

deeplearning.ai笔记（一）

1, 训练、验证、测试集的划分

数据量较小时

无验证集的情况：70% / 30%

有验证集的情况：60% / 20% / 20%

数据量大：

100万数据量：98% / 1% / 1%

超百万数据量：99.5% / 0.25% / 0.25%（或者99.5% / 0.4% / 0.1%）

（注：

建议验证集和测试集来自于同一个分布，这样可以使得机器学习算法变得更快；如果不需要用无偏估计来评估模型的性能，则可以不需要测试集。）

2, 解决High bias 和High variance 的过程

1) 是否存在**High bias**？

增加网络结构，如增加隐藏层数目；

训练更长时间；

寻找合适的网络架构，使用更大的NN结构；

2) 是否存在**High variance**？

获取更多的数据；

正则化；

寻找合适的网络结构；

3, 正交化

1) 系统在训练集上表现的好

否则，使用更大的神经网络、更好的优化算法

2) 系统在开发集上表现的好

否则，使用正则化、更大的训练集

3) 系统在测试集上表现的好

否则，使用更大的开发集

4) 在真实的系统环境中表现的好

否则，修改开发测试集、修改代价函数

4, 满足和优化指标

有时对于某一问题，对模型的效果有一定的要求，如要求模型准确率尽可能的高，运行时间在100ms以内。这里以 Accuracy 为优化指标，以 Running time 为满足指标，我们可以从中选出B是满足条件的最好的分类器。

如果要考虑N个指标，则选择一个指标为优化指标，其他N-1个指标都是满足指标

5, 改善模型的表现

1) 基本假设：

模型在训练集上有很好的表现；

模型推广到开发和测试集啥会给你也有很好的表现。

2) 减少可避免偏差

训练更大的模型

训练更长时间、训练更好的优化算法（Momentum、RMSprop、Adam）

寻找更好的网络架构（RNN、CNN）、寻找更好的超参数

3) 减少方差

收集更多的数据

正则化 (L2、dropout、数据增强)

寻找更好的网络架构 (RNN、CNN)、寻找更好的超参数