第一次作业

Deadline：2022 年 10 月 17 日

1. 回归问题

加载鲍鱼数据（abalone.csv），并完成以下工作:

1) 去除 height 为 0 的两条数据
2) 将 Rings 属性加上 1.5 生成年龄 Ages 属性，并将原来 Rings 属性删除
3) 获取数据信息，数据描述统计信息、属性直方图等
4) 将数据分为训练集和测试集
5) 获得相关关系散点图，可尝试去除异常值
6) 将 x 属性和 y 属性分开
7) 将性别数据转换为独热向量（采用 Pipeline）
8) 将数值型属性标准化（采用 Pipeline）
9) 采用线性回归、决策树和 SVM 模型来拟合数据
10) 采用 k-折交叉验证来选择模型，并将最终选择的模型用在测试集上得到测试误差 RMSE

关于鲍鱼数据的描述:
数据包括 9 个属性
Sex (M-Male, F-Female, I-Infant)
Length (longest shell measurement)
Diameter (perpendicular to length)
Height (with meat in shell)
Whole weight (swhole abalone)
Shucked weight (weight of meat)
Viscera weight (gut weight after bleeding)
Shell weight (after being dried)
Rings (+1.5 gives the age in years)

2. 分类问题

加载电离层数据集（ionosphere_data.csv），应用第三章所学知识进行分类，并对结果进行评估。

关于电离层数据集的描述:
https://archive.ics.uci.edu/ml/datasets/Ionosphere
它是一个二元分类问题。每个类的观察值数量不均等，一共有 351 个观察值，34 个输入变量和 1 个输出变量（第 35 列）。

要求

1. 描述代码步骤和过程（文字+截图）

2. 描述基本任务结果（文字+截图，需以表格或者图的形式给出实验结果，并分析结论）
3. 给出总结与一些体会