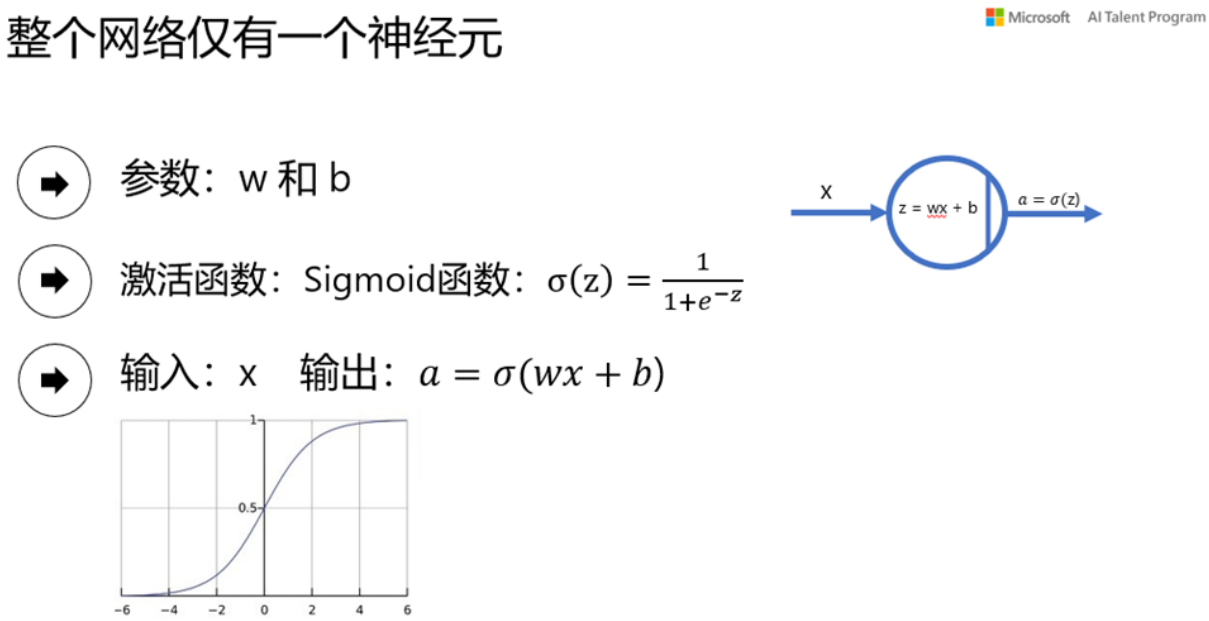
神经网络的推理和训练

该课程主要为大家讲授如下的内容：

- 全连接神经网络的前向传播计算
- 全连接神经网络的反向传播计算
- Sigmoid激活函数
- 梯度消失与梯度爆炸问题
- ReLU函数
- 深度学习的基本概念

1. 全连接神经网络的前向传播计算
   全连接神经网络的前向传播计算逐层进行。每一层的计算可以记为：
   $\alpha = \sigma(z), z = \omega x + b$



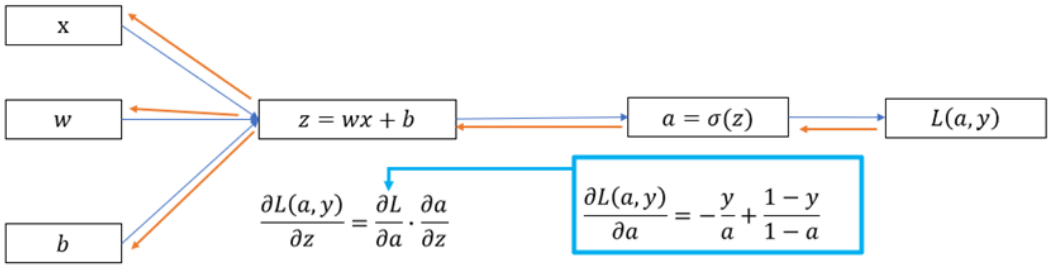
2. 全连接神经网络的反向传播计算
   在全连接神经网络的反向传播计算过程中，首先通过损失函数求取输出y与期望值a的距离，损失函数记为
   $L(a, y)$
   然后根据损失函数求取模型的参数在当前状态下与最优解：
   $w, b = argminL(a, y)$
   梯度，记为：
   $dw = x \cdot dz, db = dz | dz = \partial L(a, y)/\partial z$

最后反向传播梯度至第一层隐层，以链式求导的方式逐层向后求取权值和偏置的梯度，对第k层的梯度记为：

$$dw_k = x_k \cdot dz_k, db_k = dz_k | dz_k = \frac{\partial a_k}{\partial z_k} \cdot \frac{\partial z_{k+1}}{\partial a_k} \cdot \frac{\partial L(a,y)}{\partial z_{k+1}}$$

单神经元、单个数据的反向传播

Microsoft AI Talent Program



$$\frac{\partial L(a,y)}{\partial z} = \frac{\partial L}{\partial a} \cdot \frac{\partial a}{\partial z}$$

$$\frac{\partial L(a,y)}{\partial a} = -\frac{y}{a} + \frac{1-y}{1-a}$$

$$= \left(-\frac{y}{a} + \frac{1-y}{1-a}\right) \cdot a(1-a)$$

$$= a - y$$

令：$dz = \frac{\partial L(a,y)}{\partial z} = a - y$，

$$\frac{\partial L}{\partial w} = x \cdot \frac{\partial L(a,y)}{\partial z}$$

$$\frac{\partial L}{\partial b} = \frac{\partial L(a,y)}{\partial z}$$

$dw = \frac{\partial L}{\partial w}$，$db = \frac{\partial L}{\partial b}$

有：$dw = x \cdot dz$，$db = dz$

(c)Microsoft 2022, 仅供个人学习使用

这个迭代的过程可以如下表示。

梯度下降

Microsoft AI Talent Program

1.随机初始化w, b

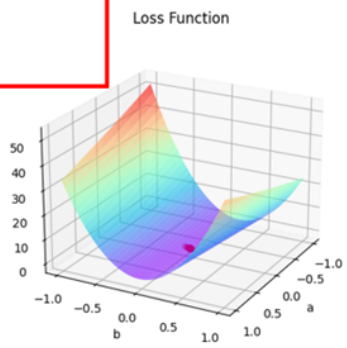
2.设置学习率γ，这是一个超参数

3.利用样本数据计算 $\mathrm{dw} = \frac{\partial \mathrm{L(w,b)}}{\partial \mathrm{w}}$ 和 $\mathrm{db} = \frac{\partial \mathrm{L(w,b)}}{\partial b}$

4.更新 $\mathrm{w}_{new} = \mathrm{w}_{old} - \gamma * dw$

$b_{new} = b_{old} - \gamma * db$

5. 回到步骤3



(c)Microsoft 2022, 仅供个人学习使用

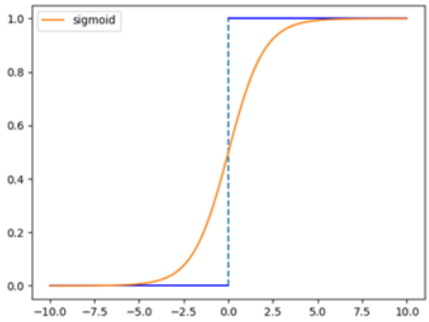
3. Sigmoid激活函数

Sigmoid函数是一种常见的激活函数，取值在(0,1)。

Sigmoid 函数的优缺点

Microsoft AI Talent Program

- 优点
  - 值域有限: (0 , 1)
  - 单调连续并且求导容易
- 缺点
  - 幂运算提高了计算成本
  - 导数特性将带来梯度消失
    - Sigmoid函数满足:$\sigma'(x)=\sigma(x)(1-\sigma(x))$



4. 梯度消失与梯度爆炸问题

梯度消失和梯度爆炸是深层神经网络中常见的问题。神经网络的层数越多，这两个问题就越容易出现。

梯度消失与梯度爆炸

Microsoft AI Talent Program

	同源异构的两个问题	不良后果
梯度消失	对激活函数求梯度，梯度较小，随着层数增多，梯度更新将以指数形式衰减，即发生**梯度消失**	模型失去学习的能力，无法有效地更新权重，甚至梯度归零，模型停滞
梯度爆炸	对激活函数求梯度，梯度很大，随着层数增多，梯度更新将以指数形式增长，即发生**梯度爆炸**	模型权重参数变化很大，导致网络不稳定，甚至权重值溢出（NaN）

这种求导结果指数级的衰减现象

(c)Microsoft 2022, 仅供个人学习使用

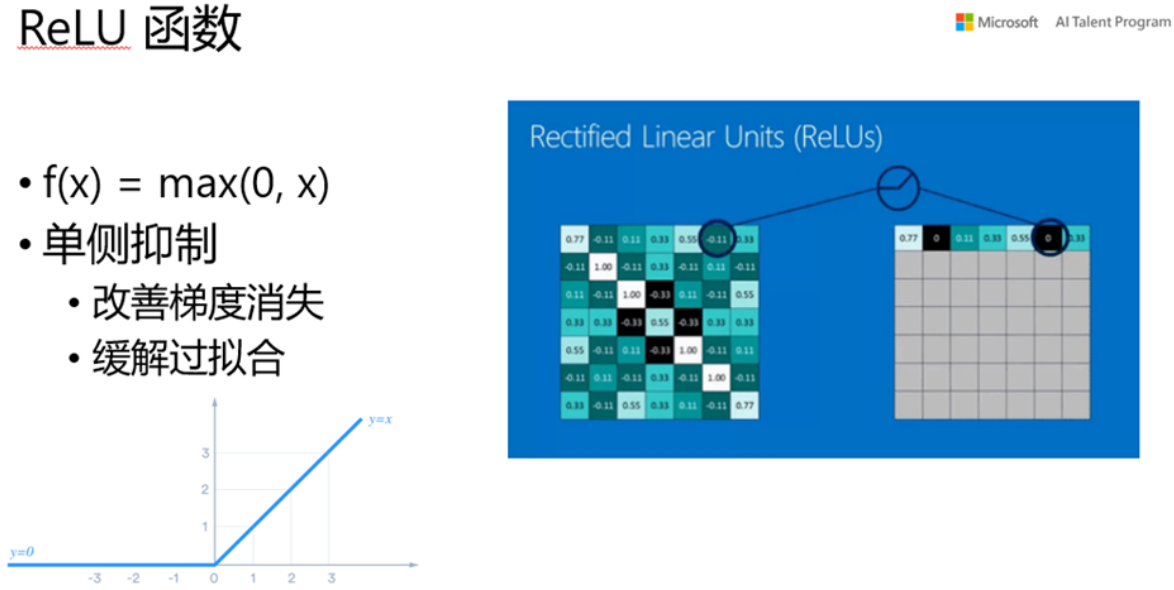
为了应对梯度消失与梯度爆炸问题的问题，研究者们给出了如下的解

决方法。

	同源异构的两个问题	不良后果
梯度消失	对激活函数求梯度，梯度较小，随着层数增多，梯度更新将以指数形式衰减，即发生**梯度消失**	模型失去学习的能力，无法有效地更新权重，甚至梯度归零，模型停滞
梯度爆炸	对激活函数求梯度，梯度很大，随着层数增多，梯度更新将以指数形式增长，即发生**梯度爆炸**	模型权重参数变化很大，导致网络不稳定，甚至权重值溢出（NaN）

5. ReLU函数

ReLU函数也是一种常见的激活函数。它的计算比起Sigmoid更简单，在同等条件下消耗的算力更少。但是ReLU函数对初始值更敏感。



6. 深度学习的基本概念

深度学习是通过加深神经网络的层数来自动提取数据特征（表征学习）的一种算法。这种算法能够大量节省知识依赖和人力，有效提取

关键特征。

深度学习

Microsoft AI Talent Program

- 以深度神经网络为架构
- 对数据进行表征学习
- 表征学习：学习如何学习
  - 将原始数据转换成为能够被计算机用来有效开发的一种形式
  - 避免了手动提取特征的麻烦
  - 允许计算机学习使用特征的同时，也学习如何提取特征