数理解析研究所講究録
1532巻 2007年 144-147

学習理論における汎化誤差の漸近挙動について

渡辺澄夫
東京工業大学　精密工学研究所
〒226-8503　横浜市緑区長津田4259 メールボックス R2-5

1　はじめに

N 次元ユークリッド空間 $\mathbf{R}^N$ 上の確率分布 $q(x)dx$ と $p(x|w)dx$ を考える。ここで dx はルベーグ測度で、w は d 次元ユークリッド空間の元であり、$p(x|w)dx$ は w をパラメータとして持つ確率分布である。$p(x|w)dx$ のことを w が与えられたもとでの x の分布と呼ぶ。$(\Omega, \mathcal{B}, P)$ を確率空間として、$X_1, X_2, ..., X_n$ を $\mathbf{R}^N$ に値を取る確率変数で、独立に $q(x)dx$ に従うものとする。確率変数の集合

$$D_n = \{X_1, X_2, ..., X_n\}$$

を学習データという。d 次元ユークリッド空間 $\mathbf{R}^d$ 上に確率分布 $\varphi(w)dw$ が与えられたとき、

$$p(x|D_n) = \frac{\int p(x|w) \prod_{i=1}^{n} p(X_i|w)\varphi(w)dw}{\int \prod_{i=1}^{n} p(X_i|w)\varphi(w)dw}$$

と定義して、確率分布 $p(x|D_n)dx$ のことを学習結果という。学習データを発生している確率分布を真の分布と呼び、$p(x|w)dx$ を学習モデルという。真の分布から学習結果までの相対エントロピー

$$G_n = \int q(x) \log \frac{q(x)}{p(x|D_n)} dx$$

を汎化誤差という。これは、学習データ D_n を用いて学習モデルが、どのくらい真の分布を正しく推測したかを表す量である。学習理論における数学的課題とは、3つの確率分布 $q(x)dx, p(x|w)dx, \varphi(w)dw$ が与えられたとき、確率変数 G_n の挙動を解明することである。特に n が無限大に近づくとき、G_n が従う確率分布とその平均値の漸近挙動を明らかにすることが問題である。この論文では、一定の条件のもとで nG_n がある確率変数に法則収束することと、$E[nG_n]$ が定数に収束することを示す。

注意. 本論文では、$p(x|w)dx$ と $q(x)dx$ が同じサポートを持つ場合だけを考える。そのサポート上で関数 $f(x,w)$ を

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

とおいて

$$K_n(w) = \sum_{i=1}^{n} f(X_i, w)$$

と書くことにすると、汎化誤差は

$$G_n = E_X\left[-\log \frac{\int \exp(-f(X,w)) \exp(-K_n(w))\varphi(w)dw}{\int \exp(-K_n(w))\varphi(w)dw}\right]$$

と表される。$E_X[\cdot]$ は X についての積分 $\int q(x)dx$ を表す。

2 定理

定理が成り立つための条件を述べる。

条件 (1). $\varphi(w)$ はコンパクトサポートである。集合

$$W_0 = \{w \in \mathbf{R}^d; \int q(x)f(x,w)dx = 0,\ \varphi(w) > 0\}$$

は空集合ではない。

条件 (2). W_0 の任意の元 w_0 に対して w_0 を含む十分小さい近傍 $\mathcal{U}$ を取れば、ある d 次元実多様体 $\mathcal{M}$ と $\mathcal{M}$ から $\mathcal{U}$ へのプロパーな解析写像 $w = g(u)$ と $\mathcal{U}$ の座標 U 毎にある関数 $a(x,u)$ および $b(u)$ が存在して

$$\begin{aligned} f(x,g(u)) &= u^k a(x,u) \\ E_X[a(X,u)] &= u^k \\ \varphi(g(u))du &= u^h b(u)du \end{aligned}$$

が成り立つようにできる。ここで k,h は d 次元の多重指数であり、$a(x,u)$ は、(x,u) の可測関数で、U に含まれる任意のコンパクト集合 K について

$$M(x) = \sup_{u \in K} |a(x,u)|$$

とおくとき、

$$E[M(X)\exp(2M(X))] < \infty$$

が成り立つ。

条件 (3). 任意の有限の j について、U 上の関数

$$\psi_n(u) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \partial_u^j \{a(X_i,u) - u^k\}$$

は、正規確率過程 $\psi(u)$ に法則収束する。ここで ∂_u^j は、u の高々 j 次までの偏微分を表している。

定理. 汎化誤差の n 倍 nG_n はある確率変数 G^* に法則収束し、

$$\lim_{n\to\infty} E[nG_n] = E[G^*] = \lambda_1$$

が成り立つ。ここで λ_1 は、多重指数 $h = (h_1,...,h_d)$, $k = (k_1,...,k_d)$ から定まる値

$$\lambda_1 = \min_{i=1}^{d} \left(\frac{k_i+1}{h_i}\right)$$

である（$h_i = 0$ は min の候補から外す）。

3 証明

$0 \le \beta \le 1$ について関数 $g(\beta)$ を

$$g(\beta) = E_X\left[-\log \frac{\int \exp(-\beta\, f(X,w))\ \exp(-K_n(w))\ \varphi(w)\ dw}{\int \exp(-K_n(w))\ \varphi(w)\ dw}\right]$$

と定義すると $g(0)=0, g(1)=G_n$ が成り立つ。従って $ng(1)$ の法則収束と $E[ng(1)]$ の収束を示せばよい。

$$\begin{aligned} n\,g(1) &= \int_0^1 n\,g'(\beta)\,d\beta \\ &= \int_0^1 d\beta\; E_X\Big[\frac{\int n\,f(X,w)\,\exp(-\beta\,f(X,w)-K_n(w))\,\varphi(w)\,dw}{\int \exp(-\beta\,f(X,w)-K_n(w))\,\varphi(w)\,dw}\Big] \end{aligned}$$

そこで、この分子と分母の確率変数を

$$\begin{aligned} A_n(\beta) &= \int nf(X,w)\,\exp(-\beta\,f(X,w)-K_n(w))\,\varphi(w)\,dw \\ B_n(\beta) &= \int \exp(-\beta\,f(X,w)-K_n(w))\,\varphi(w)\,dw \end{aligned}$$

とおく。このとき $A_n(\beta)=-nB_n(\beta)'$ である。

$$ng(1)=-\int_0^1 d\beta\; E_X\Big[\frac{nB_n(\beta)'}{B_n(\beta)}\Big]$$

が成り立つ。$\varphi(w)$ のサポートがコンパクトなので、積分 dw は、局所ごとの積分の有限和で書くことができる。局所ごとの変換 $w=g(u)$ を用いて、$[0,1]^d$ 上の積分に書き換えることができ、

$$K_n(g(u))=n\,u^{2k}+\sqrt{n}\,u^k\,\psi_n(u)$$

であるから

$$\begin{aligned} A_n(\beta) &= \sum_\alpha \int_{[0,1]^d} nu^k a(X,u)\,e^{-\beta\,u^k a(X,u)-n\,u^{2k}-\sqrt{n}\,u^k\,\psi_n(u)}\,b(u)\,u^h\,du \\ B_n(\beta) &= \sum_\alpha \int_{[0,1]^d} e^{-\beta\,u^k a(X,u)-n\,u^{2k}-\sqrt{n}\,u^k\,\psi_n(u)}\,b(u)\,u^h\,du \end{aligned}$$

と書くことができる。ここで $a(x,u),b(u),k,h$ はすべて局所座標 α に依存するが、標記が複雑になるため依存を表す添え字は省略して書いてある。中間値の定理から、ある $0<\beta^*<\beta$ が存在して

$$\begin{aligned} \frac{A_n(\beta)}{B_n(\beta)} &= -n\frac{B_n(\beta)'}{B_n(\beta)} \\ &= -n\frac{B_n(0)'}{B_n(0)}-\beta n\Big(\frac{B_n(0)'}{B_n(0)}\Big)'-\frac{\beta^2 n}{2}\Big(\frac{B_n(\beta^*)'}{B_n(\beta^*)}\Big)'' \\ &= -n\frac{B_n(0)'}{B_n(0)}-\beta n\frac{B_n(0)''}{B_n(0)}+\beta n\Big(\frac{B_n(0)'}{B_n(0)}\Big)^2-\frac{\beta^2 n}{2}\Big(\frac{B_n(\beta^*)'}{B_n(\beta^*)}\Big)'' \end{aligned}$$

である。積分 $\int d\beta E_X$ を実行すると、$E_X[a(X,u)]=u^k$ より、最初の2項の和は0である。従って、

$$ng(1)=\frac{1}{2}E_X\Big[\Big(\frac{\sqrt{n}B_n(0)'}{B_n(0)}\Big)^2\Big]-\int_0^1 d\beta\frac{\beta^2 n}{2}E_X\Big[\Big(\frac{B_n(\beta^*)'}{B_n(\beta^*)}\Big)''\Big]$$

である。上の式の第2項は0に法則収束する。実際

$$\Big(\frac{B_n(\beta^*)'}{B_n(\beta^*)}\Big)''=\frac{B_n(\beta^*)'''}{B_n(\beta^*)}-3\frac{B_n(\beta^*)'B_n(\beta^*)''}{B_n(\beta^*)^2}+2\frac{(B_n(\beta^*)')^3}{B_n(\beta^*)^3}$$

であるから条件 (2) を用いて、

$$\Big|\frac{\beta^2 n}{2}E_X\Big[\Big(\frac{B_n(\beta^*)'}{B_n(\beta^*)}\Big)''\Big]\Big|\le\frac{Const.}{n^{1/2}}\exp(Const.\sup_u|\psi_n(u)|^2)$$

である。次に

$$E_X\Big(\frac{\sqrt{n}B_n(0)'}{B_n(0)}\Big)^2 = E_X\Big(\frac{\sum_\alpha \int du\, a(X,u)\, \sqrt{n}\, u^k\ \ e^{-nu^{2k}+\sqrt{n}u^k\psi_n(u)}\, b(u)u^h}{\sum_\alpha \int du\, e^{-nu^{2k}+\sqrt{n}u^k\psi_n(u)}\, b(u)u^h}\Big)^2$$

について、参考文献 [1] の定理４６の証明と同様に、λ, μ をとる。

$$\begin{aligned} A_n^0 &= \sum_\alpha c_0 \int dt \int dy\, y^\mu\, a(X,y)\, t^\lambda e^{-t+\sqrt{t}\psi_n(0,y)} \\ B_n^0 &= \sum_\alpha c_0 \int dt \int dy\, y^\mu\, t^{\lambda-1} e^{-t+\sqrt{t}\psi_n(0,y)} \end{aligned}$$

と定義し、$a_n = (\log n)^{r-1}/n^\lambda$ とおいて、

$$\begin{aligned} C_n &= (\sqrt{n}B_n(0)' - a_n A_n^0)\log n \\ D_n &= (B_n(0) - a_n B_n^0)\log n \end{aligned}$$

によって C_n, D_n を定義すると

$$\begin{aligned} \Big|E_X\Big(\frac{\sqrt{n}B_n(0)'}{B_n(0)}\Big)^2 - E_X\Big(\frac{A_n^0}{B_n^0}\Big)^2\Big| \le E_X\Big[\Big|\frac{\sqrt{n}B_n(0)'B_n^0 + A_n^0 B_n(0)}{B_n(0)B_n^0}\Big| \\ \times\Big(\frac{1}{\log n}\Big|\frac{C_n}{B_n(0)}\Big| + \frac{1}{\log n}\Big|\frac{A_n^0 D_n}{B_n^0 B_n(0)}\Big|\Big)\Big] \end{aligned}$$

が成り立つ。参考文献 [1] の評価式 (4.21) と定理 45 を適用すると、$C_n/B_n(0)$ および $A_n^0 D_n/B_n(0)B_n^0$ は $\exp((\alpha/2)\sup_w \psi_n(u)^2)$ および $\sup_w(\partial\psi)_n(u)^2$ 以下であるから 0 に法則収束する。$E_X\Big(\frac{\sqrt{n}B_n(0)'}{B_n(0)}\Big)^2$ は ψ_n の連続関数であるから、ある確率変数に法則収束する。従って $ng(1)$ も同じ確率変数に法則収束する。最後に $ng(1)$ が漸近一様可積分であることを示す。$0 \le \beta \le 1$ のとき $A_n(\beta)/B_n(\beta) \le A_n(0)/B_n(0)$ が成り立つので、

$$ng(1) \le E_X[\frac{A_n(0)}{B_n(0)}] = Y(\gamma)'|_{\gamma=1}$$

である。ここで

$$Y(\gamma) = -\log \sum_\alpha \int_{[0,1]^d} e^{-\gamma n u^{2k}+\sqrt{n}u^k\psi_n(u)} b(u)u^h du$$

は $Y(\gamma)'' \le 0$ を満たすので、任意の $\gamma < 1$ について

$$Y'(1) \le \frac{Y(1)-Y(\gamma)}{1-\gamma}$$

が成り立つが、この式の右辺は、$\sup_u |\psi_n(u)|^2$ の定数倍でバウンドされる。

参考文献

[1] 渡辺澄夫, "代数幾何と学習理論," 共立出版, 2006.

[2] S.Watanabe,"Algebraic geometry of singular learning machines and symmetry of generalization and training errors," Neurocomputing, Vol.67,pp.198-213,2005.