

习新手都忍不住尝试很多不同的算法来进行改进。然而，往往收集更多的数据比改进学习算法要见效得多。

怎样判断是否要收集更多的数据？ 首先，确定训练集上的性能是否可接受。 如果训练集上的性能差，学习算法还不能在训练集上学习出良好的模型，那么就没必要收集更多的数据。 反之，可以尝试增加更多的网络层或每层增加更多的隐藏单元，增加模型的规模。 此外，也可以尝试调整学习率等超参数来改进学习算法。 如果更大的模型和仔细调试的优化算法没有效果，那么问题可能源自训练数据的质量。 数据可能含太多噪扰，或是可能不包含预测输出所需的正确输入。这意味着需要重新开始，收集更干净的数据或是收集特征更丰富的数据集。

② 如果训练集上的性能是可接受的，那么度量测试集上的性能。 如果测试集上的性能也是可以接受的，那么就顺利完成了。 如果测试集上的性能比训练集的要差得多，那么收集更多的数据是最有效的解决方案之一。



