3.5.2 汎化誤差と経験誤差.

Thm.5

(仮定の下.)

MLE or MAP 推定量

汎化損失: $L(\hat{w}) = L(w_0) + \frac{1}{2n}\|\xi_n\|^2 + o_p(1/n)$

経験損失: $L_n(\hat{w}) = L_n(w_0) - \frac{1}{2n}\|\xi_n\|^2 + o_p(1/n)$

$$
\begin{cases}
E[L(\hat{w})] = L(w_0) + \frac{1}{2n}\mathrm{tr}(IJ^{-1}) + o(1/n) \\
E[L_n(\hat{w})] = L(w_0) - \frac{1}{2n}\mathrm{tr}(IJ^{-1}) + o(1/n)
\end{cases}
$$

§3.6 サンプルから計算する方法.

サンプルから { 自由エネルギー $F_n(\beta)$ / 汎化損失 G_n / 経験〃 T_n } を計算したい.

3.6.1 自由エネルギー

- $p(w|X^n) = \frac{1}{Z_n(\beta)} \underline{\varphi(w) \prod_i p(X_i|w)^\beta}$
- $Z_n(\beta) = \int_W \cdots \, dw$.

§3.5と同じ

- $\mathcal{L}(w) := -\frac{1}{n}\sum_i \log p(X_i|w) - \frac{1}{n\beta}\log\varphi(w)$ $\left(= -\frac{1}{n\beta}\log \underline{p(w|X^n)} - \frac{1}{n\beta}\log Z_n(\beta) \right)$

  SSを満たす時には $\mathcal{N}(w_0 + \frac{1}{\sqrt{n}}\xi_n, (n\beta J)^{-1/2})$ Lem 15.

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

  $= -\frac{1}{\beta}\log\int \exp(-n\beta\mathcal{L}(w))dw$.

- $\hat{w} \in \operatorname{argmin} \mathcal{L}(w) \longrightarrow \mathcal{L}(w) \simeq \mathcal{L}(\hat{w}) + \frac{1}{2}(w-\hat{w})\cdot\nabla^2\mathcal{L}(\hat{w})(w-\hat{w})$

- $Z_n(\beta) = \int \exp(-n\beta\mathcal{L}(w))dw \simeq \cdots = \exp(-n\beta\mathcal{L}(\hat{w}))\left(\frac{2\pi}{n\beta}\right)^{d/2}\det(\nabla^2\mathcal{L}(\hat{w}))^{-1/2}$

  これも近似なはず.

- $F_n(\beta) = n\beta\mathcal{L}(\hat{w}) - \frac{1}{\beta}\cdot\frac{d}{2}\log\left(\frac{2\pi}{n\beta}\right) + \frac{1}{2\beta}\log\det(\nabla^2\mathcal{L}(\hat{w})) + o_p(1)$

  $= \underline{n\mathcal{L}(\hat{w}) + \frac{d}{2\beta}\log n + \frac{d}{2\beta}\log\frac{\beta}{2\pi} + \frac{1}{2\beta}\log\det(\nabla^2\mathcal{L}(\hat{w})) + o_p(1)}$

  q が未知でも計算可能.

[Thm 2 (p 64) より

$F_n(\beta) = nL_n(\hat{w}) + \frac{d}{2\beta}\log n + \frac{d}{2\beta}\log\frac{\beta}{2\pi} + \frac{1}{2\beta}\log\det J - \frac{1}{2}\|\xi_n\|^2 - \frac{1}{\beta}\log\varphi(w_0) + o_p(1)$.

ここで、$\nabla^2\mathcal{L}(\hat{w}) = J + o_p(1)$ より p 78. L13 が従う.]

- $F_n(1) = n\mathcal{L}(\hat{w}) + \frac{d}{2}\log n + \frac{d}{2}\log\frac{1}{2\pi} + \frac{1}{2}\log\det(\nabla^2\mathcal{L}(\hat{w})) + o_p(1)$.

  $= \boxed{-\sum_i \log p(X_i|\hat{w}) + \frac{d}{2}\log n} - \log\varphi(\hat{w}) + \frac{d}{2}\log\frac{1}{2\pi} + \frac{1}{2}\log\det(\nabla^2\mathcal{L}(\hat{w})) + o_p(1)$.

- $\mathrm{BIC} = \underbrace{-\sum_i \log p(X_i|\hat{w})}_{n\times\text{経験対数損失 } L_n(\hat{w})} + \frac{d}{2}\log n$

BICは自由エネルギーの近似.

※これを2倍して

$$\mathrm{BIC} = -2\times\text{対数尤度} + d\log n$$

とする方が一般的。

※ BICが小さい ⟺ 自由エネルギーが小さい ⟺ 分配関数が大きい。

$$\int \varphi(w)\prod_i p(x_i|w)\,dw.$$

この値の大きさは「事前分布と確率モデルがどれだけ観測データをよく説明するか」を表しているため、しばしば「エビデンス」と呼ばれるが、この値が大きい程モデルとして適切であると考えられる。
エビデンスは経験ベイズにおいても重要な役割を果たす。

階層ベイズ： $p(x) = \int p(x|w)\,p(w)\,dw$ （記号は乱用）

ここにも prior を与える.

$$= \int p(x|w)\left[\int p(w|\theta)\,p(\theta)\,d\theta\right]dw \qquad \theta \to w \to x$$

経験ベイズ： θ を点推定 $(p(\theta) = \delta(\theta-\theta^*))$. → このとき $\int p(w|\theta)p(\theta)d\theta = p(w|\theta^*)$

$$\theta^* = \underset{\theta}{\mathrm{argmin}} \underbrace{\int p(w|\theta)\prod_i p(x_i|w)\,dw}_{\text{エビデンス}}.$$

3.6.2 汎化損失と経験損失.

・ $G_n = L(w_0) + \frac{1}{n}\left(\frac{d}{2\beta} + \frac{1}{2}\|\xi_n\|^2 - \frac{1}{2\beta}\mathrm{tr}(IJ^{-1})\right) + o_p(1/n)$ （Thm 3. p70）

$L(\hat{w}) - \frac{1}{2n}\|\xi_n\|^2 + o_p(1/n)$ （Thm 5. p75）

$$= L(\hat{w}) + \frac{1}{n}\left(\frac{d}{2\beta} - \frac{1}{2\beta}\mathrm{tr}(IJ^{-1})\right) + o_p(1/n)$$

・ $T_n = L_n(\hat{w}) + \frac{1}{n}\left(\frac{d}{2\beta} - \frac{1}{2\beta}\mathrm{tr}(IJ^{-1})\right) + o_p(1/n)$ （↑と同様）

I, J：計算可能

・ T_n は計算可能，$L(\hat{w})$ が求まらないので G_n はわからない．

→サンプルに関する期待値の意味で G_n に等しい量を定義する

・ $\mathrm{RIC} := T_n + \frac{1}{n}\mathrm{tr}(IJ^{-1})$.

$$= L_n(\hat{w}) + \frac{1}{n}\left(\frac{d}{2\beta} + \left(1 - \frac{1}{2\beta}\right)\mathrm{tr}(IJ^{-1})\right) + o_p(1/n)$$

- $E[R2C] = E[L_n(\hat{w})] + \frac{1}{n}\left(\frac{d}{2\beta} + \left(1 - \frac{1}{2\beta}\right) tr(IJ^{-1})\right) + o(1/n)$

$$= L(w_0) - \frac{1}{2n} tr(IJ^{-1}) + \frac{1}{n}\left(\frac{d}{2\beta} + \left(1 - \frac{1}{2\beta}\right) tr(IJ^{-1})\right) + o(1/n) \quad (\text{Thm.5})$$

$$= L(w_0) + \frac{1}{2n} tr(IJ^{-1}) + \frac{1}{n}\left(\frac{d}{2\beta} - \frac{1}{2\beta} tr(IJ^{-1})\right) + o(1/n)$$

- $E[G_n] = E[L(\hat{w})] + \frac{1}{n}\left(\frac{d}{2\beta} - \frac{1}{2\beta} tr(IJ^{-1})\right) + o(1/n)$

$$= L(w_0) + \frac{1}{2n} tr(IJ^{-1}) + \frac{1}{n}\left(\frac{d}{2\beta} - \frac{1}{2\beta} tr(IJ^{-1})\right) + o(1/n)$$

R2Cはベイズ汎化損失を推測

- $\beta = 1$ のとき,
  - $G_n - L(w_0) = \frac{1}{n}\left(\frac{d}{2} + \frac{1}{2}\|\xi_n\|^2 - \frac{1}{2} tr(IJ^{-1})\right) + o_p(1/n)$ (Thm.3)
  - $R2C - L_n(w_0) = T_n - L_n(w_0) + \frac{1}{n} tr(IJ^{-1})$

$$= \frac{1}{n}\left(\frac{d}{2} - \frac{1}{2}\|\xi_n\|^2 + \frac{1}{2} tr(IJ^{-1})\right) + o_p(1/n)$$

$$(G_n - L(w_0)) + (R2C - L_n(w_0)) = \frac{d}{n} + o_p(1/n).$$

注意.30

- 竹内情報量基準: $TIC := L_n(\hat{w}) + \frac{1}{n} tr\left(I_n(\hat{w}) J_n^{-1}(\hat{w})\right)$. [竹内 1976] [Stone 1977]

$\to E[TIC] = E[L_n(\hat{w})] + \frac{1}{n} tr\left(IJ^{-1}\right) + o\left(\frac{1}{n}\right)$

$$= L(w_0) + \frac{1}{2n} tr(IJ^{-1}) + o(1/n) \quad (\text{Thm3})$$

$$= E[L(\hat{w})] + o(1/n)$$

TICは$\hat{w}$についての汎化損失を推測　※AICも同様.

- $\underbrace{(L(\hat{w}) - L(w_0))}_{= \frac{1}{2n}\|\xi_n\|^2 + o_p(1/n)} + \underbrace{(TIC - \overbrace{L_n(w_0)}^{L_n(\hat{w}) + \frac{1}{2n}\|\xi_n\|^2 + o_p(1/n)})}_{= \frac{1}{n} tr(IJ^{-1}) - \frac{1}{2n}\|\xi_n\|^2 + o_p(1/n)} = \frac{1}{n} tr(IJ^{-1}) + o_p(1)$

← nは誤植.

- $q(x)$ が $p(x|w)$ で実現可能なとき、$tr(IJ^{-1}) = d$ に留意し、

$$AIC := L_n(\hat{w}) + d/n$$ で定める.

$$
\begin{aligned}
E[AIC] &= E[L_n(\hat{w})] + d/n \\
&= E[L(\hat{w})] - d/n + o(1/n) + d/n \qquad \text{(Thm 5)} \\
&= E[L(\hat{w})] + o(1/n)
\end{aligned}
$$

(E[TIC] と同じ).

例9

$p_0(x) = N(0, 1) \quad (d=0)$

$\rightarrow \log p_0(x) = -\frac{x^2}{2} - \frac{1}{2}\log(2\pi)$

$L_n = -\frac{1}{n}\sum_i \left(-\frac{x_i^2}{2} - \frac{1}{2}\log(2\pi)\right)$

$= \frac{1}{2n}\sum_i \left(x_i^2 + \log 2\pi\right)$

$p_1(x|m) = N(m, 1) \quad (d = 1)$

$\rightarrow \log p_1(x|m) = -\frac{1}{2}(x-m)^2 - \frac{1}{2}\log(2\pi)$

$L_n(m) = \frac{1}{2n}\sum_i \left((x_i - m)^2 - \frac{1}{2}\log 2\pi\right)$.

(i) AIC を用いるとき.

$p_0 \rightarrow \frac{1}{2n}\sum_i \left(x_i^2 + \log 2\pi\right) + \frac{0}{n}$.

$p_1 \rightarrow \min_m \frac{1}{2n}\sum_i \left((x_i - m)^2 - \frac{1}{2}\log 2\pi\right) + \frac{1}{n}$.

(ii) BIC → $0/n$, $1/n$ が変わるだけ.

p_0 の下の方は,

$$\left(\frac{1}{\sqrt{n}}\sum_i x_i\right)^2 < 2 \iff \left(\sum_i x_i\right)^2 < 2n$$

の期待値をとり,

$$E\left[\left(\sum_i x_i\right)^2\right] < 2n \iff n(m_0^2 + 1) + n(n-1)m_0^2 < 2n$$

$$\iff n + n^2 m_0^2 < 2n$$

$$\iff n m_0^2 < 1.$$

($E[x_i^2] = E[x_i]^2 + V[x_i] = m_0^2 + 1$.
$E[x_i x_j] = E[x_i]E[x_j] = m_0^2$)

ということ. BIC, 検定も同じ.

注意31

感度について。

注意32

$\xi_n \rightsquigarrow N(0, I_d)$ のとき $\|\xi_n\|^2 \rightsquigarrow \chi^2_d$.

注意33

(1) モデル選択の一致性について.
(2) 一致性と汎化損失の最小化.

§3.7 質問と回答

→ Chap.4 の理論を用いよ.

$$V_n = \sum_i \left\{ E_w[(\log p(X_i|w))^2] - E_w[\log p(X_i|w)]^2 \right\}$$

: 汎関数分散. WAIC において中心的な役割を果たす量.