

《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

成绩考察：

- 实验成绩（90%）+考勤（10%）
- 5 次实验作业的综合成绩权重分别为：20,20,20,20,20。

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

[illegible]

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

相关性	$a'_k = (a_k - \text{mean}(A)) / \text{std}(A)$ $b'_k = (b_k - \text{mean}(B)) / \text{std}(B)$ $\text{correlation}(A, B) = A' \bullet B'$
-----	--------------------------------------------------------------------------------------------------------------------------------------------

实验二 《数据统计和可视化》

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。