**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** | 李子轩 | **学号** | 1806300106 |
| **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：objectFile.csv（为实验一中完成数据集成和清洗后导出的文件）

进行z-score归一化后得到的矩阵：zscoreFile.csv

91x91的相关矩阵：correlationMatrixFile.csv

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：Matrix.txt

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。

**具体的实验过程：**

1. 调用实验一最后输出的文件objectFile.csv，由于分数区间为[0,100]，故设定原始数据有100个点，设置从0开始，间隔为1，设置标题为“Scatter Diagram”，横坐标名称为“C1-grade”，纵坐标名称为“Constitution-grade”，以课程1成绩为x轴，体能成绩为y轴，通过matplotlib.pyplot模块画出散点图，并将结果保存在文件夹中命名为“散点图.png”；
2. 与上一步骤相似，读取objectFile.csv，绘制课程1成绩的直方图，将直方图颜色改为绿色，分数区间为[0,100]，故设定原始数据有100个点，设置从0开始，间隔为5，设置标题为“Histogram”，横坐标名称为“C1-grade”，纵坐标名称为“Population”，通过matplotlib.pyplot模块画出直方图，并将结果保存在文件夹中命名为“直方图.png”；
3. 首先调用了上一个实验中转换体能测试成绩为数值型的方法consGrade()以及计算平均值和标准差的方法avg()和std()，将体测成绩进行转换，便于后面的计算，将需要进行z-score归一化操作的属性名提取出来，即数据中每门成绩名，随后进行z-score归一化处理，此处不详述，最后将结果打印出来并导入一个csv文件中保存下来，命名为“zscoreFile.csv”；
4. 调用了上一个实验中转换体能测试成绩为数值型的方法consGrade()，此处的相关性只用考虑属性为成绩的，并且由于实验数据中C10项均为0，所以不纳入计算的访问中，然后计算每个学生成绩的平均值和标准差以及每两个学生之间的标准差，最后利用上述数据算出各个相关系数，打印出一个91×91的相关矩阵，共有91个学生的数据，相关矩阵中的第i行第j列代表了第i个学生和第j个学生之间成绩的相关系数，最后将这个相关矩阵保存在一个csv文件中，命名为“correlationMatrixFile.csv”，而可视化方法选择使用热力图进行显示；
5. 读取出correlationMatrixFile.csv中的数据，即91×91的相关矩阵的数据，遍历每一行数据，利用sort\_values()按降序对数据进行排序，找到其中元素除自身以外最大的三个，易知排序的第一个元素为自身与自身的相关系数，所以挑选出排序后的第2、3、4个数据，需要注意的是找到的值为对应的ID号，输出处理后的91x3矩阵，并将其结果保存在一个txt文件中，以\t,\n间隔，命名为“Matrix.txt”。
6. **难题与解决**

在实验过程中，虽然对于相关的计算公式如相关系数、z-score标准化等比较熟悉 ，但是在具体是实际运用环节中仍出现了不少麻烦，这主要还是对于python编程的不熟悉，但是在不断的debug过程中将原本十分粗燥的代码逐渐的完善，最终完成了实验的要求，甚至在后续的学习中将新的思路运用到之前旧的算法上，大大地提高了代码运行的效率和正确性，并且进行了不断的查漏补缺。

1. **总结**

在实验过程中由于在之前的实验中已经学习和使用过相关的可视化绘图模块和方法，已经积累了一定的经验，所以在本次的实验中上手速度比较快，大大地提高了实验的效率，但是后续的相关题目中，在具体的公式算法实现上在初期还是出现了比较大的纰漏，这点需要引起比较大的重视，但是在这个不断改进、纠错的过程中也大大地扩展了我对于问题思考的思路，让我不再在一个问题上钻牛角尖，我相信通过后续对于本门课程的不断学习，也将不断增强我对机器学习以及数据挖掘的相关知识兴趣，不断推动我对于相关问题的更加深入的思考。