**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11日**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **学院** | **计算机科学与网络工程学院** | **年级/专业/班** | **软件183** | **姓名** | 肖小恩 | **学号** | 1806300128 |
| **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** | 实验一《多源数据集成、清洗和统计》 | | | | | **指导老师** | 彭伟龙 |

**实验一**

**一、组员信息**

* 肖小恩（单干）

**二、作业题目和内容**

## 实验二 《数据统计和可视化》

#### 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。

**三、作业环境**

（1）文件说明：数据源：xxe\_two（数据源）.xlsx

第3题：归一化后列表数据.xls

第4题：各行之间相关系数.xls

第4题：混淆矩阵需导入的数据.xlsx

第5题：距离某样本最激动呢另三个样本ID.txt

（看名字便可一目了然，此处不多做说明了。）

（2）函数说明：numpy模块的std()求标准差函数，seaborn模块的heatmap()函数画热力图，基本没有调用其他的函数了。

（3）调用的函数库：std()，heatmap()

（4）涉及哪些技术“

* excel文件数据导入python中，存入列表
* 列表中数据导出到excel、txt文件
* 索引访问列表元素，对列表的行或列的数据进行求均值、标准差等操作，对行与行之间的数据求协方差，协方差是一步步算的，未调用库函数。
* 运用列表数据，调用seaborn模块中的heatmap函数画热力图

**四、难题与解决**

a）第1，2，3小题比较简单，在实验一的基础上很快就完成了。

b）第4小题，在理解弄清题意上也花费了一点时间，不断跟同学讨论以及请教老师，最终顺利得出了相关矩阵及可视化出了混淆矩阵。

c）第5小题是在第4小题的基础上完成的，但在数据处理，下标索引等方面也摸索了一点时间，不过最终还是顺利完成了。

d）在实验一认真完成的基础上，实验二过很多知识点都不需要再百度，因此没有实验一刚开始时那会无从下手的状态。

e）2020年12月10号晚上，肝了一个晚上终于把第4、5小题搞完，结果在整理时手残shift+delete，没能恢复，最终第二天重写了一遍，虽然说确实挺惨的，但重写一遍并没有花很多时间，而且重写完一遍对于代码逻辑更加清晰了，因祸得福呗就😒，终于肝完了！！

**五、实验总结**

通过本次实验，我学到了Z-score归一化的公式及方法，学到了如何去画散点图、直方图、热力图以及窗口的设置，同时知道了相关矩阵如何去计算，均值、方差、相关系数等的一步步计算加深了我对公式的理解、增强了对公式的运用能力，同时对索引访问列表元素有了更加深入的掌握。实验二实验过程：初始思路形成🡪摸索散点图如何绘制🡪摸索直方图如何绘制🡪摸索归一化如何计算，列表如何转矩阵🡪摸索相关矩阵的定义，得到矩阵🡪不断尝试绘制混淆矩阵🡪找到距离每个样本最近的另三个样本，其对应的ID存入short列表🡪short列表导出到txt文件🡪大功告成！✌

第4，5小题总体花费时间较多，主要是对题目所涉及的名词的概念不了解，都得先去了解概念才入手实验，而且还经历了误删重写，虽然坎坷，但终究是完成了，还挺有成就感的哈哈！😊

至此，实验二圆满完成！感谢老师的指导、同学们的热心交流。