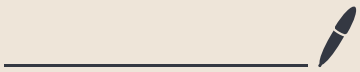
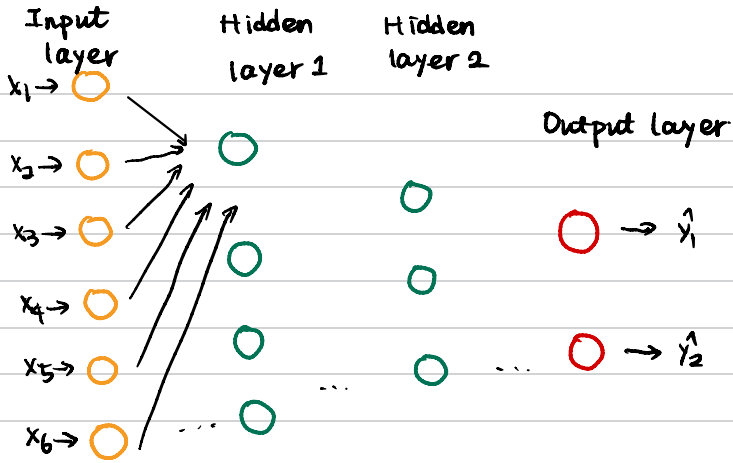
Chapter 4 Feedforward Neural Networks





$$Y = F_{w,b}(X) + \varepsilon$$

$$\hat{Y}(X) = F_{w,b}(X) = f^{(L)}_{W^{(L)},b^{(L)}} \circ \cdots \circ f^{1}_{W^{(1)},b^{(1)}}(X)$$

	notation	definition
超参数	L	神经网络的层数
	M_l	第l层神经元的个数
	$f_l(\cdot)$	第l层神经元的激活函数
参数	$W^{(l)} \in \mathbb{R}^{M_l \times M_{l-1}}$	第$l-1$层到第l层的权重矩阵
	$b^{(l)} \in \mathbb{R}^{M_l}$	第$l-1$层到第l层的偏置
活性值	$z^{(l)} \in \mathbb{R}^{M_l}$	第l层神经元的净输入(净活性值)
	$a^{(l)} \in \mathbb{R}^{M_l}$	第l层神经元的输出(活性值)

激活函数	函数	导数
Logistic函数	$f(x) = \frac{1}{1+\exp(-x)}$	$f'(x) = f(x)(1-f(x))$
Tanh函数	$f(x) = \frac{\exp(x)-\exp(-x)}{\exp(x)+\exp(-x)}$	$f'(x) = 1-f(x)^2$
ReLU函数	$f(x) = \max(0, x)$	$f'(x) = I(x>0)$
ELU函数	$f(x) = \max(0,x) + \min(0, \gamma(\exp(x)-1))$	$f'(x) = I(x>0) + I(x\leq 0)\cdot\gamma\exp(x)$

$$z^{(l)} = W^{(l)} a^{(l-1)} + b^{(l)}$$
$$a^{(l)} = f_l(z^{(l)})$$

$$a^{(l)} = f_l(W^{(l)} a^{(l-1)} + b^{(l)})$$

$$x = a^{(0)} \to z^{(1)} \to a^{(1)} \to z^{(2)} \to a^{(2)}$$
$$\cdots \to a^{(l-1)} \to z^{(l)} \to a^{(l)} = \phi(x; W, b)$$

Ex. $\sigma: \mathbb{R} \to B \subset \mathbb{R}$

$$f^{(l)}_{W^{(l)}, b^{(l)}}: \mathbb{R}^n \to \mathbb{R}^m$$

$$f(v) = W^{(l)} \sigma^{(l-1)}(v) + b^{(l)}, \quad W^{(l)} \in \mathbb{R}^{m \times n}, \quad b^{(l)} \in \mathbb{R}^m$$

eg. $f(v) = w \tanh(v) + b$

Ex. One hidden layer, $\sigma^{(1)}$ is the identity function

$$\vec{Y}(X) = W^{(2)}(W^{(1)}X + b^{(1)}) + b^{(2)} = W^{(2)}W^{(1)}X + W^{(2)}b^{(1)} + b^{(2)}$$
$$= \tilde{W}X + \tilde{b}$$

Universal Approximation Theorem

对于具有线性输出层和至少一个使用"挤压"性质的激活函数的隐藏层组成的前馈神经网络，只要其隐藏层神经元的数量足够，它可以以任意的精度来近似任何一个定义在实数空间的有界闭集函数.

多分类问题

0 1
0 2
0 3 — $p(y=c|x)$
⋮
0 C

$$P(y|x) \Rightarrow \hat{y} = \text{Softmax}(z^{(L)})$$

$$L(y, \hat{y}) = -y^T \log \hat{y}$$

$$R(W, b) = \frac{1}{N} \sum_{n=1}^{N} L(y^{(n)}, \hat{y}^{(n)}) + \frac{1}{2} \lambda \|W\|_F^2$$

$$\frac{\partial R(W, b)}{\partial W^{(l)}} \qquad \frac{\partial R(W, b)}{\partial b^{(l)}}$$ (梯度下降)

Chain Rule $y = f^5(f^4(f^3(f^2(f^1(x))))) \quad \frac{\partial y}{\partial x} = \frac{\partial f^1}{\partial x}\frac{\partial f^2}{\partial f^1}\cdots\frac{\partial f^5}{\partial f^4}$

反向传播算法，自动微分 (AD)

An arrangement of $n \geq p$ hyperplanes in $\mathbb{R}^p$ has at most

$\sum_{j=0}^{p}\binom{n}{j}$ convex regions

$$R(W,b) = \frac{1}{N}\sum_{n=1}^{N} L(y^{(n)}, \hat{y}^{(n)}) + \frac{1}{2}\lambda\|W\|_F^2$$

$$\frac{\partial R(W,b)}{\partial W^{(l)}}, \quad \frac{\partial R(W,b)}{\partial b^{(l)}}$$

$$\frac{\partial L(y,\hat{y})}{\partial W_{ij}^{(l)}}$$

$$z^{(l)} = W^{(l)} a^{(l-1)} + b^{(l)} \qquad a^{(l)} = f_l(z^{(l)})$$

$$\frac{\partial L(y,\hat{y})}{\partial W_{ij}^{(l)}} = \frac{\partial z^{(l)}}{\partial W_{ij}^{(l)}} \frac{\partial L(y,\hat{y})}{\partial z^{(l)}}$$

$$\frac{\partial z^{(l)}}{\partial W_{ij}^{(l)}} = \left[\frac{\partial z_1^{(l)}}{\partial W_{ij}^{(l)}} \cdots \frac{\partial z_i^{(l)}}{\partial W_{ij}^{(l)}} \cdots \frac{\partial z_{M_l}^{(l)}}{\partial W_{ij}^{(l)}}\right]$$

$$= [0, \cdots, a_j^{(l-1)}, \cdots, 0]$$

$$\triangleq D_i(a_j^{(l-1)}) \quad R \in \mathbb{R}^{l \times M_l}$$

$$\frac{\partial L(y,\hat{y})}{\partial b^{(l)}} = \frac{\partial z^{(l)}}{\partial b^{(l)}} \cdot \frac{\partial L(y,\hat{y})}{\partial z^{(l)}}$$

$$\frac{\partial z}{\partial b^{(l)}} = I_{M_l} \in \mathbb{R}^{M_l \times M_l}$$

$$\delta^{(l)} \triangleq \frac{\partial L(y,\hat{y})}{\partial z^{(l)}}$$

$$= \frac{\partial a^{(l)}}{\partial z^{(l)}} \cdot \frac{\partial z^{(l+1)}}{\partial a^{(l)}} \cdot \frac{\partial L(y,\hat{y})}{\partial z^{(l+1)}}$$

$$= \text{diag}(f_l'(z^{(l)})) \cdot (W^{(l+1)})^T \cdot \delta^{l+1}$$

$$= f_l'(z^{(l)}) \odot ((W^{(l+1)})^T \cdot \delta^{(l+1)}) \in \mathbb{R}^{M_l}$$

$$\Rightarrow \frac{\partial L(y, \hat{y})}{\partial w_{ij}^{(l)}} = [0, \cdots, a_j^{(l-1)}, \cdots, 0] [\delta_1^{(l)}, \cdots, \delta_i^{(l)}, \cdots \delta_{M_l}^{(l)}]^T$$

$$= \delta_1^{(l)} \cdot a_j^{(l-1)}$$

$$\frac{\partial L(y, \hat{y})}{\partial W^{(l)}} = \delta^{(l)} (a^{(l-1)})^T \in \mathbb{R}^{M_l \times M_{l-1}}$$

$$\frac{\partial L(y, \hat{y})}{\partial b^{(l)}} = \delta^{(l)} \in \mathbb{R}^{M_l}$$