Bài tập lớn Học phần Học máy INT3405

Hạn nộp bài: 23 giờ 59 phút, thứ Sáu ngày … tháng 5 năm 2025

Sinh viên ĐỌC KỸ hướng dẫn nộp bài sau và thực hiện đúng yêu cầu của bài tập lớn:

*Trong bài tập lớn cuối kỳ này, sinh viên sẽ **thực hiện toàn bộ** quá trình phân tích, làm sạch, trích xuất đặc trưng, và xây dựng mô hình dự đoán trên một tập dữ liệu thực tế. Mục tiêu là lựa chọn, triển khai, và chứng minh được mô hình mà các bạn cho là "tốt nhất" nhằm dự đoán một biến mục tiêu cụ thể.*

*Sinh viên **chọn một trong các đề tài và tập dữ liệu** phù hợp bên dưới để thực hiện bài tập lớn cá nhân của mình. Bài toán mục tiêu có thể là bài toán Phân loại (Classification) hoặc Nhận diện (Detection). Sinh viên cần xác định rõ mục tiêu bài toán bài tập lớn và mô tả chi tiết trong Báo cáo. Báo cáo cần nêu rõ lý do chọn đề tài, các đặc trưng được xem là quan trọng, các phương pháp tiền xử lý và các mô hình học máy đã áp dụng.*

Báo cáo của các bạn cần nộp một file .zip lên Github có add tài khoản của thầy, bao gồm 3 phần sau:

1. ***Báo cáo phân tích chi tiết (.pdf hoặc .docx)** với các phần sau:*
   - *(i) Giới thiệu và phân tích bộ dữ liệu ban đầu (nêu rõ đề tài, mục tiêu bài toán, nguồn dữ liệu, kích thước bộ dữ liệu thực hiện, kiểu dữ liệu, và các thống kê cơ bản)*
   - *(ii) Các bước xử lý dữ liệu số, hình ảnh, hoặc âm thanh (bao gồm gán nhãn, mã hoá, xử lý giá trị thiếu, làm sạch dữ liệu, ...).*
   - *(iii) Trích xuất và xử lý các đặc trưng (lý do lựa chọn và kỹ thuật áp dụng)*
   - *(iv) Xây dựng mô hình (thử nghiệm các mô hình, lý do chọn mô hình "tốt nhất", đánh giá hiệu năng theo các độ đo phù hợp, tối ưu tham số,...)*
   - *(v) Kết luận, tổng kết nêu bật mô hình "tốt nhất", phân tích các kết quả thu được và đề xuất hướng cải tiến (nếu có)*
2. ***Mã nguồn đầy đủ (Jupyter Notebook hoặc file Python)** kèm theo ghi chú, có thể chạy lại được quá trình xử lý và huấn luyện.*
3. ***File nén bộ dữ liệu đã làm sạch hoặc tiền xử lý** (nếu không phải bộ data gốc).*

***Lưu ý, trực quan hoá dữ liệu và kết quả là bắt buộc**, tuy nhiên sinh viên nên sử dụng biểu đồ minh họa cho các điểm chính để giúp người đọc hiểu rõ hơn về bài làm của các bạn, tránh nhồi nhét bảng biểu không cần thiết. Cuối cùng, nghiêm cấm các bạn sao chép bài của nhau, hoặc từ các mẫu có sẵn mà không giải thích, trích dẫn. Mọi gian lận học thuật sẽ bị xử lý nghiêm túc.*

Nội dung thực hiện:

Sinh viên thực hiện 2 nội dung sau:

- **Nội dung 1: Sử dụng tập dữ liệu đã gán nhãn sẵn (labeled dataset) chọn link dữ liệu 2/3/4**
  Sinh viên chọn một trong các tập dữ liệu dưới đây, trong đó đã có sẵn nhãn (label) cho bài toán. Sinh viên cần thực hiện thống kê, tiền xử lý và phân tích bộ dữ liệu đầu vào trước khi xây dựng mô hình.
- **Nội dung 2: Cùng xây dựng bộ dữ liệu và gán nhãn thủ công (manual labelling) link dữ liệu 1**
  Các sinh viên sẽ tạo một tài khoản CVAT và add thêm mail của mình. Mình sẽ chia và gửi kèm hướng dẫn một bộ data để các bạn cùng gán nhãn và xây dựng bộ dữ liệu. File dữ liệu gán nhãn cuối cùng sẽ cần được tập hợp và share với nhau để xây dựng các mô hình phân biệt, dự đoán.

Dưới đây là các đề tài và đường link dữ liệu để các bạn lựa chọn.

1) Phân vùng cải bó xôi (manual labelling)
2) Phân loại bệnh trên lá lúa (Datasets)
3) Phân loại bọ trên bẫy vàng (Datasets)
4) 3D Segment cây đậu tương (Datasets)

Kết quả chia sẻ cho tài khoản github của thầy.

Yêu cầu thực hiện chi tiết

STT	Nhiệm vụ	Mô tả chi tiết
1	Xây dựng, phân tích và mô tả bộ dữ liệu	- Gán nhãn bộ dữ liệu. - Khám phá dữ liệu: kích thước, số lượng thuộc tính, kiểu dữ liệu. - Trình bày phân bố lớp hoặc giá trị đầu ra. - Vẽ các biểu đồ.
2	Tiền xử lý dữ liệu	- Làm sạch dữ liệu - Chuẩn hoá, hiệu chỉnh nếu cần (thêm bớt, làm giàu, …) - Phát hiện và xử lý giá trị ngoại lai nếu có. - Chia dữ liệu thành các tập train/test hoặc train/val/test.
3	Trích xuất đặc trưng	- Lý giải việc chọn đặc trưng - Tạo thêm đặc trưng - Giảm chiều dữ liệu (PCA, lọc thuộc tính kém quan trọng …) - Đánh giá các đặc trưng
4	Xây dựng và đánh giá mô hình	- Chọn và huấn luyện ít nhất 3 – 5 mô hình khác nhau.

		- Tối ưu, cải tiến mô hình (nếu có thể). - Đánh giá hiệu năng bằng các chỉ số phù hợp - So sánh mô hình, chọn mô hình "tốt nhất"
5	Diễn giải mô hình và trình bày kết quả	- Giải thích lý do chọn mô hình "tốt nhất" - Trình bày biểu đồ, bảng biểu minh hoạ, - Phân tích ý nghĩa thực tế và kết quả mô hình - Nêu điểm mạnh, hạn chế, và đề xuất hướng cải tiến tiếp theo (nếu có).

Thông tin tài khoản CVAT/github: trienpm@vnu.edu.vn

CHÚC CÁC BẠN CÓ MỘT MÙA THI HIỆU QUẢ!