2022-2023 秋季《机器学习方法基础》课程第 3 次作业

1、使用 Auto（汽车）数据集进行多元线性回归：

（a）作出数据集中的所有变量的散点图矩阵。

（b）用 cor()函数计算变量之间的相关系数矩阵。需排除定性变量 name（车名）。

（c）使用 lm()函数进行多元线性回归，用除了 name（车名）之外的所有变量作为预测变量，mpg（油耗）作为响应变量。用 summary()函数输出结果并分析所得结果。

例如：

i. 预测变量和响应变量之间有关系吗？

ii. 哪个预测变量与响应变量在统计上具有显著关系？

iii. year（车龄）变量的系数说明什么？

（d）生成线性回归拟合的诊断图。请分析拟合中存在的问题。残差图表明有异常大的离群点吗？杠杆图识别出了有异常高杠杆作用的点吗？

(e) 尝试使用逐步选择法和连续正交法（successive orthogonalization）对数据集中的变量进行筛选，挑选出 3 个最佳变量进行预测。

2、使用 Hitters（击球手）数据集进行分类:

（a）对数据集中的 Salary（球员身价）变量，选择一个适当的阈值对球员进行分类，生成一个新的分类变量（例如以 300 为阈值，则 Salary 高于 300 的球员高身价球员，记为 1，反之记为 0）

（b）使用其余变量预测球员是否为高身价（注意变量筛选）。可以使用 logistics 回归或一般线性回归进行预测，给出预测模型的准确率表现。

（c）对预测错的样本进行核密度估计，并分析这部分样本在分布上的特征。