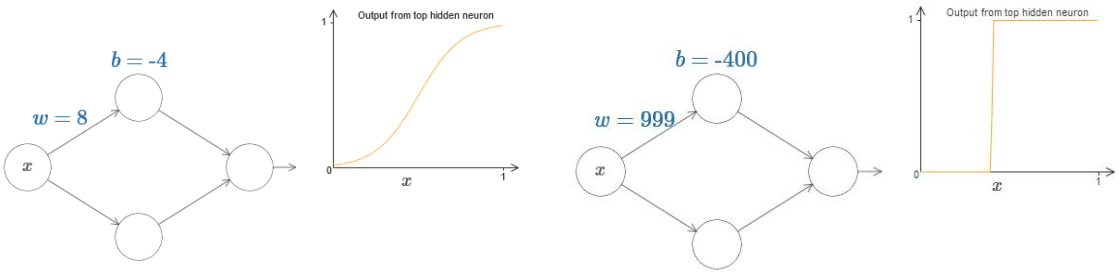
# 1 神经网络的可解释性

神经网络可解释性仍然是一个未被解决的问题，这里主要从神经网络可以拟合任意函数来解释。

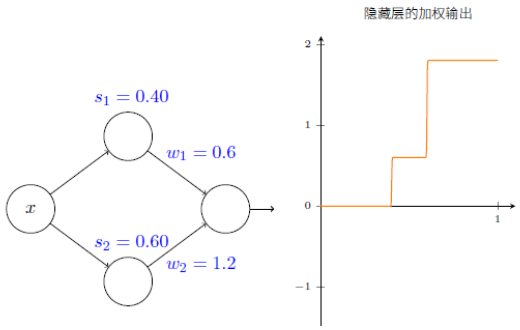
## 1.1 函数拟合的观点

神经网络能够拟合任意函数，其输出可以尽可能地接近被拟合函数的输出，但是不能精确计算出该函数的输出。只要我们拥有足够的参数，我们可以以足够的精度拟合函数。这表明神经网络具有一种普遍性，我们可以对任意所需要的目标函数进行拟合。应该注意的是，神经网络的输入是连续值，因此理论上神经网络不能拟合不连续的函数，但是连续的近似实际上也足够好，因此这通常不是一个严重的限制。

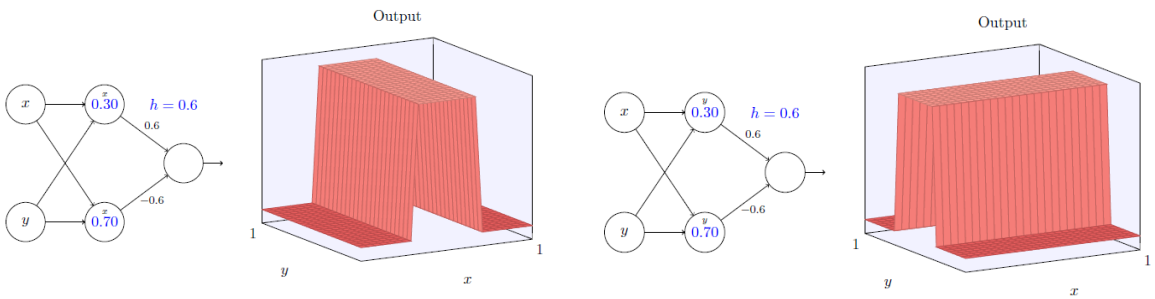


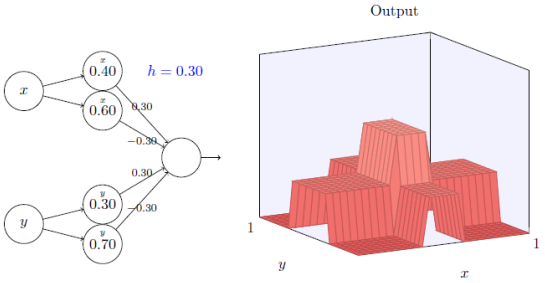

如上图所示，这是一个隐藏层含有两个神经元的普通全连接神经网络，使用 sigmoid 激活函数的神经网络的一个权重进行改变，便可以模拟阶跃函数。隐藏层的激活函数是 $\sigma(x) = \frac{1}{1+e^{-x}}$，则隐藏层的输出便是 $\sigma(wx+b)$。改变权重 W 便可以改变激活函数的输出形状，改变偏置 b 便可以改变输出图像的水平方位。

由以上两张图片我们可以知道，一个神经元配合 sigmoid 激活函数可以拟合任意形状任意位置的阶跃函数，将两个具有不同的权重 W 和不同的偏置 b 的神经元的输出进行相加，便可以拟合出矩形脉冲函数或者其他函数，如图所示




因此，若将更多的具有不同参数的神经元的输出进行叠加，我们可以近似拟合更多的函数，增加更多的神经元，我们便可以拟合任意函数。当我们继续增加隐藏层神经元的数量，我们便可以拟合任意函数。同理，对多输出函数我们也可以进行任意拟合，如图所示。






以上的例子只是简单地使用了少数的神经元，因此看起来比较粗糙，但是，我们可以通过增加隐藏层神经元个数来更加精确地模拟所需要的函数，使得其形状更为精细，这就像微积分的思想，用无限小的矩形对曲线与坐标轴的面积进行近似。

在原理上，我们知道神经网络可以拟合任意函数，但是能够拟合任意函数并不代表我们能够直接训练神经网络得到一个相应的函数。