结构化机器学习项目

为什么要结构化机器学习项目

能够快速高效的优化学习系统
- 能够少走弯路
- 系统分的分析错误

如何结构化机器学习项目

总体思路，快速建立一个简单的系统，慢慢的迭代优化，不要一开始就构建一个庞大的系统，这样可能会大幅降低效率。

设置单一数字评估指标
（类似设置一个综合评分，通过一个综合评分能够非常简单的判定出哪个机器学习系统更好）
- 制定需要的指标，自定义合理的指标，能够表现想要的结果。
  - 比如修改损失函数的权重等

要保证的是目标一致，否则南辕北辙

设置合理的训练集/开发集/测试集
- 设置合理的比例
  - 深度学习由于数据很大，通常使用一小部分作为开发/测试集，大部分用于训练集
  - 传统机器学习采用百分比方式

与人类的表现进行对比
- 如果机器学习<人类
  - 获取更多带标签的数据
  - 人工误差分析
  - **当机器学习的结果非常接近人类水平时，这个分析就不那么准确了**
  - 分析偏差/方差
    - 有哪些基础误差
      - 训练集误差
      - 人类误差
      - 开发集误差
    - 偏差
      - 人类误差-训练集误差
        - 减小偏差的方法
          - 选择更大的模型
          - 选择更好的优化器
            - 例如Adam，RMSpop等
          - 更换深度学习的结构
            - 例如RNN等
    - 方差
      - 训练集误差-开发集误差
        - 减小方差的方法
          - 更多的训练数据
          - 正规化
            - l2
            - dropout
          - 寻找更好的神经网络参数
  - **通过偏差与方差进行对比，找到主要问题，看哪个进步空间更大，优先改善哪个更好。**
- 如果机器学习>人类

误差分析
- 建立一个表格，选择一部分错误的例子，拿出来进行统计，看他们是怎么出错的，所占比例是怎么样的，得到一些结果来指导如何纠正错误。

针对不同的集合，设置不同的正交化的控制旋钮
- 训练集
- 训练-开发集
  - 当训练集与测试机分布不同时，可以设置训练开发集进行误差分析
  - 起结果与开发集进行比对能得出数据不匹配对结果影响有多大
- 开发集
- 测试机
- 真实数据

针对数据不匹配的一些措施
- 人工合成数据
  - 但是需要谨慎的是，只使用了少量的实际数据，例如人工合成的汽车图片，
  - 在语音识别领域有不错的效果
- 少什么添什么，缺什么补什么

其他学习方式

- 迁移学习
  - 通过大数据A学习到的东西应用到小数据B上来
- 多任务学习
- 端到端学习
  - 优点
    - 让数据说话，省略中间过程，比如直接由语音得到结果
    - 较少的手工设计
  - 缺点
    - 需要大量数据才能得到较好的结果
    - 有时候端到端解决不了的问题，可分为多步简单问题分别求解。