数据分析及实践

Analysis and Practice of the Data

实验四

刘 淇

Email: qiliuql@ustc.edu.cn

课程主页：

http://staff.ustc.edu.cn/~qiliuql/AD2022.html

目录

实验四-Part1

- 实验四的任务基于实验三，是在实验三实现的数据分析基础上的拓展与延伸；
- 同样使用PISA2015数据集；
- 在Part1中，同学们须手动实现一种分类算法（例如：决策树、KNN、朴素贝叶斯或者感知机、集成算法等）；
- 参考实验三中的特征工程，测试算法在 PISA2015数据集上的预测性能，并撰写实验报告。

实验四-Part1

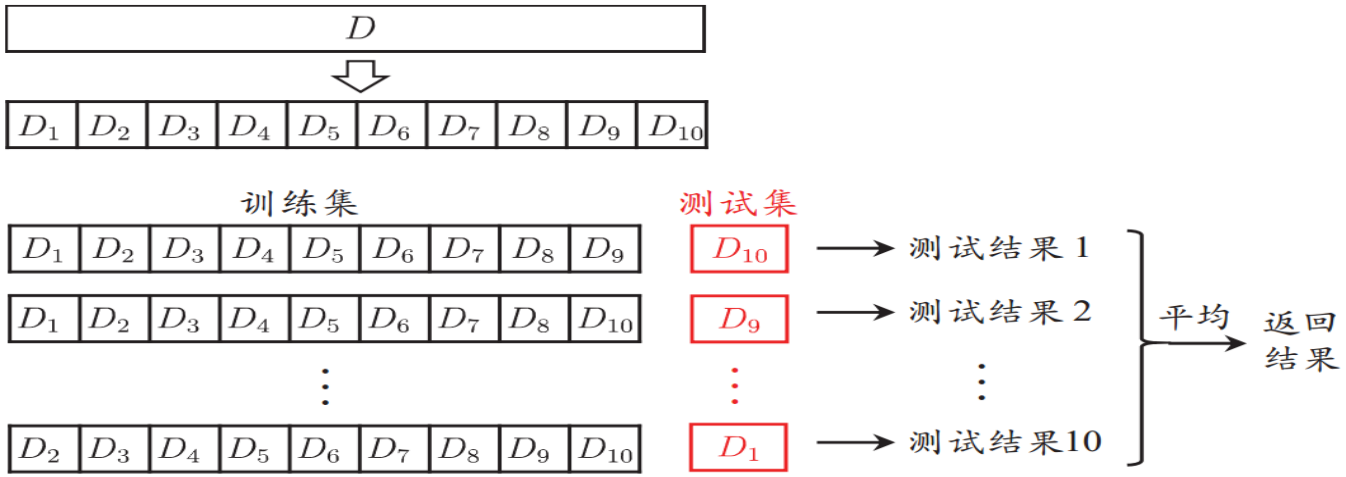
- 具体要求：
  - 代码实现只允许使用 numpy、pandas库和 python内置库，**不允许使用现有的机器学习库**。
  - 预测任务与实验三一致（实验三只是围绕预测目标进行数据分析和特征工程），即预测学生**是否会选择复读(REPEAT)**，并以**准确率(ACC)**作为评价指标（也可以使用其它指标）。
  - 请自行在 PISA2015数据集上划分训练集和测试集(4:1比例、**交叉验证**)，汇报算法在测试集上的性能 。
  - 实验报告需包括实现算法的主要流程、关键技术以及算法的性能。实验报告请用 word或者 PDF格式。

REPEAT	Grade Repetition	NUM	1.0	669	0 - 1		
						0	Did not repeat a <grade>
						1	Repeated a <grade>
						9 / .M	No Response
						SYSTEM MISSING	Missing

数据挖掘基础

- 分类——模型验证方法：
  - 交叉验证法（Cross Validation）：

将数据集分层采样划分为k个大小相似的互斥子集，每次用k-1个子集的并集作为训练集，余下的子集作为测试集，最终返回k个测试结果的均值，k最常用的取值是10.



10 折交叉验证示意图

实验四-Part2

- 同学们须使用至少一种预测算法；
- Part2实验同样使用PISA2015数据集，但预测任务有变；
- 可以利用开源工具包，并结合实验三的数据分析与特征工程，撰写实验报告；
- 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

- 具体要求
  - 助教会发布数据集中一部分样本的标签，作为对应的训练集标签，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学生的**数学能力水平(MATH)**；
  - 以**均方误差 (MSE)**作为评价指标；
  - 预测结果用 **csv**格式保存，具体格式如下：

	index	MATH
0	0	314.902
1	1	488.771
2	2	335.528
3	3	437.947
4	4	492.855
...	...	...

实验四

- 评分标准：
  - 模型效果如何
  - 代码是否逻辑清楚，能否完整运行
  - 格式是否规范，提交是否及时
  - 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
  - 实验结果的展示、数据分析是否全面
  - 实验报告是否逻辑清晰

实验四

- 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：18251859960@163.com
  - 邮件标题格式: 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.csv
    - 如：张三_PB20111111 _exp4.csv
  - 提交截止日期：**5月12日**

参考资料

- 参考资料：
  - kaggle、天池等网站的初学者教程
  - 《机器学习》-周志华
  - 《统计学习方法》-李航