mlp-s2-2021 / **mlp-proj1**

Mlp-s2-2021/mlp-proj1

0 stars 星星 1 fork 叉子

Star 明星 Watch 看着

Code 代码 Issues 问题 Pull requests 调出请求 Actions 行动 Projects 项目

master 主人

mlp-proj1 / **README.md**

rundel 兰德尔 Initial 首字母 History 历史

1 contributor 贡献者

Raw 生食 Blame 责备

97 lines (46 sloc) 97行(46 sloc) 7.33 KB

Machine Learning in Python - Project 1
Python 中的机器学习-项目1

Due Friday, March 5th by 5 pm (UK local time)

3月5日(星期五)下午5点(英国当地时间)

Data 境监察及审核资料

We will be using data about the American TV show The Office. These data are provided as `the_office.csv` in this repository and are meant to give you a reasonable starting point for this assignment. The dataset includes the following columns:

我们将使用有关美国电视节目《办公室》的数据。这些数据作为办公室提供。在这个存储库中的 csv，目的是为您提供一个合理的起点，为这个任务。数据集包括以下列表:

- `season` - Season number of the episode

  第一季-第一集

- `episode` - Episode number within a season

  第集——一季中的第集

- `episode_name` - Episode name

  Episode _ name-Episode name

- `director` - Episode director(s), names are separated by `;`

  导演——集导演，名字分开;

- `writer` - Episode writer(s), names are separated by `;`

  作家-插曲作家(s) ，名称分开;

- `imdb_rating` - Episode rating on IMDB

  IMDB _ rating-在 IMDB 上的插曲等级

- `total_votes` - Number of ratings for episode on IMDB

  总票数-在 IMDB 上收视人数

- `air_date` - Original air date of episode

  播出日期-第集播出日期

- `n_lines` - Number of spoken lines in episode

  N _ 行-集中的口语台词数

- `n_directions` - Number of lines containing a stage direction

  N _ 方向-包含舞台方向的行数

- `n_words` - Number of dialog words in episide

  N _ words-episide 中的对话词数

- `n_speak_char` - Number of different characters with spoken lines in episode

  N _ speak _ char-集中不同人物的口语台词数

- `main_chars` - Main characters appearing in episode (main characters were determined to be characters appearing in more than 1/2 of the episodes)

  主要角色——出现在剧集中的主要角色(主要角色是出现在超过1/2集中的角色)

These data were derived from the data available in the `schrutepy` package. The package provides a data frame containing the entire text transcripts from all episodes of the show.

这些数据来自 schrutepy 包中的可用数据。这个包提供了一个数据框架，其中包含了所有剧集的全部文本记录。

Assignment 转让

For the purposes of the project, consider yourself a Data Scientist contractor who has been hired by NBC Universal to advise on the creation of a special reunion episode of The Office. Your employers are particularly interested in understanding what made some episodes more popular than others. As such, your task is to use these data (or any other) to build a predictive model that captures the underlying relationships between these features and the audience ratings, and then use the insights you gain from this model to advise what NBC Universal should do to produce the highest rated reunion episode possible.

为了这个项目的目的，你可以把自己当作一个数据科学家承包商，NBC 环球公司聘请你为《办公室》的一个特别重组插曲的制作提供建议。你的雇主特别想知道是什么让某些剧集比其他剧集更受欢迎。因此，你的任务是利用这些数据(或任何其他数据)建立一个预测模型，捕捉这些特征和收视率之间的潜在关系，然后利用你从这个模型中获得的洞察力，建议 NBC 环球应该做些什么，以尽可能制作出收视率最高的重播剧集。

In other words, you need to develop an *understandable validated* model for The Office's `imdb_rating` as the outcome of interest using features derived from the data provided and any additional sources you would like. It is important that this model be accurate **and** reliable and any conclusions you draw well supported and sound. We explicitly do not want a blackbox model - you should be able to explain and justify your modeling choices and your model's predictions.

换句话说，你需要开发一个可以理解的验证模型，将 Office 的 imdb 评级作为感兴趣的结果，使用来自所提供数据的特性和任何你想要的其他来源。重要的是，这个模型是准确和可靠的，任何结论，你得出的支持和健全。我们明确地不想要一个黑盒模型-你应该能够解释和证明你的建模选择和你的模型的预测。

Your model may use as few or as many of the provided features, and you may transform and manipulate these features in any way that you want to generate additional features.

您的模型可以使用少量或多量提供的特性，您可以以任何您想要生成附加特性的方式转换和操作这些特性。

We have covered a number of models and modeling approaches in the lectures and workshops and you should explore a variety of different approaches for this particular task. However, your ultimate goal is to deliver a **single** model. These are competing interests and it is up to you to find a reasonable balance between the two - some of your marks will be based on how well you accomplish this.

我们已经在讲座和研讨会中介绍了一些模型和建模方法，您应该为这个特定任务探索各种不同的方法。然而，您的最终目标是交付一个单一的模型。这些都是相互竞争的利益，这取决于你在两者之间找到一个合理的平衡点——你的一些分数将取决于你完成这项工作的程度。

Required Structure & Formatting 所需的结构和格式

We have provided a templated Jupyter notebook called `project1.ipynb` which includes the required sections along with brief instructions on what should be included in each section. Your completed assignment must follow this structure - you should not add or remove any of these sections, if you feel it is necessary you may add addition subsections within each. Please remove the instructions for each section in the final document.

我们提供了一个名为 project1.ipynb 的模板化 Jupyter 笔记本，其中包括所需的部分以及关于每个部分应该包含哪些内容的简要说明。你完成的作业必须遵循这个结构——你不应该添加或删除这些部分中的任何一个，如果你觉得有必要，你可以在每个部分中添加附加子部分。请删除最终文档中每个部分的说明。

All of your work must be contained in the `project1.ipynb` notebook, we will only mark what is included in this file.

你所有的工作必须包含在项目1.ipynb 笔记本中，我们将只标记包含在这个文件中的内容。

Our expectation is that most projects will be roughly 10-15 pages in length (text & figures, excluding code). Your notebook must include all of your work, but make sure that you are only retaining required components, e.g. remove unused code and figures (if a figure is not explicitly discussed in the text it should not be in the final document).

我们的预期是，大多数项目将大约10-15页的长度(文本和数字，不包括代码)。你的笔记本必须包括你所有的工作，但是要确保你只保留必需的组件，例如删除未使用的代码和图形(如果一个图形在文本中没有明确讨论，那么它就不应该出现在最终文档中)。

Overall, your project will be partially assessed on your organization / presentation of the document - it should be as polished and streamlined as possible. We highly recommend that you check the appearance of your rendered PDF before submitting, as its appearance can differ significantly from the notebook.

总的来说，你的项目会在你的组织/文档的展示中得到部分的评估——它应该尽可能的优化和简化。我们强烈建议您在提交之前检查您呈现的 PDF 的外观，因为它的外观可能与笔记本电脑有很大的不同。

Group work 小组活动

This project may be completed by a team of up to 4 students. We will not be assigning or forming teams, if you are a team that is looking for more members or someone looking for a team please use the pinned post on Piazza to find each other.

这个项目最多可由4名学生组成的团队完成。我们不会分配或组成团队，如果你是一个团队，正在寻找更多的成员或有人寻找一个团队，请使用广场上的帖子找到彼此。

After the assignment is completed we will distribute a brief peer evaluation survey - members who contributed significantly less than their peers will potentially have their overall mark penalized.

在任务完成后，我们将发布一个简短的同行评估调查-成员谁贡献明显少于他们的同行将有可能受到处罚他们的总分。

Submission 提交

This project is due Friday, March 5th by 5 pm (UK local time). You are expected to submit your completed work as follows:

这个项目是由于星期五，3月5日下午5时(英国当地时间)。你须提交以下已完成的工作:

- Submit a PDF of your report (generated from a Jupyter notebook) to the Project 1 assignment on Gradescope.

  提交一份 PDF 格式的报告(从一个 Jupyter 笔记本生成)到 Gradescope 上的项目1作业。

- Submit your `ipynb` notebook file and any supplementary data and script files to the Project 1 assignment on Learn.

  提交您的 ipynb 笔记本文件和任何补充数据和脚本文件的项目1作业的学习。

Both submission steps are necessary for your work to be considered submitted. Standard late penalties will apply if either piece is not submitted by the deadline.

这两个提交步骤是必要的，您的工作被认为是提交。如果任何一件作品没有在截止日期前提交，标准的逾期处罚将适用。

For a team submission - all contributors should be added to the assignment. Only one team member needs to submit the ipynb notebook on Learn.

对于一个团队提交-所有贡献者应该被添加到任务。只有一个团队成员需要在 Learn 上提交 ipynb 笔记本。

Marking Rubric 标记规则

The project will be marked out of 100, and we will be using the following rubric to roughly guide the marking:

该项目将以100分为标记，我们将使用以下标题粗略指导标记:

- **>90**: The code runs without errors. Models are implemented, fit, and assessed correctly. The final model achieves a high level of predictive accuracy and is well documented and described in the writeup. There is significant and creative additional investigation of the problem including the use of addition data sources for features. Potentially could be used as a model answer. Write-up evidences deep understanding of the data and the model(s).

> 90: 代码运行时没有错误。模型被正确地实现、匹配和评估。最终的模型实现了高水平的预测准确性，并在记录中得到了良好的文档化和描述。对这个问题进行了重要和创造性的补充调查，包括对特性使用添加数据源。可能会被用作一个模型答案。写作证明了对数据和模型的深刻理解。

- **80-89**: The code runs without errors. Models are implemented, fit, and assessed correctly. The write up is generally good and the code is appropriately commented. The final model achieves a reasonable level of predictive accuracy and is well documented and described in the writeup. There is moderate additional investigation of the problem. Write-up evidences good understanding of the data and the model(s).

  80-89: 代码运行时没有错误。模型被正确地实现、匹配和评估。编写工作通常很好，代码也有适当的注释。最终的模型达到了一个合理的预测准确性水平，并在记录中得到了很好的记录和描述。对这个问题还有适度的补充调查。写作证明对数据和模型有很好的理解。

- **70-79**: The code runs without errors. Models are implemented, fit, and assessed correctly with only minor issues. The write-up is reasonable but could be better. Write-up evidences adequate understanding of the data and the model(s).

  70-79: 代码运行没有错误。模型的实现、匹配和评估都是正确的，只有一些小问题。这种报道是合理的，但可能会更好。写作证明对数据和模型有充分的理解。

- **60-69**: The code runs without errors. Models are implemented, fit, and assessed correctly with only moderate issues. The write-up is ok but could be better, includes some moderate errors or omissions. Write-up evidences adequate understanding of the data and the model(s).

  60-69: 代码运行时没有错误。模型的实施、适应和评估都是正确的，只有中度的问题。这篇文章写得还可以，但可以更好，包括一些适度的错误或遗漏。写作证明对数据和模型有充分的理解。

- **50-59**: The code runs with some errors. Models are implemented, fit, and assessed but with some significant issues in implementation and or understanding. The write-up is marginal and includes some significant errors or omissions. Write-up evidences an incomplete understanding of the data and the model(s).

  50-59: 代码运行时有一些错误。模型是实现的、适合的和评估的，但是在实现和理解过程中存在一些重要的问题。这些报道是边缘性的，包括一些重大错误或遗漏。写作证明对数据和模型的理解是不完整的。

- **<49**: Significant issues with the code, model(s), and/or the write up. Write-up evidences an incomplete understanding of the data and the model(s).

  < 49: 代码、模型和/或编写的重大问题。写作证明对数据和模型的理解是不完整的。

Using these criteria, specific rubrics will be used to assess each of the 4 required sections of the project.

使用这些标准，具体的规则将用于评估项目的4个必要部分。