数据分析及实践

Analysis and Practice of the Data

实验课

刘 淇
Email: qiliuql@ustc.edu.cn

课程主页：
http://staff.ustc.edu.cn/~qiliuql/AD2021.html

实验四-Part1

- 实验四任务基于实验三，是实验三的拓展与延伸。Part1中，手动实现一种分类算法（例如，决策树、 KNN或者朴素贝叶斯）。并参考实验三特征工程，测试算法在 LoL 数据集上的预测性能，撰写实验报告。
- 具体要求：
  - 代码实现只允许使用 numpy、pandas库和 python内置库，不允许使用现有的机器学习库。
  - 预测任务与实验三一致，以准确率作为评价指标。自行在 LoL数据集上划分训练集和验证集（4:1比例、交叉验证），汇报算法在验证集上的性能 。
  - 实验报告需介绍实现算法的主要流程。
- 评分标准：
  - 代码是否逻辑清楚，能否完整运行
  - 模型的性能不作为评分依据（可以采用不同的评价指标）

实验四-Part2

- 大家利用所学的知识，进行一场数据科学实战。数据集仍然使用实验三的数据集，但预测任务有变。**可以利用开源工具包**，也可以参考实验三的数据分析与特征工程。
- 实验报告需要记录最终的方案流程，也鼓励大家记录每一次失败的尝试。

实验四-Part2

- 详细说明
  - 助教会发布数据集中一部分样本的标签，作为训练集，而另一部分样本作为测试集。
  - 同学们需要预测测试集中每个样本的比赛持续时间 gameDuration。
  - 以均方误差 mean-square error作为评价指标
  - 每位同学可以提交两份预测结果，助教会取最好的一份作为最终成绩
  - 预测结果用 csv格式保存，具体格式如下：

```
1 index,gameDuration
2 0,1000
3 1,1200
4 2,900
5 3,950
6 4,1300
```

实验四-Part2

- 提交要求
  - 将Part1的代码、Part2的代码、 Part2预测结果和实验四实验报告打包发送给助教：apdata2021@163.com
  - 邮件标题: 姓名_学号_exp4
    压缩文件命名格式：姓名_学号_exp4.zip (rar)
    预测结果格式: 姓名_学号_exp4_第几份.csv;
    如：张三_PB19111111 _exp4_2.csv
  - 截止日期：**5月30日**
- 评分标准：
  - 格式是否规范，提交是否及时
  - 实验报告是否逻辑清晰
  - 是否尝试了多种算法、是否对算法进行调参
  - 是否尝试了不同的特征组合
  - 方案的性能、有没有进行可视化分析展示等等

参考资料

- 参考资料：
  - kaggle、天池网站的初学者教程
  - 《机器学习》-周志华