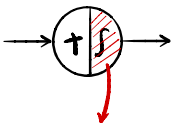
激活函数（Activation Function）

(1) 激活函数的本质



将加权求和后的值，通过非线性关系映射到新的值域上
从而实现非线性化

⇓

人工神经网络的本质，是用多个非线性函数的嵌套来逼近真实的非线性函数

↓

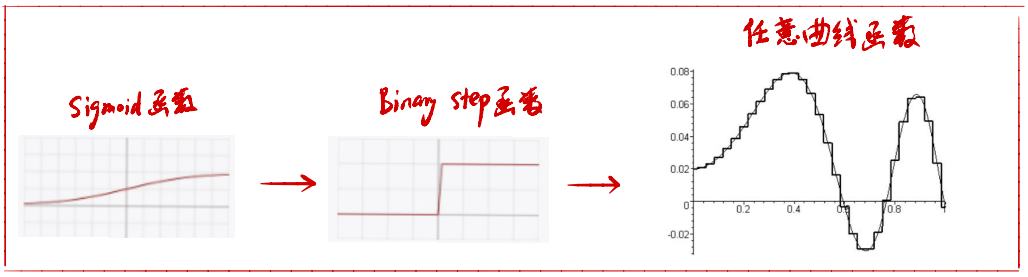
① 对非线性可分的数据进行分类
② 对维度间有着复杂关系的数据进行预测

⇓

那么，用"多个简单的非线性函数来模拟真实世界中的各种复杂函数"的理论依据是什么？

↓

万能逼近定理
（Universal approximation theorem）



以Sigmoid函数为例，它可以近似表达Binary Step函数
而多个Binary Step函数可以模拟任意曲线

Q：为什么不直接使用Binary Step函数？
A：因为Binary Step函数不可导，无法求梯度

Q：在一个Hidden Layer中使用足够多的激活函数不就可以了吗？为什么还要构建多层神经网络？
A：单层网络（即：仅有一个Hidden Layer）

↓ 若要表示复杂函数

大量参数

↓

参数过多，导致
- 无法学习到正确的值（没学到）
  或者
- 容易过拟合，难以泛化（学到）

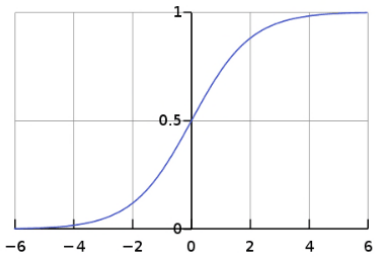
→ 其实，多层网络也存在着同样的问题

↓

但可以用Eary Stop方法或Drop Out方法进行处理

(1) 常用激活函数及它们的优缺点

① Sigmoid函数

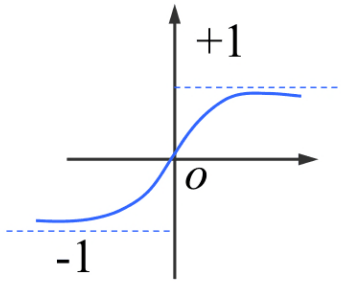


优点：值域在 (0,1)，可以直接作为概率值使用

缺点：当 $s \le -6$ 或 $s \ge 6$ 时，Sigmoid函数的导数几乎为零

当多层神经网络求梯度下降时，其后的Hidden Layer的权重将无法更新

② 双曲正切函数（Tanh）

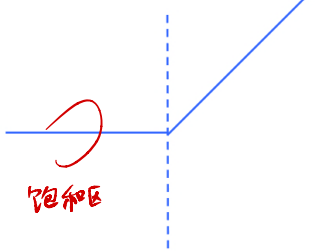


— 优、缺点与Sigmoid一样

— 值域在 [-1,1]

③ Rectified Linear Unit (ReLU)

$$g(s)=\begin{cases} s & s \ge 0 \\ 0 & s < 0 \end{cases}$$



优点：

因为梯度不会消失，所以网络层数再多也不怕！

缺点：

① 落入此区域，导致对应的权重无法更新，称为"神经元的死亡"

② 输出值的平均值 > 0

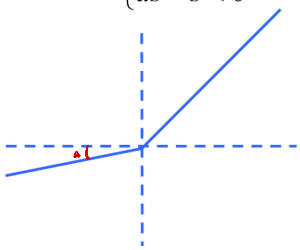
⇓

[ReLU的3个变种]

因为"饱和区"的梯度消失造成神经元死亡 → 那么，我们给饱和区留有一个微小的梯度，不就可以解决这个问题了吗？这种方法，被称为LeakyReLU

超参数a，手工设定

$$g(s)=\begin{cases} s & s \ge 0 \\ as & s < 0 \end{cases}$$



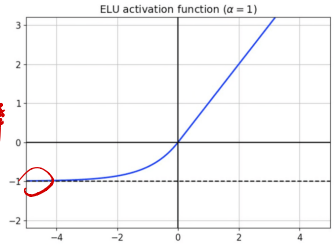
↓

让a的值通过学习得到，随着数据的变化而变化

↓

PReLU (Parametric ReLU)

对饱和区的另一种改造（下图），称为 Exponential Linear Units（ELU）



$$g(x)=\begin{cases} x, & x\geq 0 \\ \alpha(e^x-1), & x<0 \end{cases}$$

缺点：有指数运算，计算量大

② 对 <0 的异常值，鲁棒性有所增强。（因为值域被限制在 [0,-1]）

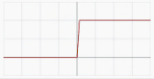
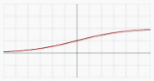
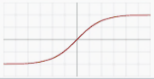
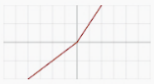
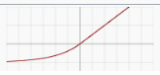
↓

与 LeakyReLU 的不同在于，其均值在零附近

↓

优点：① 收敛速度快

(3) 附录

Name	Plot	Equation	Derivative (with respect to x)	Range
Identity		$f(x)=x$	$f'(x)=1$	$(-\infty,\infty)$
Binary step		$f(x)=\begin{cases}0 & \text{for } x<0\\ 1 & \text{for } x\geq 0\end{cases}$	$f'(x)=\begin{cases}0 & \text{for } x\neq 0\\ ? & \text{for } x=0\end{cases}$	$\{0,1\}$
Logistic (a.k.a. Sigmoid or Soft step)		$f(x)=\sigma(x)=\frac{1}{1+e^{-x}}$[1]	$f'(x)=f(x)(1-f(x))$	$(0,1)$
TanH		$f(x)=\tanh(x)=\frac{(e^x-e^{-x})}{(e^x+e^{-x})}$	$f'(x)=1-f(x)^2$	$(-1,1)$
Rectified linear unit (ReLU)[15]		$f(x)=\begin{cases}0 & \text{for } x<0\\ x & \text{for } x\geq 0\end{cases}$	$f'(x)=\begin{cases}0 & \text{for } x<0\\ 1 & \text{for } x\geq 0\end{cases}$	$[0,\infty)$
Leaky rectified linear unit (Leaky ReLU)[17]		$f(x)=\begin{cases}0.01x & \text{for } x<0\\ x & \text{for } x\geq 0\end{cases}$	$f'(x)=\begin{cases}0.01 & \text{for } x<0\\ 1 & \text{for } x\geq 0\end{cases}$	$(-\infty,\infty)$
Parameteric rectified linear unit (PReLU)[18]		$f(\alpha,x)=\begin{cases}\alpha x & \text{for } x<0\\ x & \text{for } x\geq 0\end{cases}$	$f'(\alpha,x)=\begin{cases}\alpha & \text{for } x<0\\ 1 & \text{for } x\geq 0\end{cases}$	$(-\infty,\infty)$[2]
Exponential linear unit (ELU)[20]		$f(\alpha,x)=\begin{cases}\alpha(e^x-1) & \text{for } x\leq 0\\ x & \text{for } x>0\end{cases}$	$f'(\alpha,x)=\begin{cases}f(\alpha,x)+\alpha & \text{for } x\leq 0\\ 1 & \text{for } x>0\end{cases}$	$(-\alpha,\infty)$