## 机器学习与数据挖掘实验1

**题目**：广州大学某班有同学100人，现要从两个数据源汇总学生数据。第一个数据源在数据库中，第二个数据源在txt文件中，两个数据源课程存在缺失、冗余和不一致性，请用C/C++/Java程序实现对两个数据源的一致性合并以及每个学生样本的数值量化。

* 数据库表：ID (int), 姓名(string), 家乡(string:限定为Beijing / Guangzhou / Shenzhen / Shanghai), 性别（string:boy/girl）、身高（float:单位是cm)）、课程1成绩（float）、课程2成绩（float）、...、课程10成绩(float)、体能测试成绩（string：bad/general/good/excellent）；其中课程1-课程5为百分制，课程6-课程10为十分制。
* txt文件：ID(string：6位学号)，性别（string:male/female）、身高（string:单位是m)）、课程1成绩（string）、课程2成绩（string）、...、课程10成绩(string)、体能测试成绩（string：差/一般/良好/优秀）；其中课程1-课程5为百分制，课程6-课程10为十分制。

#### 参考：

数据库中Stu表数据

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| ID | Name | City | Gender | Height | C1 | ... | C10 | Constitution |
| 1 | Sun | Beijing | boy | 160 | 87 |  | 9 | good |
| 2 | Zhu | Shenzhen | girl | 177 | 66 |  | 8 | excellent |
| ... | ... | ... | ... | ... | ... | ... | ... | ... |

student.txt中

|  |
| --- |
| ID Name City Gender Height C1 。。。 C10 Constitution  202001 Sun Beijing male 1.80 87 。。。 9 good  202003 Tang Hanghai male 1.56 91 。。。 10 general  ... ... ... .. ... .. ... ... ... |
|  |

两个数据源合并后读入内存，并统计：

1. 学生中家乡在Beijing的所有课程的平均成绩。
2. 学生中家乡在广州，课程1在80分以上，且课程10在9分以上的男同学的数量。
3. 比较广州和上海两地女生的平均体能测试成绩，哪个地区的更强些？
4. 学习成绩和体能测试成绩，两者的相关性是多少？

#### 提示：

参考数据结构：

Student{

int id;

string name;

vector<float> data;

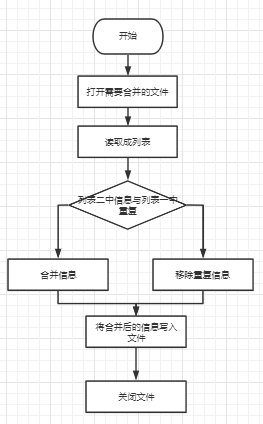
}

可能用到的公式：

|  |  |
| --- | --- |
| 均值公式 |  |
| 协方差公式 |  |
| z-score规范化 |  |
| 相关性 |  |

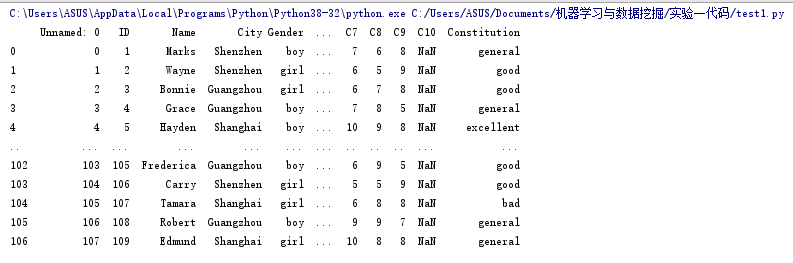
实验流程：

1. 先考虑如何合并两个数据源，具体思路如下：

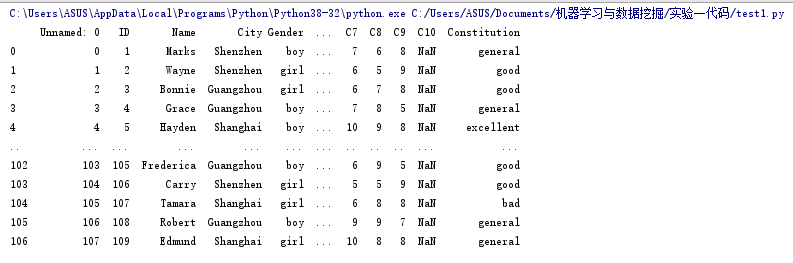


实验结果截图：

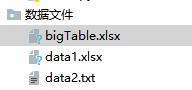
清洗后的数据：



数据分析结果：



使用到的文件：



实验心得:

实验1最难的部分我认为是清洗数据，我花了不少时间整理思路。

* 最开始遇到的问题是如何提取两个不同类型的文件数据，最后通过查找资料，我把两个数据文件都提取成列表，然后转化成dataframe类型。
* 其次是清洗数据。我的想法是用merge方法合并两个dataframe，然后去重。遇到的问题是两个dataframe的某些列的格式不一样，无法合并，要通过格式转换一系列步骤解决。
* 最后把合并成功的表写入excel文件中，完成数据清洗。
* 然后再进行数据计算分析。