《机器学习与数据挖掘实验》

指导老师：彭伟龙

课程设置

课程目标：

- 掌握数据集成，数据清洗，样本数据构建基础方法
- 掌握数据统计，条件汇总方法，数据度量，可视化基础方法
- 掌握常见机器学习原理和算法使用方法

实验课成绩考察：

- 实验成绩（90%）+考勤（10%）
- 5 次实验作业，成绩权重：20%,20%,20%,20%,20%。

理论课成绩考察：

- 实验成绩（70%）+考勤（30%）
- 自由选择 3 次实验作业，成绩权重：1/3,1/3,1/3。

作业提交内容：

1. 在 github 上建立一个仓库，按文件夹存放每次作业的内容。每次作业内容包括一下几点：
2. 代码+数据
3. 运行结果截图/文件
4. 说明文档 REDME.md（不会用 markdown 语法的同学可以用"实验报告.docx/.doc"代替），包含以下信息：

- 组员信息：标明**组长**，组员的分工信息
- 作业题目和内容
- 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
- 难题与解决
- 总结

验收提交方式：

- 分组完成作业，建议 3 人一组，原则上不超过 3 人。
- 对于理论课同学，请将组号（按照顺序编号），组员信息和 github 链接统计到共享 excel 文档：
  - 《机器学习与数据挖掘》-2020-2021-1-软件 181-183 作业汇总
- 对于实验课同学，请将组号（已指派），组员信息和 github 链接统计到共享 excel 文档：
  - 《机器学习与数据挖掘实验》-2020-2021-1-软件 182 作业汇总
- 注意：不允许 copy 抄袭，如有发现，不及格处理。如果确实不会而参考了其他组的作业完成，请在 README.md 里注明参照的 github 链接，并致谢。

课程结束后作业存档：

- **每个同学撰写个人的总结报告，将几次作业的代码/数据+运行结果+实验报告.docx，打包为一个 zip 文件，学号+姓名.zip，并提交。**

实验一　　《多源数据集成、清洗和统计》

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   180          87  。。。  9   good
202003   Tang   Hanghai male   156          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 9 在 9 分以上的男同学的数量。(备注：该处做了修正，课程 10 数据为空，更改为课程 9)
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？（九门课的成绩分别与体能成绩计算相关性）

提示

参考数据结构：

```
Student{
    int id;
    string id;
    vector<float> data;
}
```

可能用到的公式：

均值公式	$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$
协方差公式	$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{1}{n-1}[\sum_{i=1}^{n} x_i^2 - \frac{1}{n}(\sum_{i=1}^{n} x_i)^2$
z-score 规范化	$z_{if} = \frac{x_{if} - m_f}{s_f}$

数组 A 和数组 B 的相关性	$a'_k = (a_k - mean(A)) / std(A)$ $b'_k = (b_k - mean(B)) / std(B)$ $correlation(A,B) = A' \bullet B'$ 这里 A=[a_1, a_2,...a_k,..., a_n], B=[b_1, b_2,...b_k,..., b_n], mean(A)代表 A 中元素的平均值 std 是标准差，即对协方差的开平方。 点乘的定义：$A \cdot B = \sum_i a_i \cdot b_i$

注意：计算部分不能调用库函数；画图/可视化显示可以用可视化 API 或工具实现。

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. 计算协相关矩阵，并画出混淆矩阵。

5. 根据协相关矩阵，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。

实验三　《k-means 聚类算法》

题目

用 C++实现 k-means 聚类算法，

1. 对实验二中的 z-score 归一化的成绩数据进行测试，观察聚类为 2 类，3 类，4 类，5 类的结果，观察得出什么结论?
2. 由老师给出测试数据，进行测试，并画出可视化出散点图，类中心，类半径，并分析聚为几类合适。

注意

除文件读取外，不能使用 C++基础库以外的 API 和库函数。