Bài tập lớn Học máy (CO3117)
Học kỳ: 2, Năm học: 2024-2025

Võ Thanh Hùng

1/2025

1 Giới thiệu

Trong khuôn khổ môn học Machine Learning, bài tập lớn là một phần quan trọng giúp sinh viên áp dụng các lý thuyết và kỹ thuật học được vào thực tế. Mục tiêu của bài tập là phát triển khả năng xây dựng và triển khai các mô hình học máy, từ việc xử lý dữ liệu, lựa chọn thuật toán phù hợp, đến việc đánh giá và cải tiến hiệu suất của mô hình (nếu có). Qua đó, sinh viên sẽ hiểu rõ hơn về các quy trình trong Machine Learning, bao gồm việc xử lý dữ liệu thô, lựa chọn đặc trưng, và đánh giá mô hình thông qua các chỉ số như độ chính xác, độ lỗi hay độ phức tạp. Bài tập lớn này không chỉ giúp sinh viên nắm vững các kiến thức nền tảng mà còn khuyến khích khả năng sáng tạo và giải quyết vấn đề thực tế trong các bài toán của ngành học.

2 Yêu cầu về nội dung

Nhóm sinh viên được yêu cầu thực hiện các bước sau:

1. Lựa chọn tập dữ liệu để bắt đầu thực hiện (có thể nằm trong hoặc ngoài danh sách được gợi ý)
2. Thực hiện tìm hiểu về bài toán, về dữ liệu tương ứng.
3. Sử dụng các thuật toán ML có sẵn hoặc tự hiện thực, các phương pháp đã/sẽ được giới thiệu trong môn học. Nhóm sinh viên cần thực hiện ít nhất 2 phương pháp khác nhau.
4. Thực nghiệm và so sánh, đánh giá các kết quả.
5. Viết báo cáo tổng kết.
6. Soạn slide trình bày (một số nhóm sẽ được lựa chọn để trình bày).

3 Một số nguồn dữ liệu tham khảo

Phần này giới thiệu một số nguồn dữ liệu để các nhóm tham khảo. Sinh viên có thể chọn dữ liệu từ các nguồn này, hoặc các dữ liệu khác, miễn sao phù hợp với môn học là được. Trong trường hợp nhóm chọn dữ liệu khác và không chắc chắn phù hợp, vui lòng liên hệ giảng viên để được hỗ trợ.

1. The UC Irvine Machine Learning Repository https://archive.ics.uci.edu/. Trong này chứa rất nhiều dataset hữu ích.

2. https://www.kaggle.com/datasets

3. Tham khảo thêm các nguồn hữu ích tại đây: https://en.wikipedia.org/wiki/List_of_datasets_for_learning_research

4 Các yêu cầu khác

4.1 Các thông tin chung

- Bài tập này được làm theo nhóm, theo danh sách đăng ký nhóm của sinh viên, mọi điều chỉnh nhóm cần phải được báo cáo lại để tránh sai sót lúc nhập điểm.
- Việc đánh giá dựa trên cơ sở đánh giá đóng góp của từng thành viên trong nhóm. Báo cáo tổng kết nên có phụ lục phân công khối lượng công việc của từng thành viên.
- Không giới hạn môi trường hay ngôn ngữ lập trình cụ thể.

4.2 Nộp bài

- Sinh viên nén toàn bộ các file/thư mục (bao gồm source code, data, ...) vào một file nén theo dạng MSSV1-MSSV2-....zip, trong đó MSSV chính là mã số sinh viên của sinh viên, đầy đủ MSSV của toàn bộ các thành viên của nhóm. Sinh viên không nén file theo các định dạng khác.
- Trong thư mục của mỗi sinh viên sẽ có một file README.md bao gồm các thông tin về bài tập lớn cũng như các ghi chú khác về thực thi ứng dụng nếu cần.

4.3 Giới hạn và xử lý gian lận

- Đây là bài tập nhóm, nhóm sinh viên phải TỰ MÌNH làm bài.
- Được phép sử dụng các thư viện **PUBLIC** của bên thứ ba không hạn chế. Tuy nhiên, sinh viên phải tự deploy, không được phép sử dụng API có sẵn.
- Có thể trao đổi ý tưởng với các sinh viên khác trong và ngoài lớp, tuy nhiên, không được share/chép lại code, kết quả.
- Mọi hình thức GIAN LẬN nếu bị phát hiện sẽ bị xử lý NGHIÊM KHẮC theo quy định học vụ.