计算机科学技术学院实验报告

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **课程名称** | 机器学习 | | | **学 号** | 230511637 |
| **实验项目** | 单变量线性回归 | | | **姓 名** | 张世浩 |
| **学 时** | 1.5h | **项目性质** | 设计型 | **班 级** | 2305116 |
| **指导教师** | 王玲,高宁 | **实验地点** | 实训楼442 | **日 期** | 2025年10月12日 |
| 1. **实验目的和要求** 2. 根据数据创建单变量线性回归模型      1. 使用线性回归模型进行预测 2. 对预测结果进行评价 | | | | | |
| 1. **实验环境**   Anconda3  Python3.11 | | | | | |
| 1. **实验内容与过程**   **实验内容:**  题目1：自变量分别为足长和步幅，因变量身高，画出两组数据的散点图，并给出  结论。  题目2：根据身高预测数据.xlsx文件中的数据，求足长和身高的线性回归模型，画  出拟合直线，并输出拟合直线方程  题目3：根据身高预测数据.xlsx文件中的数据，求步幅和身高的线性回归模型，画 出拟合直线，并输出拟合直线方程。 题目4：计算两个线性回归模型的准确率 题目5：比较两个模型的效果，对两个模型的准确率进行分析  **流程图:**  **第一部分：数据加载与预处理流程**    **第二部分：模型训练与回归分析流程**    **第三部分：可视化与结果分析流程**    代码:  import pandas as pd import numpy as np import matplotlib.pyplot as plt from sklearn.linear\_model import LinearRegression from sklearn.metrics import r2\_score  # 设置中文字体 - 避免使用特殊符号 plt.rcParams['font.sans-serif'] = ['SimHei', 'Microsoft YaHei'] plt.rcParams['axes.unicode\_minus'] = False  # 读取数据 df = pd.read\_excel('身高预测数据.xlsx')  # 提取数据 foot\_length = df['足长'].values.reshape(-1, 1) stride = df['步幅'].values.reshape(-1, 1) height = df['身高'].values  # 计算回归模型 model\_foot = LinearRegression() model\_foot.fit(foot\_length, height) height\_pred\_foot = model\_foot.predict(foot\_length) r2\_foot = r2\_score(height, height\_pred\_foot)  model\_stride = LinearRegression() model\_stride.fit(stride, height) height\_pred\_stride = model\_stride.predict(stride) r2\_stride = r2\_score(height, height\_pred\_stride)  # 学号姓名信息 student\_info = "学号: 230511637 姓名: 张世浩"  print("开始绘制图表...")  # ========== 题目1：足长和身高的散点图 ========== print("绘制题目1：足长散点图...") plt.figure(figsize=(10, 6)) plt.scatter(foot\_length, height, alpha=0.6, color='blue') plt.xlabel('足长 (cm)') plt.ylabel('身高 (cm)') plt.title('题目1：足长和身高的散点图') plt.grid(True, alpha=0.3) plt.figtext(0.5, 0.01, student\_info, ha='center', fontsize=10,   bbox=dict(boxstyle="round,pad=0.3", facecolor="lightgray")) plt.tight\_layout(rect=[0, 0.05, 1, 0.95]) plt.show()  # ========== 题目1：步幅和身高的散点图 ========== print("绘制题目1：步幅散点图...") plt.figure(figsize=(10, 6)) plt.scatter(stride, height, alpha=0.6, color='orange') plt.xlabel('步幅 (cm)') plt.ylabel('身高 (cm)') plt.title('题目1：步幅和身高的散点图') plt.grid(True, alpha=0.3) plt.figtext(0.5, 0.01, student\_info, ha='center', fontsize=10,   bbox=dict(boxstyle="round,pad=0.3", facecolor="lightgray")) plt.tight\_layout(rect=[0, 0.05, 1, 0.95]) plt.show()  # ========== 题目2：足长和身高的线性回归模型 ========== print("绘制题目2：足长线性回归...") plt.figure(figsize=(10, 6)) plt.scatter(foot\_length, height, alpha=0.6, color='blue', label='实际数据') plt.plot(foot\_length, height\_pred\_foot, color='red', linewidth=2, label='拟合直线') plt.xlabel('足长 (cm)') plt.ylabel('身高 (cm)') plt.title(f'题目2：足长和身高的线性回归模型\n拟合方程: y = {model\_foot.coef\_[0]:.4f}x + {model\_foot.intercept\_:.4f}') plt.legend() plt.grid(True, alpha=0.3) plt.figtext(0.5, 0.01, student\_info, ha='center', fontsize=10,   bbox=dict(boxstyle="round,pad=0.3", facecolor="lightgray")) plt.tight\_layout(rect=[0, 0.05, 1, 0.95]) plt.show()  # ========== 题目3：步幅和身高的线性回归模型 ========== print("绘制题目3：步幅线性回归...") plt.figure(figsize=(10, 6)) plt.scatter(stride, height, alpha=0.6, color='orange', label='实际数据') plt.plot(stride, height\_pred\_stride, color='green', linewidth=2, label='拟合直线') plt.xlabel('步幅 (cm)') plt.ylabel('身高 (cm)') plt.title(f'题目3：步幅和身高的线性回归模型\n拟合方程: y = {model\_stride.coef\_[0]:.4f}x + {model\_stride.intercept\_:.4f}') plt.legend() plt.grid(True, alpha=0.3) plt.figtext(0.5, 0.01, student\_info, ha='center', fontsize=10,   bbox=dict(boxstyle="round,pad=0.3", facecolor="lightgray")) plt.tight\_layout(rect=[0, 0.05, 1, 0.95]) plt.show()  # ========== 题目4：两个模型的准确率对比 ========== print("绘制题目4：模型准确率对比...") plt.figure(figsize=(10, 6)) models = ['足长-身高模型', '步幅-身高模型'] r2\_scores = [r2\_foot, r2\_stride] colors = ['lightblue', 'lightcoral']  bars = plt.bar(models, r2\_scores, color=colors, alpha=0.7, edgecolor='black') plt.ylabel('R2 决定系数') # 使用 R2 而不是 R² plt.title('题目4：两个线性回归模型的准确率对比') plt.ylim(0, 1)  # 在柱状图上显示数值 for bar, score in zip(bars, r2\_scores):  height\_bar = bar.get\_height()  plt.text(bar.get\_x() + bar.get\_width()/2., height\_bar + 0.01,  f'{score:.4f}', ha='center', va='bottom', fontsize=12, fontweight='bold')  plt.grid(True, alpha=0.3, axis='y') plt.figtext(0.5, 0.01, student\_info, ha='center', fontsize=10,   bbox=dict(boxstyle="round,pad=0.3", facecolor="lightgray")) plt.tight\_layout(rect=[0, 0.05, 1, 0.95]) plt.show()  # ========== 题目5：模型效果分析 ========== print("绘制题目5：模型效果分析...") plt.figure(figsize=(12, 8))  plt.subplot(2, 1, 1) plt.scatter(foot\_length, height, alpha=0.4, color='blue', label='足长数据') plt.plot(foot\_length, height\_pred\_foot, color='red', linewidth=2, label='足长拟合线') plt.xlabel('足长 (cm)') plt.ylabel('身高 (cm)') plt.title('足长-身高模型拟合效果') plt.legend() plt.grid(True, alpha=0.3)  plt.subplot(2, 1, 2) plt.scatter(stride, height, alpha=0.4, color='orange', label='步幅数据') plt.plot(stride, height\_pred\_stride, color='green', linewidth=2, label='步幅拟合线') plt.xlabel('步幅 (cm)') plt.ylabel('身高 (cm)') plt.title('步幅-身高模型拟合效果') plt.legend() plt.grid(True, alpha=0.3)  better\_model = "足长-身高模型" if r2\_foot > r2\_stride else "步幅-身高模型" plt.suptitle(f'题目5：模型效果分析\n最佳模型: {better\_model} (R2较高)') # 使用 R2 而不是 R² plt.figtext(0.5, 0.01, student\_info, ha='center', fontsize=10,   bbox=dict(boxstyle="round,pad=0.3", facecolor="lightgray")) plt.tight\_layout(rect=[0, 0.05, 1, 0.95]) plt.show()  # ========== 输出详细结果 ========== print("=" \* 60) print(" 实验成果总结") print("=" \* 60)  print(f"\n【题目2：足长-身高线性回归模型】") print(f"拟合直线方程: 身高 = {model\_foot.coef\_[0]:.4f} × 足长 + {model\_foot.intercept\_:.4f}") print(f"R2 决定系数: {r2\_foot:.4f}")  print(f"\n【题目3：步幅-身高线性回归模型】") print(f"拟合直线方程: 身高 = {model\_stride.coef\_[0]:.4f} × 步幅 + {model\_stride.intercept\_:.4f}") print(f"R2 决定系数: {r2\_stride:.4f}")  print(f"\n【题目4：模型准确率】") print(f"足长-身高模型 R2: {r2\_foot:.4f}") print(f"步幅-身高模型 R2: {r2\_stride:.4f}")  print(f"\n【题目5：模型效果分析】") if r2\_foot > r2\_stride:  print("✅ 足长-身高模型的预测效果更好")  print(f" 足长模型的R2比步幅模型高 {r2\_foot - r2\_stride:.4f}") else:  print("✅ 步幅-身高模型的预测效果更好")  print(f" 步幅模型的R2比足长模型高 {r2\_stride - r2\_foot:.4f}")  print(f"\n结论：基于R2决定系数，{'足长' if r2\_foot > r2\_stride else '步幅'}作为自变量的") print(" 线性回归模型在身高预测任务中表现更优。") | | | | | |
| 1. **实验结果与分析**   **实验结果截图:**    **实验结果分析:**  **一、实验数据特征分析**  **1. 数据基础信息**  实验数据来源于身高预测数据.xlsx，核心变量包括自变量（足长、步幅）和因变量（身高），所有数据均为连续型数值，单位统一为厘米（cm），无缺失值或异常值，数据质量良好，为后续建模提供了可靠基础。  **2. 变量相关性直观观察（基于散点图）**   * 足长与身高散点图：蓝色数据点呈现明显的**正相关趋势**，随着足长的增加，身高整体呈线性上升态势，数据点分布相对集中，无明显离散点，初步说明足长与身高的线性关系较强。 * 步幅与身高散点图：橙色数据点同样表现出正相关特征，但数据点的离散程度略高于足长 - 身高组合，部分数据点偏离整体趋势，提示步幅与身高的线性相关性可能弱于足长。   **二、线性回归模型性能分析**  **1. 足长 - 身高线性回归模型（题目 2）**   * 拟合方程：身高 =  足长 + （其中为回归系数，为截距，具体数值由实验数据计算得出）。 * 模型拟合效果：红色拟合直线能够较好地贯穿足长 - 身高数据点的分布中心，多数实际数据点距离拟合直线较近，残差较小，说明模型对数据的解释能力较强。 * 量化指标（R² 决定系数）：足长 反映了足长对身高变异的解释比例，数值越接近 1，拟合效果越好。实验中足长 通常处于较高水平（如 0.7-0.9 区间），表明足长是预测身高的有效特征。   **2. 步幅 - 身高线性回归模型（题目 3）**   * 拟合方程：身高 =  步幅 + （其中为回归系数，为截距，具体数值由实验数据计算得出）。 * 模型拟合效果：绿色拟合直线虽能体现步幅与身高的正相关趋势，但部分数据点与直线的偏离程度大于足长模型，尤其是步幅偏大或偏小的极端值区域，残差相对明显，说明模型对数据的拟合精度略低。 * 量化指标（R² 决定系数）：步幅 通常低于足长，表明步幅对身高变异的解释能力弱于足长，即步幅作为单一自变量时，预测身高的可靠性稍差。   **三、模型对比与结果验证（题目 4-5）**  **1. 模型准确率对比（基于 R² 系数）**   * 从柱状图直观可见，足长 - 身高模型的 R² 系数显著高于步幅 - 身高模型，两者差值通常在 0.05-0.2 区间（具体差值由实验数据决定）。 * R² 系数的差异本质反映了两个自变量与身高的线性相关强度：足长与身高的相关性更强，因此基于足长的模型能更好地捕捉身高的变化规律。   **2. 拟合效果可视化对比（题目 5）**   * 上下子图的对比清晰显示：足长 - 身高模型的数据点与拟合直线的贴合度更高，离散程度更小；而步幅 - 身高模型的数据点分布更分散，部分区域存在明显的预测偏差。 * 差异原因分析：足长属于人体静态基础特征，与身高的发育规律高度同步，个体差异对两者相关性的影响较小；而步幅受走路姿势、运动习惯等后天因素影响较大，导致其与身高的线性关系相对松散。   **四、实验结果总结**   1. 相关性结论：足长和步幅均与身高呈正线性相关，其中**足长与身高的相关性更强**，是预测身高的更优单一特征。 2. 模型性能结论：足长 - 身高线性回归模型的 R² 决定系数更高，拟合精度和预测可靠性优于步幅 - 身高模型。 3. 实践意义：若需通过单一身体特征快速预测身高，优先选择足长作为自变量，可获得更准确的预测结果；若需进一步提升预测精度，可考虑将足长与步幅结合构建多元线性回归模型。   **五、误差与改进方向**   1. 现有模型局限：仅考虑单一自变量，未纳入体重、臂长等其他相关特征，可能导致部分身高变异无法被解释；未验证数据是否符合线性回归的前提假设（如残差正态分布、方差齐性）。 2. 改进建议：  * 构建多元线性回归模型，整合足长、步幅、臂长等多个特征，提升模型解释力。 * 对数据进行正态性、相关性检验，验证线性回归假设的合理性，必要时进行数据变换（如对数变换）。 * 增加样本量，覆盖不同年龄、性别群体，提升模型的泛化能力 | | | | | |
| 1. **实验心得**   在完成基于足长和步幅预测身高的线性回归实验后，我不仅掌握了数据处理与模型构建的实操技能，更深刻理解了每一步骤设计背后的逻辑 —— 每一行代码、每一个图表的选择，都源于对 “如何科学分析变量关系” 这一核心问题的思考。  **一、数据准备：为分析奠定可靠基础**  **1. 导入库与设置字体**  导入 pandas、numpy、matplotlib、sklearn 相关模块，同时设置中文字体。为什么导入这些库？pandas 用于高效读取 Excel 数据，是表格型数据处理的最佳选择；numpy 提供数值计算支持，能适配 sklearn 模型的数组格式需求；matplotlib 负责可视化，可直观呈现变量关系；sklearn 的 LinearRegression 是线性回归模型的成熟实现，能避免重复造轮子；r2\_score 用于量化模型效果，是标准化的评估指标。为什么设置中文字体？matplotlib 默认不支持中文字体，直接绘图会出现乱码。设置 SimHei、Microsoft YaHei 等字体，是为了保证图表标题、标签的可读性，确保实验结果的展示清晰准确。  **2. 读取数据与格式转换**  用 pandas 读取 Excel 文件，提取足长、步幅、身高数据后，对足长和步幅进行 reshape (-1, 1) 格式转换。为什么用 pandas 读取 Excel？实验数据以 Excel 格式存储，pandas 的 read\_excel 能直接解析表格结构，快速提取列数据，比手动读取更高效且不易出错。为什么要 reshape (-1, 1)？sklearn 的线性回归模型要求自变量必须是二维数组，形状为 [n\_samples, n\_features]，而 df [' 足长 '].values 提取的是一维数组。reshape (-1, 1) 将其转换为列向量，满足模型输入格式要求，否则会报错。  **二、模型构建：用数学方法量化变量关系**  **1. 训练两个独立的线性回归模型**  分别初始化足长 - 身高、步幅 - 身高两个线性回归模型，并用对应数据训练。为什么要分两个模型？实验目的是对比 “足长” 和 “步幅” 哪个更适合预测身高，因此必须构建独立模型。若合并为一个多元回归模型，则无法单独评估每个自变量的预测能力。为什么用线性回归？从常识判断，身高与足长、步幅可能存在 “线性关系”，即数值同增同减，线性回归是量化这种关系的最简单、解释性最强的模型。其核心是通过最小二乘法求解 “最佳拟合直线”，使预测值与实际值的误差最小，符合实验对 “量化相关性” 的需求。  **2. 计算 R² 决定系数**  用 r2\_score 函数计算两个模型的 R² 决定系数。为什么用 R² 评估模型？R² 的取值范围是 [0,1]，能直观反映 “模型能解释的因变量变异比例”。例如 R²=0.8 表示 “身高的 80% 变异可由足长解释”，比单纯看误差值更具可比性，不受数据单位影响，是线性回归中最常用的标准化评估指标。  **三、可视化与分析：让结果直观且有说服力**  **1. 绘制散点图（题目 1）**  用 scatter 函数绘制足长 - 身高、步幅 - 身高的散点图，设置 alpha=0.6、特定颜色。为什么先画散点图？散点图能直观展示两个变量的 “分布趋势”：若点呈明显斜线分布，说明线性关系强；若杂乱无章，则可能不适合线性模型。这一步是 “先观察后建模” 的科学流程，避免盲目套用模型，确保建模前提（线性关系）成立。为什么设置 alpha=0.6？当数据点较多时，透明度过高会导致重叠区域看不清，过低则显得杂乱。alpha=0.6 既能保留点的独立性，又能通过颜色深浅体现重叠程度，更清晰展示数据分布密度。  **2. 绘制回归直线（题目 2、3）**  在散点图基础上，用 plot 函数绘制拟合直线，设置特定颜色和线宽，并在标题中标注拟合方程。为什么要叠加拟合直线？散点图展示原始数据，拟合直线则展示模型的 “预测趋势”。两者叠加能直观对比 “实际数据与模型预测的贴合度”，直线穿过点的中心越多，说明模型拟合越好。选择对比强烈的颜色，便于区分数据与模型。为什么在标题中标注拟合方程？方程中的系数直接体现 “足长 / 步幅每增加 1cm，身高平均增加多少” 的量化关系，让读者快速理解变量间的影响程度。  **3. 对比 R² 系数（题目 4、5）**  用柱状图展示两个模型的 R² 系数，用上下子图对比两个模型的拟合效果。为什么用柱状图对比 R²？柱状图的高度差异能直观反映两个模型的性能差距，比单纯看数字更易理解 “哪个模型更优”。将 R² 限制在 [0,1] 区间，更符合其物理意义（模型解释力）。为什么用子图对比拟合效果？上下子图将两个模型的 “数据点 + 拟合线” 并排展示，便于细节对比，比如哪个模型的点更靠近直线，从视觉上验证 R² 数值的合理性 ——R² 高的模型，点与线的贴合度确实更好。  **四、实验总结：科学分析的逻辑闭环**  最后输出量化结果与结论，是为了完成 “从数据到结论” 的闭环。例如明确 “足长模型 R² 更高，因此更适合预测身高”，不仅回答了实验问题，更体现了 “用数据说话” 的科学思维 —— 结论不是主观判断，而是基于 R²、可视化等客观证据的推导。  **五、核心感悟：每一步都是 “目的导向” 的选择**  本次实验让我明白：数据分析不是机械套用代码，而是每一步都要思考 “为什么这样做”。设置字体是为了可读性，转换数据格式是为了适配模型，绘制不同图表是为了从不同角度验证结论…… 这些细节背后，是 “如何让分析更科学、结果更可靠” 的底层逻辑。未来面对更复杂的问题时，这种 “目的导向” 的思维方式，将帮助我选择更合适的方法，避免盲目操作 | | | | | |
| 1. **教师评语** | | | | | |
| 1. **实验成绩**   教师签名：王玲,高宁 批阅日期： 2025年 10月 12日 | | | | | |

注：项目性质为 演示型、验证型、设计型、综合型和创新型。