卧龙岗大学
计算机与信息技术学院 CSCI946 大数据分析
2024 年秋

作业 3

(20 分)

截止时间：北京时间 2024 年 11 月 29 日 23:59

目标

本作业旨在为学生提供使用 R 或 Python 编程语言进行大数据分析实验的基本经验。在本作业中，您应该

- 按照大数据分析生命周期进行大数据分析、
- 适当选择、应用和评估核心模型/算法和分析技术，以完成分析任务、
- 理解并整合本学科所学的知识和技能，包括大数据分析生命周期、数据准备、聚类、分类、回归、关联规则、数据/模型评估、数据可视化和文本/图像/网络数据处理。

小组合作：你们要以小组为单位完成这项作业。每个小组应独立于其他小组完成本作业。小组和小组成员在 Moodle 上有规定。您可以自行组建小组。每个学生只能加入一个小组。每个小组成员不得超过 5 人，不得少于 3 人。所有小组成员都应为本次作业做出贡献。请在开始作业前做好计划，然后为每个小组成员保存一份详细的数字工作日志和时间表。本作业的所有答案必须附有理由和/或解释。每个小组只能提交一份作业。

处罚：如果小组成员未能做出最低贡献，该成员将被记零分。声称贡献较少或没有贡献的成员应提供工作日志等证据。抄袭作业中的任何部分都将导致整个小组得零分。

序言

通读第 1 - 12 周的讲座幻灯片、实验指导和推荐读物。进行相关背景研究。您应该使用 R 或 Python 来完成本作业中的任务。您可以使用任何可公开访问的 R 和 Python 工具箱库。您提交的作业必须包括源代码文件，该文件在运行时将重新创建您的所有结果。

关于数据集和原始项目

作业 3 使用的金融交易数据集涉及 Kaggle.(https://www.kaggle.com/datasets/ealaxi/paysim1?resource=download)。原始数据集为 493.53 MB，个人电脑可能无法处理。A3 提供了一个调整大小的数据集 - **A3dataset.csv**。在新兴的移动支付交易领域，金融数据集对许多研究人员都很重要，尤其是对我们进行欺诈检测领域的研究。Kaggle 提供了一个使用名为 PaySim 的模拟器生成的合成数据集，作为解决此类问题的一种方法。
PaySim 使用私人数据集的汇总数据生成一个合成数据集，该数据集类似于正常交易操作，并注入恶意行为，以便日后评估欺诈检测方法的性能。

PaySim 根据从一个非洲国家实施的移动支付服务一个月的财务日志中提取的真实交易样本模拟移动支付交易。原始记录由一家跨国公司提供，该公司是移动金融服务的提供商，目前在全球超过 14 个国家开展业务。

这个合成数据集缩小了原始数据集的 1/4，是专为 Kaggle 创建的。出于练习的目的，我们将 A3 的比例进一步缩小了 85%。

注意：作业 3 不同于任何公共项目。抄袭任何公共项目将导致作业 3 得零分。

作业 3 的基本任务

作业 3 的目的是检测欺诈性金融交易。您的基本任务包括以下任务 1-4：

任务 1：按照大数据分析生命周期设计一个大数据分析项目。(3 分) **任务 2：**根据文章 "银行交易数据集及其与无标度网络的比较分析 "进行统计分析。尝试编程并制作图 1 - 图 16 等图形和表 I - 表 IV 等表格。需要对每个步骤和每个发现进行介绍和说明。没有介绍和解释的表格和图形将被扣分。(10 分）
任务 3：根据**任务 2** 的结果，应用至少两个必须使用聚类和分类方法的核心模型。最好应用神经网络。A3 不限制算法的选择和应用算法的数量。(5 分） **任务 4：**研究影响欺诈检测的因素，并从金融服务提供商的角度提出金融预警建议。(2 分）

要求在报告中以条理清晰的方式总结任务 1-4，并在写作中引用参考文章和编程资源。任务 1-4 可能需要 R/Python 编程来支持您的分析。

提交：

作业 3 的提交链接在本学科的 Moodle 网站上。每个小组只能提交一份作业。**提交的材料必须是两个文件。一个是以 "A3.pdf "命名的报告（必须提交）；另一个是以 "A3.zip "命名的压缩文件，小于 200 MB，包含代码（必须提交）和视频演示（可选）。**以下两种方式均可接受：

1. `pdf` 格式的报告，以及 `.R` 或 `.py` 或 `.ipynb` 格式的代码文件

可选择 `.mp4` 格式的视频演示或保存在 `.txt` 文件中的共享链接。

重要：

- 报告必须为单个文件，格式为 `.pdf` 或 `.ipynb`。扉页必须列出小组所有成员的全名和学号。清楚地标明没有做出最低贡献的成员。
- 报告没有页数限制。
- 报告将通过 Moodle 网站上的 **Turnitin** 系统进行**剽窃**检测。
- 描述不完整或含糊不清将被扣分。
- 应在代码中提供充分、适当和清晰的注释，使其易于理解。代码不整洁、难以阅读、无法运行或无法在报告中再现结果的代码将被扣分。

注意：代码运行失败可能会得零分。剽窃代码中的任何部分或报告中的任何部分都将导致本次作业得零分。小组有责任确保您提交的报告不包含抄袭材料。您可能会被要求在报告中演示和解释您的程序或解释您的答案。如果您无法解释，将被扣分。正确的设计、实施、风格、完整性和合理性都会得到分数。

---------- **END**----------