数据仓库与数据挖掘:第一次作业

南京农业大学

April 10, 2023

本次作业截止时间为 2023 年 4 月 20 日 24 时,请提交 pdf 文件和压缩后的代码文件(必须为 zip 格式), PDF 以"姓名 + 学号"命名,代码压缩包以"姓名 + 学号 + 代码"命名,例:"张三 +11519113.pdf""张三 +11519113+代码.zip"。请注意代码压缩包中只允许存在py 和 java 为后辍名的文件,且不允许存在文件夹。上传至提交地址,每人仅限上传一次,以第一次上传的结果为准。

所有编程题请使用 Python (3.9 以上版本) 或 JAVA (JDK 11 以上版本) 完成,且 仅限使用标准库。

任何雷同或与互联网相关资料重复均会触发双倍扣分惩罚。

任何不符合上述所有要求的作业均无效。

1 第一题

附件中的数据集 data.online.scores 是一门课学生考试成绩。每行数据代表一名学生的考试成绩,不同列的数据用制表符分隔,第一列数据代表学生的 id,第二列数据代表学生的期中考试成绩,第三列代表学生的期末考试成绩。

对于 a,b,c 三问,请编程解决,在 PDF 中给出结果,并提交相应的源代码,d 问请在 PDF 中回答。

- (a) 计算期末考试成绩的上四分位数、中位数、下四分位数
- (b) 计算期末考试成绩的平均值
- (c) 计算期末考试成绩的众数
- (d) 对于学生期末成绩的分布,数据是正倾斜的还是负倾斜的?给出理由。

2 第二题

对于 a,b,c 三题,请在 PDF 中给出详细计算步骤 (仅有结果不给分)和解释。对于 d,请编程解决,在 PDF 中给出结果,并提交相应的源代码。

(a) 给定两个对象 Obj1 和 Obj2, 每个对象都有 200 个二元属性, 表1是这两个对象的列联 表。请计算 obj1 和 obj2 的 Jaccard 系数。

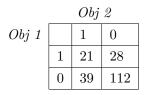


Table 1: Obj1 和 Obj 2 的列联表

- (b) 给出三维空间中的两个点 $A=(3,\ 1,\ 2)$ 和 $B=(-1,\ 0,\ 8)$ 。请计算他们的 Euclidean 距离、Manhattan 距离以及 $h=\infty$ 的 Minkowski 距离。
- (c) 对于空间中随机的两个点 C 和 D, 请解释为何它们的 Euclidean 距离总是小于等于 Manhattan 距离。
- (d) 附件 vectors.txt 中给出了两个向量 A 和 B, 每个向量有 100 个属性 (制表符分隔), 请分别计算 h=2 和 h=3 的 Minkowski 距离。

3 第三题

请使用 Z-score 规范化 data.online.scores 数据集中的期中考试成绩,请提交相应代码,并回答以下问题.

- (a) 计算比较规范化前后的平均值和方差。
- (b) 原始分数 90 规范化后是多少?

4 第四题

ChiMerge(自行搜索相关资料)是监督的、自底向上的(即基于合并的)数据离散化方法。它依赖于 χ^2 分析: 具有最小 χ^2 值得相邻区间合并在一起,直到满足确定的停止标准。

- (a) 请简述 ChiMerge 如何工作。
- (b) 取鸢尾花数据集作为待离散化的数据集合,鸢尾花数据集可以从 UCI 机器学习数据 库https://archive-beta.ics.uci.edu/ml/datasets/iris下载。使用 Chimerge 方法,对四个数值属性分别进行离散化。(令停止条件为: max-interval=6). 你需要写一个小程序,以避免麻烦的数值计算。提交你的简要分析和检验结果:分裂点、最终的区间以及源代码。