机器学习策略上

http://blog.csdn.net/red\_stone1/article/details/78519599

改善深度学习模型的方法很多，哪个最有效不确定，调试时间和算力成本有很高，因此正确的调试策略应当考虑

正交化调试方法，每次只调试一个参数（超参数），保持其他参数不变，来观察模型某一性能的改变。（生物学中单一变量原理似乎哪里都适用）

Early stop不推荐，因为不符合正交化调试原则

评估指标可以分为优化指标和满意指标。前者力求最好，后者达标就行。

训练集，交叉验证集和测试集应当来自同一分布。即使不是，混合后再随机分配也比完全不处理好。其实更常用效果更好的方法，是将train set与dev/test set一部分混合作为新train set，剩下dev/test set继续作为验证。原因本篇下章会讲。

贝叶斯误差，人类（转件或普通人，根据应用要求而定）能够达到的最低误差水平。

Avoidable bias用于定义模型**训练集**误差与贝叶斯误差之间的差值。反应欠拟合情况

Variance bias用于表达模型交叉验证集或测试集误差与训练集误差的差值。反映了过拟合情况

解决avoidable bias和variance bias的常见方法



机器学习模型接近人类表现是相对容易的，但是超越却很难。

机器学习策略下

http://blog.csdn.net/red\_stone1/article/details/78600255

误差分析

追踪具体错误原因，以及其在所有错误中的比例。该比例的高低，决定了解决该类错误可以带来的性能改善天花板

样本错误标记问题。如果是正态分布的错误标记，影响不大。但如果是带有明显偏向特征的错误，需要费力气纠正。

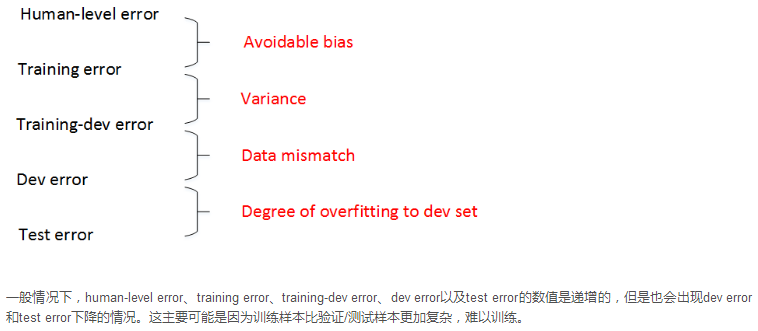
如何解决train set和dev/test set来自不同分布

一种方法是将train set和dev/test set完全混合，然后随机分配为新的train set和dev/test set。优点是实现了分布一致，缺点是新dev/test set中，原始train set比重太大，测试结果不太能反应最终场景，通常表现一般。

二种方法是**将train set和一部分dev/test set混合为新的train set，剩下的dev/test set继续作为dev set和test set**。最为常用，表现更好

在train set和dev set明显分布不一致（比如图像清晰度差别明显）时，可能出现dev error明显比train error大，却不是模型variance问题的情况。排除方法是设置train-dev set，从原来的train set中分割出一部分作为train-dev set。其不用于训练，而是和dev set一起用于验证

各样本集合各误差该如何分析，data mismatch是指dev set和train set分布明显不同



迁移学习transfer learning，将已经训练好模型的一部分知识（知识即网络结构和参数）应用到另一类似场景中。根据拥有的样本数据量，选择再训练该模型的最后一层或几层

多任务学习，单一神经网络同时完成多个任务。

端到端学习，即使用单个神经网络实现从样本数据到最终目标的映射。

能够使用端到端，很多时候取决于可获取样本数量及其代价。如果样本少，代价高，就需要经过中间过程处理。

