**数据科学导论第二章实验报告**

|  |  |  |  |
| --- | --- | --- | --- |
| **姓 名:** | 队长 | **学 号:** |  |
| **组 号:** |  | | |
| **组 员:** | **组员姓名（组内成员不多于四人）** | | |
| **学 院:** |  | | |

**南京理工大学基础前沿交叉中心**

**年 月 日**

**实验要求**

* **数值试验题应该同时提交书面报告和程序(打包), 其中书面报告有详细的推导和数值结果及分析.**
* **实验报告分组提交. 自由组合, 不超过4人一组. 多人组队, 请明确说明每人负责的部分和内容.**
* **迟交一天(24小时)打折10%, 不接受晚交4天的作业和项目, 任何时候理由都不接受.**
* **鼓励开源交流, 禁止直接抄袭. 有讨论或从其它任何途径取得帮助, 请列出来源.**

**目 录**

[案例一 混凝土强度数据标准化处理 1](#_Toc60132448)

[【1】案例背景 1](#_Toc60132449)

[【2】方法陈述 1](#_Toc60132450)

[【3】实验代码 1](#_Toc60132451)

[（1） 1](#_Toc60132452)

[（2） 1](#_Toc60132453)

......

[【4】结果分析 1](#_Toc60132461)

[案例二 XX实验](#_Toc60132462) **[错误!未定义书签。](#_Toc60132462)**

[【1】案例背景](#_Toc60132463) **[错误!未定义书签。](#_Toc60132463)**

[【2】方法陈述](#_Toc60132464) **[错误!未定义书签。](#_Toc60132464)**

[【3】实验代码](#_Toc60132465) **[错误!未定义书签。](#_Toc60132465)**

[（1）](#_Toc60132466) **[错误!未定义书签。](#_Toc60132466)**

[（2）](#_Toc60132467) **[错误!未定义书签。](#_Toc60132467)**

[【4】结果分析](#_Toc60132470) **[错误!未定义书签。](#_Toc60132470)**

# 案例二 皮肤病数据缺失值填补

## 【1】案例背景

本题提供了一个红斑鳞状疾病患者的基本信息数据集，含22个组织病理学特征和12个临床特征。但是出于某些原因，本题数据集中年龄这一信息存在缺失值，缺失值部分用“？”代替。因此本题要求我们将有问号的部分用均值填补法对其进行填补。

## 【2】方法陈述

利用python中的pandas库中的函数，将数据库的数据部分和表头部分进行分别导入，将表头部分和数据部分整合，筛选出年龄这一数据列并用均值计算函数对其进行均值填补。

## 【3】实验代码

### 步骤一 将原始数据整理为csv文件

import pandas as pd

data = pd.read\_table('dermatology.data',sep=',',header = None,index\_col= None) # 加载本地数据

data.to\_csv('dermatology.csv', sep=',',header=True,index=False)

# data转成csv并保存为dermatology.csv文件

csv = pd.read\_csv('dermatology.csv',sep=',')

# 读取生成的dermatology.csv文件进行验证

print(csv.head(15)) # 打印前15行

### 步骤二 将数据文件和数据表的表头合并

# 创建一个空字典来存储列名及其对应的序号

column\_names\_dict = {}

# 打开 .names 文件以读取列名

with open('dealt\_column\_name.names', 'r') as file:

for line in file:

column\_name = line.strip() # 去除前导/尾随空格

if column\_name: # 判断行是否非空

# 提取序号和列名

parts = column\_name.split('. ', 1)

if len(parts) == 2:

index, name = parts

# 使用 strip() 去除前导空格并尝试将序号转换为整数

index = int(index.strip())

# 将序号和列名添加到字典中

column\_names\_dict[index] = name

# 对列名按照序号进行排序

sorted\_column\_names = [column\_names\_dict[i] for i in sorted(column\_names\_dict.keys())]

# 打印排序后的列名

print(sorted\_column\_names)

# 读取数据文件并保存为CSV文件，使用逗号作为分隔符

data = pd.read\_csv('dermatology.csv',sep=',',index\_col=None)

df = data

df.columns = sorted\_column\_names

df.to\_csv('merged\_dermatology.csv', sep=',', index=False)

# 打印DataFrame

df

### 步骤三 对缺失数据进行缺失值填补

# 列名

column\_to\_fill = 'Age (linear)' # 选择要填补缺失值的列

# 将'?'替换为NaN

df[column\_to\_fill] = df[column\_to\_fill].replace('?', pd.NA)

df[column\_to\_fill] = pd.to\_numeric(df[column\_to\_fill], errors='coerce').round()

# 计算该列的均值

mean\_value = df[column\_to\_fill].mean()

# 均值填补函数

def fill\_missing\_with\_mean(df, column, mean\_value):

df[column] = df[column].fillna(mean\_value)

return df

# 使用均值填补指定列中的缺失值

df = fill\_missing\_with\_mean(df, column\_to\_fill, mean\_value)

df.to\_csv('mean\_substitution\_dermatology.csv',sep=',',index=False)

print(df)

## 【4】结果分析

成功将原始杂乱的数据添加表头，重新整理为csv数据表，并且将表中含有问号的部分进行平均值填补。其中数据集一开始的整理部分难度明显高于后面的平均值填补部分，可能是由于数据集一开始给的形式并不是表格形式，需要自己手动整理。本段代码参考了ChatGPT的修改意见进行了优化。