Bài tập lớn Học máy (CO3117)
Học kỳ: 1, Năm học: 2025-2026

Võ Thanh Hùng

9/2025

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 đã học vào các bài toán thực tế. Mục tiêu của bài tập là phát triển khả năng xây dựng và triển khai các mô hình học máy, từ khâu xử lý dữ liệu, lựa chọn thuật toán, huấn luyện mô hình, đến đánh giá và cải tiến hiệu suất. Thông qua quá trình này, sinh viên sẽ nắm vững các bước trong một quy trình học máy đầy đủ: tiền xử lý dữ liệu thô, lựa chọn đặc trưng, chọn mô hình, tối ưu tham số, và đánh giá bằng nhiều chỉ số (accuracy, precision, recall, F1-score, RMSE, v.v.). Bài tập cũng khuyến khích sinh viên tư duy sáng tạo, thử nghiệm nhiều phương pháp, và giải quyết vấn đề một cách thực tiễn.

2 Yêu cầu về nội dung

Mỗi nhóm sinh viên cần thực hiện các bước sau:

1. Lựa chọn một tập dữ liệu phù hợp (có thể từ danh sách gợi ý hoặc nguồn khác, nhưng phải liên hệ giảng viên để duyệt nếu không chắc chắn).

2. Phân tích bài toán và tìm hiểu đặc điểm dữ liệu: dạng dữ liệu, số lượng mẫu, phân bố, các thách thức có thể gặp.

3. Áp dụng ít nhất 2 thuật toán học máy khác nhau (ví dụ: Decision Tree, SVM, Random Forest, Logistic Regression, Neural Networks, Gradient Boosting, v.v.) để giải quyết bài toán.

4. Tiến hành thực nghiệm, so sánh và phân tích kết quả của các phương pháp đã chọn.

5. Nếu có thể, thử nghiệm thêm các kỹ thuật nâng cao như: lựa chọn đặc trưng (feature selection), tối ưu tham số (hyperparameter tuning), ensemble learning, hoặc sử dụng baseline DL để đối chiếu.

6. Viết báo cáo tổng kết, trình bày rõ ràng mục tiêu, phương pháp, kết quả, thảo luận và kết luận.

7. Chuẩn bị slide thuyết trình (một số nhóm sẽ được lựa chọn để trình bày trên lớp).

3 Một số nguồn dữ liệu tham khảo

Sinh viên có thể tham khảo các nguồn dữ liệu sau hoặc chọn nguồn khác phù hợp:

1. UC Irvine Machine Learning Repository: https://archive.ics.uci.edu/
2. Kaggle Datasets: https://www.kaggle.com/datasets
3. Danh sách tổng hợp: https://en.wikipedia.org/wiki/List_of_datasets_for_machine-learning_research

3.1 Nộp bài

- Nén toàn bộ thư mục bài làm (source code, data, models, báo cáo, slide, Dockerfile nếu có) thành một file duy nhất theo dạng MSSV1-MSSV2-....zip (danh sách MSSV đầy đủ của nhóm).
- Trong thư mục nộp cần có file README.md mô tả cách chạy code, môi trường, và các ghi chú cần thiết.
- Khuyến khích đính kèm Dockerfile hoặc môi trường ảo (environment.yml, requirements.txt) để dễ dàng tái hiện kết quả.

3.2 Giới hạn và xử lý gian lận

- Đây là bài tập nhóm, các thành viên phải tự mình tham gia thực hiện.
- Được phép sử dụng thư viện open-source, nhưng không được sử dụng API thương mại có sẵn để thay thế toàn bộ pipeline.
- Không được chia sẻ hoặc sao chép code giữa các nhóm. Mọi hình thức gian lận sẽ bị xử lý theo quy định học vụ.