小组项目

INT3086 数据挖掘的数据结构和算法

2023-24
权重占期末考试的 40

1. 导言

本项目旨在将您在本课程中学到的知识应用到实际应用中。您需要使用数据挖掘技术对现实世界的数据集进行数据分析，并根据数据源对自我定义的有意义的应用进行研究和分析。

2. 组建小组

每个小组最多由 5 名成员组成。不建议少于 3 人的小组。

3. 过程

研究互联网上的以下资源，找到你们小组感兴趣的数据集，并以此为基础进行分析。

日期集（选择并下载到您的项目）

https://www.kaggle.com/datasets?datasetsOnly=true

GitHub - awesomedata/awesome-public-datasets：以主题为中心的总部开放数据集列表。

其他高级数据挖掘项目

https://favtutor.com/blogs/data-mining-projects
https://sites.google.com/site/datathinkingpractice/assignments/student-projects

[注意：这些资源中的项目说明仅供参考。您不需要按照此专业标准水平完成您的工作]。

您的一般流程

1. 选择您的小组感兴趣的数据集。确定要关注的主题。

2. 根据项目的重点，编写 python 代码来分析和可视化您的发现。
3. 使用 python 程序提取合适的数据，并根据分析重点展示数据。
4. 提供一份报告，支持您的计划成果，讨论任何发现。报告字数不超过 2000 字，但不包括附录。(例如，将不重要的数据作为附录）。

并非必须使用参考文献中提供的数据集。您可以搜索其他网站来选择数据集。但是，您的结果必须是原创的，并且与其他组别不同。应清楚解释结果的相关性。

4. 评估标准

数据分析结果（20）

您应该根据一些真实世界的数据集来解决问题。你的项目结果应该有证据支持，并应提供与小组收集的数据结果相关的有趣、有用和有意义的信息。(注意：本项目可选择提供预测）。

编程技能（20）

要实现分析目的，计划必须有效。结果要清晰、有意义。能在设计中应用良好的数据结构，并能证明（通过分析和讨论）在时间复杂性方面的有效性能。

报告（20）

内容包括项目的所有重要组成部分，格式布局合理、一致。可读性强，无重大语法问题。易于理解，并提供适当的图形、表格和截图。

难度（30）

我们将根据项目的难度和您所展示的知识对您的项目进行评估。为了获得更高的分数，你可以解释为什么你的数据分析工作是复杂的，以及你使用了什么技术来解决你遇到的问题。您还可以解释数据挖掘技术背后的原理，以证明您的理解能力。

小组展示（约 20 分钟录音）（10%）。

演示文稿应简明扼要地解释和展示项目成果。因此，请尽量在演示文稿中清楚地解释*为什么*你认为你的项目应该获得高分。如果你在陈述中没有给出相关的理由和证据，指导教师可能不会给你的项目足够的学分。

5. 建议的报告组成部分

您可以考虑在您的项目中包含以下部分内容：（注意，不限于这些内容。您可以根据自己的项目设计报告，也可以与下列内容不同）。

- 解释所选数据集提供的信息。例如，它如何为实际应用提供价值。
- 提供用于分析数据集的数据挖掘算法的实施方案，以便在所选数据集上实现你的数据挖掘想法。
- 解释数据挖掘算法背后的设计原则。
- 报告数据挖掘算法的结果。
- 分析算法设计的时间复杂性。如果有，找出其中的缺陷，并就算法设计提出解决方案

。

- 提出结论和改进建议。

6. 提交

每个小组都必须在到期日或之前（将在 Moodle 上设置）提交以下内容。**一个小组只需提交一名成员**。逾期提交可能无法获得分数。

划界案内容概要：

1. 所选数据集（如 CVS 文件）
2. 源代码（如所有 Python 程序文件）
3. 项目报告（如 pdf 文件格式）
4. Powerpoint 演示文稿，并在第一张幻灯片上提供访问演示文稿录音的链接（如有密码，请提供）。

注意：

压缩所有文件（不包括演示视频）和文件夹 **提交到 Moodle 的[小组项目提交]链接下**。

对于录音，您可以将其存储在 Google Drive 或 youTube 中，并在报告中提供共享/访问链接。切勿将录音上传到 Google 的私人区域，因为这需要访问权限。

您的报告应为 pdf 格式，源代码应为 Python 源代码文件。如果有多个 Python 程序文件，需要将它们放在一个名为 Python 源代码的文件夹中。

7. 剽窃政策

您的项目应代表您所在小组的工作成果。不要在项目中包含任何不是由小组成员编写的代码。您可以使用任何外部 Python 库和数据集，但必须充分说明其来源，并清楚解释项目中哪些部分不是由您所在小组完成的。

[一般概念适用于剽窃，例如，您导入一个 Python 库使用，这不会被视为剽窃。但是，如果你从一些现有程序中复制并粘贴了一组代码，你可能需要解释为什么以及如何使用这些代码，并声明它们不是你的代码]。