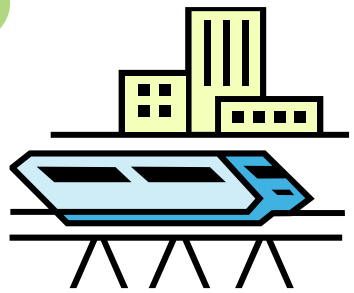
ベイズ統計入門　(10)

目標　　一般理論

東京工業大学
渡辺澄夫

旅の地図

(1) ベイズ統計の定義
(2) 密度と条件つき密度
(3) 混合正規分布＋ギブスサンプラー
(4) 神経回路網＋ランジュバン方程式
(5) 真とモデルの関係
(6) 正則モデルの漸近理論
(7) AIC と BIC
(8) ハイパーパラメータ最適化
(9) 一般モデルの漸近理論
(10) 一般モデルの漸近理論
(11) 一般モデルの選択
(12) 条件つき独立 高次元
(13) 階層ベイズ
(14) 相転移
(15) まとめ



1　　自由エネルギー

事後分布の漸近挙動

以上で述べてきたことを統合することにより、事後分布の漸近挙動を次のように導出することができます。ここが学習理論の核心部分です。

$$\exp(-nK_n(w))\,\varphi(w)\,dw$$

$$= \exp(- nu^{2k} + n^{1/2}u^k\xi_n(u))\,\varphi(g(u))|g'(u)|\,du$$

$$= \int dt\,\delta(t-nu^{2k})\,u^h\,b(u)\exp(- t + t^{1/2}\xi_n(u))\,du$$

$$\rightarrow \quad \frac{(\log n)^{m-1}}{n^{\lambda}} \int dt\,t^{\lambda-1}\exp(- t + t^{1/2}\xi(u))\,D(u)\,du$$

☆　特異点解消定理を用いてパラメータの空間を w から u に移行することにより超関数と経験過程の漸近挙動を、どちらも数学的に扱うことが可能になりました。

☆　パラメータ空間を複数の座標に分割したとき、各座標ごとに上記の漸近挙動が得られます。λが一番小さく、m　が一番大きい座標が事後分布の主要項を与えます。

事後分布を二つに分けることができた

事後分布が定義する測度

$$\exp(-nK_n(w))\,\varphi(w)\,dw$$

$$= \frac{(\log n)^{m-1}}{n^{\lambda}} \int dt\, t^{\lambda-1} \exp\left(- t + t^{1/2}\xi_n(u) \right) D(u)du$$

$n \to \infty$で
零になる速さ

確率的に揺らいでいる部分

☆　事後分布の挙動が解明できたので、後は計算の問題になります。

自由エネルギーの漸近挙動の導出

自由エネルギー

$$F = nL_n - \log \int \exp(- nK_n(w)) \varphi(w)\, dw$$

に事後分布の漸近挙動

$$\exp(-nK_n(w))\, \varphi(w)\, dw$$

$$= \frac{(\log n)^{m-1}}{n^{\lambda}} \int dt\, t^{\lambda-1} \exp(- t + t^{1/2}\xi_n(u))\, D(u)du$$

を代入すればよい。∫ は局所座標の和の積分で書けるが、
一番大きなオーダーのところだけ残る。次の定理が得られた。

定理1 : 自由エネルギーの漸近挙動

定理.1． f(X,w) が相対的に有限な分散を持つとする。
自由エネルギーの漸近挙動は

$$F = n L_n + \lambda \log n - (m-1) \log\log n + O_p(1).$$

主定理1は、自由エネルギーの理論的な挙動を示したものである。真の分布が学習モデルに対して正則であれば、λ＝d/2, m=1であるが、一般にはそうではない。なお、一般に(λ,m) は真の分布に依存するので，真の分布がわからない場合には、主定理1を直接に使って自由エネルギーを計算することはできないが、この性質を利用して F を近似計算する方法（WBIC)や、真の分布の推定とモデルの選択を同時に行なうアルゴリズム(sBIC, Drton & Plummer 2017, Journal of Royal Statistical Society, series B)が提案されている。若い研究者のみなさんの新しい発想を期待します。

自由エネルギーの漸近挙動の例

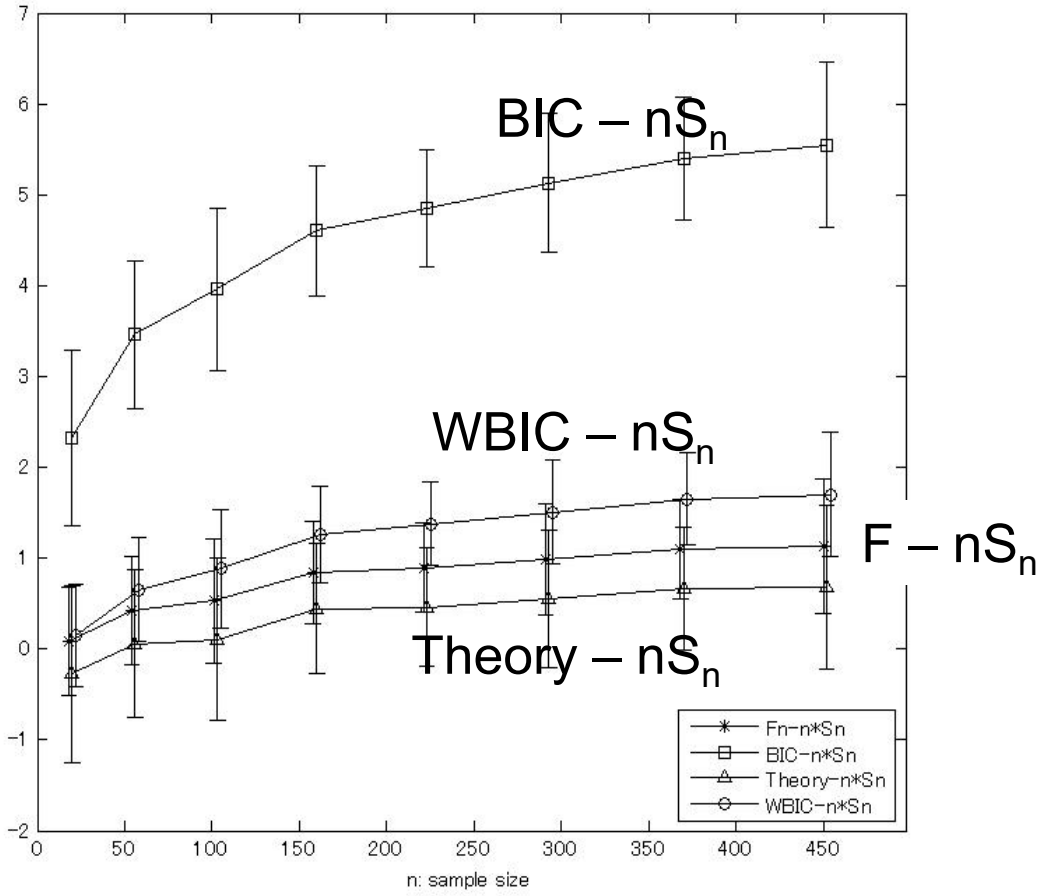
統計モデル　$p(y|x,a,b) = (1/2\pi)^{1/2} \exp(-(1/2)(y-a\tanh(bx))^2)$
事前分布　　$\varphi(a,b) \propto 1$
真の分布　　$q(y|x)=p(y|x,0,0)$, X の分布は[-2,2]上の一様分布

この場合　λ=1, m=2.

n =20, …, 450 まで

BIC
WBIC
F
Theory

を比較した



２　　汎化損失

繰り込まれた事後分布の定義

事後分布は $\exp(-nK_n(w))\,\varphi(w)\,dw$

$= \quad (n\text{の関数}) \times \int dt\, t^{\lambda-1} \exp(- t + t^{1/2}\xi_n(u))\, D(u)du$

定義：繰り込まれた事後分布による平均

$$< \quad > = \frac{\int dt \int du\, D(u) (\quad) t^{\lambda-1}\exp(-t+t^{1/2}\xi(u))}{\int dt \int duD(u)\, t^{\lambda-1}\exp(-t+t^{1/2}\xi(u))}$$

補題 $< t > = \lambda+(1/2)<t^{1/2}\xi(u) >$

(分子を t で部分積分すると得られる)

スケーリング関係の導出

変数 t による積分には $\delta(t-nu^{2k})$ があるからスケーリング関係

$$f(x,g(u)) = a(x,u)\, u^{k} = a(x,u)\, (t/n)^{1/2}$$

$$u^{2k} = t/n$$

補題　任意の $s \geqq 0$ で、次の法則収束が成立

$$n^{s/2}\, E_w [\ f(x,w)^s\ ] \rightarrow < t^{s/2}\, a(x,u)^s >$$

$$n^{s}\, E_w [\ K(w)^s\ ] \rightarrow < t^{s} >$$

事後分布による f(x,w) の平均は繰り込まれた事後分布による平均で表すことができる。

特異ゆらぎの定義

定義. 特異ゆらぎを繰り込まれた分布で定義する

$$\mathrm{Fluc}(\xi) = E_x[\ \langle t\, a(X,u)^2 \rangle - \langle t^{1/2} a(X,u) \rangle^2\ ]$$

$$2\nu \equiv E_\xi[\ \ \mathrm{Fluc}(\xi)\ ]$$

定義. 汎関数分散 を事後分布で定義する

$$V = (1/n)\ \Sigma\ \ V_w[\ \log p(X_i|w)]$$

記号　$f(x,w) = \log(p_0(x)/p(x|w))$

スケール関係　$f(x,g(u)) = a(x,u)\ (t/n)^{1/2}$

から次の補題が示される。

特異ゆらぎと汎関数分散の関係

補題　f(X,w) が相対的に有限な分散を持つとき、

$$\lim_{n \to \infty} E[\, n\, V\,] = 2\nu$$

$$E_\xi[\, \langle t^{1/2}\xi(u) \rangle\,] = 2\nu$$

独立性の仮定①から $E[G_{n-1}]=E[C_n]$ が成り立つので

$$L_0+E\langle t\rangle/n - \nu/n = L_0 + E\langle t/n - t^{1/2}\xi(u)/n\rangle + \nu/n + o(1/n)$$

従って　$E\langle t^{1/2}\xi(u)/n\rangle = 2\nu/n + o(1/n)$　（証明終）

☆　独立性がないときは①が成り立ちませんが、条件つき独立ならば、$\xi(u)$が正規確率過程であることから $E_\xi[\, \langle t^{1/2}\xi(u)\rangle\,] = E_\xi[\ \mathrm{Fluc}(\xi)\,]$ を示すことができます。

☆　真の分布が統計モデルで実現可能であり、かつ正則であれば $\nu = d/2$ です。正則であって実現可能でない場合には　$\nu = \mathrm{tr}(IJ^{-1})/2$　になります。正則でない場合には ν の値はまだ分かっておりません。

定理2：汎化誤差の漸近挙動

主定理.2.　f(X,w) が相対的に有限な分散を持つとき

$$G_n = L_0 + (1/2n)\{ 2\lambda + \langle t^{1/2}\xi_n(u)\rangle - \mathrm{Fluc}(\xi_n)\} + o_p(1/n),$$
$$T_n = L_n + (1/2n)\{ 2\lambda - \langle t^{1/2}\xi_n(u)\rangle - \mathrm{Fluc}(\xi_n)\} + o_p(1/n),$$
$$C_n = L_n + (1/2n)\{ 2\lambda - \langle t^{1/2}\xi_n(u)\rangle + \mathrm{Fluc}(\xi_n)\} + o_p(1/n),$$
$$W_n = L_n + (1/2n)\{ 2\lambda - \langle t^{1/2}\xi_n(u)\rangle + \mathrm{Fluc}(\xi_n)\} + o_p(1/n).$$

平均値については

$$E[G_n] = L_0 + \lambda / n + o(1/n),$$
$$E[T_n] = L_0 + (\lambda - 2\nu) / n + o(1/n),$$
$$E[C_n] = L_0 + \lambda / n + o(1/n),$$
$$E[W_n] = L_0 + \lambda / n + o(1/n),$$

交差損失が汎化損失と漸近的に同じ平均を持つ事は定義からわかりますが
確率変数としてどのくらい異なるかは理論によって初めてわかります。

(証明1) 汎化損失の解析

キュムラント母関数　$G(\alpha) \equiv \alpha L_0 - E_x[\log E_w[\exp(-\alpha f(X,w))]]$

を用いると　$G_n = G(1) = G(0) + G'(0) + G''(0)/2 + O_p(1/n^{3/2})$

それぞれの平均値が計算できる。

$G(0) = 0$

$G'(0) = L_0 + E_x E_w[f(X,w)] = L_0 + E_w[K(w)]$

$= L_0 + \langle t \rangle / n = \lambda + (1/2n)\langle t^{1/2}\xi(u) \rangle$

$G''(0) = -E_x V_w[f(x,w)] = -(\mathrm{Fluc}(\xi))/n + o_p(1/n)$

従って

$G_n = L_0 + \lambda + (1/2n)\langle t^{1/2}\xi(u)\rangle - (1/2n)(\mathrm{Fluc}(\xi)) + o_p(1/n)$

$E[G_n] = L_0 + \lambda/n + O(1/n^{3/2})$

(証明2)学習損失と交差損失

キュムラント母関数　$T(\alpha)= \alpha L_n - (1/n)\,\Sigma \log E_w[\exp(-\alpha f(X_i,w))]$

を用いると

$$T_n = T(1) = T(0) + T'(0) + T''(0)/2 + O_p(1/n^{3/2})$$

$$C_n = -T(-1) = -T(0) + T'(0) - T''(0)/2 + O_p(1/n^{3/2})$$

$$W_n = -T(-1) = -T(0) + T'(0) - T''(0)/2 + O_p(1/n^{3/2})$$

ここで

$$T(0) = 0$$

$$T'(0) = L_n + E_w[(1/n)\Sigma f(X_i,w)] = L_n + E_w[K_n(w)]$$

$$= L_n + \langle t/n - t^{1/2}\xi(u)/n \rangle = L_n + \lambda - (1/2n)\langle t^{1/2}\xi(u) \rangle$$

$$T''(0) = -(1/n)\Sigma V_w[f(X_i,w)] = -Fluc(\xi)/n + o_p(1/n)$$

これを代入すれば定理前半が得られる。平均を取ると後半が得られる。

（証明終）

何がわかったか

自由エネルギー

$$F = n L_n + \lambda \log n - (m-1) \log\log n + O_p(1).$$

汎化誤差

$$G_n = L_0 + (1/2n) \{ 2\lambda + \langle t^{1/2} \xi_n(u) \rangle - \mathrm{Fluc}(\xi_n) \} + o_p(1/n),$$
$$C_n = L_n + (1/2n) \{ 2\lambda - \langle t^{1/2} \xi_n(u) \rangle + \mathrm{Fluc}(\xi_n) \} + o_p(1/n),$$
$$W_n = L_n + (1/2n) \{ 2\lambda - \langle t^{1/2} \xi_n(u) \rangle + \mathrm{Fluc}(\xi_n) \} + o_p(1/n).$$

$\mathbf{E} \langle t^{1/2} \xi_n(u) \rangle = \mathbf{E}\, \mathrm{Fluc}(\xi_n)$ が成り立つ。

３　実対数閾値

実対数閾値の性質(1)

定理. 実対数閾値を λ とすると(-λ) は下記のゼータ関数の最大極である。
（このゼータ関数の極は全て実数であることが分かっている）。

ゼータ関数　　$\zeta(z) = \int K(w)^z \varphi(w)\, dw$

メリン変換

状態密度関数　　$v(t) = \int \delta(t-K(w))\, \varphi(w)\, dw$

ラプラス変換

分配関数　　$Z(n) = \int \exp(-nK(w))\, \varphi(w)\, dw$

☆　メリン変換とラプラス変換は実質的な逆変換を持つので上記の3つの関数は、どれかひとつ分かれば残りも分かる。

実対数閾値の性質(2)

定理. 実対数閾値を求めれば汎化損失の漸近挙動がわかる。

$$E[G_n] = S + \lambda/n + o(1/n).$$

実対数閾値の性質

1. $K_1(w) \leqq K_2(w)$ かつ $\varphi_1(w) \leqq \varphi_2(w)$なら $\lambda_1 \leqq \lambda_2$.

2. $K(w)=K_1(w_1)+K_2(w_2)$ かつ$\varphi(w)=\varphi_1(w_1)\varphi_2(w_2)$なら$\lambda= \lambda_1+\lambda_2$.

3. $\{f_k(w)\}$から生成されるイデアルと$\{g_k(w)\}$から生成されるイデアルが等しければ $K(w)=\Sigma_k f_k(w)^2$, $H(w)=\Sigma_k g_k(w)^2$の実対数閾値も等しい。

☆　いろいろな確率モデルの実対数閾値が求められている。与えられた統計モデルと事前分布に対して実対数閾値を求めるという研究課題がある。

汎化誤差の例

縮小ランク回帰　$X \in R^M$, $Y \in R^N$, $B \in R^{HM}$, $A \in R^{NH}$

統計モデル　$p(y|x,A,B) = (1/2\pi)^{N/2} \exp(-(1/2)|| y-BAx||^2)$

事前分布　$\varphi(a,s) \propto 1$

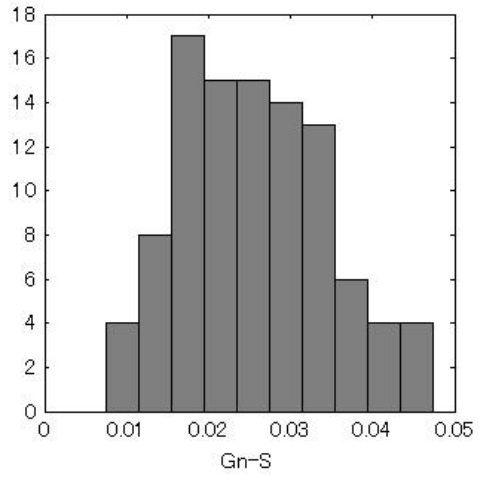
真の分布　$q(y|x)=p(y|x,A_0,B_0)$, X の分布は標準正規分布の直積

$M+N+H+H_0$ が偶数のとき $\lambda = (1/8)(2(H+H_0)(M+N)-(M-N)^2-(H+H_0)^2)$

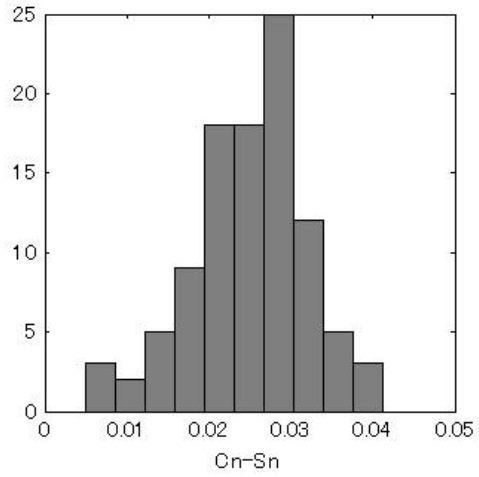
$M+N+H+H_0$ が奇数のとき $\lambda = (1/8)(2(H+H_0)(M+N)-(M-N)^2-(H+H_0)^2+1)$

実験 n=500,　M=N=H=5, H_0=3,　λ / n =0.024　（理論は M.Aoyagi, 2006）

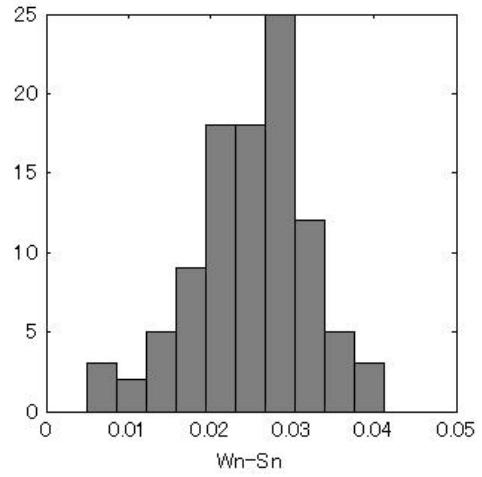
$E[G_n]$=0.0256　　$E[C_n]$=0.0247　　$E[W_n]$=0.0247



汎化損失



交差損失



WAIC