**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/电子信息楼  **2020 年10 月 1日**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **学 院** | 计算机科学与网络工程学院 | **年级/专业/班** |  | **姓名** |  | **学号** |  |
| **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2** | | | | | **指导老师** | 彭伟龙 |

(\*\*\*（1）报告只能为文字和图片；（2）实验项目名称不能有任何改动；（3）必须填写正确的姓名和学号；否则会影响最后成绩的统计，后果自负\*\*\*)

1. **组员：**

汤韬1806300183，李炜煜1806300182，张洪鑫1806300003

**二、实验目的及要求**

1. 掌握对多维数据统计、数据立方的理解

2. 掌握基本数据归一化、散点可视化、分布可视化方法

3. 按照既定格式书写实验报告

**三、实验设备与平台**

1. 实验设备：计算机；

2. 平台：Windows 10操作系统

**四、实验内容**

**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，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

**实验实施**：

（在此详述平台，技术栈，思路，处理逻辑等等）

平台：python

用到的函数包：cmath，pandas，numpy，matplotlib.pyplot，seaborn

camth开方，pandas用来读取数据，numpy用来处理数据，合成两个数据表，matplotlib.pyplot和seaborn用来画图

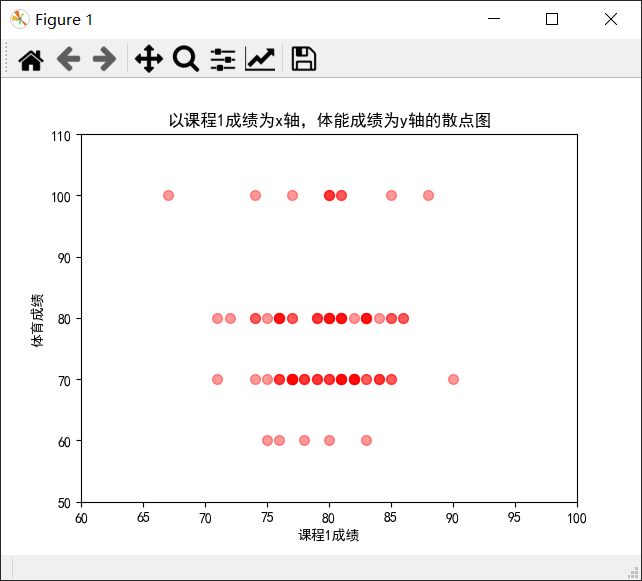
数据：文件123.txt和123.xlsx是老师提供的两个数据，data.xlsx是两个数据合成后经过处理的数据，1.py是实验一的代码

**运行结果**：

（运行结果、现象的说明与截图）

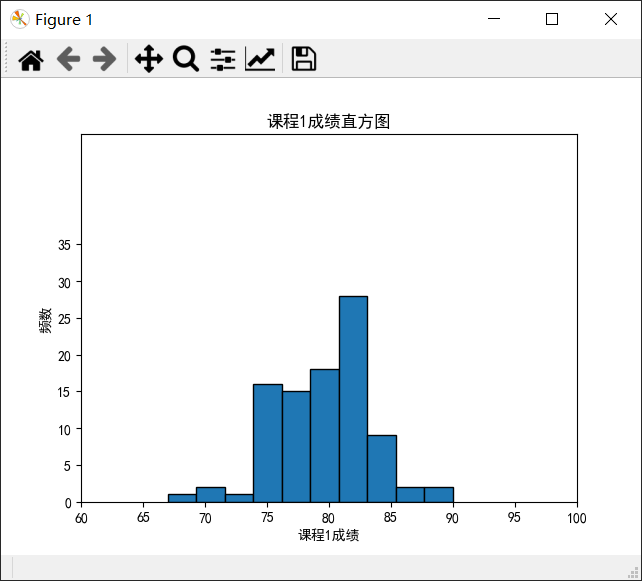
1. 请以课程1成绩为x轴，体能成绩为y轴，画出散点图。

**散点图.png**



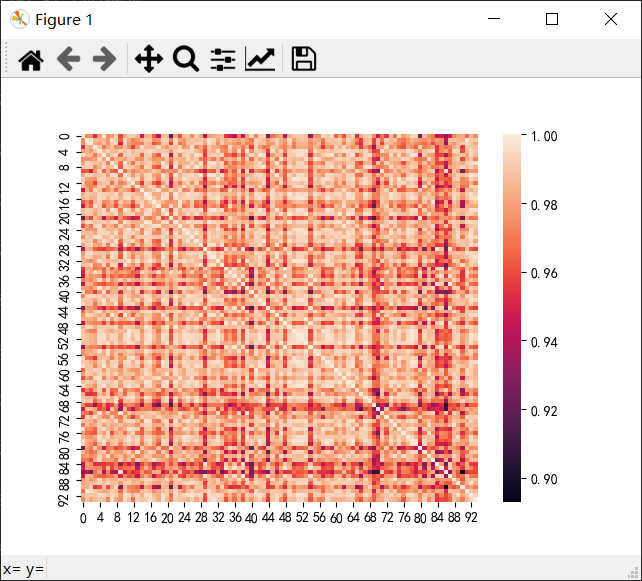
1. 以5分为间隔，画出课程1的成绩直方图。

**直方图.png**

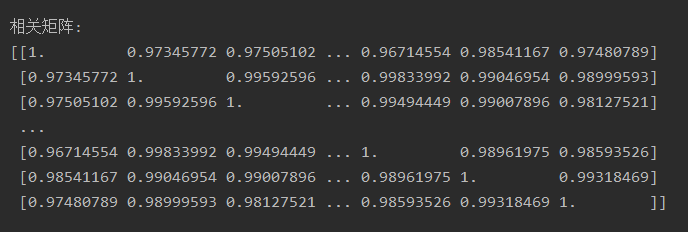


1. 对每门成绩进行z-score归一化，得到归一化的数据矩阵。

## ****混淆矩阵图.png****



1. 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）



## **完整的数据：归一化数据矩阵.txt已经放在github的文件夹里了**

1. 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

完整的相关矩阵数据已经放在github的文件夹里了

**完整的数据：100x3矩阵.txt已经放在github的文件夹里了**

**实验总结**：

（遇到的问题，有什么感想和收货等等）

学会了如何用pandas去读取数据，怎么在python画图，怎么用numpy来处理数据，去百度python如何使用，在这过程中学到了很多有关于机器学习的知识，在组员之间的相互配合下，我们终于完成了这个实验。