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]

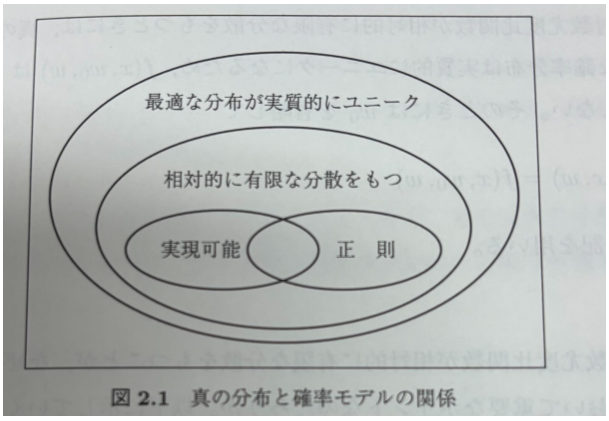

図 1　真の分布と確率モデルに対する関係（渡辺ベイズ p35）

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] 「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。）

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：

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

式 (6) の n の増大に対する振る舞いを見るためには、式 (7) を計算する必要がある。しかしこの積分を解析的に行うことは不可能であるため、ラプラス近似を用いて n に対するオーダー評価として計算することを考える。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、$K_n(w)$ は $f(x,w)$ のサンプル平均である。

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、第三章で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。なお、$n \to \infty$ 極限における $K_n(w)$ の収束先を平均誤差関数 $K(w)$ と呼ぶ：

$$K_n(w) \to K(w) \coloneqq \mathbb{E}_X[f(X,w)] = \int f(x,w) q(x) \, \mathrm{d}x \ 。 \tag{10}$$

$K(w)$ は真の分布 $q(x)$ とモデル設計方法 $p(x|w)$ に依存して決まる関数である。

1.3 期待値計算の評価において一般的に成り立つ事実

期待値計算における積分の分割

事後分布による期待値計算は以下の形で表される：

$$\langle O(x,w)\rangle_{w|X^n} = \frac{\int_W O(x,w)\exp(-n\beta K_n(w))\varphi(w)\ \mathrm{d}w}{\int_W \exp(-n\beta K_n(w))\varphi(w)\ \mathrm{d}w}\ 。 \tag{11}$$

分母は分子の積分において $O(x,w)=1$ を代入した形であるため、分子の n に対する振る舞いを見れば全体の振る舞いがわかる。分子の積分領域を 2 つに分割し、n に対する主要項（第一項目）、非主要項（第二項目）へと分割する：[a] [b]

$$\int_{K(w)<\epsilon_n} O(x,w)\exp(-n\beta K_n(w))\varphi(w)\ \mathrm{d}w + \int_{K(w)\geq\epsilon_n} O(x,w)\exp(-n\beta K_n(w))\varphi(w)\ \mathrm{d}w\ 。 \tag{12}$$

ただし、値 $\epsilon_n > 0$ は n の単調減少関数であり、

$$\lim_{n\to\infty}\epsilon_n = 0, \quad \lim_{n\to\infty}\sqrt{n}\epsilon_n = \infty \tag{13}$$

を満たすように選ぶ。このとき、非主要項（第二項目）は正則理論の仮定なしで、

$$\int_{K(w)\geq\epsilon_n} O(x,w)\exp(-n\beta K_n(w))\varphi(w)\ \mathrm{d}w = o_p(\exp(-\sqrt{n})) \tag{14}$$

になることが示される。この事実を示すために必要なのは $K_n(w)$ の n に対する振る舞いの調査のみであり、正則理論の仮定は必要なく常に成り立つ。したがって、

$$\langle O(x,w)\rangle_{w|X^n} = \frac{\int_{K(w)<\epsilon_n} O(x,w)\exp(-n\beta K_n(w))\varphi(w)\ \mathrm{d}w}{\int_{K(w)<\epsilon_n} \exp(-n\beta K_n(w))\varphi(w)\ \mathrm{d}w} + o_p(\exp(-\sqrt{n}))\ 。 \tag{15}$$

が常に成立する。

[a] 主要項、非主要項という言葉は、各項のオーダー評価が意味をなすくらい n が大きな領域ということを前提に使っている。また、二項目が非主要項であることは、第一項目の評価を行って初めて分かる事実ではあるが、それは認めてください。

[b] $O(x,w)$ は n に依存していない関数であることを前提としており、これが n に依存していればオーダー評価も変わってくる。

以降は主要項

$$\int_{K(w)<\epsilon_n} O(x,w)\exp(-n\beta K_n(w))\varphi(w)\ \mathrm{d}w \tag{16}$$

のオーダー評価を行う。$K(w)$ に性質の良い仮定を課してオーダー評価を行っているのが正則理論（第三章）であり、そうでない場合は一般理論（第四章）である。[*2]

[*2] $K_n(w)$ ではなく $K(w)$ に仮定を課していることには注意。正則理論の枠組みでは、$K(w)<\epsilon_n$ の領域内部において

$$\exp(-n\beta K_n(w)) \propto \mathcal{N}\Big(w_0 + \hat{\xi}_n/\sqrt{n}, (n\beta J(w_0))^{-1}\Big)(1+o_p(1)) \tag{17}$$

と表される。なお、$J(w_0)$ は正定値行列で、$J(w)=\nabla^2 K(w)$ と定義されている。また、$\hat{\xi}_n$ は $K_n(w_0)$ の確率的な振る舞いを担う確率変数である。積分領域はこの分布の広がりよりも十分に大きく、$\mathbb{R}^d$ 全体での積分に置き換えて Gauss 積分が適用できる。

2　正則理論

2.1　正則理論での仮定

正則理論とは、真の分布とモデル設計（と、サンプルの数）が、以下の 3 つの条件を満たすことを仮定する理論である[*3]：

1. $K(w)$ を最小にするパラメータが唯一 w_0 である。これは、最適なパラメータの集合が $W_0 = \{w_0\}$ と表されることと同じ意味である。
2. $K(w)$ の Hessian：$J(w) = \nabla^2 K(w)$ に対し、$J \coloneqq J(w_0)$ が正定値である。
3. サンプルの数 n が十分に大きいこと。

これらの仮定によって、

- 主要項の積分範囲を考える際には $w = w_0$ の近傍のみでよくなる。
- $w = w_0$ 近傍で主要項の被積分関数は Gaussian として扱える。

という嬉しい性質が成り立つ。結論だけ述べると、この仮定の元で主要項のオーダー評価は次のようになる。

期待値のオーダー評価（正則理論の場合）

正則理論という仮定を課している場合、以下のオーダー評価が成立する：

$$\langle O(x,w)\rangle_{w|X^n} = \frac{\int_{\mathbb{R}_d} O(x,w)\exp\left(-\frac{n\beta}{2}\left\|J^{1/2}\left(w - w_0 - \frac{\hat{\xi}_n}{\sqrt{n}}\right)\right\|^2\right)\mathrm{d}w}{\int_{\mathbb{R}_d}\exp\left(-\frac{n\beta}{2}\left\|J^{1/2}\left(w - w_0 - \frac{\hat{\xi}_n}{\sqrt{n}}\right)\right\|^2\right)\mathrm{d}w}(1 + o_p(1))\,。 \tag{20}$$

2 章の残りでは、この計算途中で大事だと思われる式変形について説明する。

[*3] 本では $K(w)$ ではなく $L(w)$ についての仮定をしていたが、これらは基準点を変えただけであり、

$$L(w) = L(w_0) + K(w) \tag{18}$$

によって結びつけられるため、$K(w)$ に関する仮定として考えても良い。なお、$L(w)$ は平均対数損失であり、

$$L(w) = -\mathbb{E}_X[\ln p(X|w)] \tag{19}$$

と定義される。

2.2 正則理論で用いる文字の定義

正則理論の仮定の元での主要項の振る舞いの結果を述べる前に、少しだけ文字の定義を行う。まずは $K_n(w)$ を次のように変形する：

$$K_n(w) = K(w) - (K(w) - K_n(w)) = K(w) - \frac{1}{\sqrt{n}}\eta_n(w)\ 。 \tag{21}$$

ここで $\eta_n(w)$ は $K_n(w)$ の確率的な振る舞い方を担う確率過程であり、次のように定義される：

$$\eta_n(w) \coloneqq \frac{1}{\sqrt{n}}\sum_{i=1}^{n}(K(w) - f(X_i, w))\ 。 \tag{22}$$

この確率過程は経験過程と呼ばれるものであり、$n \to \infty$ でとある確率過程に法則収束するという良い性質を持っている。この性質があることで、収束先の確率分布からのズレとして $K_n(w)$ をオーダー評価することが可能になる。必要となる確率変数 $\xi_n, \hat{\xi}_n$ を

$$\xi_n = J^{-1/2}(w_0)\nabla\eta_n(w_0)\ 、 \tag{23}$$

$$\hat{\xi}_n = J^{-1/2}(w_0)\xi_n = J^{-1}\nabla\eta_n(w_0) \tag{24}$$

と定義する。$\nabla\eta_n(w)$ がすべての w で $\mathcal{N}(0, I(w))$ に法則収束するという性質を持っているため、

- 確率変数 ξ_n は正規分布 $\mathcal{N}(0, J^{-1/2}IJ^{-1/2})$ に法則収束する。
- 確率変数 $\hat{\xi}_n$ は正規分布 $\mathcal{N}(0, J^{-1}IJ^{-1})$ に法則収束する。

ここで $I = I(w_0)$ であり、$I(w)$ は以下で定義される：

$$I(w) \coloneqq \mathbb{E}_X\left[\nabla f(X, w)(\nabla f(X, w))^\top\right] - \nabla K(w)(\nabla K(w))^\top\ 。 \tag{25}$$

2.3 正則理論における主要項のオーダー評価

正則理論の仮定のもとでは、$w = w_0$ 近傍の $K_n(w)$ の振る舞いを見れば良い。$K_n(w)$ を平方完成して Gauss 積分として扱いたいため、平均値の定理を利用して $w = w_0$ 周りの展開を考える。そのため

$$K_n(w) = K(w) - \frac{1}{\sqrt{n}}\eta_n(w) \tag{26}$$

の右辺の項にそれぞれ平均値の定理を用いると、w により定まる 2 つの値 w^*, w^{**} が存在し、

$$K(w) = \frac{1}{2}(w - w_0) \cdot J(w^*)(w - w_0)\ , \tag{27}$$

$$\eta_n(w) = (w - w_0) \cdot \nabla\eta_n(w^{**})\ 。 \tag{28}$$

が成立する。$K(w)$ の一次項が出てこないのは、$\nabla K(w_0) = 0$ が成立しているためである。これらの結果から、$K_n(w)$ は w の二次式として表される：

$$K_n(w) = \frac{1}{2}(w - w_0) \cdot J(w^*)(w - w_0) - \frac{1}{\sqrt{n}}(w - w_0) \cdot \nabla\eta_n(w^{**})\ 。 \tag{29}$$

なお、平方完成を行うためには $J(w^*)$ が正則（対角化可能）であることが必要である。本では、w^* が w_0 近傍では $J(w^*)$ が常に正則ということを仮定しているが、その妥当性がどこから来ているのかは不明。正則であることを認めたうえで、$K_n(w)$ は

$$nK_n(w) = \frac{n}{2}\left\| J(w^*)^{1/2}\left(w - w_0 - \frac{1}{\sqrt{n}}J(w^*)^{-1}\nabla\eta_n(w^{**})\right)\right\|^2 - \frac{1}{2}\left\|J(w^*)^{-1/2}\nabla\eta_n(w^{**})\right\|^2 \tag{30}$$

と平方完成できる。

また、$n \to \infty$ のとき $\epsilon_n \to 0$ であり、正則理論の仮定 1 のおかげで、$w^*, w^{**} \to w_0$ が成立する。したがって

$$J(w^*) = J + o_p(1), \quad \nabla\eta_n(w^{**}) = \nabla\eta_n(w_0) + o_p(1) \tag{31}$$

が成立する。このオーダー評価の結果を用いると、主要項は次のように表せる：

$$\begin{aligned}
&\int_{K(w)<\epsilon_n} O(x,w)\exp(-n\beta K_n(w))\varphi(w)\,\mathrm{d}w \\
&= \exp\left(\frac{\beta}{2}\|\xi_n\|^2\right)\varphi(w_0)\int_{K(w)<\epsilon_n} O(x,w)\exp\left(-\frac{n\beta}{2}\left\|J^{1/2}\left(w - w_0 - \frac{\hat{\xi}_n}{\sqrt{n}}\right)\right\|^2\right)\mathrm{d}w\,(1+o_p(1)) \\
&= \exp\left(\frac{\beta}{2}\|\xi_n\|^2\right)\varphi(w_0)\int_{\mathbb{R}^d} O(x,w)\exp\left(-\frac{n\beta}{2}\left\|J^{1/2}\left(w - w_0 - \frac{\hat{\xi}_n}{\sqrt{n}}\right)\right\|^2\right)\mathrm{d}w\,(1+o_p(1))\,。
\end{aligned} \tag{32}$$

最後の等式は、積分範囲を $\mathbb{R}^d$ 全体に広げてもよいということを利用している。これにより、先ほど提示したオーダー評価の形が得られる：

期待値のオーダー評価（正則理論の場合）

正則理論という仮定を課している場合、以下のオーダー評価が成立する：

$$\langle O(x,w)\rangle_{w|X^n} = \frac{\int_{\mathbb{R}_d} O(x,w)\exp\left(-\frac{n\beta}{2}\left\|J^{1/2}\left(w - w_0 - \frac{\hat{\xi}_n}{\sqrt{n}}\right)\right\|^2\right)\mathrm{d}w}{\int_{\mathbb{R}_d}\exp\left(-\frac{n\beta}{2}\left\|J^{1/2}\left(w - w_0 - \frac{\hat{\xi}_n}{\sqrt{n}}\right)\right\|^2\right)\mathrm{d}w}(1+o_p(1))\,。 \tag{33}$$