Project Guides for Python for Data Analysis

Quách Đình Hoàng

2022/04/05

Thông tin chung về project môn học

- Project môn học sẽ bao gồm các phân tích trên tập dữ liệu do mỗi nhóm lựa chọn. Tập dữ liệu có thể đã tồn tại hoặc mỗi nhóm có thể tự thu thập dữ liệu bằng cách sử dụng các khảo sát hoặc bằng cách tiến hành một thí nghiệm. Mỗi nhóm có thể chọn dữ liệu dựa trên mối quan tâm của nhóm. Mục tiêu của project này là để các nhóm thể hiện sự thành thạo trong các kỹ thuật đã được đề cập trong môn học này (hoặc hơn thế nữa) và áp dụng chúng vào tập dữ liệu mới theo cách có ý nghĩa.
- Mục tiêu của project không phải là thực hiện phân tích dữ liệu toàn diện, tức là không tính toán mọi thống kê và kiểm định đã học cho mọi biến, mà là để GV biết rằng các nhóm thành thạo trong việc đặt những câu hỏi có ý nghĩa và trả lời chúng bằng kết quả phân tích dữ liệu, rằng các nhóm thành thạo trong việc sử dụng Python, trong việc diễn giải và trình bày kết quả. Mỗi nhóm cần tập trung vào các kiểm định thống kê giúp trả lời các câu hỏi nghiên cứu đặt ra của nhóm. Các nhóm không cần phải áp dụng mọi phương pháp thống kê đã học. Ngoài ra, mỗi nhóm cần có nhận xét về các kiểm định và đưa ra các đề xuất để cải thiện phân tích. Các vấn đề liên quan đến độ tin cậy và hợp lệ của dữ liệu cũng như tính thích hợp của các phân tích, kiểm định thống kê nên được thảo luận.
- Project có thể có kết thúc mở. Mỗi nhóm nên tạo một số loại biểu đồ trực quan hấp dẫn về tập dữ liệu đã lựa chọn dùng Python. Không có giới hạn về công cụ hoặc thư viên nào mỗi nhóm có thể sử dụng. Tuy nhiên, các nhóm nên sử các thư viện sau: `numpy`, `pandas`, `mathplotlib`, `seaborn`, `scikit-learn`. Các nhóm không cần phải trực quan hóa tất cả dữ liệu cùng một lúc. Một biểu đồ chất lượng cao sẽ nhận được điểm cao hơn nhiều so với nhiều biểu đồ chất lượng kém. Chất lượng bài thuyết trình cũng quan trọng. Bài trình bày súc tích và rõ ràng sẽ được đánh giá cao. Tất cả các phân tích phải được thực hiện bằng notebook (.ipynb) sử dụng Python và một IDE (Jupyter Lab hay VS Code).
- **Teamwork**: Các thành viên phải hoàn thành project với tư cách là một nhóm. Tất cả các thành viên trong nhóm phải có trách nhiệm đóng góp như nhau (cả về chất lượng và số lượng) vào việc hoàn thành project. Các thành viên trong nhóm nên dành thời gian để làm việc cùng nhau. Bất kỳ ai bị đánh giá là không có đóng góp đủ cho sản phẩm cuối cùng sẽ bị phạt điểm. Mặc dù các thành viên trong nhóm có thể có nền tảng và khả năng khác nhau, nhưng trách nhiệm của mọi thành viên trong nhóm là phải hiểu cách thức và lý do tại sao tất cả các mã chương trình hoạt động và cách tiếp cận trong project là hợp lý.

Dữ liệu (Data)

- Để đạt được kết quả tốt nhất với project môn học, điều quan trọng là các nhóm phải chọn một tập dữ liệu có thể quản lý được. Điều này có nghĩa là dữ liệu phải dễ truy cập và đủ lớn để có thể khám phá nhiều mối quan hệ. Tập dữ liệu của mỗi nhóm phải có ít nhất 150 quan sát và từ 10 đến 20 biến và nên ở dạng csv (có thể có ngoại lệ nhưng bạn phải trao đổi với GV trước). Các biến trong tập dữ liệu nên bao gồm các biến phân loại, các biến số rời rạc và các biến số liên tục.
- Dưới đây là danh sách các nguồn dữ liệu mà các nhóm có thể xem xét để lựa chọn. Các nhóm không bị giới hạn ở những nguồn dữ liệu này. Nhưng đây là các nguồn hữu tích để các nhóm có thể tìm thấy tập dữ liệu phù hợp cho vấn đề mà nhóm quan tâm:

- Kaggle datasets: https://www.kaggle.com/datasets
- OpenIntro datasets: http://openintrostat.github.io/openintro/reference/index.html
- Awesome public datasets: https://github.com/awesomedata/awesome-public-datasets
- UC Irvine machine learning datasets: https://archive.ics.uci.edu/ml/index.php

Những thứ cần nộp

- **Proposal**: hạn chót: **23h59, 17/04/2022**.
- **Milestone**: hạn chót: **23h59, 22/05/2022**.
- **Presentation**: hạn chót: **23h59, 12/06/2022**.
- **Report**: hạn chót: **23h59, 19/06/2022**

Đề xuất (Proposal)

Đề xuất cần bao gồm các phần sau:

Phần 1 - Giới thiệu

- Phần này nên giới thiệu câu hỏi nghiên cứu chung và dữ liệu của nhóm (nó đến từ đâu, thu thập như thế nào, các trường hợp xảy ra là gì, các biến số là gì, v.v.). Phần này không quá 1 trang (không kể số liệu).

Phần 2 - Dữ liệu

- Đặt dữ liệu của nhóm vào thư mục `data`, cho biết số chiều (số dòng, số cột) và giải thích các biến trong file `README` trong thư mục đó. Thực hiện phần EDA để in ra vài dòng đầu và cuối trong tập dữ liệu của nhóm (nên dùng thư viện `pandas` để load và thao tác trên tập dữ liệu).

Phần 3 - Kế hoạch phân tích dữ liệu

- Các biến kết quả (phản hồi, `Y`) và dự đoán (giải thích, `X`) mà nhóm sẽ sử dụng để trả lời câu hỏi của mình.
- Các nhóm so sánh nhóm sẽ sử dụng, nếu có.
- Phân tích dữ liệu sơ bộ, bao gồm một số thống kê tóm tắt và biểu đồ trực quan hóa, cùng với một số giải thích về cách chúng giúp nhóm hiểu thêm về tập dữ liệu. (Nhóm có thể cập nhật những thông tin khi hoàn thành project.)
- (Các) phương pháp thống kê mà nhóm tin rằng sẽ hữu ích trong việc trả lời (các) câu hỏi đặt ra. (Nhóm có thể cập nhật những thông tin này khi hoàn thành project.)
- Kết quả nào từ các phương pháp thống kê cụ thể này là cần thiết để hỗ trợ cho giả thuyết của nhóm?

Bài thuyết trình (Presentation)

- Mỗi nhóm có **tối đa 10 phút** để trình bày và mỗi thành viên trong nhóm nên nói điều gì đó quan trọng. Các nhóm có thể trình bày trực tiếp trong buổi báo cáo của mình hoặc ghi âm trước và gửi video để phát trong buổi báo cáo. Ngay cả khi đã ghi video trước, mỗi nhóm vẫn phải có mặt ở buổi báo cáo để xử lý các vấn đề phát sinh và trả lời các câu hỏi.
- Mỗi nhóm cần chuẩn bị một bản trình chiếu (slides) bằng Jupyter notebook với cú pháp Markdown. Không có giới hạn về số lượng slide trình bày mỗi nhóm có thể sử dụng, chỉ có giới hạn thời gian (tối đa 10 phút). Mỗi thành viên trong nhóm nên có cơ hội phát biểu trong buổi thuyết trình. Bản trình bày của nhóm không nên chỉ là tất cả những gì nhóm đã thử (chúng tôi đã làm điều này, sau đó chúng

tôi đã làm điều này, v.v.), thay vào đó nó phải truyền tải những lựa chọn mà nhóm đã thực hiện, tại sao và những gì nhóm thấy thú vị.

- Lịch trình thuyết trình: Các bài thuyết trình sẽ diễn ra trong buổi báo cáo cuối cùng. Mỗi nhóm có thể chọn trình bày trực tiếp bài trình bày của mình hoặc ghi âm trước. Trong buổi báo cáo, mỗi nhóm sẽ xem các bài thuyết trình từ các nhóm khác và cung cấp phản hồi dưới hình thức đặt câu hỏi và đánh giá ngang hàng. Thứ tự trình bày sẽ được GV lựa chọn ngẫu nhiên.

Báo cáo (Report)

- Cùng với các slide trình bày, mỗi nhóm cần cung cấp một báo cáo về project của nhóm. Báo cáo này cũng được viết dùng Jupyter notebook với cú pháp Markdown. File báo cáo là file .html được xuất ra từ file .ipynb.
- Báo cáo này là bản tóm tắt về project của nhóm. Nó cung cấp thông tin về tập dữ liệu nhóm đang sử dụng, (các) câu hỏi nghiên cứu, phương pháp luận và kết quả.

Tổ chức project

Project của nhóm cần được tổ chức theo cấu trúc sau:

- **Thư mục `data`**: gồm tập dữ liệu (dataset) ở dạng .csv được sử dụng cùng với file `README.md` để mô tả về tập dữ liệu.
- **Thư mục `image` (nếu có)**: gồm các file ảnh được sử dụng.
- **Đề xuất**: gồm `proposal.ipynb` và `proposal.html`.
- **Bài thuyết trình**: gồm `presentation.ipynb` và `presentation.html`.
- **Báo cáo**: gồm `report.ipynb` và `report.html`.

Ghi chú: Cách trình bày (proposal, presentation, report) cũng được tính điểm. Các nhóm nên trình bày chúng sao cho rõ ràng. Code cũng nên được định dạng hợp lý.

Cách tính điểm

- **Proposal**: 10%
- **Milestone**: 10%
- **Presentation**: 20%
- **Report**: 40%
- **Đánh giá của các bạn trong nhóm**: 10%
- **Hỏi đáp**: 10%

Các tiêu chí chấm điểm

- **Nội dung**: Chất lượng của câu hỏi nghiên cứu và mức độ liên quan của dữ liệu với những câu hỏi đó.
- **Tính đúng đắn**: Các thủ tục thống kê có được thực hiện và giải thích một cách chính xác không?
- **Viết và trình bày**: Chất lượng của bài trình bày, báo cáo và giải thích.
- **Sự sáng tạo và tư duy phản biện**: Project có được suy nghĩ cẩn thận không? Những hạn chế có được xem xét cẩn thận? Thời gian và công sức đã dành cho việc lập kế hoạch và thực hiện dự án?

Đánh giá giữa các thành viên trong nhóm

- Mỗi thành viên sẽ được yêu cầu cung cấp một đánh giá cá nhân trong đó mỗi bạn đánh giá sự đóng góp và tinh thần đồng đội của từng thành viên trong nhóm trên thang điểm 10. Mỗi thành viên cần báo cáo những công việc mà các thành viên khác đã thực hiện và tỷ lệ phần trăm đóng góp của mỗi thành viên trong nhóm. Đánh giá cá nhân là điều kiện tiên quyết để nhận được điểm cho project. Nếu bạn cho rằng một cá nhân nào trong nhóm đã làm ít hơn 20% công việc, hãy cung cấp một số lời giải thích. Nếu bất kỳ cá nhân nào nhận được điểm đánh giá trung bình cho thấy rằng họ đã làm ít hơn 15% công việc, người này sẽ nhận được một nửa số điểm của những người còn lại trong nhóm. Đánh giá của bạn trong nhóm chiếm 10% điểm project. Nói chúng, giáo viên sẽ tôn trọng ý kiến của các thành viên trong nhóm về mức độ đóng góp. Tuy nhiên, nếu có những điểm mà giáo viên thấy chưa hợp lý, đánh giá này có thể được xem xét và điều chỉnh bởi giáo viên.