GỢI Ý ĐỀ TÀI BÀI TẬP LỚN MÔN MACHINE LEARNING

Các ý tưởng dưới đây chỉ là gợi ý, các nhóm có thể chọn những đề tài khác, tuy nhiên cần đảm bảo tương đương về mặt khối lượng công việc; về mức độ bao quát các kiến thức đã học; về kích thước (số lượng mẫu và độ phức tạp) của dữ liệu.

Mỗi nhóm đề tài có tối đa 03 thành viên, không có đề tài độc lập (để xét điểm về khả năng làm việc nhóm). Tất cả thành viên cần phải thực hiện ít nhất 01 mô hình/phương pháp đã học và nắm vững về nó, phân tích – giải thích kết quả tương ứng của phần đó. Các thành viên cần phối hợp với nhau để tiền xử lý dữ liệu; đưa ra đánh giá, nhận xét – so sánh trên kết quả giữa các phương pháp cùng mục đích và/hoặc so sánh kết quả của phương pháp trước/sau khi xử lý dữ liệu bằng các phương pháp, ví dụ như giảm chiều.

Kết quả cần có sẽ gồm:

- 01 báo cáo (report): Trình bày về phương pháp/dữ liệu/kết quả thực nghiệm một cách chi tiết. Hình thức: Khoảng 15-20 trang khổ A4, font chữ Times New Roman 13,căn lề giàn đều hai bên, khoảng cách dòng 1.3 đến 1.5, có đủ các đề mục, chia làm các phần giới thiệu bài toán và phương pháp, kiến thức lý thuyết của các phương pháp, dữ liệu và kết quả thực nghiệm, kết luận.
- Bản trình bày
- Chương trình nguồn của các phương pháp thực nghiệm
- File text hướng dẫn về liên kết lấy dữ liệu, cách tổ chức dữ liệu và các kịch bản thực nghiệm.

Chú ý không gửi kèm dữ liệu, trong trường hợp dữ liệu tự xây dựng, các bạn cần chia sẻ trên google drive và gửi link. Các nhóm khi đăng ký đề tài cần đưa luôn thông tin về dữ liệu. Nếu nhiều hơn 01 nhóm chọn cùng bài toán, cùng dữ liệu nào đó, thì nhóm nào đăng ký trước sẽ được ưu tiên, nhóm đăng ký sau sẽ phải thay đổi bài toán hoặc dữ liệu.

1. NHÓM BÀI TOÁN HỒI QUY

Hãy tìm một tập dữ liệu với đầu vào có ít nhất 15 thuộc tính bao gồm cả dạng số và dạng lựa chọn (category), 300 mẫu – đầu ra có dạng giá trị số (đại lượng liên tục - numeric. Gợi ý: Có thể tìm dữ liệu theo từ khóa “High dimensional large scale free dataset for regression analysis”).

Thực hiện các yêu cầu sau:

1.1. **Tiền xử lý dữ liệu**:

- Đọc dữ liệu vào, mô tả cấu trúc của dữ liệu (các trường, số bản ghi…), thống kê sơ bộ về dữ liệu
- Chuẩn hóa dữ liệu thông qua xử lý dữ liệu lỗi (thiếu trường, sai định dạng dữ liệu, giá trị không hợp lệ …)
- Chuyển đổi dữ liệu về dạng phù hợp (ví dụ chuyển từ category sang onehot coding hoặc dạng số khác).
- Chuẩn hóa giá trị nếu cần thiết.
- Mô tả dữ liệu sau chuẩn hóa.

1.2. **Phân tích và trực quan hóa dữ liệu**: Các ý dưới đây cần thực hiện với ít nhất 02 phương pháp giảm chiều dữ liệu khác nhau.

- Phân tích các tham số thống kê của dữ liệu (theo từng trường hoặc tổng thể)
- Chuẩn hóa dữ liệu và đánh giá các thành phần chính (của dữ liệu gốc hoặc sau khi phân thích thành phần chính) theo các tham số thống kê.
- Thực hiện hiển thị trực quan đối với dữ liệu theo từng cặp 02 thành phần chính, áp dụng cho khoảng 4 – 6 thành phần chính.
- Xác định lượng thông tin được bảo tồn theo phương sai giải thích (explained variances) trong mỗi trường hợp ở trên.
- Thực hiện việc trực quan hóa mối quan hệ của một số chiều dữ liệu chính với đầu ra để xem xét khả năng có tương quan tuyến tính.
- So sánh đánh giá giữa các phương pháp phân tích, giảm chiều nói trên.

1.3. Phân cụm dữ liệu: Mỗi nhóm đề tài chọn thực hiện ít nhất hai trong các phương pháp dưới đây

(i) K-Means

(ii) GMM (thuật toán cực đại hóa kỳ vọng)

(iii) DBScan

Thực hiện phân cụm dữ liệu sau khi bỏ qua trường đầu ra, với số cụm được chọn phù hợp.

- Nhận xét về mối quan hệ giữa các mẫu dữ liệu đầu vào trong các cụm; đánh giá quan hệ giữa các đầu ra tương ứng trong các cụm. Các đánh giá trên cần dựa vào các độ đo định lượng.
- Thực hiện trực quan hóa dữ liệu và đánh dấu phân biệt các mẫu dữ liệu thuộc mỗi cụm thông qua màu sắc.

1.4. **Phân tích hồi quy**

(a) Hãy thực hiện ít nhất 02 phương pháp hồi quy trong số các phương pháp đã học (ví dụ K-NN, Hồi quy tuyến tính, Multi Layers Perceptron…) trên tập dữ liệu đã tiền xử lý và với tỉ lệ train:validation được chia khác nhau (ít nhất gồm train:validation = 4:1; 7:3; 6:4).

- Thực hiện với dữ liệu gốc
- Thực hiện với dữ liệu đã giảm chiều
- So sánh kết quả của các trường hợp và đưa ra nhận xét, đánh giá về các kết quả đó. Nhận xét xem phương pháp có xẩy ra overfit hay không.
- Áp dụng các biện pháp hiệu chỉnh (regularization) phù hợp để giảm mức độ overfit.

(b) Trực quan hóa và đánh giá tương quan giữa phần dư (sai lệch dự đoán – thực tế) và bản thân đầu vào. Từ đó đánh giá việc sử dụng mô hình tương ứng có phù hợp hay không.

(c) Kết hợp giữa kết quả phân cụm ở đầu ra và số lượng các giá trị trong các khoảng, hãy chia đầu ra thành 3 đến 4 khoảng sao cho số lượng mẫu là xấp xỉ nhau. Đưa ra các ngưỡng ứng với các khoảng chia đó và chuyển bài toán về dạng phân loại (classification) với các nhãn tương ứng. Thực hiện yêu cầu sau với dữ liệu gốc và dữ liệu giảm chiều sao cho số chiều còn lại bằng 1/3 số chiều ban đầu. Dữ liệu chia train : test như ý (a).

- Thực hiện phân loại bằng phương pháp naïve bayes phù hợp và ít nhất 01 phương pháp phân loại khác.
- Đưa ra kết quả trong mỗi trường hợp.
- Đánh giá, so sánh các kết quả thực nghiệm trong mỗi trường hợp. Giải thích xem tại sao lại như vậy.

2. NHÓM BÀI TOÁN PHÂN LOẠI

Hãy tìm một tập dữ liệu với đầu vào có ít nhất 15 thuộc tính bao gồm cả dạng số và dạng lựa chọn (category), 300 mẫu – đầu ra có dạng nhãn phân loại (Categories. Gợi ý: Có thể tìm dữ liệu từ internet như nhóm bài toán hồi quy hoặc tìm dữ liệu dạng hình ảnh, chữ cái viết tay tiếng Anh (chú ý không dùng lại dữ liệu chữ số viết tay) – với hai dạng dữ liệu này không nhất thiết bao gồm 2 kiểu thuộc tính số và lựa chọn).

Thực hiện các yêu cầu sau:

2.1. **Tiền xử lý dữ liệu**:

- Đọc dữ liệu vào, mô tả cấu trúc của dữ liệu (các trường, số bản ghi…), thống kê sơ bộ về dữ liệu
- Chuẩn hóa dữ liệu thông qua xử lý dữ liệu lỗi (thiếu trường, sai định dạng dữ liệu, giá trị không hợp lệ …)
- Chuyển đổi dữ liệu về dạng phù hợp (ví dụ chuyển từ category sang onehot coding hoặc dạng số khác).
- Chuẩn hóa giá trị nếu cần thiết.
- Mô tả dữ liệu sau chuẩn hóa.

2.2. **Phân tích và trực quan hóa dữ liệu**: Các ý dưới đây cần thực hiện với ít nhất 02 phương pháp giảm chiều dữ liệu khác nhau trên dữ liệu đầu vào.

- Phân tích các tham số thống kê của dữ liệu (theo từng trường hoặc tổng thể)
- Chuẩn hóa dữ liệu và đánh giá các thành phần chính (của dữ liệu gốc hoặc sau khi phân thích thành phần chính) theo các tham số thống kê.
- Thực hiện hiển thị trực quan đối với dữ liệu theo từng cặp 02 thành phần chính, áp dụng cho khoảng 4 – 6 thành phần chính.
- Xác định lượng thông tin được bảo tồn theo phương sai giải thích (explained variances) trong mỗi trường hợp ở trên.
- Thực hiện việc trực quan hóa mối quan hệ của một số chiều dữ liệu chính với đầu ra để xem xét khả năng có tương quan hay hình thành cụm dữ liệu.
- So sánh đánh giá giữa các phương pháp phân tích, giảm chiều nói trên.

2.3. Phân cụm dữ liệu: Mỗi nhóm đề tài chọn thực hiện ít nhất một trong các phương pháp dưới đây

(i) K-Means

(ii) GMM (thuật toán cực đại hóa kỳ vọng)

(iii) DBScan

Thực hiện phân cụm dữ liệu sau khi bỏ qua trường đầu ra, với số cụm được chọn phù hợp.

- Nhận xét về mối quan hệ giữa các mẫu dữ liệu đầu vào trong các cụm; đánh giá quan hệ giữa các đầu ra tương ứng trong các cụm. Các đánh giá trên cần dựa vào các độ đo định lượng.
- Thực hiện trực quan hóa dữ liệu và đánh dấu phân biệt các mẫu dữ liệu thuộc mỗi cụm – cũng như thuộc về mỗi nhãn - thông qua màu sắc và hình dạng biểu thị mẫu.

2.4. Phân loại

(a) Hãy thực hiện ít nhất 03 phương pháp phân loại trong số các phương pháp đã học, bao gồm các cách tiếp cận khác nhau (sinh vs. phân biệt) – gồm cả tuyến tính và phi tuyến lấy từ 03 nhóm sau: Nhóm 1: K-NN, cây quyết định, Naïve Bayes; Nhóm 2: Hồi quy SoftMax - Logistic, Multi Layers Perceptron; Nhóm 3: SVM) trên tập dữ liệu đã tiền xử lý và với tỉ lệ train:validation được chia khác nhau (ít nhất gồm train:validation = 4:1; 7:3; 6:4).

- Thực hiện với dữ liệu gốc
- Thực hiện với dữ liệu đã giảm chiều
- So sánh kết quả của các trường hợp và đưa ra nhận xét, đánh giá về các kết quả đó. Nhận xét xem phương pháp có xẩy ra overfit hay không.
- Áp dụng các biện pháp hiệu chỉnh (regularization) phù hợp để giảm mức độ overfit nếu mô hình phù hợp.

(b) Trực quan hóa và đánh giá tương quan giữa dự đoán – thực tế.

- Từ đó đánh giá việc sử dụng mô hình tương ứng có phù hợp hay không.
- Giải thích những nhận định trong ý ngay trên.

(c) Xét phương pháp trong Nhóm 2 và Nhóm 3, lựa chọn 1 phân lớp và dựa vào giá trị của hàm quyết định cho phân lớp đó (ví dụ hàm softmax – logistic hoặc hàm đánh giá score trong SVM) để chuyển bài toán về dạng hồi quy. Dữ liệu chia train : test như ý (a).

- Thực hiện ít nhất 02 mô hình hồi quy trên tập dữ liệu với đầu ra mới xây dựng.
- Đưa ra kết quả trong trường hợp dữ liệu đầu vào nguyên bản và dữ liệu giảm về còn 1/3 số chiều.
- Đánh giá, so sánh các kết quả thực nghiệm trong mỗi trường hợp. Giải thích xem tại sao lại như vậy.

3. NHÓM CÁC BÀI TOÁN KHÁC

Các nhóm có thể tự đề xuất các đề tài khác nếu đáp ứng đủ các yêu cầu như đã đề cập ở phần đầu.

YÊU CẦU ĐẾN BÀI THI GIỮA KỲ

(a) Cần có đủ các tài liệu theo yêu cầu như đã đề cập ở phần đầu

(b) Tất cả đề tài đều đã lựa chọn được dữ liệu và thực hiện xong phần tiền xử lý dữ liệu

(c) Các phần giảm chiều và trực quan hóa dữ liệu; phân tích hồi quy hoặc phân loại: thực hiện xong ít nhất 02 mô hình cho bài toán tương ứng.