神经网络基础

全连接神经网络

$$f(x;\theta) = f_L \circ \phi_\theta(x)$$
$$= W^L \circ \phi_\theta(x) + b^L$$

随机特征回归：

$$\sum c_i \phi_i(x) \Rightarrow W^L \phi(x) + b^L$$

GELU：Gaussian Error Linear Unit

$$\mathrm{GELU}(x) = x\,\Phi(x) = x\,P(X \le x)$$
$$= x\int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-\frac{X^2}{2}} dx$$
$$= x\left(\int_0^x \frac{1}{\sqrt{2\pi}} e^{-\frac{X^2}{2}} dx + \frac{1}{2}\right) \quad \boxed{\frac{X}{\sqrt{2}} = z}$$

$$\mathrm{erf}(z) = \frac{2}{\sqrt{\pi}}\int_0^z e^{-t^2} dt$$

$$= x\left(\int_0^{\frac{x}{\sqrt{2}}} \frac{1}{\sqrt{2\pi}} e^{-z^2} d\sqrt{2}z + \frac{1}{2}\right)$$
$$= \frac{x}{2}\left(\int_0^{\frac{x}{\sqrt{2}}} \frac{2}{\sqrt{\pi}} e^{-z^2} dz + 1\right)$$

silu (Sigmoid Linear Unit)

$$\mathrm{silu}(x) = x \cdot \mathrm{sigmoid}(x)$$

Wierstrass 逼近定理

$$(x+1-x)^n = \sum_k C_n^k x^k (1-x)^{n-k}$$

考虑 $\sum C_n^k x^k (1-x)^{n-k} f(\frac{k}{n}) \to f(x)$ in $[0,1]$

通用逼近定理 (universal approximation)

引理：g 是 $R \to R$ 的 p-Lipschitz 函数，2层 $\lceil \frac{p}{\varepsilon} \rceil$ 个神经元

使用 $\mathbb{1}_{x \geq 0}$ 的神经网络，能对 g 在 $[0,1]$ 上达到 ε 近似。

证明

$$\hat{g} = \sum_{i=1}^{m} a_i \mathbb{1}(x - b_i), \text{ 其中}$$

$$b_i = \frac{(i-1)\varepsilon}{p}, \quad m = \left\lceil \frac{p}{\varepsilon} \right\rceil, \quad \text{取}$$

$$a_1 = g(b_1), \quad a_2 = g(b_2) - g(b_1) \cdots, \quad a_m = g(b_m) - g(b_{m-1})$$

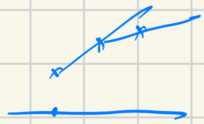
对 $b_j \leq x < b_{j+1}$

$$\hat{g}(x) = \sum_{i=1}^{j} a_i \mathbb{1}(x - b_i)$$

$$= \sum_{i=1}^{j} a_i = g(b_j)$$

$$|\hat{g}(x) - g(x)| = |g(b_j) - g(x)| \leq (b_j - x) p = \frac{\varepsilon}{p} \cdot p = \varepsilon$$

$\Rightarrow$ sigmoid, ReLU ......



分片线性函数

引理：g 是 $\mathbb{R}^d \to \mathbb{R}$ 的 P-lipschitz 函数，2层神经元使用 $\mathbb{1}_{x\in R_i}$ 作为激活函数的神经网络，能对 g 在 $[0,1]^d$ 上达到 ε 近似

其中 $R_1, R_2 \cdots R_m$ 是 $[0,1]^d$ 上的边长不超过 δ 的小方块

选取 $x_i \in R_i$

$$f(x) = \sum_{i=1}^{D} \alpha_i \mathbb{1}_{x\in R_i} \qquad \alpha_i = g(x_i)$$

那么

$$\sup \left| g(x) - f(x) \right| = \sup_i \sup_{x\in R_i} \left| g(x) - f(x) \right|$$

$$\leq \sup_i \sup_{x\in R_i} \underbrace{\left| g(x) - g(x_i) \right|}_{\leq \delta} + \underbrace{\left| g(x_i) - f(x) \right|}_{0}$$

激活函数去近似 $\mathbb{1}_{x\in R_i}$ $(\delta \sim \text{error})$

$$\Rightarrow \quad D = \left(\frac{1}{\delta}\right)^d \qquad \text{维度灾难}$$

目标：指数收敛 $\quad \text{error} = e^{-\left(\frac{1}{\delta}\right)^d}$ 计算量 $\left(\frac{1}{\delta}\right)^d = -\log \text{error}$

直观理解：Fourier transform

$$\hat{f}(x) = \frac{1}{(2\pi)^d} \int_{R^d} f(x) e^{-iw^T x} dx$$

$$f(x) = \int \hat{f}(w) e^{iw^T x} dw \qquad \left(\hat{f}(w) = |\hat{f}(w)| e^{ib(w)} \right)$$

$$= \int |\hat{f}(w)| e^{ib(w) + iw^T x} dw$$

$$= \int |\hat{f}(w)| \cos\left(b(w) + w^T x\right) dw$$

定义： $C_f = \int |\hat{f}(w)| dw$ ， $p(w) = \frac{\hat{f}(w)}{C_f}$

概率密度函数

$$f(x) = C_f E_{w\sim p} [\cos b(w) + w^T x]$$

定义 $f_m = \frac{1}{m} \sum_{j=1}^{m} C_f \cos(w_j^T x + b(w_j))$ $\quad w_j \sim p$

$$Z_j = C_f \cos(w_j^T x + b(w_j))$$

$$E_w[Z_j - f(x)] = 0$$

$$E_w[(Z_j - f(x))^2] = E_w Z_j^2 - f(x)^2$$

$$\leq E_w Z_j^2$$

$$\leq C_f^2$$

对任意概率密度函数 p，我们有

$$E_w \| f_m - f \|^2_{L_2(P)}$$

$$= E_w \left\| \frac{1}{m} \sum_j (Z_j - f) \right\|^2_{L_2(P)} \qquad Z_j \text{ 独立}$$

$$= E_w \int \frac{1}{m^2} \left(\sum_j (Z_j - f)\right)^2 P(x)\, dx$$

$$= \int \frac{m}{m^2} E_w (Z_j - f)^2 \; P(x)\, dx$$

$$= \frac{1}{m^2} E_{x\sim P}\left[m E_w (Z_j - f)^2 \right]$$

$$\le \frac{C_f^2}{m}$$

那么$\exists\, w$，使

$$\| f_m - f \|^2_{L_2(P)} \le \frac{C_f^2}{m}$$

与维度无关

i) $C_f = \mathrm{poly}(d)$

ii) cos 一般不用于激活函数

iii) 一般考虑紧支集

Barron space 近似

$$f_e|_\Omega = f \qquad (\text{假设 } 0 \in \Omega)$$

$$f_e(x) = \int e^{iw^Tx} \hat{f}_e(w)\,dw$$

$$f(x) - f(0) = \int (e^{iw^Tx} - 1) \hat{f}_e(w)\,dw$$

$$= \int \frac{e^{iw^Tx} - 1}{\|w\|} \|w\| \hat{f}_e(w)\,dw$$

$$= \int \underbrace{\frac{\cos(w^Tx + b(w)) - \cos(b(w))}{\|w\|}}_{g(x,w)} \|w\| |\hat{f}_e(w)|\,dw$$

其中 $\|w\| = \sup_{x\in\Omega} |w^Tx|$

定义 $h(t; w) = \dfrac{\cos(\|w\|^2 t + b(w)) - \cos(b(w))}{\|w\|}$

$$h\left(\frac{w^Tx}{\|w\|}; w\right) = g(x; w) \qquad \left(\hat{w} = \frac{w}{\|w\|}\right)$$

由 $\|w\|$ 的定义，$\dfrac{w^Tx}{\|w\|} \in [-1, 1]$

若 $\int \|w\| |\hat{f}_e(w)| dw = C_f' < +\infty$

$p \sim \|w\| |\hat{f}_e(w)| / C_f'$

$f(x) - f(0) = C_f' E_{w\sim p}[g(x,w)]$

我们有 $g(x,w) = h(\hat{w}^T x; w)$

其中 $\hat{w}^T x \in [-1, 1]$

注意 $h(-1; w)$, $h'(t; w) \in [-1, 1]$，因为

$$h(t;w) = \frac{\cos(\|w\| t + b(w)) - \cos(b(w))}{\|w\|}$$

$$= \frac{-2\sin\left(\frac{\|w\| t}{2}\right) \sin\left(\frac{\|w\| t}{2} + b(w)\right)}{\|w\|}$$

$$h'(t;w) = -\sin(\|w\| t + b(w))$$

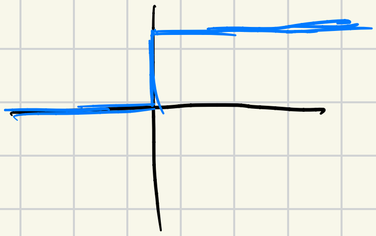
我们有

$$h(t;w) = h(-1;w) + \int_{-1}^{t} h'(s;w)\, ds$$

$$= h(-1;w) + \int_{-1}^{1} h'(s;w)\, H(t-s;w)\, ds$$

Heaviside:

$$H(t) = 1 \quad (t \geq 0)$$



$$f(x) = f(0) + C_f' E_{w\sim p}[h(-1;w)]$$ 与 x 无关的常数

$$+ C_f' E_{w\sim p} E_{s\sim U[-1,1]}[h'(s;w) H(\hat{w}^T x - s)]$$

$$f_D(x) = a_0 + C_f' \frac{1}{D}\sum_{j=1}^{D} \underbrace{h'(s_j;w_j) H(\hat{w}_j^T x - s_j)}_{Z_j}$$

$$E_w \| f_D - f \|^2_{L_2(P)}$$

$$E = E Z_j$$

$$= E_w \left\| C_f' \left(\frac{1}{D}\sum_{j=1}^{D} Z_j - E \right) \right\|^2_{L_2(P)}$$

$$= \frac{C_f'^2}{D} E_{x\sim P} E_w (Z_j - E)^2$$

$$\leq \frac{C_f'^2}{D} E Z_j^2 \leq \frac{C_f'}{D}$$

非凸优化

$$y_k = x_k + \beta(x_k - x_{k-1})$$ (Polyak's heavy-ball method)

$$x_{k+1} = y_k - \alpha \nabla f(x_k)$$

$$y_k = x_k + \beta(x_k - x_{k-1})$$ (Nesterov's accelerated gradient descent)

$$x_{k+1} = y_k - \alpha \nabla f(y_k)$$

In Adam, bias correction

$m_0 = v_0 = 0$

当 $g_t = g$

$m_1 = (1-\beta_1) g \qquad m_2 = (1-\beta_1) g \cdot \beta_1 + (1-\beta_1) g$

$\qquad\qquad\qquad\qquad = (1-\beta_1^2) g$

$m_k = (1-\beta_1^k) g$