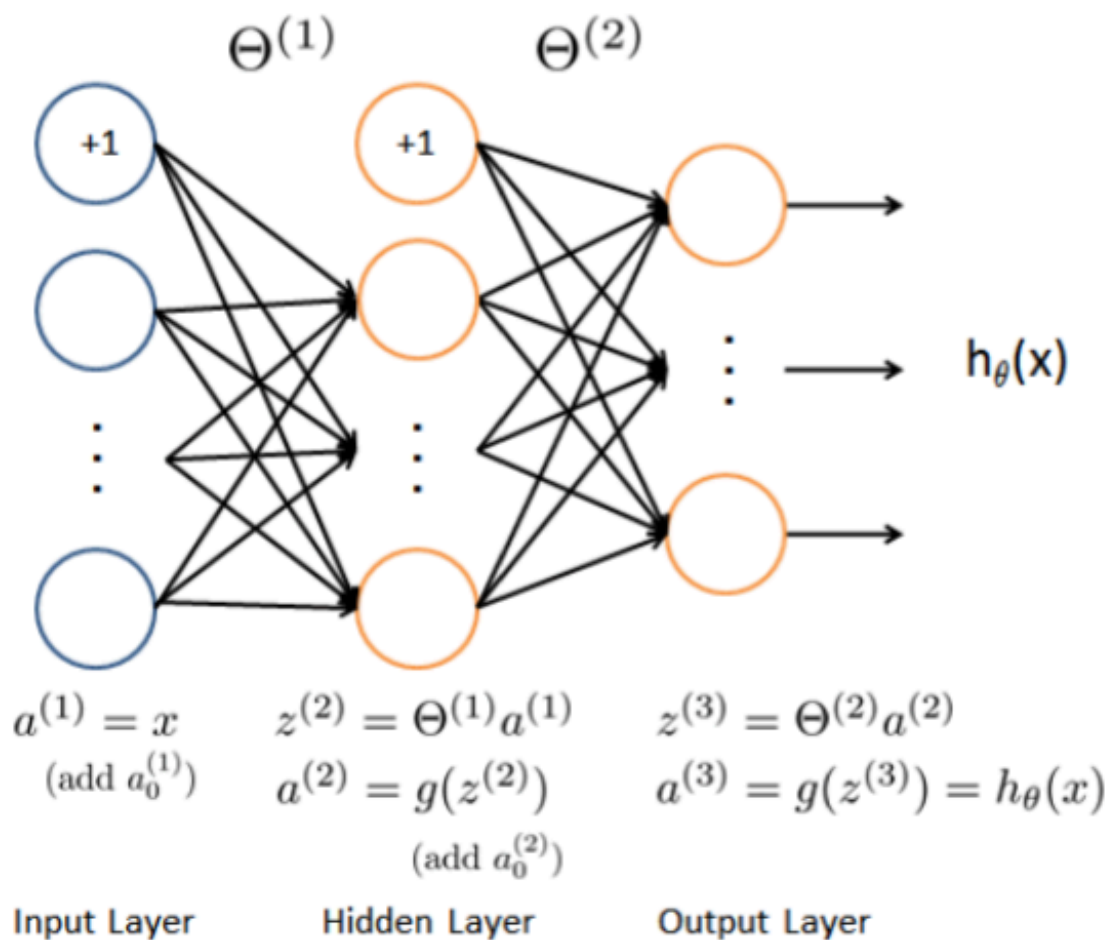


神经网络公式推导

前向传播：

网络结构如图所示：



前向传播公式推导

- 符号含义说明：
 - $z^{(l)}$ 表示第 l 层网络激活函数的输入
 - $a^{(l)}$ 表示第 l 层激活函数的输出
 - $\Theta^{(l)}$ 表示第 l 层与第 $l+1$ 层之间的参数
- 以四层神经网络为例：

1. 输入层

$$a^{(1)} = x$$

2. 第二层

$$z^{(2)} = \Theta^{(1)} a^{(1)}$$

$$a^{(2)} = g(z^{(2)})$$

3. 第三层

$$z^{(3)} = \Theta^{(2)} a^{(2)}$$

$$a^{(3)} = g(z^{(3)})$$

4. 输出层

$$z^{(4)} = \Theta^{(3)} a^{(3)}$$

$$h_{\theta}(x) = a^{(4)} = g(z^{(4)})$$

参数推导说明:

假设l层有m个神经元，l+1层有n个神经元。对于第l+1层来说， $z^{(l+1)}$ 计算用到的参数 $\Theta^{(l)}$ 的维度为n*(m+1)，加一的原因是要考虑到每一层的偏置。矩阵表示如下：

$$\begin{bmatrix} \Theta_{0,1} & \Theta_{1,1} & \dots & \Theta_{m,1} \\ \Theta_{0,2} & \Theta_{1,2} & \dots & \Theta_{m,2} \\ \dots & \dots & \dots & \dots \\ \Theta_{0,n} & \Theta_{1,n} & \dots & \Theta_{m,n} \end{bmatrix}$$

反向传播推导:

主要思想:

1. 计算每一层的误差项 δ^l : 含义为 $\frac{\partial J}{\partial z^l}$, 即损失函数到当前层l输入的导数
2. 计算每一层输入对每个参数的导数: 含义为 $\frac{\partial z^l}{\partial \Theta_{j,i}^l}$, 其中 $\Theta_{j,i}^l$ 表示前一层的第j个单元与当前层i单元的
参数
3. 综合1、2, 得到 Θ^l 的导数 $\frac{\partial J}{\partial \Theta^l}$ 为: $\delta^l a^{(l-1)T}$ (默认求导为分母布局, $\frac{\partial J}{\partial \Theta^l}$ 的维度为(n* m+1); δ^l 维度为(n,1), $a^{(l-1)T}$ 为(1,m+1))

关于 δ^l 的推导过程:

已知 $\delta^{l+1} = \frac{\partial J}{\partial z^{l+1}}$, 并且已知 $\delta^l = \frac{\partial J}{\partial z^{l+1}} \cdot \frac{\partial z^{l+1}}{\partial z^l}$ 那么只需要求出 $\frac{\partial z^{l+1}}{\partial z^l}$ 即可。根据前向传播的公式推导, 我们很容易得到 z^{l+1} 与 z^l 之间的关系:

$$z^{l+1} = \Theta^l \cdot g(z^l)$$

那么很容易求得 $\frac{\partial z^{l+1}}{\partial z^l}$:

$$\frac{\partial z^{l+1}}{\partial z^l} = \Theta^l \cdot \text{diag}\left(\frac{\partial g(z^l)}{\partial z^l}\right)$$

其中, $\frac{\partial z^{l+1}}{\partial z^l}$ 维度为(n * m, 向量对向量求导遵循分子布局); Θ^l 维度为 n * m; $\text{diag}\left(\frac{\partial g(z^l)}{\partial z^l}\right)$ 维度为 m*m, 由于激活函数指针对单个变量进行计算, 所以求导后的矩阵是个对角矩阵。

由此可以推导出 δ^l :

$$\delta^l = \frac{\partial J}{\partial z^{l+1}} \cdot \frac{\partial z^{l+1}}{\partial z^l} = \frac{\partial z^{l+1}{}^T}{\partial z^l} \cdot \delta^{l+1} = \text{diag}\left(\frac{\partial g(z^l)}{\partial z^l}\right) \cdot \Theta^{lT} \cdot \delta^{l+1} = \Theta^{lT} \cdot \delta^{l+1} \odot \frac{\partial g(z^l)}{\partial z^l}$$

其中, Θ^l 维度为 $n * m$; δ^{l+1} 维度为 $n * 1$; 则前两项的结果为 $m * 1$; 而 $\text{diag}(\frac{\partial g(z^l)}{\partial z^l})$ 的维度为 $m * m$, 由于它是对角矩阵, 所以最终的结果可以看作 $m * 1$ 维对角元素与 前两项结果的 Hadamard 积。