deeplearning.ai

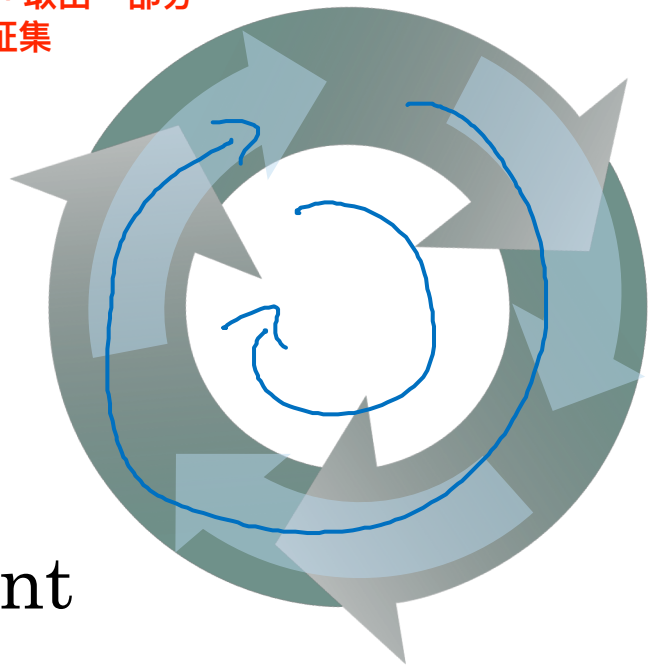
Setting up your ML application

Train/dev/test sets

Applied ML is a highly iterative process

机器学习的应用是相当反复的 迭代的过程 你只需要将这个循环进行许多次 就有希望能为你的应用中的网络找出好的参数 所以有一件事能决定你能多快地取得进展 那就是你进行迭代过程时的效率 而恰当地将你的数据集分为训练集 开发集和测试集 就能让你的迭代效率更高 假设这是你的训练数据 我们把它画成一个大矩形 那么传统的做法是你可能会从所有数据中 取出一部分用作训练集 然后再留出一部分作为hold-out交叉验证集

- # layers
- # hidden units
- learning rates
- activation functions
- ...

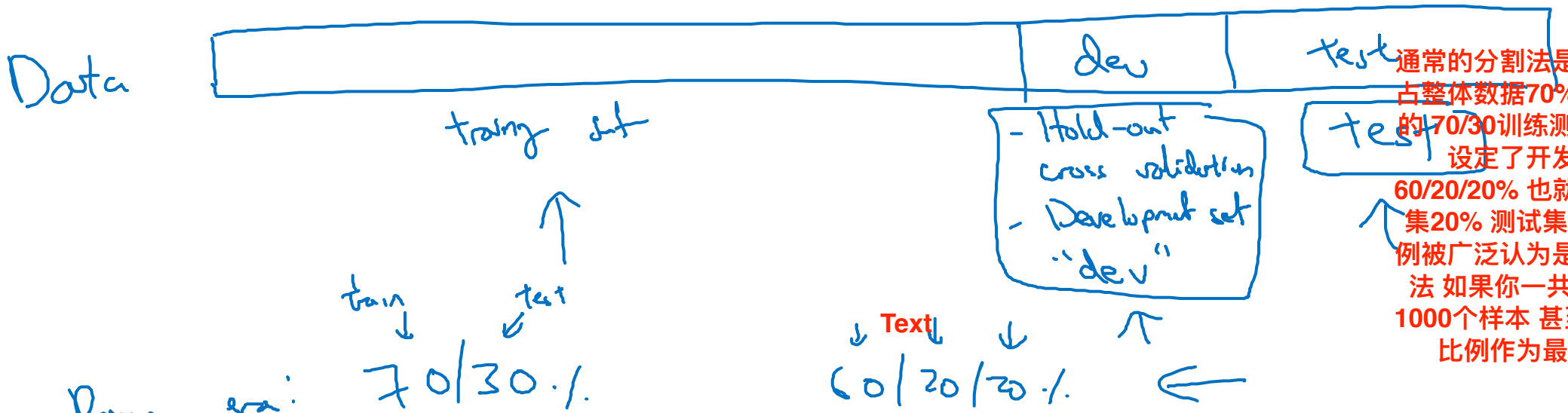


NLP, Vision, Speech, Structured data

Ads Search Security logistic ....

Train/dev/test sets

所以开发集只要足够大到 能够用来在评估两种不同的算法 或是十种不同的算法时快速选出较好的一种 达成这个目标可能不需要多达20%的数据 所以如果你有100万个训练样本 可能开发集只要1万个样本就足够了 足够用来评估两种算法中哪一种更好 与开发集相似 测试集的主要功能是 对训练好的分类器的性能 给出可信度较高的评估



通常的分割法是 训练集和测试集分别占整体数据70%和30% 也就是人们说的70/30训练测试分割 如果你明确地设定了开发集 那比例可能是60/20/20% 也就是测试集占60% 开发集20% 测试集20% 数年以前这个比例被广泛认为是 机器学习中的最佳方法 如果你一共只有100个样本 也许1000个样本 甚至到1万个样本时 这些比例作为最佳选择都是合理的

我还看见过一些应用 这些应用中样本可能多于100万个 分割比率可能会变成99.5/0.25/0.25% 或者开发集占0.4% 测试集占0.1% 所以总结起来 当设定机器学习问题时 我通常将数据分为训练集 开发集和测试集 如果数据集比较小 也许就可以采用传统的分割比率 但如果数据集大了很多 那也可以使开发集 和测试集远小于总数据20% 甚至远少于10%

Big data: 1,000,000 10,000 10,000

98 / 1 / 1 %

99.5 / .4 / .1 %

Andrew Ng

Mismatched train/test distribution

Cats

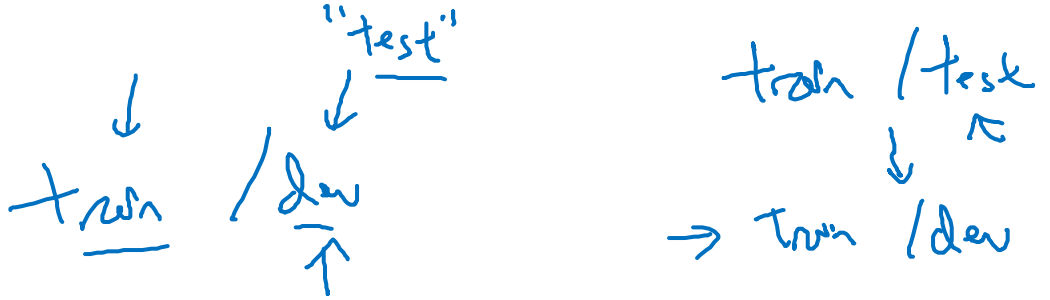
当前的深度学习中还有一个趋势是 有越来越多的人的训练集与测试集的数据分布不匹配

The rule of thumb I'd encourage you to follow in this case is to make sure that the dev and test sets come from the same distribution.

Training set:
Cat pictures from webpages

Dev/test sets:
Cat pictures from users using your app

→ Make sure dev and test come from same distribution.



Not having a test set might be okay. (Only dev set.)