机器学习中有个非常有意思的问题，叫做过拟合，overfit，意思是对于训练集合中的数据产生了过大的依赖，而在测试集上的结果产生了很大的偏差。

导致过拟合的一种典型的问题，就是数据平衡。

数据分布相对于判断阈值的位置，这个位置如果偏差较小，就是数据平衡的，反之，训练数据大量偏离判断阈值，则是数据不平衡。

ы тр[ез Тот СО 
Тот ст 

详细公式和推导可以参考文献，这里简单说明。上面是两个高斯分布（mentioned in ELEC2600），一个均值0，一个均值2，用来训练正负判断网络。

那么有人生经验的同学肯定知道，均值0训练的神经网络，由于训练集比较合理，输出结果错误的可能性就比较小了；而对于右边的均值2的数据，由于90%的数据都是正数，神经网络权值向容易产生正数判断的方向倾斜。

这非常容易理解，如果你看到的90%的数都是正数，你就是全猜正，也有90%的可能性是对的，而猜对它是负数的准确性，就大大降低了。

这就是数据集影响神经网络认知的原理。

参考文献：

<https://zhuanlan.zhihu.com/p/56960799>