deeplearning.ai笔记（二）

1，训练集和开发集、测试集来自于不同的分布怎么做

（eg：从网上获取大量的高清晰的猫的图片去做分类，如200000张，但是只能获取少量利用手机拍摄的不清晰的图片，如10000张。但是我们系统的目的是应用到手机上做分类。）

训练集均是来自网上下载的20万张高清图片，也可以加上5000张手机非高清图片；对于开发和测试集都是手机非高清图片。
好处：开发集全部来自手机图片，瞄准目标；
坏处：训练集和开发、测试集来自不同的分布。
从长期来看，这样的分布能够带来更好的系统性能。

2，**方差和分布原由分析**

若我们模型的误差为：
Training error： 1%
Dev error：10%
那么我们如何去确定是由于分布不匹配的问题导致开发集的误差，还是由于算法中存在的方差问题所致？

设立Training-dev dataset：
如果最终，我们的模型得到的误差分别为：
Training error：1%
Training-dev error： 9%
Dev error： 10%
那么，由于**训练开发集**尽管和训练集来自同一分布，但是却有很大的误差， 模型无法泛化到同分布的数据，那么说明我们的模型存在**方差问题**。

但如果我们的模型得到的误差分别为：
Training error： 1%
Training-dev error：1.5%
Dev error： 10%
那么在这样的情况下，我们可以看到，来自同分布的数据，模型的泛化能力强，而开发集的误差主要是来自于**分布不匹配**导致的。

3， 解决数据分布不匹配问题

1）进行人工误差分析，尝试去了解训练集和开发测试集的具体差异在哪里。如：噪音等；
2）尝试把训练数据变得更像开发集，或者收集更多的类似开发集和测试集的数据，如增加噪音；

4，多任务学习

（eg：自动驾驶）在自动驾驶的例子中，我们需要检测的物体很多，如行人、汽车、交通灯等等。对于现在的任务，我们的目标值变成了一个向量的形式向量中的每一个值代表检测到是否有如行人、汽车、交通灯等，一张图片有多个标签。