○ ニューラルネットワークの近似理論

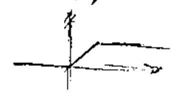
- ニューラルネットワークは広い範囲の関数を任意の精度で近似できる。
- しかも、最適なレートで近似できることが多い。

"任意の関数"を"任意の精度"で近似できる能力 ⟹ **万能近似能力**

基本的方針 → 調和解析, wavelet, 多項式スプライン近似へ帰着

例：1. $\sigma(x) = \frac{1}{1+e^{-x}}$

2. $\sigma(x) = \max(0, \min(1, x))$



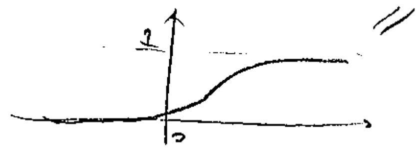
1. 連続関数 on $[0,1]^d$ の近似定理 (Cybenko, 1989)

$\sigma : \mathbb{R} \to \mathbb{R}$ を活性化関数とする。 $I^d := [0,1]^d$ とする。

Def

σ がシグモイド的 $\iff$ $\sigma(x) \longrightarrow \begin{cases} 1 & (x \to \infty) \\ 0 & (x \to -\infty) \end{cases}$ //

(- σ の単調性は仮定しない)



$$g(x) = \sum_{i=1}^{N} \alpha_i \sigma(a_i^T x + b_i) \qquad (x \in [0,1]^d,\ \alpha_i \in \mathbb{R},\ a_i \in \mathbb{R}^d,\ b_i \in \mathbb{R})$$

なる形の関数を用いて、連続関数を近似する。

Def

σ が識別的 $\iff$ I^d 上の任意の有界正則な符号付き測度 μ に対し
($M(I^d)$ と書く)

$$\int_{I^d} \sigma(a^T x + b)\, d\mu(x) = 0 \qquad (\forall a \in \mathbb{R}^d,\ b \in \mathbb{R})$$

ならば、$\mu = 0$ である。 //

Lem

σ を連続な識別的関数であるとする。すると、$\forall f \in C(I^d)$ (連続関数の集合), $\forall \varepsilon > 0$ に対し、
ある $g(x) = \sum_{i=1}^{N} \alpha_i \sigma(a_i^T x + b)$ が存在して、

$$\sup_{x \in I^d} |f(x) - g(x)| \leq \varepsilon$$

が成り立つ。 //

証明

S を g の形で書ける関数全体の集合とする。これは σ が連続なので $C(I^d)$ の線形部分空間である。S の閉包が $C(I^d)$ であれば o.k. そうでないと仮定する。
すると、Hahn-Banach の定理より $C(I^d)$ 上の有界線形汎関数 $L : C(I^d) \to \mathbb{R}$ が存在して、$\forall f \in \bar{S}$ に対し $L(f) = 0$ で $\exists f \notin \bar{S}$ で $L(f) \neq 0$ とできる。

すると、Rieszの表現定理より (Rudin, Thm 6.19)

$$L(h) = \int_{I^d} h(x)\, d\mu(x) \qquad \left(\text{ただし } \underline{\mu \in M(I^d)}\right)$$

ここに符号性入ってる。 (これが有界)

と書くことができる。特に、h として $h(x) = \sigma(a^T x + b)$ を用いると、

$$L(h) = \int_{I^d} \sigma(a^T x + b)\, d\mu(x) = 0 \qquad (\forall a \in \mathbb{R}^d,\ b \in \mathbb{R})$$

が成り立つ。しかし、今 σ は識別的であると仮定しているので、$\mu = 0$ である。

つまり、$L(f) = 0$ $(\forall f \in C(I^d))$ で L の構成に矛盾する。よって $\bar{S} = C(I^d)$ である。 //

あとは、連続なシグモイド的関数が識別的であることを示せば良い。

Lem.

任意の連続なシグモイド的関数は識別的である。

(<u>有界なシグモイド的関数に</u>拡張可能)。

Proof. シグモイド的関数の定義より

$$\sigma(\eta(a^T x + b) + \theta) \xrightarrow[\eta \to \infty]{} \begin{cases} 1 & (a^T x + b > 0) \\ 0 & (a^T x + b < 0) \\ \sigma(\theta) & (a^T x + b = 0) \end{cases}$$

である。今 μ を $\int \sigma(a^T x + b)\, d\mu(x) = 0$ $(\forall a, b)$ なる $\mu \in M(I^d)$ とすると、

σ が有界であることより、ルベーグの優収束定理より、

$$\begin{aligned} 0 &= \lim_{\eta \to \infty} \int_{I^d} \sigma(\eta(a^T x + b) + \theta)\, d\mu(x) \\ &= \int_{I^d} \lim_{\eta \to \infty} \sigma(\ \cdots\ )\, d\mu(x) \\ &= \sigma(\theta)\, \mu(\Pi_{a,b}) + \mu(H_{a,b}) \qquad (\forall \theta, a, b) \end{aligned}$$

である。ただし、$\Pi_{a,b} = \{x \mid a^T x + b = 0\}$, $H_{a,b} = \{x \mid a^T x + b > 0\}$ である。

θ は任意なので、$\underset{(\theta \to \infty)}{0 = \mu(\Pi_{a,b}) + \mu(H_{a,b})}$, $\underset{(\theta \to -\infty)}{0 = \mu(H_{a,b})}$ でもある。

ここで $a \in \mathbb{R}^d$ を任意にとって固定する。このとき、有界な可測関数 h に対し、

$$F_a(h) = \int_{I^d} h(a^T x)\, d\mu(x)$$

とする。h として $h = \mathbb{1}_{(-b,\infty)}$ や $h = \mathbb{1}_{[-b,\infty)}$ をもって来れば

$$F_a(h) = 0$$

が上の議論よりわかる。

よって、F_aの線形性より、任意の単関数 h に対しても $F_a(h)=0$ が成り立つ。 ☐

すると、優収束定理から

$$F_a(\cos) = \int_{I^d} \cos(a^T x)\, d\mu(x) = 0$$

$$F_a(\sin) = \int_{I^d} \sin(a^T x)\, d\mu(x) = 0$$

がわかる。つまり、 $\int_{I^d} e^{i a^T x} d\mu(x) = 0$ である。 $a \in \mathbb{R}^d$ は任意なので

$\mu = 0$ が得られる。

//

Thm σ が連続なシグモイド関数なら任意の $f \in C(I^d)$ と任意の $\varepsilon > 0$ に対し、

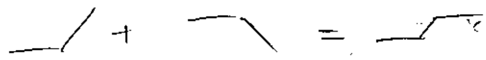
ある $g(x) = \sum_{i=1}^{n} \alpha_i \sigma(a_i^T x + b_i)$ が存在して

$$\|f - g\|_\infty \leq \varepsilon$$

とできる。つまり $C(I^d)$ の中で、万能近似能力をもつ。

//

※ σ が不連続でも L_1 内での稠密性が示せる。

※ ReLUに対しても、  とできるので（シグモイド型を一度で作れる）万能近似能力が示せる。

2. 積分表現

(Ridgelet transform) (Murata, '96; Sonoda and Murata, '15; Kostadinova et al., '14)

η: activation function (e.g. ReLU) $\eta \in \mathcal{S}'$ (緩増加関数)

$\exists \psi \in \mathcal{S}$ (急減少関数) で、

$$K_{\psi,\eta} = (2\pi)^{d-1} \int_{\mathbb{R}} \frac{\overline{\hat{\psi}(\zeta)}\,\hat{\eta}(\zeta)}{|\zeta|^d} d\zeta < \infty \quad (\text{許容条件})$$

とする。

Thm (再構成 formula)

かつ η が連続有界

$f \in L^1(\mathbb{R}^d)$, $\hat{f} \in L^1(\mathbb{R}^d)$ かつ $(\psi, \eta) \in \mathcal{S}(\mathbb{R}) \times \mathcal{S}_0'(\mathbb{R})$ が許容条件を満たすなら、

$$\mathcal{R}_\psi f(a, b) := \int_{\mathbb{R}^d} f(x)\, \overline{\psi(x^T a - b)}\, dx$$

$$\mathcal{R}_\eta^\dagger T(x) = \int_{a \in \mathbb{R}^d} \int_{b \in \mathbb{R}} T(a, b)\, \eta(a^T x - b)\, db\, da$$

に対し、

$$\mathcal{R}_\eta^\dagger \mathcal{R}_\psi f(x) = K_{\psi,\eta} f(x) \quad (\text{a.e.})$$

//

(略証) 左辺 $= \int_{\mathbb{R}^d \times \mathbb{R}} \left[\int_{\mathbb{R}^d} f(x') \overline{\psi(a^T x' - b)}\, dx' \right] \eta(a^T x - b)\, da\, db$

$= \int_{\mathbb{R}^d} f(x') \left[\int_{\mathbb{R}^d \times \mathbb{R}} \overline{\psi(a^T x' - b)}\, \eta(a^T x - b)\, da\, db \right] dx'$

$= \int_{\mathbb{R}^d} f(x') \left[\underbrace{\int_{\mathbb{R}^d} \tilde{\psi} * \eta(a^T(x - x'))\, da}_{\shortparallel} \right] dx'$

$(2\pi)^{-1} \int_{\mathbb{R}^d} \left(\int_{\mathbb{R}} \overline{\hat{\psi}(\zeta)}\, \hat{\eta}(\zeta)\, e^{i\zeta a^T(x-x')} d\zeta \right) da = (2\pi)^{-1} \int_{\mathbb{R}^d} \overbrace{\left(\int_{\mathbb{R}} \frac{\overline{\hat{\psi}(\zeta)}\,\hat{\eta}(\zeta)}{|\zeta|^d} d\zeta \right)}^{(2\pi)^{-(d-1)} K_{\psi,\eta}} e^{i a^T (x - x')} da$ (変数変換)

$= (2\pi)^{-d} K_{\psi,\eta} \times \int_{\mathbb{R}^d} e^{i a^T x} \left(\int_{\mathbb{R}^d} f(x') e^{-i a^T x'} dx' \right) da$

$= (2\pi)^{-d} \int_{\mathbb{R}^d} e^{i a^T x} \hat{f}(a)\, da \cdot K_{\psi,\eta}$

$= f(x) \cdot K_{\psi,\eta}$

//

3. ソボレフ空間での近似理論（無限回微分可能な σ）

$\Omega = I^d$ として

$$\|f\|_p := \begin{cases} \left(\int_\Omega |f(x)|^p dx\right)^{\frac{1}{p}} & (1 \le p < \infty) \\ \operatorname{ess\,sup}_{x\in\Omega} |f(x)| & (p=\infty) \end{cases}$$

とおく。$k=(k_1,\dots,k_d) \in \mathbb{Z}^d$ に対し、a.e. で s 回微分可能な関数 f に対して

$$\|f\|_{W_p^s(\Omega)} := \sum_{0\le k\le s} \|D^k f\|_p \qquad (s\in\mathbb{Z},\ s\ge 0,\ 1\le p\le\infty)$$

とおく。ただし $0\le k\le s$ とは $0 \le \overset{=|k|}{\sum_{j=1}^d |k_j|} \le s$ を意味し、

$$D^k f = \frac{\partial^{|k|} f}{\partial x_1^{k_1}\cdots\partial x_d^{k_d}}$$

とおく。

$$\Pi_N := \left\{ \sum_{j=1}^N \alpha_j \sigma(a_j^T x + b_j) \;\middle|\; \alpha_j\in\mathbb{R},\ a_j\in\mathbb{R}^d,\ b_j\in\mathbb{R} \right\} \quad (\text{ニューラルネットの集合})$$

目標：$f \in W_p^s(\Omega)$ を $g\in\Pi_N$ で近似する。その近似精度を求める。

Thm

σ がある開区間で無限回連続微分可能であるとする。また、その開区間のある点 $b\in\mathbb{R}$ において

$$\frac{\partial^k}{\partial x^k}\sigma(b) \ne 0 \qquad (\forall k\in\mathbb{Z},\ k\ge 0)$$

とおく。すると、$\forall f\in W_p^s(\Omega)$ に対し、$\exists g\in\Pi_N$ で

$$\|f-g\|_p \le c\, \underline{N^{-\frac{s}{d}}}\, \|f\|_{W_p^s}$$

である。

※ この近似精度は $O(N)$ 個の基底を用いた近似法の中で最適なオーダーである（非線形も含む）

※ $\sigma(x) = \frac{1}{1+e^{-x}}$（シグモイド関数）は条件を満たす。

※ なめらかな関数はより近似しやすい。

（略証）チェビシェフ多項式 $T_k(x)$ $(k=0,1,2,\dots)$, $T_0(x)=1$, $T_1(x)=x$, $T_k(x)=2xT_{k-1}(x)-T_{k-2}(x)$
$(T_k(\cos(t)) = \cos(kt))$ を用いて

$$T_k(x) = \prod_{j=1}^d T_{k_j}(x_j) \qquad (k=(k_1,\dots,k_d),\ x=(x_1,\dots,x_d))$$

とすると、$m\ge1$ に対し、ある $P_m(f) = \sum_{0\le k\le 2m} \overset{\text{係数}}{c_k(f)}\, T_k(x)$ が存在して、

$$\left.\begin{aligned} &\|f-P_m(f)\|_p \le \frac{c}{m^s}\|f\|_{W_p^s} \\ &\sum_{0\le k\le 2m} |c_k(f)| \le c\, m^\alpha \|f\|_{W_p^s} \qquad \left(\alpha = \frac{s}{p\wedge 2}\right) \end{aligned}\right\} \text{——}(*)$$

とできることが知られている。

よって、あとは多項式 $P_m(f)$ を Π_N の元で近似すればよい。

実は、$\forall \varepsilon > 0$ に対し、$\exists g_{k,\varepsilon} \in \Pi_{(2m+1)^d}$ で $\|T_k - g_{k,\varepsilon}\|_\infty \leq \varepsilon$ とできる。(ただし $|k| \leq 2m$)
そのネットワークを構成する。

今 σ が $[b-\delta, b+\delta]$ 上で無限回連続微分可能とする。

すると、$\sigma_k(a, x) := \dfrac{\partial^{|k|}}{\partial a_1^{k_1} \cdots \partial a_d^{k_d}} \sigma(a^T x + b) = x^k \sigma^{(|k|)}(a^T x + b)$ $(k = k_1, \dots, k_d)$

である。よって、$x^k = \dfrac{1}{\sigma^{(|k|)}(b)} \cdot \sigma_k(0, x)$ を得る。($x^k = x_1^{k_1} \cdots x_d^{k_d}$ とする)

一方で、
$$\rho_{k,h}(x) := h^{-|k|} \sum_{0 \leq r \leq k} (-1)^{|r|} \binom{k}{r} \sigma\left(h(\underbrace{2r - k}_{\text{ベクトル}})^T \cdot x + b\right) \quad (r, k \in \mathbb{Z}^d)\ (0 < h, h \in \mathbb{R})$$
$$\binom{k}{r} = \prod_{j=1} \binom{k_j}{r_j}$$

とすると、
$$\|\rho_{k,x} - \sigma_k(0, x)\|_\infty \leq \underset{\text{定数}}{M_k} \cdot h^2$$
が示せる。つまり、$\left\| x^k - \dfrac{\rho_{k,h}}{\sigma^{(|k|)}(b)} \right\|_\infty \leq \dfrac{M_k}{\sigma^{(|k|)}(b)} h^2$

である。このことから h を十分に小さくとれば $T_k(x) = \sum_{0 \leq r \leq k} t_{k,r} x^r$ を、
$g_{k,\varepsilon}(x) = \sum_{0 \leq r \leq k} \dfrac{t_{k,r}}{\sigma^{(|k|)}(b)} \rho_{r,h}$ を用いて、誤差 ε 以内で近似できる。

なお、$g_{k,\varepsilon} \in \Pi_{(2m+1)^d}$ もみたす。

さらに、(*) から、$(2m+1)^d \leq N$ として、ε を十分小さくとると、

$G_N(x) = \sum_{0 \leq k \leq 2m} c_k(f) g_{k,\varepsilon}(x)$, $G_N \in \Pi_N$ を用いて、

$$\|P_m(f) - G_N\|_\infty \leq c \cdot m^{-\rho} \|f\|_{W_p^\rho}$$

とできる。さらに (a) より $\|f - P_m(f)\|_p \leq O(m^{-\rho})$ より 全体で

$$\begin{aligned}
\|f - G_N\|_p &\leq \|f - P_m(f)\|_p + \|P_m(f) - G_N\|_\infty \quad (\because \Omega = [0,1]^d) \\
&\lesssim m^{-\rho} \cdot \|f\|_{W_p^\rho} \\
&\simeq N^{-\frac{\rho}{d}} \|f\|_{W_p^\rho} \quad (N \simeq m^d \text{ とする})
\end{aligned}$$

//

この解析ではσとして十分滑らかな関数を用いていた。

よって、現在主流であるReLU活性化関数 $\sigma(x) = \max(x, 0)$ は含まれない。

しかし、スパースな(深層)ネットワークを用いれば、ReLUを用いても、

$$\|f - G_N\|_{p} \le N^{-\frac{s}{d}} \log(N)$$

→ スパースなネットワーク

をみたすようにできる（幅：$O(N)$，深さ：$O(\log(N))$，パラメータ数：$O(N \log N)$）→ 深層にしなければ(浅層なら) $s > 2d$ とて sub-optimal となる。

→ x^2

誤差 2^{-m}（m層）

$\Rightarrow$ $(x+y)^2 - x^2 - y^2 = 2xy$ が近似可能 → 区分的多項式が近似可能

$\Rightarrow$ ソボレフ空間の元をうまく近似できる。