Đồ án cuối kì - Bảo mật tính riêng tư cho mô hình hồi quy tuyến tính

1 Tập dữ liệu về xe

1.1 Giới thiệu

- Cho tập dữ liệu về giá bán xe hơi cũ. Ta chia tập dữ liệu thành hai phần X và Y như sau:
  - X: Các thông tin về xe được bán.
    + Make: Thương hiệu của xe.
    + Model: Mô hình xe.
    + Year: Năm sản xuất.
    + Kilometer: Số cây số đã chạy.
    + Fuel Type: Loại nhiên liệu cho xe.
    + Tranmission: Loại xe (auto = xe số, manual = xe sàn).
    + Location: địa điểm bán
    + Color: Màu xe.
    + Owner: Người chủ thứ bao nhiêu của chiếc xe.
    + Seller Type: Người bán/ đơn vị bán.
    + Engine: Mã lực động cơ
    + Max Power: Công suất tối đa
    + Max Torque: Mô-men xoắn tối đa
    + DriveTrain: Loại Hệ thống truyền động.
    + Length, Width, Height: Kích thước xe.
    + Seating Capacity: Số lượng chỗ ngồi.
    + Fuel Tank Capacity: Dung lượng thùng xăng.
  - Y: Giá xe được bán.

1.2 Tải tập dữ liệu

- Tải từ moodle.
- Thư mục dữ liệu bao gồm file:
  - train.csv: Gồm những dữ liệu dùng để huấn luyện mô hình.
  - val.csv: Gồm những dữ liệu để đánh giá mô hình sau khi train.

1.3 Yêu cầu

Trong phần này, nhóm cần hoàn thành các yêu cầu sau:

- Tải xuống và đọc được toàn bộ tập dữ liệu.
- In ra một số thông tin của cả file train.csv: Số dòng, tên các cột.
- Đọc dữ liệu từ file và in ra **5** dòng đầu tiên của tập dữ liệu huấn luyện.

2 Tiền xử lý dữ liệu

2.1 Yêu cầu

Trong phần này, nhóm cần hoàn thành các yêu cầu sau:

- **Chọn** các cột dữ liệu trong X để phục vụ cho việc huấn luyện mô hình. Lưu ý: Giá trị của các cột có thể không phải là số thì các bạn nên làm như sau:
  - Dữ liệu gồm số + chữ: Có thể loại bỏ chữ và giữ lại số nếu tất cả các giá trị cùng đơn vị đo.
  - Dữ liệu chỉ gồm chữ: Các bạn có thể đổi thành số bằng cách đánh số thứ tự cho mỗi loại giá trị tương ứng. (Ví dụ: A thành 1, B thành 2,...)

3 Đề bài chính

3.1 Mô hình tuyến tính

Ở lab này, nhóm được áp dụng các kiến thức về hồi quy tuyến tính (Linear Regression) để huấn luyện mô hình dựa vào tập dữ liệu train.csv và dự đoán giá xe ở phần Y.

Nhóm có thể tự quyết định mô hình như thế nào là phù hợp và có đủ đầu vào cho các cột dữ liệu đã chọn ở phần tiền xử lý dữ liệu.

File train.csv được cung cấp cho nhóm chỉ được dùng cho mục đích huấn luyện mô hình. Mô hình của nhóm sẽ được đánh giá trên cả tập huấn luyện và tập kiểm thử.

3.2 Yêu cầu

Trong phần này, nhóm cần hoàn thành các yêu cầu sau:

- Định nghĩa và viết ra những cột dữ liệu nào được sử dụng và các công thức hồi quy mình sẽ sử dụng. (Nên viết ra trong markdown block của jupiter notebook).
- Công thức hồi quy tối thiểu phải có 4 loại phương trình tuyến tính khác loại với nhau. Giả sử như nhóm chọn 4 cột lần lượt là x_1, x_2, x_3, x_4, một số ví dụ về những loại khác nhau:
  - $y = a_1x_1 + a_2x_2 + a_3x_3 + a_4x_4$
  - $y = a_1x_1^2 + a_2x_2 + a_3x_3^2 + a_4x_4$
  - $y = a_1(x_1 + x_2) + a_3x_3^2 + a_4x_4$
  - $y = a_1x_1x_2 + a_3x_3^2$

  Trong báo cáo cần nêu rõ lý do chọn các loại công thức này về mặt ý tưởng (Nếu có quan sát hoặc chứng minh thì càng tốt).
- Huấn luyện mô hình theo công thức được định nghĩa (Lưu ý: không sử dụng thư viện sklearn)
- In ra độ chính xác (accuracy) của mô hình trên tập huấn luyện: Ở đây, nhóm có thể sử dụng Mean Square Error (MSE) hoặc mean average error (MAE).
- Viết một code block cho phép đọc một file csv từ một đường dẫn và in ra độ chính xác của mô hình trên file được nhập vào (Lưu ý là dùng cùng độ đo ở bước trên).

3.3 Ứng dụng Differential Privacy

Dựa vào các công thức tuyến tính đã chọn phía trên, nhóm hãy chọn ra một công thức tốt nhất và ứng dụng Differential Privacy để bảo vệ thông tin của các cột dữ liệu đã chọn tương ứng.

Trong phần này, nhóm cần hoàn thành các yêu cầu sau:

- Chọn ra một trong 2 hướng tiếp cận sau đây:
  1. Áp dụng DP ở **user-level**: Ở trường hợp này, người dùng (user) không muốn đưa thông tin thật của mình cho máy chủ nên ta sẽ thả nhiễu vào data trước khi cho mô hình mấy học học dữ liệu huấn luyện. Có thể tham khảo tạp chí How to DP-fy ML: A Practical Guide to Machine Learning with Differential Privacy.
  2. Áp dụng DP ở **server-level**: Ở trường hợp này, dữ liệu được huấn luyện cho mô hình là dữ liệu có giá trị chính xác. Tuy nhiên, để chống lại các tấn công thiên về tính riêng tư của dữ liệu, ta có thể thả nhiễu vào câu trả lời của mô hình trước khi đưa cho người đang truy vấn mô hình. Có thể tham khảo bài báo One Parameter Defense - Defending against Data Inference Attacks via Differential Privacy.
- **Cách sinh nhiễu:** Nhóm có thể sử dụng một trong ba cơ chế tạo nhiễu đã học: Guassian Mechanism, Laplace mechanism và Exponential mechanism. Lưu ý là chứng minh cho Guassian Mechanism chỉ đúng với $\epsilon < 1$.
- Tạo ra **kế toán riêng tư (privacy accountant)** để tính lượng mất mát tính riêng tư (privacy loss) trong quá trình huấn luyện (trường hợp 1) hoặc trong quá trình trả lời câu hỏi (trường hợp 2). Cách tinh tổ hợp cho quỹ riêng tư nhóm có thể sử dụng định lý tổ hợp cơ bản (basic composition theorem) hoặc định lý tổ hợp nâng cao (advanced composition theorem).
- **Các mục tiêu chính được đánh giá:**
  1. Đảm bảo được tính riêng tư cho dữ liệu ở hướng tiếp cận đã chọn: cần giải thích được cách thả nhiễu và mô tả được ý tưởng thực hiện thông qua việc áp dụng các định lý đã học (không cần chứng minh lại các định lý).
  2. Thực hiện phân tích hiệu suất (hiệu năng của mô hình ở trường hợp 1 hoặc số lượng câu hỏi/loại câu hỏi và độ chính xác của câu trả lời với trường hợp 2) với các mức của quỹ riêng tư sau đây:
     (a) $\epsilon < 1$.
     (b) $1 \leq \epsilon \leq 10$.
     (c) $\epsilon > 10$.
  3. (Bonus 1 điểm) Nếu nhóm có thể lập trình được privacy preserving cho mô hình Federated Learning.

4 Các yêu cầu khác

- Ngôn ngữ sử dụng bắt buộc là Python, không được phép sử dụng ngôn ngữ khác. Để tiện cho việc lập trình và chấm bài, nhóm nên sử dụng Jupiter Notebook.
- Giới hạn thư viện: nhóm chỉ được sử dụng các thư viện cho các tác vụ nằm ngoài việc huấn luyện mô hình (ví dụ: pandas, numpy,...) và không được sử dụng các thư việc cho tác vụ này (ví dụ: sklearn,...)
- Các nhóm cần kiểm tra mã nguồn trước khi nộp. Nếu mã nguồn không chạy được mà không phải do nguyên nhân khách quan (thiếu thư viện, lỗi do thư viện gây ra, sử dụng thư viện sai phiên bản,...) thì sẽ bị 0 điểm đồ án.

- Bài nộp phải gồm có 2 phần:
  + Report: Chứa các file báo cáo.
  + Source: Chứa các file mã nguồn.
- Trong các file nộp, nhóm cần ghi rõ thông tin về các thành viên gồm họ tên và MSSV. Riêng đối với mã nguồn, nhóm có thể ghi thông tin trên dưới dạng comment trong code của nhóm.
- Bài nộp sẽ được đặt trong thư mục có tên `MSSV01[_MSSV02[_MSSV03[...]]]` và được nén lại bằng định dạng ZIP với cùng tên như trên. Ví dụ đặt tên nhóm có 1 nhóm là `MSSV01`, nhóm có 2 nhóm là `MSSV01_MSSV02`.
- Nghiêm cấm các hành vi gian lận, không trung thực trong học tập như sao chép bài làm giữa các nhóm với nhau, sao chép bài làm của các nhóm khóa trước hoặc các nhóm lớp khác trường khác, nhờ người làm hộ. Nếu phát hiện các hành vi trên thì cả nhóm sẽ bị 0 điểm và xử lý theo quy định của Khoa và Trường.