**广州大学学生实验报告**

**开课学院及实验室：**计算机科学与网络工程学院软件实验室 **2020年 12 月10日**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **学院** | **计算机科学与网络工程学院** | **年级/专业/班** | **18级软件工程3班** | **姓名** | 阮逸栋 | **学号** | 1806300142 |
| **实验课程名称** | **机器学习与数据挖掘实验** | | | | | **成绩** |  |
| **实验项目名称** | 数据统计和可视化 | | | | | **指导老师** | 彭伟龙 |

**实验二**

1. **组员信息**

李子轩（小组共1人）

1. **作业环境（文件说明，函数说明，调用的函数库以及涉及哪些技术）**

**文件说明：**

数据源1：一.数据源1.xlsx

数据源2：一.数据源2-逗号间隔.txt

实验题目5中最后输出的距离每个样本最近的三个样本，得到一个91x3的矩阵（91 为学生的数量，3为3个样本ID），并导入一个txt文件中

实验一中的相关代码文件：exp01.py

实验题目1中生成的以课程1成绩为x轴，体能成绩为y轴的散点图.png

实验题目2中生成的以5分为间隔，画出课程1的成绩的直方图.png

**函数说明：**

在代码中均有详细标识和说明，在此不重复描述，主要调用了pandas下处理数据的函数和方法，调用了如matplotlib、seaborn等进行可视化以及绘图的相关操作，并且调用了上一个实验中计算平均值、标准差以及转换体能测试成绩为数值型的方法，减少实验的重复性。

**调用的数据库以及涉及的技术：**

未调用数据库。

涉及的技术主要包括基于matplotlib、seaborn等模块对给定的数据源进行可视化的操作，其中数据计算均未调用库函数进行计算，均通过对相关计算公式的复原完成操作。

1. **作业题目和内容**

**题目：**基于实验一中清洗后的数据练习统计和视化操作，100个同学（样本），每个同学有11门课程的成绩（11维的向量）；那么构成了一个100x11的数据矩阵。以你擅长的语言C/C++/Java/Python/Matlab，编程计算：

1. 请以课程1成绩为x轴，体能成绩为y轴，画出散点图。

2. 以5分为间隔，画出课程1的成绩直方图。

3. 对每门成绩进行z-score归一化，得到归一化的数据矩阵。

4. 计算出100x100的相关矩阵，并可视化出混淆矩阵。（为避免歧义，这里“协相关矩阵”进一步细化更正为100x100的相关矩阵，100为学生样本数目，视实际情况而定）

5. 根据相关矩阵，找到距离每个样本最近的三个样本，得到100x3的矩阵（每一行为对应三个样本的ID）输出到txt文件中，以\t,\n间隔。

提示：

计算部分不能调用库函数；画图/可视化显示可可视化工具或API实现。

1. **难题与解决**

在实验过程中，我对于矩阵可视化等函数不够熟悉，可能是由于之前接触并学习python的过程中并没有了解到相关的内容，在多次百度，参考csdn上的代码与同学的帮助下，我最终完成了想要得到的结果。