A2-QA

Q1

1. **前向传播矩阵形式**

   输入层到隐藏层:

   输入 $x = [x_1, x_2]$，隐藏层有 3 个节点，故权重矩阵为 W_1 为 2×3，偏置项 b_1 为 1×3，因此，隐藏层输入为 $Z_1 = x \cdot W_1 + b_1$，其中得到 Z_1 为 1×3 的向量矩阵，通过激活函数输出为 $A_1 = \sigma(Z_1)$

   隐藏层到输出层:

   权重矩阵 W_2 为 3×1，偏置矩阵 b_2 为 1×1，隐藏层输出为 $Z_2 = A_1 \cdot W_2 + b_2$，结果为 $A_2 = \sigma(Z_2)$

2. **交叉熵损失函数如何衡量预测概率分布与真实分布的差异？写出二分类问题的交叉熵公式**

   交叉熵损失函数可以衡量预测值和真实值之间的差异

   $$L = -(y \log \hat{y} + (1 - y) \log(1 - \hat{y}))$$

   其中 y 是真实标签，$\hat{y}$ 是预测概率

   $t = 1$ 时，损失为 $-\log y$，如果 y 接近 1，损失小，接近 0，损失大

   $t = 0$ 时，损失为 $-\log(1 - y)$，如果 y 接近 0，损失小，接近 1，损失大

Q2

1. **在前馈神经网络中，所有的参数能否被初始化为0？如果不能，能否全部初始化为其他相同的值？原因是什么？**

   不可以全部初始化为 0，也不能全部初始化为相同的值

   如果参数初始化为相同值，则同一层的所有神经元在正向传播时会生成相同的结果，反向传播的时候也是会得到相同的梯度更新，会致使神经元之间无法学习差异化的特征，降低模型的表达能力

2. **计算权重并说明梯度下降的更新方向**

梯度计算: $L(w) = w^2+2w+1$

$$\left.\frac{dL}{dw}\right|_{w=3} = 8$$

∴更新方向与梯度方向相反

$$w' = w - 8\cdot\eta$$

其中η为学习率. 更新方向沿负方向

Q3

1. **证明**:

1. 证: $1-\sigma(x) = 1 - \frac{1}{1+e^{-x}} = \frac{e^{-x}}{1+e^{-x}}$

$$\sigma(-x) = \frac{1}{1+e^{x}} = \frac{e^{-x}}{1+e^{-x}}$$

$$\therefore 1-\sigma(x) = \sigma(-x)$$

2. **证明**:

2. $\sigma'(x) = \frac{d}{dx}(1+e^{-x})^{-1} = (-1)(e^{-x})\left(-\frac{1}{(1+e^{-x})^2}\right)$

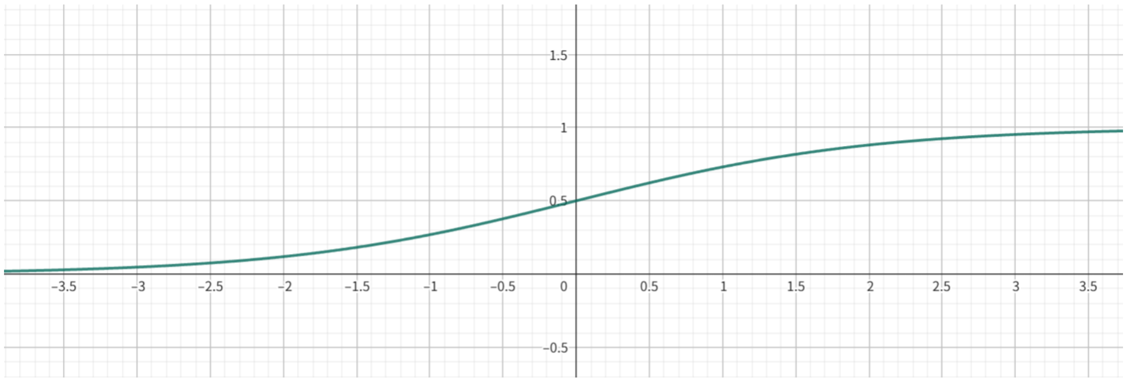
$$= \frac{e^{-x}}{(1+e^{-x})^2}$$

$$\sigma(x)(1-\sigma(x)) = \frac{e^{-x}}{(1+e^{-x})^2}$$

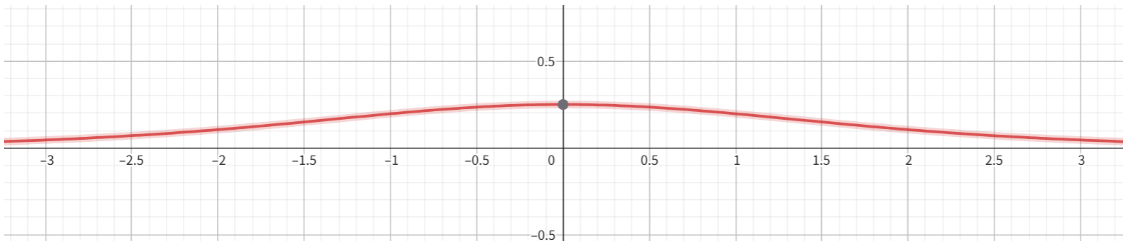
$$\therefore \sigma'(x) = \sigma(x)(1-\sigma(x))$$

画图:

$\sigma(x)$



$\sigma'(x)$



3. **证明**

第 i 层输出结果

$$\text{线性变换} \quad u^{(i)} = W^{(i)} y^{(i-1)} + b^{(i)}$$
$$\text{激活函数} \quad y^{(i)} = f^{(i)}(u^{(i)})$$

我们需要求的是该层参数 $\theta^{(i)}$ 对于损失函数 L 的梯度 $\frac{\partial L}{\partial \theta^{(i)}}$

首先，梯度可以拆分为

$$\frac{\partial L}{\partial \theta^{(i)}} = \frac{\partial L}{\partial y^{(i)}} \frac{\partial y^{(i)}}{\partial \theta^{(i)}}$$

前者记录为“损失对本层输出”的灵敏度

后者记录为“本层参数对本层输出”的雅可比矩阵

进而进行链式法则递推

$$\frac{\partial L}{\partial y^{(i)}} = \frac{\partial L}{\partial y^{(D)}} \prod_{k=i+1}^{D} \frac{\partial y^{(k)}}{\partial y^{(k-1)}}$$

对于第 i 层

$$\frac{\partial y^{(i)}}{\partial W^{(i)}} = \frac{\partial y^{(i)}}{\partial u^{(i)}} \frac{\partial u^{(i)}}{\partial W^{(i)}} = \left[f^{(i)\,\prime}(u^{(i)}) \right] \left[y^{(i-1)} \right]^{\top}$$
$$\frac{\partial y^{(i)}}{\partial b^{(i)}} = f^{(i)\,\prime}(u^{(i)})$$

合并后 $\theta^{(i)} = [W^{(i)}, b^{(i)}]$，即得到了 $\frac{\partial y^{(i)}}{\partial \theta^{(i)}}$

可以发现对于每个 k，$\frac{\partial y^{(k)}}{\partial y^{(k-1)}}$ 中都包含对于 σ 的导数

由于

$$0 < \sigma^{'}(u^{(k)}) \leq \frac{1}{4}$$

因此当网络很深时

$$\prod_{k=i}^{D} \sigma^{'}(u^{(k)}) \leq (\frac{1}{4})^{L-i+1}$$

随着 D 深度的增大，其值呈指数级衰减，导致梯度消失