2-2 只有一个隐藏层的神经网络 II

（多个训练样本）

王中雷

厦门大学王亚南经济研究院和经济学院, 2025

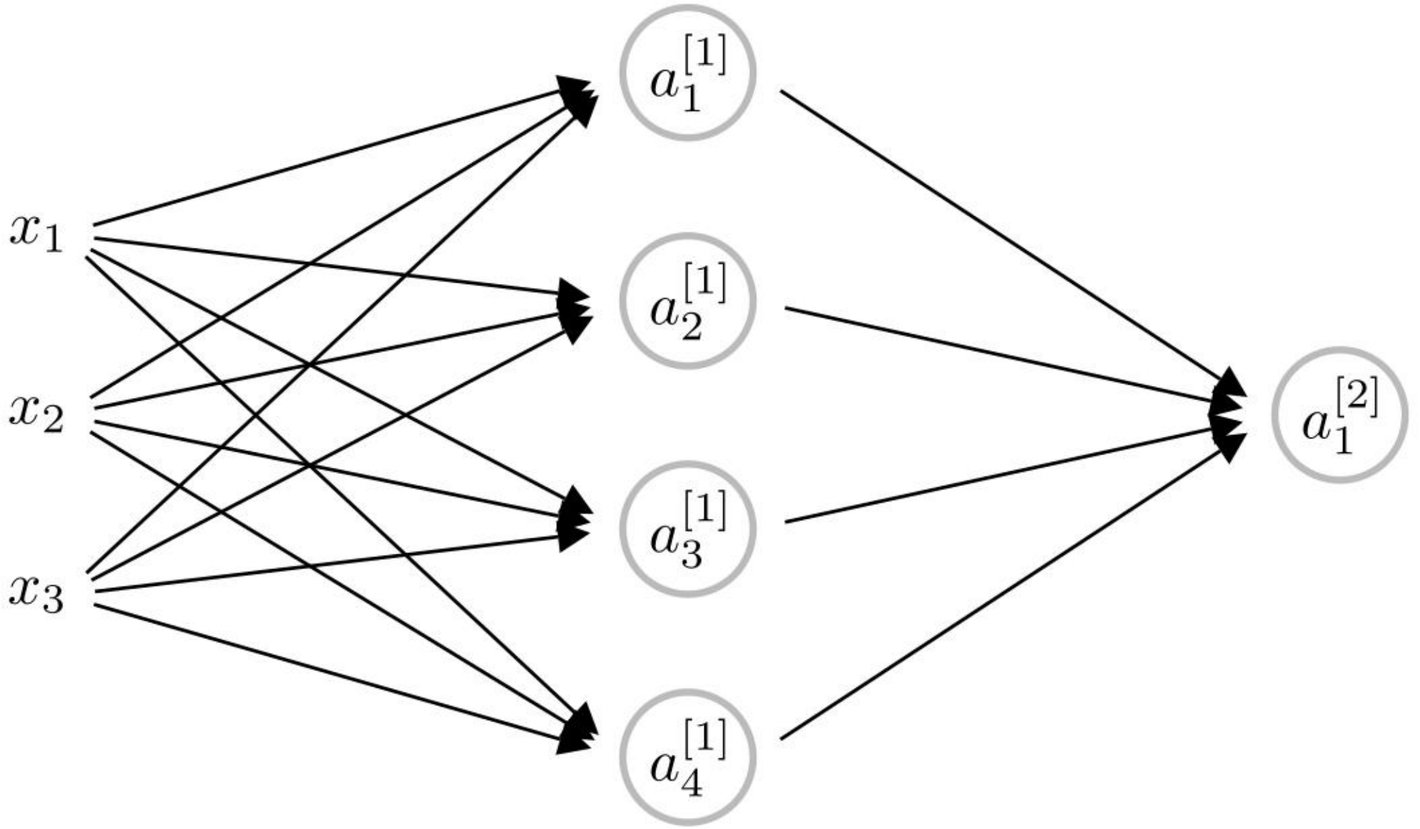
内容摘要

1. 前向传播

2. 后向传播

3. 批量梯度下降法

回顾



输入层($d^{[0]} = 3$)　　第一隐藏层($d^{[1]} = 4$)　　输出层($d^{[2]} = 1$)

回顾

1. 回顾

- L：神经网络模型的层数
- $d^{[l]}$：第l 层神经元的个数 $\quad (l = 0, \ldots, L)$
- $\boldsymbol{a}^{[l]} = (a_1^{[l]}, \ldots, a_{d^{[l]}}^{[l]})^{\mathrm{T}} \in \mathbb{R}^{d^{[l]} \times 1}$
- $\boldsymbol{W}^{[l]} = (\boldsymbol{w}_1^{[l]}, \ldots, \boldsymbol{w}_{d^{[l]}}^{[l]})^{\mathrm{T}} \in \mathbb{R}^{d^{[l]} \times d^{[l-1]}}$
- $\boldsymbol{b}^{[l]} = (b_1^{[l]}, \ldots, b_{d^{[l]}}^{[l]})^{\mathrm{T}} \in \mathbb{R}^{d^{[l]} \times 1}$

回顾

1．如果我们只有一个训练样本$(\boldsymbol{x}, y)$，前向传播的计算如下

$$\boldsymbol{z}^{[1]} = \boldsymbol{b}^{[1]} + \boldsymbol{W}^{[1]}\boldsymbol{x}$$

$$\boldsymbol{a}^{[1]} = \sigma(\boldsymbol{z}^{[1]})$$

$$z^{[2]} = b^{[2]} + \boldsymbol{W}^{[2]}\boldsymbol{a}^{[1]}$$

$$a^{[2]} = \sigma(z^{[2]})$$

2．代价函数为

$$\mathcal{J} = \mathcal{L} = -\left\{y \log a^{[2]} + (1-y) \log\left(1 - a^{[2]}\right)\right\}$$

前向传播

1. 假设我们有n 个训练样本$\{(\boldsymbol{x}_i, y_i) : i = 1, \ldots, n\}$

2. 记$\boldsymbol{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)^{\mathrm{T}} \in \mathbb{R}^{n \times d}$

3. 基于向量化的前向传播

$$\boldsymbol{Z}^{[1]} = (\boldsymbol{b}^{[1]})^{\mathrm{T}} + \boldsymbol{X}(\boldsymbol{W}^{[1]})^{\mathrm{T}} \in \mathbb{R}^{n \times d^{[1]}}$$

$$\boldsymbol{A}^{[1]} = \sigma(\boldsymbol{Z}^{[1]}) \in \mathbb{R}^{n \times d^{[1]}}$$

$$\boldsymbol{Z}^{[2]} = b^{[2]} + \boldsymbol{A}^{[1]}(\boldsymbol{W}^{[2]})^{\mathrm{T}} \in \mathbb{R}^{n \times d^{[2]}}$$

$$\boldsymbol{A}^{[2]} = \sigma(\boldsymbol{Z}^{[2]}) \in \mathbb{R}^{n \times d^{[2]}}$$

- Python 中的广播机制被用于计算$\boldsymbol{A}^{[1]}$ 以及$\boldsymbol{A}^{[2]}$

前向传播

1. 对于以上例子，我们有$d^{[0]}=3, d^{[1]}=4, d^{[2]}=1$

2. 因此，我们有

$$\boldsymbol{Z}^{[1]}=(\boldsymbol{b}^{[1]})^{\mathrm{T}}+\boldsymbol{X}(\boldsymbol{W}^{[1]})^{\mathrm{T}}\in\mathbb{R}^{n\times 4}$$

$$\boldsymbol{A}^{[1]}=\sigma(\boldsymbol{Z}^{[1]})\in\mathbb{R}^{n\times 4}$$

$$\boldsymbol{Z}^{[2]}=b^{[2]}+\boldsymbol{A}^{[1]}(\boldsymbol{W}^{[2]})^{\mathrm{T}}\in\mathbb{R}^{n\times 1}$$

$$\boldsymbol{A}^{[2]}=\sigma(\boldsymbol{Z}^{[2]})\in\mathbb{R}^{n\times 1}$$

前向传播

1．代价函数为

$$\mathcal{J}=-\frac{1}{n}\sum_{i=1}^{n}\left\{y_i\log a_i^{[2]}+(1-y_i)\log\left(1-a_i^{[2]}\right)\right\}$$

- $a_i^{[2]}$：对应于特征向量$\boldsymbol{x}_i$ 的“激活值”

后向传播

1. 记$\mathrm{d}\cdot = \dfrac{\partial \mathcal{J}}{\partial \cdot}$

2. 回顾: 对于一个样本点$(\boldsymbol{x}, y)$，我们有

$$\mathrm{d}b^{[2]} = a^{[2]} - y, \quad \mathrm{d}\boldsymbol{W}^{[2]} = (a^{[2]} - y)(\boldsymbol{a}^{[1]})^{\mathrm{T}}$$

$$\mathrm{d}\boldsymbol{b}^{[1]} = (a^{[2]} - y)\boldsymbol{D}(\boldsymbol{W}^{[2]})^{\mathrm{T}}, \quad \mathrm{d}\boldsymbol{W}^{[1]} = (a^{[2]} - y)\boldsymbol{D}(\boldsymbol{W}^{[2]})^{\mathrm{T}}\boldsymbol{x}^{\mathrm{T}}$$

- $\boldsymbol{D} = \mathrm{diag}(\{a_j^{[1]}(1 - a_j^{[1]}) : j = 1, \ldots, d^{[1]}\})$

3. 我们有

$$\mathrm{d}z^{[2]} = a^{[2]} - y, \quad \mathrm{d}\boldsymbol{z}^{[1]} = (a^{[2]} - y)\boldsymbol{D}(\boldsymbol{W}^{[2]})^{\mathrm{T}} = \mathrm{d}z^{[2]}\boldsymbol{D}(\boldsymbol{W}^{[2]})^{\mathrm{T}}$$

后向传播

1．此外，我们还有

$$\boldsymbol{D}(\boldsymbol{W}^{[2]})^{\mathrm{T}} = (\boldsymbol{W}^{[2]})^{\mathrm{T}} \circ \sigma'(\boldsymbol{z}^{[1]})$$

- $\sigma'(\boldsymbol{z}^{[1]}) = (a_1^{[1]}(1-a_1^{[1]}), \ldots, a_{d^{[1]}}^{[1]}(1-a_{d^{[1]}}^{[1]}))^{\mathrm{T}}$
- $\circ$：Hadamard 乘积（矩阵对应元素相乘）

2．我们有

$$\mathrm{d}z^{[2]} = a^{[2]} - y, \quad \mathrm{d}b^{[2]} = \mathrm{d}z^{[2]}, \quad \mathrm{d}\boldsymbol{W}^{[2]} = \mathrm{d}z^{[2]}(\boldsymbol{a}^{[1]})^{\mathrm{T}}$$

$$\mathrm{d}\boldsymbol{z}^{[1]} = \mathrm{d}z^{[2]}\boldsymbol{D}(\boldsymbol{W}^{[2]})^{\mathrm{T}}, \quad \mathrm{d}\boldsymbol{b}^{[1]} = \mathrm{d}\boldsymbol{z}^{[1]}, \quad \mathrm{d}\boldsymbol{W}^{[1]} = \mathrm{d}\boldsymbol{z}^{[1]}\boldsymbol{x}^{\mathrm{T}}$$

后向传播

1. 一般地，如果我们有n 个训练样本，则我们有

$$\mathrm{d}\boldsymbol{\theta} = \frac{1}{n}\sum_{i=1}^{n}\frac{\partial \mathcal{J}(y_i, a_i)}{\partial \boldsymbol{\theta}}$$

2. 因此，为了得到$\mathrm{d}\boldsymbol{\theta}$，我们只需要将对应的偏导数求均值即可

3. 利用向量化

$$\sum_{k=1}^{K} a_k b_k = \boldsymbol{a}\boldsymbol{b}$$

- $\boldsymbol{a} = (a_1, \ldots, a_K)$ 以及$\boldsymbol{b} = (b_1, \ldots, b_K)^{\mathrm{T}}$
- a_k 一个数值或者一个列向量
- b_k 一个数值或者一个行向量

后向传播

1．基于向量化，我们有

$$\mathrm{d}\boldsymbol{Z}^{[2]} = n^{-1}(\boldsymbol{A}^{[2]} - \boldsymbol{Y}) \quad \mathrm{d}\boldsymbol{b}^{[2]} = (\mathrm{d}\boldsymbol{Z}^{[2]})^{\mathrm{T}}\mathbf{1} \quad \mathrm{d}\boldsymbol{W}^{[2]} = (\mathrm{d}\boldsymbol{Z}^{[2]})^{\mathrm{T}}\boldsymbol{A}^{[1]}$$

$$\mathrm{d}\boldsymbol{Z}^{[1]} = \mathrm{d}\boldsymbol{Z}^{[2]}\boldsymbol{W}^{[2]} \circ \sigma'(\boldsymbol{Z}^{[1]}) \quad \mathrm{d}\boldsymbol{b}^{[1]} = (\mathrm{d}\boldsymbol{Z}^{[1]})^{\mathrm{T}}\mathbf{1} \quad \mathrm{d}\boldsymbol{W}^{[1]} = (\mathrm{d}\boldsymbol{Z}^{[1]})^{\mathrm{T}}\boldsymbol{X}$$

- $\sigma'(z)$：对应于$\sigma(z)$ 的偏导数
- $\sigma'(\boldsymbol{Z}^{[1]})$：利用广播机制，计算$\boldsymbol{Z}^{[1]}$ 每个元素对应的导数值

批量梯度下降法

步骤1. 随机初始化$\boldsymbol{\theta}^{(0)}$

步骤2. 基于$\boldsymbol{\theta}^{(t)}$ 计算

$$\nabla \mathcal{J}\left(\boldsymbol{\theta}^{(t)}\right)=\frac{\partial \mathcal{J}}{\partial \boldsymbol{\theta}}\left(\boldsymbol{\theta}^{(t)}\right)$$

步骤3. 更新参数

$$\boldsymbol{\theta}^{(t+1)}=\boldsymbol{\theta}^{(t)}-\alpha \nabla \mathcal{J}\left(\boldsymbol{\theta}^{(t)}\right)$$

步骤4. 回到步骤 2 直至收敛