

KDD Lab 作业 2

1. 作业上交时间:

4/10 14:00~ 4/19 22:00 通过 blackboard 上传作业文件, 本次作业仍不支持补交或缓交。逾期不收! 逾期不收! 逾期不收!

2. 作业题目:

第二次作业共有**两个小题**, 主要是让同学们学会熟练的使用 python 对给定数据集寻找合适的分类器, 进行分类。

题目 1: 给定数据集 HW2_1_data.csv, 要求使用两种不同的分类器对葡萄酒的品质进行预测, 使得分类器在测试数据上的 $\text{accuracy} > 0.85$, 代码运行时间 $< 2\text{min}$ 。(学生需自己划定训练集和测试集, 比例为 8:2)。

该数据集为 3 种品质葡萄酒的详细参数数据, 包含内容为:

Input variables:

- 1 - fixed acidity: 非挥发性酸
- 2 - volatile acidity: 挥发性酸
- 3 - citric acid: 柠檬酸
- 4 - residual sugar: 残留糖
- 5 - chlorides: 氯化物
- 6 - free sulfur dioxide: 游离二氧化硫
- 7 - total sulfur dioxide: 总二氧化硫
- 8 - density: 密度
- 9 - pH: 酸碱度
- 10 - sulphates: 硫酸盐
- 11 - alcohol: 酒精

Quality (品质) : 5, 6, 7

题目 2: 使用的数据为 HW2_2_data.csv, 该数据包含 3 个属性信息(X_1, X_2, X_3) 和 1 个类别信息(Y), 要求基于该数据, 使用 1 种合适的分类器进行分类, 使得分类器在测试数据上的 $\text{accuracy} > 0.85$, 代码运行时间 $< 2\text{min}$ 。(学生需自己划定训练集和测试集, 比例为 8:2)

3. 作业要求:

按照上述 2 小题要求完成代码，在作业的上交期限之前，上交 ipynb 文件至 blackboard, 作业命名规则：HW2_学生学号.ipynb

备注：

- (1) 为方便作业批改，尽量提交 ipynb 格式文件，提交前先清除所有输出，减少 ipynb 文件的大小
- (2) 为方便作业批改，请读取在同一个目录下的 HW2_1_data.csv 和 HW_2_2_data.csv（比如在 read csv 时使用相对路径）
- (3) 请备注标明所有作业中没有明确指示的额外代码，比如数据预处理过程中的细节；以及标注题目号与分类器名称
- (4) 请勿写入源文件 HW2_1_data.csv 和 HW_2_2_data.csv！！

4. 作业计分：

题目 1：6 分（两个分类器各 3 分）； 题目 2：4 分；总分 10 分扣完即止。

(1) 内容扣分点：

题目 1：

- 分类器 1 无法实现： -2 分
- 分类器 2 无法实现： -2 分
- 分类器 1 实现，但精度不满足要求： -1 分
- 分类器 1 实现，但运行时间不满足要求： -1 分
- 分类器 2 实现，但精度不满足要求： -1 分
- 分类器 2 实现，但运行时间不满足要求： -1 分
- 没有打印预测精度(accuracy): -1 分
- 满分 6 分，扣完即止。

题目 2：

- 无法实现分类器： -2 分
- 分类器的精度不满足要求： -1 分
- 分类器的运行时间不满足要求： -1 分
- 没有打印预测精度(accuracy): -1 分
- 满分 4 分，扣完即止。

(2) 其他扣分点：

迟交： -10 分

不按照命名规则命名：-1 分

完全不注意备注信息：-1 分

题目 1 程序无法执行：-2 分

题目 2 程序无法执行：-2 分

并没有提交 ipynb 文件，而是提交了其他文件：第一次-5 分，第二次-10 分