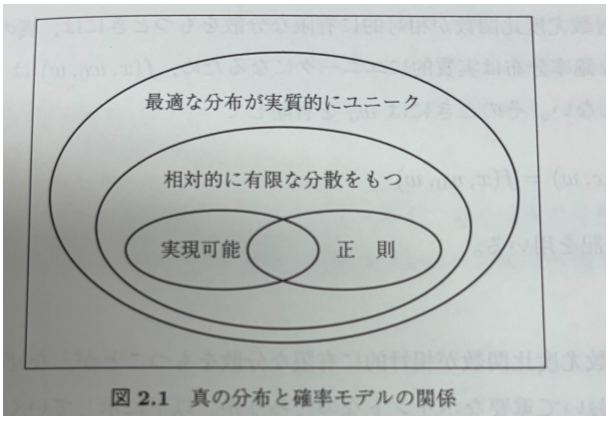

図 1　真の分布と確率モデルに対する関係（渡辺ベイズ p35）

以下の説明は、「相対的に有限な分散を持つ」ことが達成されている、すなわち「最適な分布が実質的にユニーク」な状況を前提にしている。したがって、真の分布 $q(x)$ に KL divergence の意味で最も近い確率モデルは唯一 $p_0(x)$ に定まっている。

1.2　Bayes 理論の目的を達成するために必要なこと

Bayes 理論の目的は、$D_{\mathrm{KL}}[q(x) \parallel p_n^*(x)]$ の n に対する振る舞いを調べることである。以下では計算の都合上、式 (3) の両辺から $D_{\mathrm{KL}}[q(x) \parallel p_0(x)]$（$n$ にはよらない定数）を引いた値の振る舞いを考える：

$$D_{\mathrm{KL}}[q(x) \parallel p_n^*(x)] - D_{\mathrm{KL}}[q(x) \parallel p_0(x)] = -\int q(x) \ln \frac{p_n^*(x)}{p_0(x)} \, \mathrm{d}x \ 。 \tag{5}$$

予測分布の構成方法として Bayes 推測を用いた場合、右辺は

$$-\int q(x) \ln \frac{p_n^*(x)}{p_0(x)} \, \mathrm{d}x = -\mathbb{E}_X \left[\ln \left\langle e^{\ln \frac{p(X|w)}{p(X|w_0)}} \right\rangle_{w|X^n} \right] \tag{6}$$

と表される。なお、$\langle \cdots \rangle_{w|X^n}$ は事後分布による期待値を表す：

$$\langle O(x,w) \rangle_{w|X^n} \coloneqq \int O(x,w) p(w|X^n) \, \mathrm{d}w \ 。 \tag{7}$$

[*1] 「相対的に有限な分散を持つ」ことが達成されていない場合、後で定義する重要な量である $K_n(w)$ の分散がその平均でバウンドできなくなるため、事後分布がサンプルに応じて大きく変動してしまうことが問題となるらしいです。これ以上はよくわからないです。（渡辺ベイズ p36, 注意 12 を参照。）

式 (6) の n の増大に対する振る舞いを見るためには、式 (7) の n に対する振る舞いを調べることがすべてである。その調査を行えるようにするため、$p(w|X^n)$ を天下り的だが（w の積分計算にラプラス近似を用いたいため）次のように変形する：

$$p(w|X^n) \propto \varphi(w) \prod_{i=1}^{n} p(X_i|w)^{\beta} \propto \varphi(w) \exp(-n\beta K_n(w)) \,。 \tag{8}$$

ここで $K_n(w)$ は経験誤差関数であり、

$$K_n(w) \coloneqq \frac{1}{n} \sum_{i=1}^{n} f(X_i, w) \coloneqq \frac{1}{n} \sum_{i=1}^{n} \ln \frac{p(X_i|w_0)}{p(X_i|w)} \tag{9}$$

と定義される。なお、$f(x, w)$ は対数尤度比関数と呼ばれる。このように変形することで、事後分布の確率的な振る舞いを $K_n(w)$ に押し付けることができた。さらに指数関数の肩が $-n$ に比例しているため、n が十分に大きな場合、式 (7) の積分計算に寄与する領域は $K_n(w)$ の最小値付近のみと考えることができる。したがって、

1. $K_n(w)$ の最小値を取る点が、パラメータ空間 W 全体でどのような分布をしているか
2. $K_n(w)$ の最小値周りにおける関数の局所的な振る舞い

を調べることが重要である。$K_n(w)$ の w に対する振る舞い方は、真の分布とモデル設計に依存して決まるため、まずは正則理論という枠組みの元で $K_n(w)$ の振る舞いを考える。

2　正則理論

正則理論とは、以下の 3 つの条件を真の分布とモデル設計に課す理論である：

1. 最適なパラメータの集合は唯一の元からなる：

$$W_0 = \{w_0\} \,。 \tag{10}$$

   これは $K(w)$ を最小にするパラメータが唯一 w_0 であることと同値である。
2. $K(w)$ のヘシアン $\nabla^2 K(w)$ の $w = w_0$ における値が正定値である。
3. サンプルの数 n が非常に大きいこと。

正則理論の

3　正則理論の場合

天下り的ではあるが、$K_n(w)$ を次のように変形する：

$$K_n(w) = K(w) - (K(w) - K_n(w)) = K(w) - \frac{1}{\sqrt{n}} \eta_n(w) \,。 \tag{11}$$

ここで

事後分布の振る舞い（正則理論の場合）

正則理論という仮定を課している場合、式 (7) の振る舞いは次のように表される：

$$\langle O(x,w)\rangle_{w|X^n} = \frac{\int O(x,w)\exp\left(-\frac{n\beta}{2}\left|J^{1/2}(w-w_0-\frac{\hat{\xi_n}}{\sqrt{n}})\right|\right)dw}{\int \exp\left(-\frac{n\beta}{2}\left|J^{1/2}(w-w_0-\frac{\hat{\xi_n}}{\sqrt{n}})\right|\right)dw}(1+o_p(1)) \tag{12}$$