ニューラルネットワークの汎化誤差
– 汎化誤差解析の一般論

$x \in \mathcal{X} \longmapsto y \in \mathcal{Y}$ の入出力関係を学習したい.

予: モデル ($\mathcal{X} \to \mathbb{R}$ なる関数からなる集合)

例: ニューラルネットワーク

再生核ヒルベルト空間の単位球

目標: 予の中から汎化誤差の小さい良い元を推定したい.

$l : \mathcal{Y} \times \mathbb{R} \to \mathbb{R}$ を損失関数とする.

e.g. 二乗損失: $l(y, f(x)) = (y - f(x))^2$

ロジスティック損失: $l(y, f(x)) = \log(1 + \exp(-y f(x)))$

$L(f) = E_{X,Y}[l(Y, f(X))]$ 汎化誤差, 期待損失

$\hat{L}(f) = \frac{1}{n}\sum_{i=1}^{n} l(y_i, f(x_i))$ 訓練誤差, 経験損失

$Pf = E_{X,Y}[f(X,Y)]$, $P_n f = \frac{1}{n}\sum_{i=1}^{n} f(x_i, y_i)$ とおく.

($L(f) = P(l \circ f)$, $\hat{L}(f) = P_n(l \circ f)$)

$f^* := \underset{f \in \mathcal{F}}{\operatorname{argmin}}\, L(f)$ とすると

推定量 $\hat{f}$ に対し $L(\hat{f}) - L(f^*)$ を Excess risk (残余誤差) と言う

これを汎化誤差ということも多い

※ $L(\hat{f}) - \hat{L}(\hat{f})$ のことを汎化誤差ということも多いが

より正確には汎化ギャップと言うがさておこう.

⟶ 基本的考え方.

$$L(\hat{f}) - L(f^*) = \underbrace{L(\hat{f}) - \hat{L}(\hat{f})}_{\text{汎化ギャップ}} + \underbrace{\hat{L}(\hat{f}) - \hat{L}(f^*)}_{\hat{f}\text{が訓練誤差を最小化していれば} \le 0} + \underbrace{\hat{L}(f^*) - L(f^*)}_{\text{大数の法則 } O_p(\frac{1}{\sqrt{n}})}$$

よって, 基本的に汎化ギャップをおさえれば汎化誤差をおさえられる.

$\hat{f}$ が訓練誤差を最小化しているなら $\hat{L}(\hat{f}) - \hat{L}(f^*) \le 0$ であり, このとき

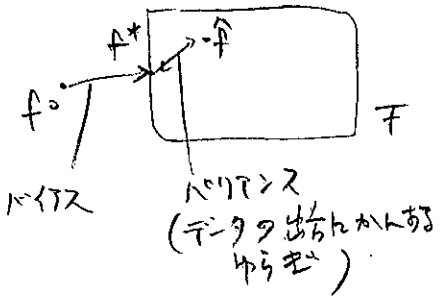
$$0 \le L(\hat{f}) - L(f^*) \le L(\hat{f}) - L(f^*) - (\hat{L}(\hat{f}) - L(f^*))$$
$$= (P - P_n)(l \circ \hat{f} - l \circ f^*) \quad < \quad ?$$

- バイアス-バリアンスのトレードオフ.

$f^\circ := \operatorname{argmin} L(f)$
f:可測関数

f^*との汎化誤差の差:

$$L(\hat{f}) - L(f^\circ) = \underbrace{L(\hat{f}) - L(f^*)}_{\text{バリアンス}} + \underbrace{L(f^*) - L(f^\circ)}_{\text{(モデル)バイアス}}$$



$\mathcal{F}$が大きい → バイアス小
バリアンス大

$\mathcal{F}$が小さい → バイアス大
バリアンス小

適度な大きさの$\mathcal{F}$を用いて、バイアスとバリアンスのバランスを取る必要がある.

- バリアンスのバウンド

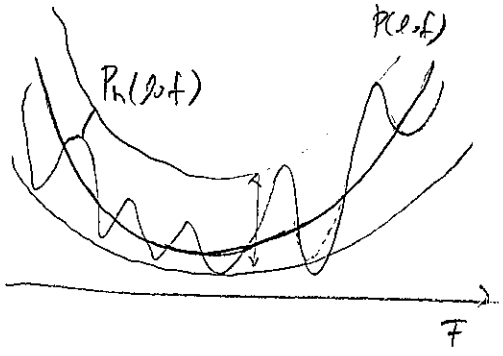
とりあえずバイアスは考えず、バリアンスをおさえる.

$$L(\hat{f}) - L(f^*) \leq (P - P_n)(l_{\hat{f}} - l_{f^*})$$

この右辺を

$$\sup_{f \in \mathcal{F}} (P - P_n)(l_f - l_{f^*})$$

でおさえる.



(注) $\hat{f}$はデータに依存するので、単純な大数の法則ではダメ.

$z = (x, y)$で$g(z) = g(x, y)$とおく

$g := l_f - l_{f^*}$, $\mathcal{G} := \{g \mid l_f - l_{f^*},\ f \in \mathcal{F}\}$ とすれば

$\sup_{g \in \mathcal{G}} (P - P_n) g$ を抑える問題になる.

※ ここで、 $\sup_{g \in \mathcal{G}} P_n g$ と $\sup_{g \in \mathcal{G}} \frac{1}{n} \sum_{i=1}^n \xi_i g(z_i)$ ($\{0,1\}$-value)

らは以後可測であると仮定する. 可測性がない場合は以下の議論は成り立たない (Fubiniの定理が成り立たない. $E_X E_\xi [\cdot]^* \neq E_\xi E_X [\cdot]^*$)

- van der Vaart & Wellner: Weak convergence and Empirical processes (1996) Sec.2-3
- Dudley "Uniform Central Limit theorems". Suslin image admissible.

Rademacher complexity

仮定：・ℓ は 1-リプシッツ連続とする $(\ell(y, f(x)) - \ell(y, f'(x)) \leq |f(x) - f'(x)|$ $(\forall y))$
・$\sup_z |g(z)| \leq M$ $(\forall g \in G)$

Def

σ_i $(i=1, \ldots, n)$ を $P(\sigma_i = 1) = P(\sigma_i = -1) = \frac{1}{2}$ なる i.i.d. sequence とする．
(Rademacher 変数)

$$\hat{R}_n(G) := E_{\sigma_i}\left[\sup_{g \in G} \frac{1}{n}\sum_{i=1}^{n} \sigma_i g(z_i) \,\middle|\, z_1, \ldots, z_n \right]$$

(経験 Rademacher 複雑さ)

$$R_n(G) := E_{z_i}\left[\hat{R}_n(G) \right]$$

(Rademacher 複雑さ) //

Thm (対称化)

$$E_{z_i}\left[\sup_{g \in G} (P - P_n) g \right] \leq 2 R_n(G)$$

//

(直接 $\sup(P-P_n)g$ を評価するより $R_n(G)$ の方が評価しやすい.)

Proof

z_i' は z_i と同じ i.i.d sequence

$$左辺 = E_{z_i}\left[\sup_{g \in G} \left\{ E_{z_i'}\left[\frac{1}{n}\sum_{i=1}^{n} g(z_i') \right] - \frac{1}{n}\sum_{i=1}^{n} g(z_i) \right\} \right]$$

$$\leq E_{z_i, z_i'}\left[\sup_{g \in G} \frac{1}{n}\sum_{i=1}^{n} (g(z_i') - g(z_i)) \right] \quad \text{(Jensen の不等式)}$$

$$= E_{\sigma_i}\left[E_{z_i, z_i'}\left[\sup_{g \in G} \frac{1}{n}\sum_{i=1}^{n} \sigma_i (g(z_i') - g(z_i)) \right] \right]$$

(z_i' と z_i は同じ分布)

σ_i と $-\sigma_i$ は同じ分布

$$\leq E_{\sigma_i} E_{z_i'}\left[\sup_{g \in G} \frac{1}{n}\sum_{i=1}^{n} \sigma_i g(z_i') \right] + E_{\sigma_i} E_{z_i}\left[\sup_{g \in G} \frac{1}{n}\sum \sigma_i g(z_i) \right]$$

$$= 2 R_n(G)$$

($\because$ Fubini の定理) ← 有界性より可積分性は O.K.
可測性がここで必要 //

Thm

$$P\left(\sup_{g \in G} (P - P_n) g \geq 2 R_n(G) + M\sqrt{\frac{2\log(1/\delta)}{n}} \right) \leq \delta \quad (\forall \delta > 0)$$

Proof

Lem. (McDiamid の不等式)

$X_1, \ldots, X_n$ を $\mathcal{X}$ 上の独立な (同一とは限らない) 確率変数とする．

$f: \mathcal{X}^n \to \mathbb{R}$ を可測関数とし，

$$|f(x_1, \ldots, x_i, \ldots, x_n) - f(x_1, \ldots, x_i', \ldots, x_n)| \leq C_i$$

が $\forall x_1, \ldots, x_n, x_i' \in \mathcal{X}$ に対して成り立つとき，

$$P\left(f(X_1, \ldots, X_n) - E[f(X_1, \ldots, X_n)] \geq \varepsilon \right) \leq \exp\left(-\frac{2\varepsilon^2}{\sum C_i^2} \right) \quad (\forall \varepsilon > 0)$$ //

今 $\|g\|_\infty \le M$ $(\forall g \in G)$ より

$$f(z_1,\dots,z_n) = \sup_{g\in G} \left\{ E[g(z)] - \frac{1}{n}\sum_{i=1}^n g(z_i) \right\}$$

とおく.

$$\left| f(z_1,\dots,z_n) - f(z_1,\dots,z_i',\dots z_n) \right| \le \frac{2}{n} M$$

をみたす. あとは McDiamid の不等式と先の対称化により示される. //

Thm (Rademacher 複雑さの性質)

(1) $G \subset G'$ なら $R_n(G) \le R_n(G')$

(2) $\phi_i : \mathbb{R} \to \mathbb{R}$ が $|\phi_i(x) - \phi_i(y)| \le D|x-y|$ なら

$$E_{z_i,\sigma_i}\left[\sup_{g\in G} \frac{1}{n}\sum_{i=1}^n \sigma_i \phi_i(g(z_i)) \right] \le D R_n(G) \quad \text{(contraction ineq.)}$$

(3) G の凸包 $\mathrm{conv}(G) = \left\{ \sum_{i=1}^m \lambda_i g_i \;\middle|\; \sum_{i=1}^m \lambda_i = 1,\ \lambda_i \ge 0,\ g_i \in G,\ m = 1,2,\dots \right\}$

$$R_n(G) = R_n(\mathrm{conv}(G))$$

(4) $R_n(cG) = |c| R_n(G) \quad (c \in \mathbb{R})$ //

Ex. (1) 線形モデル.

- $F = \{ f(x) = x^T\omega \mid \|\omega\|_2 \le 1 \}$
- $|\ell(y,f(x)) - \ell(y,f'(x))| \le |f(x) - f'(x)|$
- $E[\|x_i\|_2^2] \le I$

なら $R_n(G) \le R_n(F)$ (contraction ineq.)

$$= E_{z_i,\sigma_i}\left[\sup_{\omega:\|\omega\|_2\le 1} \frac{1}{n}\sum_{i=1}^n \sigma_i \omega^T x_i \right]$$

$$\le E_{z_i,\sigma_i}\left[\sup_{\|\omega\|_2\le 1} \left\| \frac{1}{n}\sum_{i=1}^n \sigma_i x_i \right\|_2 \cdot \|\omega\|_2 \right]$$

$$\le \sqrt{E_{z_i,\sigma_i}\left[\frac{1}{n^2}\sum_{i,j} \sigma_i\sigma_j (x_i^T x_j) \right]}$$

$$= \sqrt{\frac{1}{n^2}\sum_{i=1}^n E[\|x_i\|^2]} \le \sqrt{\frac{I}{n}}$$ //

(2) F が有限集合のとき.

$F = \{f_1,\dots,f_M\}$; $\|f_j\|_\infty \le R$, なら

$$R_n(F) \le R\sqrt{\frac{2\log(M)}{n}} \quad \text{(Hoeffding の不等式よりでる.)}$$

(証明は略) (Massart の定理)

深層ニューラルネットワークのRademacher複雑さ

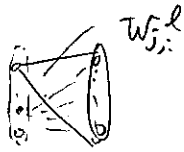
ーノルムを用いたバウンド

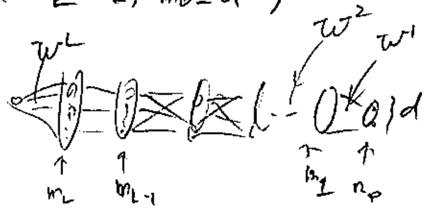
$$\mathcal{F} = \left\{ W^{(L)} \eta(W^{(L-1)} \eta(\cdots \eta(W^{(1)} x) \cdots)) \;\middle|\; W^{(\ell)} \in \mathbb{R}^{m_\ell \times m_{\ell-1}},\ \|(W^{(\ell)})^\top\|_{1,\infty} \le B_\ell \right\}$$

とする。ここで $\eta(x) = \max(x, 0)$ (ReLU)，かつ　$(m_L = 1,\ m_0 = d)$

$$\|(W^{(\ell)})^\top\|_{1,\infty} := \max_{j=1,\dots,m_\ell} \|W^{(\ell)}_{j,:}\|_1$$

とする。





$$\mathcal{F}_i = \left\{ f(x) = \sum_{j=1}^{m_{i-1}} w_j \eta(f_j(x)) \;\middle|\; f_j \in \mathcal{F}_{i-1},\ \|w\|_1 \le B_i \right\}$$

$$\mathcal{F}_1 = \left\{ f(x) = w^\top x \;\middle|\; \|w\|_1 \le B_1 \right\}$$

と定義すると，$\mathcal{F} = \mathcal{F}_L$ である。

Lem

$$\hat{R}_n(\mathcal{F}_i) \le 2 B_i \hat{R}_n(\mathcal{F}_{i-1})$$
//

Proof.

$$\begin{aligned}
\hat{R}_n(\mathcal{F}_i) &= E_{\sigma}\left[\sup_{\substack{\|w\|_1 \le B_i \\ f_j \in \mathcal{F}_{i-1}}} \frac{1}{n}\sum_{k=1}^n \sigma_k \sum_j w_j \eta(f_j(x_k)) \right] \\
&= E_{\sigma}\left[\sup_{w, f_j} \sum_j w_j \frac{1}{n}\sum_{k=1}^n \sigma_k \eta(f_j(x_k)) \right] \\
&\le E_{\sigma}\left[\sup_{w, f_j} \underbrace{\|w\|_1}_{\le B_i} \cdot \max_{1 \le j \le m_{i-1}} \left| \frac{1}{n}\sum_{k=1}^n \sigma_k \eta(f_j(x_k)) \right| \right] \\
&\le B_i E_{\sigma}\left[\sup_{f \in \mathcal{F}_{i-1}} \left| \frac{1}{n}\sum_{k=1}^n \sigma_k \eta(f(x_k)) \right| \right]
\end{aligned}$$

(ここで

$$\begin{aligned}
E_{\sigma}\left[\sup \left| \frac{1}{n}\sum_k \sigma_k g(x_k) \right| \right] &= E_{\sigma}\left[\max\left\{ \sup \frac{1}{n}\sum_k \sigma_k g(x_k),\ -\inf \frac{1}{n}\sum_k \sigma_k g(x_k) \right\} \right] \\
&= E_{\sigma}\left[\max\left\{ \sup \frac{1}{n}\sum_k \sigma_k g(x_k),\ \sup \frac{1}{n}\sum_k (-\sigma_k) g(x_k) \right\} \right] \quad (-\sigma_k \text{ は } \sigma_k \text{ と同分布}) \\
&\le 2 E_{\sigma}\left[\sup \frac{1}{n}\sum_k \sigma_k g(x_k) \right] \quad (g \text{ と } -g \text{ もモデルに入るので})
\end{aligned}$$
)

ηについて

$$\begin{aligned}
&\le 2 B_i E_{\sigma}\left[\sup_{f \in \mathcal{F}_{i-1}} \frac{1}{n}\sum_{k=1}^n \sigma_k \eta(f(x_k)) \right] \\
&\le 2 B_i \hat{R}_n(\mathcal{F}_{i-1}) \quad (\because \text{contraction ineq.})
\end{aligned}$$
//

ここで $\|x\|_\infty \le 1$ (a.s.) なら，

$$\hat{R}_n(\mathcal{F}_1) = E_{\sigma}\left[\sup_w \frac{1}{n}\sum_{k=1}^n \sigma_k w^\top x_k \right] \le E_{\sigma}\left[\sup_w \underbrace{\|w\|_1}_{\le B_1} \max_{j=1,\dots,d} \left| \frac{1}{n}\sum_{k=1}^n \sigma_k x_{k,j} \right| \right]$$

$$\le B_1 E_{\sigma}\left[\max_{j=1,\dots,d} \left| \frac{1}{n}\sum \sigma_k x_{k,j} \right| \right] \le \sqrt{\frac{2\log(2d)}{n}} \quad (\text{前より})$$

以上をまとめると、以下を得る。

Thm. $\|x\|_\infty \leq 1$ なら

$$R_n(\mathcal{F}) \leq 2^L \cdot \prod_{i=1}^{L} B_i \cdot \sqrt{\frac{2\log(d)}{n}}$$

//

よって、適切に各層のノルムが制御されていれば、汎化誤差は $O_p\left(\frac{1}{\sqrt{n}}\right)$ で減少することがわかる。

このバウンドはパラメータ数に依存せずノルムでおさえられている ⟶ 横幅が広くても使える。

他のバウンド

- Spectrally Normarized Bound (Bartlett et al., 17)

$$\prod_{k=1}^{L} \|W^k\|_{op} \left(\sum_{j=1}^{L} \frac{\|W^j - M^j\|_{1,2}^{2/3}}{\|W^j\|_{op}}\right)^{3/2} \cdot \frac{1}{\sqrt{n}}$$

ただし、$\|W^j\|_{1,2} = \sum_{i=1}^{m_j} \|W^j_{i,:}\|_2$ で、M^jは任意の行列である。

、特に M^j を初期値とすれば、最適化によってパラメータがあまり大きく動かないならば、汎化性能も良いということも示唆する。

、$\|W^j_{i,:}\|_2$ を用いているので、スパースでなくてもバウンドが大きくなりにくい。

、実験結果をよく説明する。

- (Golowich et al. 18)

$$\min\left\{ \prod_{k=1}^{L} \|W^k\|_F \cdot \frac{1}{n^{1/4}},\ \prod_{k=1}^{L} \|W^k\|_F \cdot \sqrt{\frac{L}{n}} \right\}$$

ただし、$\|W^k\|_F = \sqrt{\sum_{i,j} \left(W^k_{i,j}\right)^2}$

。2^L がない。

。$\|W^k\|_F$ は横幅の影響を受けにくい。

実際のネットワークは、各層の固有値が速く減衰することが多いので、$\|W^k\|_F$ も小さいことが多い。