Python进阶

一、引言

套用一下大神们对机器学习的定义，机器学习研究的是计算机怎样模拟人类的学习行为，以获取新的知识或技能，并重新组织已有的知识结构使之不断改善自身。简单一点说，就是计算机从数据中学习出规律和模式，以应用在新数据上做预测的任务。近年来互联网数据大爆炸，数据的丰富度和覆盖面远远超出人工可以观察和总结的范畴，而机器学习的算法能指引计算机在海量数据中，挖掘出有用的价值，也使得无数学习者为之着迷。

二、解决的问题

1. 分类问题（有限个类别）
    - 垃圾邮件分类
    - 文本情感分类
    - 图像内容识别
2. 回归问题
    - 电影票房预测
    - 房价预测
3. 聚类问题（相近\相关的样本抱团）
    - 用户群体划分
    - 电影划分

实际应用

4. 计算机视觉
    - **人脸识别**、**车牌识别**、**扫描文字**
5. 自然语言处理
    - **搜索引擎匹配**、**文本理解**、**文本情绪判断**
6. 社会网络分析
    - **用户画像**、**欺诈作弊发现**、**热点发现**
7. 产品推荐
    - 音乐的**"歌曲推荐"**、某宝/某东的**"商品推荐"**

三、学习任务

1. 可能会用到的工具
    - 网页爬虫（之前学过的）主要是获取数据
    - pandas：模拟R，进行数据浏览与预处理
    - numpy：数组运算
    - scipy：高效的科学计算
    - matplotlib：非常方便的数据可视化工具（各种曲线图）
    - sklearn：机器学习包，自带很多模型
    - libsvm：svm模型实现
    - TensorFlow：用于深度学习的包，很方便搭建神经网络，配上tensorboard食用更佳

    - nltk：自然语言处理相关的功能
    - ipython notebook：这个类似于编辑器，其实推荐安装一个Anconda，自带python2.7/3.6，所以可能会和之前安装过的python冲突
2. 可能会用到的数学知识
    - 微积分
    - 线性代数
    - 概率与统计

四、基本工作流程

这个自己在后面学习之后就会知道，在这里我可以先总结一下：

1. 分析问题，抽象成数学问题
2. 获取数据集
3. 特征预处理、特征选择
4. 模型建立、训练、调优
5. 判断准确度、诊断

五、入门资源

内容可能会有点多，但是可以先一点一点的学，相关还没学到的数学知识也可以先跳过直接记住结论。

- Coursera 吴恩达的机器学习（这个可能需要科学上网，有条件可以看，没条件可以看B站大佬下载的）
- google 机器学习速成课程 （这个我是在看完coursera后看的，感觉还不错，提供练习，而且内容也不多，但是节奏有些快，不过入门是够了，**重点**是不用科学上网就能访问）
- 机器学习 ——周志华 （西瓜书）
- 机器学习实战 （封面是个农民，有很多实战例子）
- 统计学习 ——李航 （这本书我还没看过，听别人说挺不错的，用统计学的方法来描述机器学习，很强）

> ps:大概就这些视频跟这些书籍，这些如果能看得完，其他的什么也没必要了，如果还有空可以去**kaggle**上面逛逛。

六、暑假的学习任务

- [x] 每周写一份周记（记录自己学了什么，相当于笔记吧） 最迟也要半月一次，记录格式任意，自己之前搭的博客或者文本编辑器，只要之后能提交就可以了，记得记录日期。
- [ ] 如果觉得自己学的差不多了，可以做一个简单的验证码识别（数据集找我要） **选做** 毕竟还要以合作项目为主，以后有空可以再研究。