ベイズ統計の理論と方法-2.2 & 2.3 for 2017/5/25 講究

林　直輝

2017/5/25

1 2.2. 理論の基礎
- 重要な確率変数
- キュムラントと母関数

2 2.3. ベイズ統計理論の構造

3 おまけ

2.2.1 基礎概念

ここまで出てきた諸々の定義をまとめる。真の分布 $q(x)$ からサンプル $X^n = (X_1, ..., X_n)$ 及び X が独立に得られたものとする。

2.2.1 基礎概念

定義 (6)

平均対数損失函数は

$$L(w) = -\mathbb{E}_X[\log p(X|w)],$$

経験対数損失函数は

$$L_n(w) = -\frac{1}{n}\sum_{k=1}^{n} \log p(X_k|w).$$

2.2.1 基礎概念

$\forall k \in \{1, \ldots, n\}, X_k \sim^{i.i.d.} q(x)$ より、

$$h(X_k)\mathrm{d}X_k = h(x)q(x)\mathrm{d}x.$$

$$\begin{aligned} \therefore \quad \mathbb{E}_X[L_n(w)] &= -\frac{1}{n}\sum_{k=1}^{n}\int q(x)\log p(x|w)\mathrm{d}x \\ &= \frac{1}{n}\sum_{k=1}^{n} L(w) \\ &= L(w) \end{aligned}$$

重要な確率変数

Bayes 推測において重要な確率変数を挙げる：

- 自由エネルギー or 負の対数周辺尤度

$$F_n(\beta) = -\frac{1}{\beta} Z_n(\beta).$$

- 汎化損失

$$G_n = -\mathbb{E}_X \left[\log \mathbb{E}_w[p(X|w)]\right].$$

- 経験損失

$$T_n = -\frac{1}{n} \sum_{k=1}^{n} \log \mathbb{E}_w[p(X_k|w)].$$

2.2.2. 正規化された変量

本書の主な仮定：真の分布と確率モデル（尤度）から定義される対数尤度比函数が相対的に有限な分散を持つ。
真の分布に対する最適なパラメータの集合 W_0 は一意とは限らないが、最適な確率分布 p_0 は実質的にユニーク。
$\rightarrow \forall w_0 \in W_0, p(x|w_0) = p_0(x).$

2.2.2. 正規化された変量

対数尤度比函数は式 (2.5) の通りだが上の仮定から $w_0 \in W_0$ に依存しない：

$$f(x,w) = \log \frac{p(x|w_0)}{p(x|w)}.$$

このとき

$$p(x|w) = p(x|w_0)e^{-f(x,w)} = p_0(x)e^{-f(x,w)}.$$

すなわち、$-\log p(\cdot|w)$ と $f(\cdot,w)$ は定数（w の定数関数）分の差がある。

2.2.2. 正規化された変量

定義 (7)

$f(x,w) := \log \frac{p(x|w_0)}{p(x|w)}$ と置き、次を定義する：

- 平均誤差函数

$$K(w) := \mathbb{E}_X[f(X,w)] = \mathrm{KL}(q\|p),$$

- 経験誤差函数

$$K_n(w) := \frac{1}{n}\sum_{k=1}^{n} f(X_i,w).$$

2.2.2. 正規化された変量

$\log p(x|w) = \log p(x|w_0) - f(x, w)$ であるので

$$\begin{aligned} L(w) &= -\mathbb{E}_X[\log p(X|w)] \\ &= -\mathbb{E}_X[\log p(X|w_0) - f(X, w)] \\ &= -\mathbb{E}_X[\log p(X|w_0)] + \mathbb{E}_X[f(X, w)] \\ &= L(w_0) + K(w). \end{aligned}$$

同様にして

$$L_n(w) = L_n(w_0) + K_n(w).$$

2.2.2. 正規化された変量

$K(w) = \mathrm{KL}\left(p(\cdot|w_0)||p(\cdot|w)\right)$ 及び
$W_0 = \mathrm{argmin}_{w \in W} L(w)$ より

$$K(w) \geqq 0, K(w) = 0 \iff p(\cdot|w_0) = p(\cdot|w)$$
$$\iff w \in W_0$$

2.2.2. 正規化された変量

正規化された分配函数を

$$Z_n^{(0)}(\beta) := \int \exp(-n\beta K_n(w))\varphi(w)\mathrm{d}w$$

と定義する。

$$\begin{aligned}\prod_{k=1}^{n} p(X_k|w) &= \prod_{k=1}^{n} q(X_k)e^{-f(X_k,w)} \\ &= \left(\prod_{k=1}^{n} p(X_k|w_0)\right)\left(\prod_{k=1}^{n} e^{-f(X_k,w)}\right)\end{aligned}$$

2.2.2. 正規化された変量

$$
\begin{aligned}
&= \left(\prod_{k=1}^{n} e^{\log p(X_k|w_0)}\right) \exp\left(-\sum_{k=1}^{n} f(X_k, w)\right) \\
&= \left(\prod_{k=1}^{n} p(X_k|w_0)\right) \exp\left(-nK_n(w)\right) \\
&= \exp\left(\sum_{k=1}^{n} \log p(X_k|w_0)\right) \exp\left(-nK_n(w)\right) \\
&= \exp\left(-nL_n(w_0)\right) \exp\left(-nK_n(w)\right).
\end{aligned}
$$

2.2.2. 正規化された変量

$$\therefore \prod_{k=1}^{n} p(X_k|w_0) = \exp\left(-nL_n(w)\right)\exp\left(-nK_n(w)\right).$$

両辺のβ乗を考えることにより、分配函数と事後分布は以下のようになる：

2.2.2. 正規化された変量

- 分配函数

$$
\begin{aligned}
Z_n(\beta) &= \int_W \varphi(w) \prod_{k=1}^{n} p(X_k|w)^\beta \mathrm{d}w \\
&= \int_W \varphi(w) \exp\left(-nL_n(w_0)\right)^\beta \exp\left(-nK_n(w)\right)^\beta \mathrm{d}w \\
&= \exp\left(-\beta nL_n(w_0)\right) \int_W \exp\left(-n\beta K_n(w)\right) \varphi(w) \mathrm{d}w \\
&= \exp\left(-\beta nL_n(w_0)\right) Z_n^{(0)}(\beta).
\end{aligned}
$$

2.2.2. 正規化された変量

- 事後分布

$$
\begin{aligned}
p(w|X^n) &= \frac{1}{Z_n(\beta)}\varphi(w)\prod_{k=1}^{n} p(X_k|w)^{\beta} \\
&= \frac{\exp\left(-n\beta L_n(w_0)\right)\exp\left(-n\beta K_n(w)\right)\varphi(w)}{\exp\left(-\beta n L_n(w_0)\right) Z_n^{(0)}(\beta)} \\
&= \frac{1}{Z_n^{(0)}(\beta)}\exp\left(-n\beta K_n(w)\right)\varphi(w).
\end{aligned}
$$

2.2.2. 正規化された変量

定義 (8)

- 正規化された自由エネルギー

$$
\begin{aligned}
F_n^{(0)}(\beta) &= -\frac{1}{\beta}\log\int \exp(-n\beta K_n(w))\varphi(w)\mathrm{d}w \\
&= -\frac{1}{\beta}\log Z_n^{(0)}(\beta).
\end{aligned}
$$

定義 (8)

- 汎化誤差

$$G_n^{(0)} = -\mathbb{E}_X[\log \mathbb{E}_w[\exp(-f(X,w))]].$$

- 経験誤差

$$T_n^{(0)} = -\frac{1}{n}\sum_{k=1}^{n} \log \mathbb{E}_w[\exp(-f(X,w))].$$

重要な確率変数

この定義は以下と等価である：

$$
\begin{aligned}
G_n^{(0)} &= -\mathbb{E}_X[\log \mathbb{E}_w[\exp(-f(X,w))]] \\
&= \mathbb{E}_X\left[-\log \mathbb{E}_w\left[\frac{p(X|w)}{q(X)}\right]\right] \\
&= \mathbb{E}_X\left[-\log \frac{1}{q(X)}\mathbb{E}_w[p(X|w)]\right] \\
&= \mathbb{E}_X\left[\log \frac{q(X)}{\mathbb{E}_w[p(X|w)]}\right] = \mathrm{KL}(q\|p^*).
\end{aligned}
$$

$$
T_n^{(0)} = \frac{1}{n}\sum_{k=1}^{n} \log \frac{q(X)}{\mathbb{E}_w[p(X|w)]}.
$$

2.2.2. 正規化された変量

補題 (5)

- $F_n(\beta) = nL_n(w_0) + F_n^{(0)}(\beta)$.
- $G_n = L(w_0) + G_n^{(0)}$.
- $T_n = L_n(w_0) + T_n^{(0)}$.

Proof.

自由エネルギーについては

$$Z_n(\beta) = \exp\left(-\beta n L_n(w_0)\right) Z_n^{(0)}(\beta)$$

の対数から直ちに得られる。汎化損失と経験損失については $\log p(x|w) = \log q(x) - f(x, w)$ より

$$L(w) = L(w_0) + K(w),$$

$$L_n(w) = L_n(w_0) + K_n(w)$$

の証明と同様にして得られる。 □

2.2.2. 正規化された変量

注意 (15 なぜ正規化？)

以上より、$L(w_0)$ 及び $L_n(w_0)$ は定数であるため、自由エネルギー $F_n(\beta)$、汎化損失 G_n、経験損失 T_n の理論的挙動を解明するためにはそれぞれの正規化（$F_n^{(0)}(\beta), G_n^{(0)}, T_n^{(0)}$）について解明すればよいことがわかった。

2.2.2. 正規化された変量

注意 (16 なぜ自由エネルギー？)

周辺尤度（分配函数）も負の対数周辺尤度（自由エネルギー）も確率変数である。しかし $n \to \infty$ の極限を考えるとき、前者は自身は法則収束するが平均が発散する。一方後者は自身は法則収束し平均も収束する。
自由エネルギーは確率変数としても平均値としてもオーダーが同じ挙動（収束する）を持つ。従ってどちらも等価なものであるので挙動を見る場合に重要な確率変数は自由エネルギーとなる。

キュムラントと母函数

1. 2.2. 理論の基礎
   - 重要な確率変数
   - キュムラントと母函数
2. 2.3. ベイズ統計理論の構造
3. おまけ

2.2.3. キュムラントと母函数

3 章及び 4 章では Bayes 統計の正則理論及び一般理論を構築する。そのためのフラグとしてキュムラント母函数をここで取り扱う。

定義 (9)

- 汎化損失のキュムラント母函数

$$\mathcal{G}_n(\alpha) = \mathbb{E}_X[\log \mathbb{E}_w[p(X|w)^\alpha]].$$

- 経験損失のキュムラント母函数

$$\mathcal{T}_n(\alpha) = \frac{1}{n}\sum_{k=1}^{n} \log \mathbb{E}_w[p(X_k|w)^\alpha].$$

定義 (9)

それぞれの k 次キュムラントを、母函数の $\alpha = 0$ における k 次微分係数として定める：

$$\mathcal{G}_n^{(k)}(0) = \left(\frac{\mathrm{d}}{\mathrm{d}\alpha}\right)^k \mathcal{G}_n(0),$$

$$\mathcal{T}_n^{(k)}(0) = \left(\frac{\mathrm{d}}{\mathrm{d}\alpha}\right)^k \mathcal{T}_n(0).$$

注意 (17)

明らかに

$$G_n = -\mathcal{G}_n(1)$$

及び

$$T_n = -\mathcal{T}_n(1)$$

が成り立つ。符号の違いは慣例による。また、*0* 次キュムラントは *0* であることもすぐわかる。

定義 (10)

確率変数 A に対して以下を定義する：

$$\ell_k(A) = \frac{\mathbb{E}_w[(\log p(A|w))^k p(A|w)^\alpha]}{\mathbb{E}_w[p(A|w)^\alpha]}.$$

補題 (6)

- 汎化損失について

$$\mathcal{G}_n^{(1)}(\alpha) = \mathbb{E}_X[\ell_1(X)],$$

$$\mathcal{G}_n^{(2)}(\alpha) = \mathbb{E}_X[\ell_2(X) - \ell_1(X)^2].$$

補題 (6)

- 経験損失について

$$\mathcal{T}_n^{(1)}(\alpha) = \frac{1}{n}\sum_{k=1}^{n} \ell_1(X_k),$$

$$\mathcal{T}_n^{(2)}(\alpha) = \frac{1}{n}\sum_{k=1}^{n} [\ell_2(X_k) - \ell_1(X_k)^2].$$

（証明）
汎化損失について。x や w についての可積分性及び α についての微分可能性を仮定すれば積分と微分は交換できる。これに基づいて微分を $\mathbb{E}_X$ 中にいれることができ、その後合成函数の微分及び $\mathbb{E}_w$ と微分の交換を行う。2 階以降は C^{k+1} 級函数 g について成り立つ商の微分公式

$$\left(\frac{\mathrm{d}}{\mathrm{d}\alpha}\right)\frac{g^{(k)}(\alpha)}{g(\alpha)} = \frac{g^{(k+1)}(\alpha)}{g(\alpha)} - \frac{g^{(k)}(\alpha)}{g(\alpha)}\frac{g^{(1)}(\alpha)}{g(\alpha)}$$

を用いつつ微積分を交換する。経験損失も同様。

1 階微分について、

$$\frac{\mathrm{d}}{\mathrm{d}\alpha}\mathcal{G}_n(\alpha) = \frac{\mathrm{d}}{\mathrm{d}\alpha}\mathbb{E}_X[\log \mathbb{E}_w[p(X|w)^\alpha]]$$

$$= \mathbb{E}_X\left[\frac{\mathrm{d}}{\mathrm{d}\alpha}\log \mathbb{E}_w[p(X|w)^\alpha]\right]$$

$$= \mathbb{E}_X\left[\frac{\frac{\mathrm{d}}{\mathrm{d}\alpha}\mathbb{E}_w[p(X|w)^\alpha]}{\mathbb{E}_w[p(X|w)^\alpha]}\right]$$

キュムラントと母函数

$$
\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}\alpha}\mathcal{G}_n(\alpha) &= \mathbb{E}_X\left[\frac{\mathbb{E}_w\left[\frac{\mathrm{d}}{\mathrm{d}\alpha}p(X|w)^\alpha\right]}{\mathbb{E}_w[p(X|w)^\alpha]}\right] \\
&= \mathbb{E}_X\left[\frac{\mathbb{E}_w[(\log p(X|w))p(X|w)^\alpha]}{\mathbb{E}_w[p(X|w)^\alpha]}\right] \\
&= \mathbb{E}_X[\ell_1(X)].
\end{aligned}
$$

2 階微分について、

$$
\begin{aligned}
\frac{\mathrm{d}^2}{\mathrm{d}\alpha^2}\mathcal{G}_n(\alpha) &= \mathbb{E}_X\left[\frac{\mathrm{d}}{\mathrm{d}\alpha}\frac{\frac{\mathrm{d}}{\mathrm{d}\alpha}\mathbb{E}_w[p(X|w)^\alpha]}{\mathbb{E}_w[p(X|w)^\alpha]}\right] \\
&= \mathbb{E}_X\left[\frac{\frac{\mathrm{d}^2}{\mathrm{d}\alpha^2}\mathbb{E}_w[p(X|w)^\alpha]}{\mathbb{E}_w[p(X|w)^\alpha]}\right. \\
&\qquad \left. - \frac{\frac{\mathrm{d}}{\mathrm{d}\alpha}\mathbb{E}_w[p(X|w)^\alpha]}{\mathbb{E}_w[p(X|w)^\alpha]}\frac{\frac{\mathrm{d}}{\mathrm{d}\alpha}\mathbb{E}_w[p(X|w)^\alpha]}{\mathbb{E}_w[p(X|w)^\alpha]}\right] \\
&= \mathbb{E}_X[\ell_2(X) - \ell_1(X)^2].
\end{aligned}
$$

注意 (18 3 次以上のキュムラント)

$$\mathcal{G}_n^{(3)}(\alpha) = \mathbb{E}_X[\ell_3(X) - 3\ell_2(X)\ell_1(X) + 2\ell_1(X)^3].$$

$$\begin{aligned}\mathcal{G}_n^{(4)}(\alpha) &= \mathbb{E}_X[\ell_4(X) - 4\ell_3(X)\ell_1(X) - 3\ell_2(X)^2 \\ &\quad +12\ell_2(X)\ell_1(X)^3 - 6\ell_1(X)^4].\end{aligned}$$

経験損失についても $\mathbb{E}_X \leftarrow \frac{1}{n}\sum_{k=1}^{n}$ 及び $X \leftarrow X_k$ としたものが得られる。

Proof.

$$g(X,\alpha) = \mathbb{E}_w[p(X|w)^\alpha] = \int p(X|w)^\alpha p(w|X^n)\mathrm{d}w$$

と置くと $k \geqq 1$ のとき

$$\ell_k(X) = \frac{g^{(k)}(X,\alpha)}{g(X,\alpha)}$$

が成り立つことから上述の商の微分公式を用いて結論を得る。 □

補題6において$\alpha = 0$とすることでキュムラントを具体的に書き下すことができる。特に1次キュムラントについては

$\log p(x|w) = \log p(x|w_0) - f(x,w)$より次が成立する：

$$\begin{aligned}\mathcal{G}_n^{(1)}(0) &= \mathbb{E}_X[\mathbb{E}_w[\log p(X|w)]] \\ &= \mathbb{E}_X[\mathbb{E}_w[\log p(x|w_0) - f(x,w)]] \\ &= -L(w_0) - \mathbb{E}_w[K(w)].\end{aligned}$$

経験損失についても同様にして

$$\mathcal{T}_n^{(1)}(0) = -L_n(w_0) - \mathbb{E}_w[K_n(w)].$$

定義 (11)

確率変数 A に対して以下を定義する：

$$\mathcal{L}_k(A) = \frac{\mathbb{E}_w[(-f(A,w))^k \exp(-\alpha f(A,w))]}{\mathbb{E}_w[\exp(-\alpha f(A,w))]}.$$

補題 (7)

- 汎化損失について

$$\mathcal{G}_n^{(1)}(\alpha) = -L(w_0) + \mathbb{E}_X[\mathcal{L}_1(X)],$$

$$\mathcal{G}_n^{(2)}(\alpha) = \mathbb{E}_X[\mathcal{L}_2(X) - \mathcal{L}_1(X)^2].$$

補題 (7)

- 経験損失について

$$\mathcal{T}_n^{(1)}(\alpha) = -L_n(w_0) + \frac{1}{n}\sum_{k=1}^{n}\mathcal{L}_1(X_k),$$

$$\mathcal{T}_n^{(2)}(\alpha) = \frac{1}{n}\sum_{k=1}^{n}[\mathcal{L}_2(X_k) - \mathcal{L}_1(X_k)^2].$$

(証明)
いつもの $p(x|w) = p(x|w_0)\exp(-f(x,w))$ より、

$$
\begin{aligned}
\mathcal{G}_n(\alpha) &= \mathbb{E}_X[\log \mathbb{E}_w[p(X|w)^\alpha]] \\
&= \mathbb{E}_X[\log \mathbb{E}_w[p(x|w_0)^\alpha \exp(-\alpha f(X,w))]] \\
&= \mathbb{E}_X[\log p(x|w_0)^\alpha + \log \mathbb{E}_w[\exp(-\alpha f(X,w))]] \\
&= \alpha \mathbb{E}_X[\log p(x|w_0)] + \mathbb{E}_X[\log \mathbb{E}_w[\exp(-\alpha f(X,w))]] \\
&= -\alpha L(w_0) + \mathbb{E}_X[\log \mathbb{E}_w[\exp(-\alpha f(X,w))]].
\end{aligned}
$$

補題 6 の証明と同様にして（$p(X|w) \leftarrow -f(X,w)$ の置換をする）結論を得る。経験損失も同様。 □

注意 (19)

注意 *18* において ℓ を $\mathcal{L}$ に置き換えたものが成立する。

Proof.

Clear. Just do the math. □

補題 (8)

$c_2 = 2, c_3 = 6, c_4 = 26$ と置く。このとき、
$k = 2, 3, 4$ について

$$\left| \left(\frac{\mathrm{d}}{\mathrm{d}\alpha} \right)^k \mathcal{G}_n(\alpha) \right| \leqq c_k \mathbb{E}_X \left[\frac{\mathbb{E}_w[|f(X, w)|^k \exp(-\alpha f(X, w))]}{\mathbb{E}_w[\exp(-\alpha f(X, w))]} \right]$$

$$\left| \left(\frac{\mathrm{d}}{\mathrm{d}\alpha} \right)^k \mathcal{T}_n(\alpha) \right| \leqq \frac{c_k}{n} \sum_{i=1}^{n} \left[\frac{\mathbb{E}_w[|f(X_i, w)|^k \exp(-\alpha f(X_i, w))]}{\mathbb{E}_w[\exp(-\alpha f(X_i, w))]} \right]$$

（証明）まず補題 8 を証明する準備を行う。

$$\mathbb{E}_w^{(\alpha)}[h(A, w)] := \frac{\mathbb{E}_w[h(A, w)\exp(-\alpha f(A, w))]}{\mathbb{E}_w[\exp(-\alpha f(A, w))]},$$

$$\bar{\mathcal{L}}_k(A) := \frac{\mathbb{E}_w[|f(A, w)|^k \exp(-\alpha f(A, w))]}{\mathbb{E}_w[\exp(-\alpha f(A, w))]}$$

と置くと $\mathcal{L}_k(A) = \mathbb{E}_w^{(\alpha)}[(-f(A, w))^k]$ 及び $\bar{\mathcal{L}}_k(A) = \mathbb{E}_w^{(\alpha)}[|f(A, w)|^k] \geqq 0$ である。

キュムラントと母函数

$\mathbb{E}_w^{(\alpha)}[1] = 1$ より確かに平均操作になっている。また、一般に $\mathcal{L}_k(X) \leqq |\mathcal{L}_k(X)| \leqq \bar{\mathcal{L}}_k(X)$ が明らかに成立する。よって Hoelder の不等式より $j \leqq k$ とすると

$$\mathbb{E}_w^{\alpha}[|f(X,w)|^j] \leqq \mathbb{E}_w^{\alpha}[|f(X,w)|^k]^{j/k}$$

すなわち

$$\bar{\mathcal{L}}_j(X) \leqq \bar{\mathcal{L}}_k(X)^{j/k}.$$

これを用いて各 k について補題を示す。

$k=2$ のとき

$$
\begin{aligned}
\left|\mathcal{G}_n^{(2)}\right| &\leqq \mathbb{E}_X[|\mathcal{L}_2(X)-\mathcal{L}_1(X)^2|] \\
&\leqq \mathbb{E}_X[|\mathcal{L}_2(X)|+|\mathcal{L}_1(X)^2|] \\
&\leqq \mathbb{E}_X[\bar{\mathcal{L}}_2(X)]+\mathbb{E}_X[\bar{\mathcal{L}}_1(X)^2] \\
&\leqq \mathbb{E}_X[\bar{\mathcal{L}}_2(X)]+\mathbb{E}_X[\bar{\mathcal{L}}_2(X)] \\
&= 2\mathbb{E}_X\left[\frac{\mathbb{E}_w[|f(A,w)|^2\exp(-\alpha f(A,w))]}{\mathbb{E}_w[\exp(-\alpha f(A,w))]}\right].
\end{aligned}
$$

$k = 3$ のとき

$$
\begin{aligned}
\left|\mathcal{G}_n^{(3)}\right| &\leqq \mathbb{E}_X[|\mathcal{L}_3(X) - 3\mathcal{L}_2(X)\mathcal{L}_1(X) + 2\mathcal{L}_1(X)^3|] \\
&\leqq \mathbb{E}_X[|\mathcal{L}_3(X)| + 3|\mathcal{L}_2(X)||\mathcal{L}_1(X)| + 2|\mathcal{L}_1(X)|^3] \\
&\leqq \mathbb{E}_X[\bar{\mathcal{L}}_3(X) + 3\bar{\mathcal{L}}_2(X)\bar{\mathcal{L}}_1(X) + 2\bar{\mathcal{L}}_1(X)^3] \\
&\leqq \mathbb{E}_X[\bar{\mathcal{L}}_3(X) + 3\bar{\mathcal{L}}_3(X)^{2/3}\bar{\mathcal{L}}_3(X)^{1/3} + 2\bar{\mathcal{L}}_3(X)] \\
&= \mathbb{E}_X[\bar{\mathcal{L}}_3(X) + 3\bar{\mathcal{L}}_3(X) + 2\bar{\mathcal{L}}_3(X)] \\
&= 6\mathbb{E}_X\left[\frac{\mathbb{E}_w[|f(A,w)|^3 \exp(-\alpha f(A,w))]}{\mathbb{E}_w[\exp(-\alpha f(A,w))]}\right].
\end{aligned}
$$

$k=4$ のとき

$$
\begin{aligned}
\left|\mathcal{G}_n^{(4)}\right| &\leqq \mathbb{E}_X[|\mathcal{L}_4(X) - 4\mathcal{L}_3(X)\mathcal{L}_1(X) - 3\mathcal{L}_2(X)^2 \\
&\quad + 12\mathcal{L}_2(X)\mathcal{L}_1(X)^2 - 6\mathcal{L}_1(X)^4|] \\
&\leqq \mathbb{E}_X[|\mathcal{L}_4(X)| + 4|\mathcal{L}_3(X)||\mathcal{L}_1(X)| + 3|\mathcal{L}_2(X)^2| \\
&\quad + 12|\mathcal{L}_2(X)||\mathcal{L}_1(X)^2| + 6|\mathcal{L}_1(X)^4|] \\
&\leqq \mathbb{E}_X[\bar{\mathcal{L}}_4(X) + 4\bar{\mathcal{L}}_3(X)\bar{\mathcal{L}}_1(X) + 3\bar{\mathcal{L}}_2(X)^2 \\
&\quad + 12\bar{\mathcal{L}}_2(X)\bar{\mathcal{L}}_1(X)^2 + 6\bar{\mathcal{L}}_1(X)^4]
\end{aligned}
$$

キュムラントと母函数

$$
\begin{aligned}
\left|\mathcal{G}_n^{(4)}\right| &\leqq \mathbb{E}_X[\bar{\mathcal{L}}_4(X) + 4\bar{\mathcal{L}}_3(X)\bar{\mathcal{L}}_1(X) + 3\bar{\mathcal{L}}_2(X)^2 \\
&\quad + 12\bar{\mathcal{L}}_2(X)\bar{\mathcal{L}}_1(X)^2 + 6\bar{\mathcal{L}}_1(X)^4] \\
&\leqq \mathbb{E}_X[\bar{\mathcal{L}}_4(X) + 4\bar{\mathcal{L}}_4(X)^{3/4}\bar{\mathcal{L}}_4(X)^{1/4} + 3\bar{\mathcal{L}}_4(X)^{(1/2)*2} \\
&\quad + 12\bar{\mathcal{L}}_4(X)^{1/2}\bar{\mathcal{L}}_4(X)^{(1/4)*2} + 6\bar{\mathcal{L}}_4(X)] \\
&= \mathbb{E}_X[\bar{\mathcal{L}}_4(X) + 4\bar{\mathcal{L}}_4(X) + 3\bar{\mathcal{L}}_4(X) \\
&\quad + 12\bar{\mathcal{L}}_4(X) + 6\bar{\mathcal{L}}_4(X)]
\end{aligned}
$$

キュムラントと母函数

$$
\begin{aligned}
\left|\mathcal{G}_n^{(4)}\right| &\leqq \mathbb{E}_X[\bar{\mathcal{L}}_4(X) + 4\bar{\mathcal{L}}_4(X) + 3\bar{\mathcal{L}}_4(X) \\
&\quad + 12\bar{\mathcal{L}}_4(X) + 6\bar{\mathcal{L}}_4(X)] \\
&= \mathbb{E}_X[26\bar{\mathcal{L}}_4(X)] \\
&= 26\mathbb{E}_X\left[\frac{\mathbb{E}_w[|f(A,w)|^4 \exp(-\alpha f(A,w))]}{\mathbb{E}_w[\exp(-\alpha f(A,w))]}\right].
\end{aligned}
$$

以上より汎化損失について不等式が証明された。
経験損失についても同様である。 □

定理 (1 ベイズ統計の基礎定理)

$\left|\left(\frac{\mathrm{d}}{\mathrm{d}\alpha}\right)^3 \mathcal{G}_n(\alpha)\right| = o_p(1/n)$, $\left|\left(\frac{\mathrm{d}}{\mathrm{d}\alpha}\right)^3 \mathcal{T}_n(\alpha)\right| = o_p(1/n)$ を仮定する。このとき以下が成立する：

$$G_n = -\mathcal{G}_n(1) = -\mathcal{G}_n^{(1)}(0) - \frac{1}{2}\mathcal{G}_n^{(2)}(0) + o_p\left(\frac{1}{n}\right).$$

$$T_n = -\mathcal{T}_n(1) = -\mathcal{T}_n^{(1)}(0) - \frac{1}{2}\mathcal{T}_n^{(2)}(0) + o_p\left(\frac{1}{n}\right).$$

上の定理に関して

$$\mathcal{G}_n^{(1)}(0) = -L(w_0) - \mathbb{E}_w[K(w)].$$

$$\mathcal{G}_n^{(2)}(0) = \mathbb{E}_X\left[\mathbb{E}_w[f(X,w)^2] - \mathbb{E}_w[f(X,w)]^2\right].$$

である。経験損失についても $\mathbb{E}_X \leftarrow \frac{1}{n}\sum_{i=1}^n$ 及び $X \leftarrow X_i$ として同様。補題 7 において $\alpha = 0$ とすることで共に簡単に得られる。

キュムラントと母函数

（証明）
$\mathcal{G}_n$ は3階微分可能なので $[0,\alpha]$ に平均値の定理（3次剰余項）を適用すると、

$$\mathcal{G}_n(\alpha) = \mathcal{G}_n(0) + \alpha\mathcal{G}_n^{(1)}(0) + \frac{1}{2}\alpha^2\mathcal{G}_n^{(2)}(0) + \frac{1}{6}\alpha^3\mathcal{G}_n^{(3)}(\alpha^*)$$

を満たす $\alpha^* \in [0,\alpha]$ が存在する。$\alpha = 1$ とすると $\mathcal{G}_n(0) = 0$ と $G_n = -\mathcal{G}_n(1)$ により結論を得る。経験損失についても同様である。 □

キュムラントと母函数

サンプルが現れるとき、サンプルたちが確率変数というだけでなく現れ方も揺らいでいる。そのサンプルの現れ方に関する平均（データの取り方に関する平均）を $\mathbb{E}$ で書く。
上の定理 1 は平均する前の確率変数としての挙動であり、平均したものについては更に次の補題が成立する。以下の補題での汎化損失・経験損失の関係は Bayes 推測特有である。４章の伏線。

補題 (9)

$$\mathbb{E}[\mathcal{G}_{n-1}(\beta)] = -\mathbb{E}[\mathcal{T}_n(-\beta)].$$

が成立する。故に定理 *1* と同じ仮定の下、以下が成立：

$$\mathbb{E}\left[\mathcal{G}_{n-1}^{(1)}(0)+\frac{\beta}{2}\mathcal{G}_{n-1}^{(2)}(0)\right]=\mathbb{E}\left[\mathcal{T}_n^{(1)}(0)-\frac{\beta}{2}\mathcal{T}_n^{(2)}(0)\right]+o\left(\frac{1}{n}\right).$$

（証明）
前半を示せば後半は定理 1 のように平均値の定理を用いてすぐ得られる。前半を示す。

$$\mathbb{E}[\mathcal{G}_{n-1}(\beta)] = \mathbb{E}[\mathbb{E}_X[\log \mathbb{E}_w[p(X|w)^\beta]]]$$

$$= \mathbb{E}\left[\int q(x) \log\left(\int p(x|w)^\beta \frac{\varphi(w)\prod_{i=1}^{n-1} p(X_i|w)^\beta}{\int \varphi(w)\prod_{i=1}^{n-1} p(X_i|w)^\beta \mathrm{d}w}\mathrm{d}w\right)\mathrm{d}x\right]$$

真の分布に関して平均を取っているため確率密度 $p(x|w)$ を X_n の尤度 $p(X_n|w)$ と思っても良いので

$$
\begin{aligned}
(\text{右辺}) &= \mathbb{E}\left[\int q(x)\log\left(\frac{\int \varphi(w)\prod_{i=1}^{n} p(X_i|w)^\beta \mathrm{d}w}{\int \varphi(w)\prod_{i=1}^{n-1} p(X_i|w)^\beta \mathrm{d}w}\right)\mathrm{d}x\right] \\
&= \mathbb{E}\left[\int q(x)\log\frac{Z_n(\beta)}{Z_{n-1}(\beta)}\mathrm{d}x\right] \\
&= \mathbb{E}\left[\log\frac{Z_n(\beta)}{Z_{n-1}(\beta)}\right].
\end{aligned}
$$

ここで log 内の分母と分子を入れ替えて符号を変えると

$$
\begin{aligned}
(\text{右辺}) &= -\mathbb{E}\left[\log \frac{Z_{n-1}(\beta)}{Z_n(\beta)}\right] \\
&= -\mathbb{E}\left[\log\left(\frac{1}{Z_n(\beta)}\int p(X_n|w)^{-\beta}\varphi(w)\prod_{i=1}^{n} p(X_i|w)^{\beta}\mathrm{d}w\right)\right] \\
&= -\mathbb{E}\left[\log\left(\int p(X_n|w)^{-\beta}\frac{\varphi(w)\prod_{i=1}^{n} p(X_i|w)^{\beta}}{Z_n(\beta)}\mathrm{d}w\right)\right] \\
&= -\mathbb{E}\left[\log \mathbb{E}_w[p(X_n|w)^{-\beta}]\right].
\end{aligned}
$$

サンプルの出方について平均を取っているので右辺の $\mathbb{E}$ の中身に注目すると $\log \mathbb{E}_w[p(X_1|w)^{-\beta}] = \ldots = \log \mathbb{E}_w[p(X_n|w)^{-\beta}]$ となるため、入れ替えることができるので

$$\begin{aligned}\mathbb{E}[\mathcal{G}_{n-1}(\beta)] &= -\mathbb{E}\left[\log \mathbb{E}_w[p(X_n|w)^{-\beta}]\right] \\ &= -\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} \log \mathbb{E}_w[p(X_i|w)^{-\beta}]\right] \\ &= -\mathbb{E}[\mathcal{T}_n(-\beta)].\end{aligned}$$

よって前半が証明された。 □

Bayes 統計理論では

$$(q(x), p(x|w), \varphi(w)) \mapsto (F_n(\beta), G_n, T_n)$$

として決まる重要な確率変数；自由エネルギー、汎化損失、経験損失の挙動を解明する。ここで、$q(x)$ と $p(x|w)$ から定まる対数尤度比函数は相対的に有限な分散を持つと仮定しておく。
2.2. 節の内容より、以下の手続でこの目的は達成できる：

- 平均対数損失 $L(w)$ に対し最小化元集合 $\mathrm{argmin} L(w) = W_0$ を求める。
- 最適な確率密度 $p(x|w_0)$ $(w_0 \in W_0)$ とモデルから定まる対数尤度比函数 $f(x,w)$ から経験誤差函数 $K_n(w)$ を求める。
- 以下の一般分配函数の挙動を解明し $\mathcal{L}_k(X) = Z_{n,k}(\alpha,\beta)/Z_{n,0}(\alpha,\beta)$ を計算可能にする。

$$Z_{n,k}(\alpha,\beta) = \int (-f(X,w))^k e^{-\alpha f(X,w) - n\beta K_n(w)} \varphi(w) \mathrm{d}w.$$

- 正規化された自由エネルギーは

$$F_n^{(0)}(\beta) = -\frac{1}{\beta}\log Z_n^{(0)}(\beta) = -\frac{1}{\beta}\log Z_{n,0}(0,\beta)$$

  として計算できる。これより
  $F_n(\beta) = nL_n(w_0) + F_n^{(0)}(\beta)$ として自由エネルギーもわかる。
- 定理 1 から汎化損失 G_n 及び経験損失 T_n の挙動がわかる。1 次及び 2 次キュムラントを計算しておいて定理 1 を用いる。

以上の方針に従い、3 章と 4 章で Bayes 統計の理論を構築する。まず 3 章で事後分布が正規分布で近似可能な場合（正則統計モデル）についての理論を構築する。次に 4 章では事後分布が正規分布で近似可能とは限らない一般の場合（特異統計モデルを含む一般の場合）についての理論を構築する。

飾り文字

今回登場した $\mathcal{G}$ などはカリグラフィック体と呼ばれる字体のアルファベットである。LaTeX においては \mathcal{G} と書く。遅くとも 2013 以降の MSOffice の数式エディタ中では \scriptG としてスペースを 1 回押すと勝手に変換してくれる。

飾り文字

なぜか MSOffice ではスクリプト体として登録されているが、LaTeX におけるスクリプト \mathscr{G} は $\mathscr{G}$ である。なお、実数の集合 $\mathbb{R}$ は LaTeX では \mathbb{R} だが、MSOffice では \doubleR である。

平均・汎化と経験、積分と和

計算がほぼ同様なため、上では平均・汎化の場合だけを示し、経験の場合は省略してきた。しかし計算がほぼ同様という以外にもちょっとした理由がある。

平均・汎化と経験、積分と和

測度論の言葉を用いると、和を積分として見ることが容易くなる。個数測度νを$(\mathbb{N}, 2^{\mathbb{N}})$上の測度として次のように定める：

$$\nu(A) = \begin{cases} |A| & A : \text{finite set} \\ \infty & A : \text{infinite set} \end{cases}.$$

いろいろ頑張ると、この個数測度による積分が馴染み深い和になることがわかる：

$$\sum_{k\in A} g_k = \int_A g \mathrm{d}\nu = \int_A g_k \mathrm{d}\nu(k).$$

積分範囲が有限集合 $A = \{1, \ldots, n\}$ であれば

$$\sum_{k=1}^{n} g_k = \int_A g \mathrm{d}\nu = \int_1^n g_k \mathrm{d}\nu(k).$$

このような視点と Fubini の定理を組み合わせると無限和と積分の交換が可能であることが簡単に証明できる場合がある。ちょっと気色悪い可測集合・可測函数をいじるときに便利な気がする。今回の輪読に関して言えば、要素数 n の有限集合上の場合に限り、$(1/n)\mathrm{d}\nu$ で積分することが $q(x)\mathrm{d}\mu(x) = q\mathrm{d}m$ で積分するのと同様に何らかの分布に基づく平均操作をしていると見れる。m は Lebesgue 測度であり $q(x)\mathrm{d}m(x)$ は Lebesgue-Stieltjes 積分として知られる（と理解しているが怪しいのでマサカリポイント）。

このような、$1/n$ を一様分布と思ってこれによる平均として上述してきた有限和を見ると、経験 hoge も平均 $\mathbb{E}_X$ が上の意味の平均に置き換わったものであるとわかる。この点からも、キュムラントで頻出した平均と微分の交換なども含めて、汎化 hoge のほうで行っておけば経験 hoge も「同様にして」と説明できる（と思う）。