CSC14119 – NHẬP MÔN KHOA HỌC DỮ LIỆU

ĐỒ ÁN THỰC HÀNH

Trong môn học này, sinh viên đã được cung cấp những kiến thức và kỹ năng cơ bản về khoa học dữ liệu để có thể giải quyết những vấn đề thực tế dựa trên dữ liệu. Thông qua những chủ đề như tổng quan quy trình khoa học dữ liệu: thu thập dữ liệu bằng cách công cụ; tiền xử lý dữ liệu để làm sạch dữ liệu (thường đan xen với pha phân tích và khám phá dữ liệu); phân tích thống kê dữ liệu với những tính toán đơn giản và trực quan hóa dữ liệu để khám phá ra những thông tin hữu ích từ dữ liệu để trả lời những câu hỏi đề ra; lưu trữ và tính toán hiệu năng; mô hình hóa và đánh giá mô hình; trình bày kết quả. Phần tổng hợp lại cho môn học, đồ án thực hành, được chia thành hai phần bao gồm **Phần 1: Quy trình khoa học dữ liệu** và **Phần 2: Mô hình hóa đơn giản**. Đồ án thực hành sẽ được thực hiện trong **5 tuần**, và tiến hành theo hình thức làm việc nhóm. Các phần tiếp theo sẽ trình bày chi tiết về nội dung, cách thực hiện, yêu cầu, và hình thức bảo vệ đồ án.

PHẦN 1: QUY TRÌNH KHOA HỌC DỮ LIỆU

Sinh viên thực hiện tìm kiếm, thu thập dữ liệu về một chủ đề mà nhóm sinh viên quan tâm, hứng thú. Sinh viên thực hiện một quy trình khoa học dữ liệu đối với dữ liệu đã thu thập được: khám phá dữ liệu (trong quá trình khám phá dữ liệu, có thể dừng lại để tiền xử lý rồi mới tiếp tục khám phá dữ liệu), xác định các câu hỏi có thể trả lời được bằng dữ liệu (số lượng câu hỏi không phải yếu tố quan trọng nhất, quan trọng nhất và ưu tiên hàng đầu là các câu hỏi cần có ý nghĩa), tiền xử lý dữ liệu và phân tích để trả lời cho mỗi câu hỏi.

A Thu thập dữ liệu (Data collection)

Mỗi nhóm phải thu thập dữ liệu liên quan đến chủ đề cần giải quyết, mà nhóm quan tâm (thích). Sinh viên phải quản lý và xử lý dữ liệu theo cách thủ công bằng các công cụ hỗ, ví dụ: bằng cách sử dụng selenium, request hoặc API. **Không được phép sử dụng tập dữ liệu đã được công khai hoặc có sẵn trên Kaggle trong đồ án này**. Dữ liệu phải được cấu trúc hóa thành một bảng gồm ít nhất 5 thuộc tính (trường dữ liệu) và 1000 dòng (records).

Lưu ý: Không được phép sử dụng dữ liệu đã thu thập được để sử dụng trong môn học **Lập trình cho Khoa học Dữ liệu**. Nếu phát hiện, đồ án thực hành sẽ được **điểm 0**.

B Khám phá dữ liệu (thường đan xen với tiền xử lý dữ liệu)

Sinh viên thực hiện khám phá dữ liệu đã thu thập bằng cách sử dụng thống kê mô tả để hiểu dữ liệu tốt hơn, tức là để xác định các vấn đề về dữ liệu (dữ liệu bị thiếu giá trị, giá trị không hợp lệ, cột có kiểu dữ liệu không phù hợp để xử lý thêm, v.v.). Dưới đây là một số thông tin để xem xét:

- Mỗi dòng có ý nghĩa gì? Có vấn đề các dòng có ý nghĩa khác nhau không?
- Mỗi cột có ý nghĩa gì?
- Mỗi cột hiện đang có kiểu dữ liệu gì? Khoảng biểu diễn của kiểu dữ liệu đó ra sao? Có cột nào có kiểu dữ liệu chưa phù hợp để có thể xử lý tiếp hay không?
- Với mỗi cột, dữ liệu phân bố như thế nào?

C Đặt các câu hỏi có ý nghĩa cần trả lời

Mỗi nhóm đưa ra ít nhất 5 câu hỏi có thể được trả lời bằng dữ liệu này. Tất cả các câu hỏi phải có ý nghĩa (Lợi ích của việc tìm ra câu trả lời là gì?), Và bạn không thể trả lời chúng một cách rõ ràng.

Trong tập tin notebook, với mỗi câu hỏi, sinh viên cần trình bày:

- Câu hỏi là gì?
- Nếu trả lời được câu hỏi thì sẽ có lợi ích gì?
- Nguồn cảm hứng của câu hỏi: tự nghĩ hay tham khảo ở đâu?

Các câu trả lời cần được thể hiện bằng hình ảnh để người hướng dẫn và các nhóm khác có thể hiểu chúng mà không nghi ngờ gì cả.

PHẦN 2: MÔ HÌNH HÓA DỮ LIỆU

Với dữ liệu đã thu thập, tiền xử lý và phân tích ở bên trên, sinh viên thực hiện xây dựng mô hình cho bài toán mình quan tâm trên tập dữ liệu này.

A Mô hình hóa dữ liệu và đánh giá mô hình

Tùy bài vào bài toán, sinh viên tự mô hình bằng cách thuật toán máy học đơn giản như thuật toán hồi quy, thuật toán phân lớp, …. Sinh viên cần phân tích và lựa chọn ra những đặc trưng có mức độ liên quan cao, có tính quan trọng cho bài toán đang xem xét.

Ví dụ:

- Dự đoán giá nhà thông qua các đặc trưng của ngôi nhà (số phòng ngủ, diện tích, …)
- Phân loại người dùng (tin cậy, không tin cậy) dựa trên các đặc trưng của người dùng (số ngày mua hàng, số tiền mua hàng hàng tháng/ quý/ năm, …)

B Đánh giá mô hình

Sau khi chọn các thuật toán mô hình hóa thích hợp (nhóm cần lựa chọn ít nhất hai môn hình trở lên để so sánh), mỗi nhóm phải xác thực kỹ lưỡng các siêu tham số của mô hình bằng cách sử dụng các kỹ thuật khoa học dữ liệu (ví dụ: sử dụng tập kiểm định (validation set) hoặc kỹ thuật cross-validation) và báo cáo quá trình tinh chỉnh (fine-tuning process). Hiệu suất của các mô hình phải được đánh giá bằng cách sử dụng các độ đo đánh giá phân loại phổ biến (accuracy, precision, recall) hoặc độ đo đánh giá hồi quy (lỗi bình phương trung bình - MSE và sai số bình phương trung bình căn - RMSE).

PHẦN 3: TỔNG HỢP KẾT QUẢ

Sau khi hoàn thành đồ án, mỗi nhóm viết một báo cáo để đánh giá lại công việc như sau:

- Từng thành viên: Bạn đã gặp những khó khăn gì?
- Từng thành viên: Bạn đã học được gì?
- Nhóm của bạn: Bạn sẽ làm gì nếu có nhiều thời gian hơn?

CÁC QUY ĐỊNH

1 LÀM VIỆC NHÓM

Mỗi nhóm phải sử dụng Git và GitHub để kiểm soát phiên bản và tương tác với các thành viên khác một cách hiệu quả. Mỗi giai đoạn hoặc tác vụ phải có nhánh riêng của nó thay vì tập trung mọi thứ cho nhánh chính (main/master branch). Nhóm sinh viên cần đảm bảo các yêu cầu sau:

- Một kế hoạch cho từng nhiệm vụ được lập cẩn thận (Ai sẽ thực hiện nhiệm vụ? Mất bao lâu để giải quyết nó?)
- Khối lượng công việc được cân bằng giữa các thành viên (Lịch sử cam kết trong Github sẽ cho thấy điều đó)
- Mỗi thành viên phải hiểu tường tận công việc của các thành viên khác trong nhóm.

Kế hoạch và lịch trình phải được theo dõi bằng các công cụ như Notion và Trello. Mỗi nhóm cần thể hiện chiến lược tổng thể và công việc của từng thành viên trong slide báo cáo cuối cùng.

2 CÁC YÊU CẦU THỰC HIỆN ĐỒ ÁN

a) Tổ chức thư mục cho đồ án: Các file notebooks phải được tách biệt rõ ràng cho từng giai đoạn, từ thu thập dữ liệu, tiền xử lý dữ liệu, phân tích đến xây dựng mô hình, đánh giá và phân tích kết quả.

b) Việc trả lời câu hỏi cần được thể hiện thông qua các hình vẽ biểu đồ trực quan và giải thích có tính hợp lý và thuyết phục của sinh viên.

c) Phải có giải thích rõ ràng cho mọi cell code trong file jupyter notebook. Tức là, mỗi cell code nên có một cell markdown kèm theo để giải thích.

3 QUY ĐỊNH NỘP BÀI VÀ HÌNH THỨC VẤN ĐÁP

Mỗi nhóm sẽ thiết lập một GitHub repository trong một cài đặt riêng tư (chế độ private). Thùng chứa sẽ được công khai một ngày trước hội thảo để người hướng dẫn và các cá nhân được chọn có thể xem lại tất cả các công việc. Điều quan trọng cần lưu ý là điểm số sẽ bị ảnh hưởng nếu nhóm bỏ ra ít nỗ lực (ví dụ: ít hơn 10 commit trên GitHub repository) hoặc sử dụng các thủ thuật vào những ngày cuối cùng (ví dụ: dùng một commit duy nhất để hoàn thành đồ án, số lượng commit quá nhiều vào những ngày cuối). Các file cần nộp cho final version trước khi seminar:

- Tất cả file jupyter notebook, các mã nguồn Python nếu có
- Slide trình bày báo cáo (dạng .pdf)
- File .pdf phân công công việc của nhóm

- Dữ liệu đã thu thập (có thể sử dụng Google Drive, One Drive, … và để link trong file .txt)

Vào ngày vấn đáp, mỗi nhóm sẽ có ít hơn/ khoảng trong 15 phút để trình bày (Trợ giảng sẽ quyết định thứ tự của người trình bày) và 10 phút cho phần Hỏi & Đáp. Hơn nữa, bài thuyết trình nên tập trung vào công việc một cách rõ ràng, những kỹ thuật và phương pháp thực hiện, thay vì chỉ tập trung vào mã nguồn. **Khi phát hiện hành vi không trung thực, toàn bộ đồ án sẽ bị 0 điểm**. Mọi tài liệu trực tuyến có thể được sử dụng làm tài liệu tham khảo cho các chủ đề của bạn, nhưng cần phải trích dẫn đầy đủ. Tất cả sinh viên được tự do thảo luận về chủ đề của mình với bất kỳ nhóm/ bạn học nào trong lớp, nhưng công việc của nhóm phải được triển khai và diễn giải theo cách hiểu của riêng nhóm sinh viên.

4 TIÊU CHÍ CHẤM ĐIỂM

Tùy thuộc vào tính sáng tạo và độ khó của riêng từng bài toán, các nhóm sẽ được điều chỉnh đánh giá cho phù hợp. Căn cứ vào sản phẩm mà sinh viên đạt được, bảng dưới đây trình bày nội dung chung cho đánh giá điểm đồ án.

Nội dung chấm	**Tỷ lệ**
Quy trình làm việc với dữ liệu (thu thập, tiền xử lý, phân tích, mô hình hóa dữ liệu) phù hợp.	30%
Xây dựng các câu hỏi và cung cấp được các thông tin hữu ích quan trọng từ dữ liệu.	30%
Cài đặt và giải thích quá trình cài đặt một cách tường minh.	10%
So sánh với các phương pháp khác và phân tích ưu, nhược điểm.	10%
Trình bày báo cáo (slide, bố cụ trình bày)	10%
Vấn đáp	10%
Điểm cộng (Chủ đề thú vị, giải pháp hay, những case study thú vị, …)	10%
Tổng cộng	**110%**

5 LIÊN HỆ

Nếu nhóm sinh viên/ sinh viên có bất kỳ thắc mắc, khó khăn cần hỗ trợ, liên hệ ngay cho Trợ giảng/ Hướng dẫn thực hành qua email hoặc Hệ thống Moodle hoặc kênh Zalo.

- Trần Đại Chí: ctran743@gmail.com
- Nguyễn Bảo Long: baolongnguyen.mac@gmail.com
- Lê Nhựt Nam: lenam.fithcmus@gmail.com
- Nguyễn Thái Vũ: vunguyenthai73@gmail.com