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

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

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

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

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、

$$D_{\mathrm{KL}}[q(x)\parallel p_n^*(x)] = \mathbb{E}_X[\ln q(x)] - \mathbb{E}_X[\ln p_n^*(x)] \tag{4}$$

であり、右辺第一項は真の分布のみで定まる定数である。したがって右辺第二項の振る舞いを調べられれば良い。予測分布を Bayes 推測で構成した場合、

$$-\mathbb{E}_X[\ln p_n^*(x)] = -\mathbb{E}_X\left[\ln\langle p(X|w)\rangle_{w|X^n}\right] \tag{5}$$

となり、この値は汎化損失 G_n と呼ばれるものになっている：

$$G_n \coloneqq -\mathbb{E}_X\left[\ln\langle p(X|w)\rangle_{w|X^n}\right]。 \tag{6}$$

しかし我々は真の分布にアクセスできないため、原理的に G_n の値を求めることはできない。Bayes 理論の目的を言い換えると、この G_n の値をいかにして推定するかということになる[a]。特に

- 事後分布による期待値計算を如何にして実行するか（原理的には計算可能だが…？）
- 真の分布による期待値計算をどのような推定量として評価するか

ということが重要である。また、G_n は出現サンプルに依存した値であることにも注意。

[a] 原理的に計算可能な T_n を推定量として見るのではだめ？ G_n を推定すること自体をしたいわけではない？たぶん情報量規準はこの疑問と関係している気がする。

事後分布はモデルと事前分布から定まるため、原理的には計算可能である。実用上では、事後分布やモデルが複雑であれば積分計算が実行できないため、積分の収束先からのズレを見るオーダー評価しか行えない。オーダー評価の処方箋を与えているのが正則理論や一般理論と呼ばれる理論である。

（情報量規準についてまだ理解してないので、この段落は自分の予想）真の分布による期待値計算は、原理的に何らかの推定量として扱う必要がある。推定量として扱う方法を与えるために、情報量規準と呼ばれる量を導入する。そのため、異なる情報量規準を用いると異なる推定結果となり、評価の意味合いが変わってくる。

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：

$$W_0 \coloneqq \{w \in W \mid w = \arg\min_w D_{\mathrm{KL}}[q(x) \parallel p(x|w)]\} \tag{7}$$

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。渡辺ベイズ本では、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]

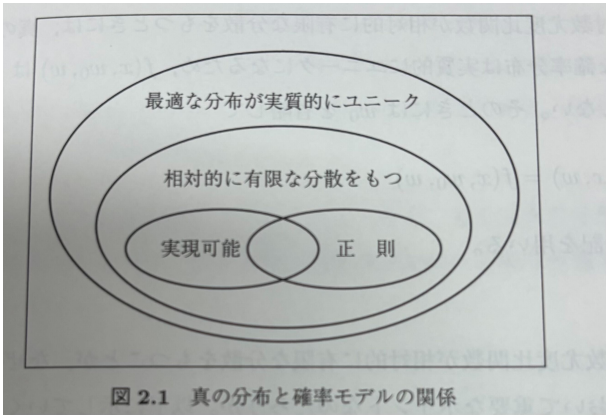

図 1　真の分布と確率モデルに対する関係（渡辺ベイズ p35）

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。

[*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 事後分布の期待値計算を行いたい

本当は、任意のモデルと事前分布の設計に対して

$$\ln \langle p(X|w) \rangle_{w|X^n} \tag{8}$$

の計算を行いたい。しかしそれすら難しいので、n に対するオーダー評価で我慢する[*2]。$p(X|w)$ は n に依存しない関数なので、事後分布が n に対してどのような振る舞いをしているかを見ればオーダー評価の意味では十分。すなわち示すべきことは、$O(x,w)$ を n に依存しない関数として、

$$\langle O(x,w) \rangle_{w|X^n} \coloneqq \int O(x,w) p(w|X^n) \, \mathrm{d}w \ 。 \tag{9}$$

のオーダー評価を行うことである。それを実行するため、$p(w|X^n)$ を天下り的だが（ラプラス近似を用いたいため）次のように変形する：

$$p(w|X^n) \propto \varphi(w) \prod_{i=1}^{n} p(X_i|w)^{\beta} \propto \varphi(w) \exp(-n\beta K_n(w)) \ 。 \tag{10}$$

ここで $K_n(w)$ は経験誤差関数であり、

$$K_n(w) \coloneqq \frac{1}{n} \sum_{i=1}^{n} f(X_i, w) \coloneqq \frac{1}{n} \sum_{i=1}^{n} \ln \frac{p(X_i|w_0)}{p(X_i|w)} \tag{11}$$

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。この変形を用いると、

$$\langle O(x,w) \rangle_{w|X^n} = \frac{\int_W O(x,w) \exp(-n\beta K_n(w)) \varphi(w) \, \mathrm{d}w}{\int_W \exp(-n\beta K_n(w)) \varphi(w) \, \mathrm{d}w} \tag{12}$$

となる。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 (12) 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：

$$K_n(w) \to K(w) \coloneqq \mathbb{E}_X[f(X,w)] = \int f(x,w) q(x) \, \mathrm{d}x \ 。 \tag{13}$$

$K(w)$ 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。

[*2] オーダー評価をしようと思うと、原理的に計算できない量 $K(w)$ を出現させる必要がある。これは収束先からのズレを見たいためしょうがないのか？

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：

$$\langle O(x,w)\rangle_{w|X^n} = \frac{\int_W O(x,w)\exp(-n\beta K_n(w))\varphi(w)\ \mathrm{d}w}{\int_W \exp(-n\beta K_n(w))\varphi(w)\ \mathrm{d}w}\ 。 \tag{14}$$

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]

$$\int_{K(w)<\epsilon_n} O(x,w)\exp(-n\beta K_n(w))\varphi(w)\ \mathrm{d}w + \int_{K(w)\geq\epsilon_n} O(x,w)\exp(-n\beta K_n(w))\varphi(w)\ \mathrm{d}w\ 。 \tag{15}$$

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

$$\lim_{n\to\infty}\epsilon_n = 0, \quad \lim_{n\to\infty}\sqrt{n}\epsilon_n = \infty \tag{16}$$

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、

$$\int_{K(w)\geq\epsilon_n} O(x,w)\exp(-n\beta K_n(w))\varphi(w)\ \mathrm{d}w = o_p(\exp(-\sqrt{n})) \tag{17}$$

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、

$$\langle O(x,w)\rangle_{w|X^n} = \frac{\int_{K(w)<\epsilon_n} O(x,w)\exp(-n\beta K_n(w))\varphi(w)\ \mathrm{d}w}{\int_{K(w)<\epsilon_n} \exp(-n\beta K_n(w))\varphi(w)\ \mathrm{d}w} + o_p(\exp(-\sqrt{n}))\ 。 \tag{18}$$

が常に成立する。なお、$K(w)$ がでてきているので、計算できない量になっているのは注意。

[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

$$\int_{K(w)<\epsilon_n} O(x,w)\exp(-n\beta K_n(w))\varphi(w)\ \mathrm{d}w \tag{19}$$

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。[*3]

[*3] $K_n(w)$ 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

$$\exp(-n\beta K_n(w)) \propto \mathcal{N}\Big(w_0 + \hat{\xi}_n/\sqrt{n}, (n\beta J(w_0))^{-1}\Big)(1+o_p(1)) \tag{20}$$

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[*4]：

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、$J := J(w_0)$ 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：

$$\langle O(x,w)\rangle_{w|X^n} = \frac{\int_{\mathbb{R}_d} O(x,w)\exp\left(-\frac{n\beta}{2}\left\|J^{1/2}\left(w-w_0-\frac{\hat{\xi}_n}{\sqrt{n}}\right)\right\|^2\right)\mathrm{d}w}{\int_{\mathbb{R}_d}\exp\left(-\frac{n\beta}{2}\left\|J^{1/2}\left(w-w_0-\frac{\hat{\xi}_n}{\sqrt{n}}\right)\right\|^2\right)\mathrm{d}w}(1+o_p(1))\,。 \tag{23}$$

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。

[*4] 渡辺ベイズ本では $K(w)$ 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、

$$L(w) = L(w_0) + K(w) \tag{21}$$

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、

$$L(w) = -\mathbb{E}_X[\ln p(X|w)] \tag{22}$$

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：

$$K_n(w) = K(w) - (K(w) - K_n(w)) = K(w) - \frac{1}{\sqrt{n}}\eta_n(w)\ 。 \tag{24}$$

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：

$$\eta_n(w) \coloneqq \frac{1}{\sqrt{n}}\sum_{i=1}^{n}(K(w) - f(X_i, w))\ 。 \tag{25}$$

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

$$\xi_n = J^{-1/2}(w_0)\nabla\eta_n(w_0)\ 、 \tag{26}$$

$$\hat{\xi}_n = J^{-1/2}(w_0)\xi_n = J^{-1}\nabla\eta_n(w_0) \tag{27}$$

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：

$$I(w) \coloneqq \mathbb{E}_X\left[\nabla f(X, w)(\nabla f(X, w))^\top\right] - \nabla K(w)(\nabla K(w))^\top\ 。 \tag{28}$$

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

$$K_n(w) = K(w) - \frac{1}{\sqrt{n}}\eta_n(w) \tag{29}$$

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、

$$K(w) = \frac{1}{2}(w - w_0)\cdot J(w^*)(w - w_0)\ , \tag{30}$$

$$\eta_n(w) = (w - w_0)\cdot\nabla\eta_n(w^{**})\ 。 \tag{31}$$

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：

$$K_n(w) = \frac{1}{2}(w - w_0)\cdot J(w^*)(w - w_0) - \frac{1}{\sqrt{n}}(w - w_0)\cdot\nabla\eta_n(w^{**})\ 。 \tag{32}$$

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。渡辺ベイズ本では、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

$$nK_n(w) = \frac{n}{2}\left\| J(w^*)^{1/2}\left(w - w_0 - \frac{1}{\sqrt{n}}J(w^*)^{-1}\nabla\eta_n(w^{**})\right)\right\|^2 - \frac{1}{2}\left\|J(w^*)^{-1/2}\nabla\eta_n(w^{**})\right\|^2 \tag{33}$$

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。したがって一般の n ではそこからのズレとして

$$J(w^*) = J + o_p(1), \quad \nabla\eta_n(w^{**}) = \nabla\eta_n(w_0) + o_p(1) \tag{34}$$

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：

$$\begin{aligned}
&\int_{K(w)<\epsilon_n} O(x,w)\exp(-n\beta K_n(w))\varphi(w)\,\mathrm{d}w \\
&= \exp\left(\frac{\beta}{2}\|\xi_n\|^2\right)\varphi(w_0)\int_{K(w)<\epsilon_n} O(x,w)\exp\left(-\frac{n\beta}{2}\left\|J^{1/2}\left(w - w_0 - \frac{\hat{\xi}_n}{\sqrt{n}}\right)\right\|^2\right)\mathrm{d}w\,(1+o_p(1)) \\
&= \exp\left(\frac{\beta}{2}\|\xi_n\|^2\right)\varphi(w_0)\int_{\mathbb{R}^d} O(x,w)\exp\left(-\frac{n\beta}{2}\left\|J^{1/2}\left(w - w_0 - \frac{\hat{\xi}_n}{\sqrt{n}}\right)\right\|^2\right)\mathrm{d}w\,(1+o_p(1))\,。
\end{aligned} \tag{35}$$

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：

$$\langle O(x,w)\rangle_{w|X^n} = \frac{\int_{\mathbb{R}_d} O(x,w)\exp\left(-\frac{n\beta}{2}\left\|J^{1/2}\left(w - w_0 - \frac{\hat{\xi}_n}{\sqrt{n}}\right)\right\|^2\right)\mathrm{d}w}{\int_{\mathbb{R}_d}\exp\left(-\frac{n\beta}{2}\left\|J^{1/2}\left(w - w_0 - \frac{\hat{\xi}_n}{\sqrt{n}}\right)\right\|^2\right)\mathrm{d}w}(1+o_p(1))\,。 \tag{36}$$

3　正則理論における汎化損失の振る舞い

正則理論においては、事後分布による期待値計算のオーダー評価が可能になった。その結果を用いると G_n は

$$G_n = -\mathbb{E}_X\Big[\ln \langle p(X|w)\rangle_{w|X^n}\Big] = L(w_0) + \frac{1}{n}\bigg(\frac{d}{2\beta} + \frac{1}{2}\|\xi_n\|^2 - \frac{1}{2\beta}\operatorname{tr}\big\{IJ^{-1}\big\}\bigg) + o_p\bigg(\frac{1}{n}\bigg) \tag{37}$$

となる。これには原理的に計算が不可能な量が入っており、これは

- 真の分布による期待値を外から取ったから
- 事後分布の期待値計算をオーダー評価するために収束先の値を見たから

という 2 つの理由がある。真の分布をしらないと計算不可能な量は

$$L(w_0) = -\mathbb{E}_X[\ln p(X|w_0)]\,, \tag{38}$$
$$\xi_n = J^{-1/2}\nabla\eta_n(w_0)\,, \tag{39}$$
$$I = I(w_0) = \mathbb{E}_X\big[\nabla f(X,w_0)(\nabla f(X,w_0))^\top\big] - \nabla K(w_0)(\nabla K(w_0))^\top\,, \tag{40}$$
$$J = J(w_0) = \nabla^2 K(w_0) \tag{41}$$

である。また、w_0 の値も真の分布をしらないと求められない。いったいこれからどうするのか。ちなみに経験損失 T_n は原理的に計算できる：

$$T_n = -\frac{1}{n}\sum_{i=1}^{n}\ln\langle p(X_i|w)\rangle_{w|X^n}\ 。 \tag{42}$$

経験損失のオーダー評価は

$$T_n = L_n(w_0) + \frac{1}{n}\bigg(\frac{d}{2\beta} - \frac{1}{2}\|\xi_n\|^2 - \frac{1}{2\beta}\operatorname{tr}\big\{IJ^{-1}\big\}\bigg) + o_p\bigg(\frac{1}{n}\bigg) \tag{43}$$

となる。

$$L_n(w_0) = -\frac{1}{n}\sum_{i=1}^{n}\ln p(X_i|w_0)\,, \tag{44}$$
$$\tag{45}$$

は ω_0 を別の量に置き換えれば計算可能ではある。また、サンプルによる期待値を取ると、

$$\mathbb{E}[G_n] = L(w_0) + \frac{1}{n}\bigg(\frac{\lambda-\nu}{\beta} + \nu\bigg) + o\bigg(\frac{1}{n}\bigg)\,, \tag{46}$$
$$\mathbb{E}[T_n] = L(w_0) + \frac{1}{n}\bigg(\frac{\lambda-\nu}{\beta} - \nu\bigg) + o\bigg(\frac{1}{n}\bigg) \tag{47}$$

と表される。ここで

$$\lambda = \frac{d}{2},\quad \nu = \frac{1}{2}\operatorname{tr}\big\{IJ^{-1}\big\} \tag{48}$$

である。

4　Bayes 推測以外の方法で予測分布を構成する

予測分布の構成方法として Bayes 推測以外の手法を用いた場合にどのようになるかを説明する。特に、最尤推定法と事後確率最大化法（MAP 推定法）が統一的に説明できることを述べる。これらはパラメータを点推定するような方法である。まず、ハイパーパラメータ $\beta > 0$ に依存する関数

$$\mathcal{L}_n^{\beta}(w) \coloneqq -\frac{1}{n}\sum_{i=1}^{n} \ln p(X_i|w) - \frac{1}{n\beta}\ln \varphi(w) \tag{49}$$

を最小にするパラメータを $\hat{w}_n^{\beta}$ とおく。このようなパラメータが唯一とは限らない。このとき、予測分布を

$$p_n^*(x) = p(X|\hat{w}_n^{\beta}) \tag{50}$$

と定義する。$\beta \to \infty$ とした場合が最尤推定[*5] であり、$\beta = 1$ とした場合が事後確率最大化法（MAP 推定）に対応する。すなわち β は事前分布と学習データをどれくらい重視するかを調整する意味を持っている。

推定量の一致性

式 (7) で定義した集合 W_0 の元が唯一 w_0 であるとき、任意の β について、$n \to \infty$ のとき $K(\hat{w}_n^{\beta})$ の値は 0 に確率収束する。また、正則理論を仮定していれば、$\hat{w}_n^{\beta} \to w_0$ に確率収束する。

これらの方法で得られた $\hat{w}_n^{\beta}$ を用いて、汎化損失と経験損失は

$$L(\hat{w}_n^{\beta}) = -\mathbb{E}_X[\ln p(X|\hat{w}_n^{\beta})] \tag{51}$$

$$L_n(\hat{w}_n^{\beta}) = -\frac{1}{n}\sum_{i=1}^{n}[\ln p(X|\hat{w}_n^{\beta})] \tag{52}$$

と定義される。

[*5] 最尤推定はデータを完全に信じてモデルの対数尤度を最大化するパラメータを選択するような推定方法であるため過学習を起こしやすい？

5　情報量規準

この章は自分の中でその意味を捉えられていないまま書いています。

- G_n そのものを情報量規準で推定したいわけではない？
- G_n のオーダー評価の結果を用いて計算できない量は微小量として扱えるようにし、主要項は計算できる量で推定したい？
- $\mathbb{E}[G_n]$ の意味で、主要項は計算できる量で推定したい？

5.1　BIC

BIC の定義は次である：

$$\mathrm{BIC} \coloneqq -\sum_{i=1}^{n} \ln p(X_i|\hat{w}_n^\beta) + \frac{d}{2}\ln n \tag{53}$$

複数のモデルの比較において BIC が小さいほど適切であると考えるとき、この値のことをベイズ情報量規準という。

BIC は、真の分布は単純なモデルで表現できるということがわかっているときに使うような指標だと思った。BIC が他の指標と違うのは、n に比例する項が入っていることであり、たとえば $n \to \infty$ を考える（真の分布が再現できるほどデータが与えられている）場合を考えると、任意の $X \sim q(x)$ に対して $p(X|\hat{w})$ がゼロにならない $\hat{w}$ が存在するくらいの表現能力を持つ d の中で、一番小さい d をもつモデルを選ぶことになる。つまり真の分布を再現できる表現能力を持つもののうち、最も単純なモデルが選ばれる

BIC について定性的に理解する

解釈をしやすくするために d が同じモデルを 2 つ持ってきて、β も固定する。データセットも同じものを用意する。事前分布も同じにする。もっともらしいパラメータ $\hat{w}_n^\beta$ を決めたとき、BIC が小さいほうが適切ということは、予測分布 $p_n^*(x)$ に関する対数尤度

$$\sum_{i=1}^{n} \ln p_n^*(X_i) \tag{54}$$

が大きいモデルのほうが適切であるということを意味する。これは最尤推定法を擬似的に利用したいということ？また、表現能力が必要以上に高すぎると過学習のおそれがあるので、同じくらいの表現能力を持っているモデルでは、よりパラメータが少ないモデルを選択している。これ以上の意味はわからない。自由エネルギーの主要項が BIC そのものらしいので、自由エネルギーの意味を理解しないといけない気がする。

5.2 RIC

RIC の定義は以下である：

$$\mathrm{RIC} \coloneqq T_n + \frac{1}{n}\operatorname{tr}\{IJ^{-1}\}\ 。 \tag{55}$$

これは計算できない量である。しかし、RIC の $1/n$ よりも大きなオーダー項はすべて計算できる量としてまとめられる：

$$\mathrm{RIC} = L_n(\hat{w}_n^\beta) + \frac{1}{n}\left(\frac{d}{2\beta} + \left(1 - \frac{1}{2\beta}\operatorname{tr}\{I_n(\hat{w}_n^\beta)J_n(\hat{w}_n^\beta)^{-1}\}\right)\right) + o_p\left(\frac{1}{n}\right)\ 。 \tag{56}$$

ここで計算できる新たな量として

$$I_n(w) \coloneqq \frac{1}{n}\sum_{i=1}^n \nabla \ln p(X_i|w)(\nabla \ln p(X_i|w))^\top \tag{57}$$

$$J_n(w) \coloneqq -\frac{1}{n}\sum_{i=1}^n \nabla^2 \ln p(X_i|w) \tag{58}$$

$$\tag{59}$$

を定義した。この量は、大数の法則と $\hat{w}_n^\beta \to w_0$ を用いて

$$I(w_0) = I_n(\hat{w}_n^\beta) + o_p(1), \quad J(w_0) = J_n(\hat{w}_n^\beta) + o_p(1) \tag{60}$$

となることが示せるらしい。

RIC では常に

$$\mathbb{E}[G_n] = \mathbb{E}[\mathrm{RIC}] + o\left(\frac{1}{n}\right) \tag{61}$$

が成立する。よって $1/n$ より大きなオーダーにおいて、汎化損失 G_n のサンプル期待値を推定したいのが目的な気がする。本当はチューニングしきったモデルの中で G_n が小さくなる方がどちらであるかを見たいんだけど、それが難しくてサンプル期待値の意味でしか見れないということ。

5.3 TIC

TIC の定義は以下である：

$$\mathrm{TIC} \coloneqq L_n(\hat{w}_n^\beta) + \frac{1}{n}\operatorname{tr}\{I_n(\hat{w}_n^\beta)J_n^{-1}(\hat{w}_n^\beta)\}\ 。 \tag{62}$$

これは Bayes 以外の推測方法の場合に適用できる RIC というような立ち位置だと思う。実際に

$$\mathbb{E}[L(\hat{w}_n^\beta)] = \mathbb{E}[\mathrm{TIC}] + o\left(\frac{1}{n}\right) \tag{63}$$

が成立する。

5.4 AIC

AIC の定義は以下である：

$$\mathrm{AIC} \coloneqq L_n(\hat{w}_n^\beta) + \frac{d}{n} 。 \tag{64}$$

この量は、モデルが真の分布が実現可能な場合、$\mathrm{tr}\{IJ^{-1}\} = d$、すなわち $\mathrm{tr}\{I_n(\hat{w}_n^\beta)J_n^{-1}(\hat{w}_n^\beta)\} = d + o_p(1/n)$ であるため、

$$\mathrm{RIC} = \mathrm{AIC} + o_p\left(\frac{1}{n}\right) \tag{65}$$

が成立する。しかし、実現可能でない場合はこの関係式が成り立たない。実現可能でない場合は RIC と AIC は異なる量であるため、AIC も大事な量だと思う。

AIC と BIC の意味合い（注意 31 の意味を考えてみる。）

AIC を規準と使う場合、サンプルの現れ方の揺れに応じて選択されるモデルが変動しやすいらしい。すなわち、AIC は感度が高い（データセットの与え方に依存して判定が変わってしまう）ような指標である。
たくさんのモデルを用意しておいて、AIC を規準として採用した場合を考える。AIC を使って生き残ったモデルは、どのようなデータセットが与えられても常に AIC の意味で良いモデルという意味合いを持つ。これは、データセットに依らず堅牢なモデルを選びたいときに使う気がする。
その一方で、BIC は AIC と比べて感度が低い。つまりデータセットを変化させても選択されるモデルは変わりにくい。仮にデータセットが信頼できない場合（データセットに外部からのノイズが乗っているようなイメージのとき）に BIC を使ってしまうと、ノイズが乗ったデータセットを正しいと思い込んだ上でモデルが選択されてしまう。
ふわっとしたイメージとして、

- AIC は、新しいデータセットに対しても（AIC の値が低くなるという意味で）良い予測ができるモデルを選択する
- BIC は、データセットが信頼できることを期待し、同じくらいの表現能力を持っている場合は、パラメータができるだけ低いモデルを選びたい（$d \ln n$ を低くしたい）

という感じで理解しました。これ以上はわからないです。RIC は AIC と BIC と別もの（$\mathbb{E}[G_n]$ を計算できる量で推定したい）という感じかなと思いました。