TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN, ĐHQG-HCM
KHOA CÔNG NGHỆ THÔNG TIN

MÔN: PHÂN TÍCH DỮ LIỆU BẢO TOÀN TÍNH RIÊNG TƯ
BÁO CÁO ĐỒ ÁN CUỐI KÌ

Sinh viên thực hiện:

18120040 – Nguyễn Đăng Khoa

18120040@student.hcmus.edu.vn

SĐT:0865316054

18120404 – Trần Hữu Khải

18120404@student.hcmus.edu.vn

Table of Contents

1. Viết chương trình tạo CSDL cho huấn luyện ngoại tuyến (Offline learning)

1.1 Ý tưởng

Lấy cảm hứng từ góc đồ người có CSDL muốn đem bán cho bên thứ 3 nhưng vẫn bảo toàn được tính riêng tư của các cá nhân trong CSDL đó. Cụ thể, cho trước một CSDL có dạng $\mathcal{D} = \{(X_1, y_1), \dots, (X_n, y_n)\}$ với $y_i \in \{0,1\}$, sau đó dùng cơ chế tung đồng xu để sinh ra 2 tập khác $\mathcal{D}' = \{(X_1, y'_1), \dots, (X_n, y'_n)\}$ trước khi công bố.

Lưu ý, ở cơ chế Đáp ứng ngẫu nhiên – Random Response (thường sử dụng Tung đồng xu) thuộc tính được áp dụng sẽ có đầu ra dạng nhị phân {0, 1}.

Thuật toán tung đồng xu là một biến thể của cơ chế đáp ứng ngẫu nhiên (randomized response). Thuật toán này cho phép người tham gia khảo sát không cần phải lo lắng khi đưa ra câu trả lời thật.

Thuật toán:

1. Tung đồng xu lần 1
2. Nếu mặt ngửa, cung cấp câu trả lời thật.
3. Nếu mặt sấp, tung đồng xu lần 2
    a. Nếu ngửa, trả lời "Có"
    b. Nếu sấp, trả lời "Không"

1.2 Thực nghiệm:

Link code:

https://colab.research.google.com/drive/1_HrrvYZn6Og13ym8eE2FkxpdcVVHDcYn?usp=sharing

Quá trình thực nghiệm:

1. Tải và tiền xử lý dữ liệu
    a. Trong báo cáo này, nhóm em sử dụng bô dữ liệu "Adult dataset" - https://archive.ics.uci.edu/ml/datasets/adult.
    b. Nhưng để đơn giản, nhóm chỉ thực hiện trên 4 cột thuộc tính "Name", "Zip", Age", "Occupation"
2. Áp dụng thuật toán đáp ứng ngẫu nhiên để sinh ra CSDL mới. Trong phần thực nghiệm này, nhóm em chọn cột "Occupation" để thử che giấu nghề nghiệp của các đối tượng, các đối tượng chỉ mang giá trị có hoặc không làm bán hàng ("Sales" hoặc "Not Sales")

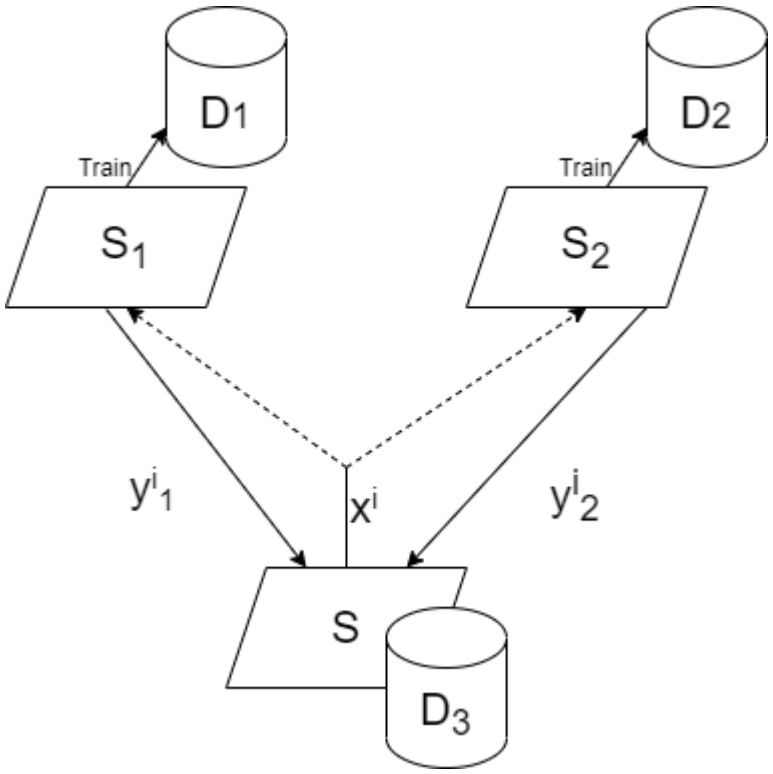
2. Mô phỏng học trực tuyến bằng hồi quy tuyến tính

2.1 Phương pháp

Mô tả bài toán: Cho trước một CSDL có dạng $\mathcal{D} = \{(X_1, y_1), \dots, (X_n, y_n)\}$. Chia tập dữ liệu trên thành 3 tập dữ liệu con $\mathcal{D}_1, \mathcal{D}_2, \mathcal{D}_3$. Xây dựng hai mô hình hồi quy tuyến tính $\mathcal{S}_1, \mathcal{S}_2, \mathcal{S}$, trong đó, $\mathcal{S}_1, \mathcal{S}_2$ huấn luyện trên tập $\mathcal{D}_1, \mathcal{D}_2$ tương ứng.

Với tập dữ liệu $\mathcal{D}_3$ thực hiện như sau:

- Khởi tạo trọng số cho mỗi mô hình được gán $w_i = 0.5$
- Lặp t = 1, 2, …, K:
  - Với mỗi record trong cơ sở dữ liệu, chọn một trong hai mô hình $\mathcal{S}_1, \mathcal{S}_2$ bằng cách chọn ngẫu nhiên theo xác suất $\frac{w_i}{\sum_{i=1}^{2} w_i}$
  - Dùng mô hình được chọn để tính record trên và trả về kết quả cho cơ sở dữ liệu $\mathcal{D}_3$ và tính $l_i^t = |y^t - y_i^t|$
  - Cập nhật mô hình $\mathcal{S}$ và trọng số $w_i = w_i \cdot e^{-\eta l_i^t}$



2.1 Thực nghiệm

Link code: https://colab.research.google.com/drive/1OQefYCFi9Av9KBpuah2l-jc0Q2iRV-_5?usp=sharing

Quá trình thực nghiệm:

3. Tải và tiền xử lý dữ liệu
    a. Trong báo cáo này, nhóm em sử dụng bô dữ liệu "California Housing Prices" để thử nghiệm
    b. Tạm loại bỏ thuộc tính "longtitude", "latitude" và mã hóa thuộc tính "ocean_proximity" thành số để phù hợp với tính toán
    c. Tách tập dữ liệu trên thành 3 tập con $\mathcal{D}_1, \mathcal{D}_2, \mathcal{D}_3$ theo tỷ lệ 4:4:2
4. Huấn luyện và đánh giá mô hình
    a. Ở từng tập con, chia thành tập train và test theo tỷ lệ 8:2
    b. Sử dụng mô hình LinearRegression() của thư viện sklearn và huấn luyện cho cả 3 mô hình
    c. Dùng độ đo MSE để đánh giá mô hình.
5. Áp dụng thuật giải đồng thuận (Online learning)

Kết quả: Tệ hơn khi huấn luyện riêng với CSDL của chính mô hình S. Nguyên nhân có thể do:

- Khi các biến X được gửi cho các "chuyên gia" và nhận về kết quả để cập nhật thì các chuyên gia ấy không thấy được lỗi sai của mình. Hay nói cách khác, ở mô hình S sai nhưng lại chỉnh theo các chuyên gia, dẫn đến sai số không những không giảm mà còn tăng
- Bộ dữ liệu còn quá nhỏ (chỉ hơn 3000 mẫu) và mô hình còn đơn giản nên chưa thể thấy được sức mạnh của online learning.

3. Phân công công việc

MSSV	Họ tên	Công việc
18120040	Nguyễn Đăng Khoa	• Phân tích, nghiên cứu thuật giải đồng thuận cho huấn luyện trực tuyến trên cơ chế lũy thừa. • Cài đặt, thực thi phương pháp ngoại tuyến và trực tuyến • Viết báo cáo

18120404	Trần Hữu Khải	Phân tích, nghiên cứu tạo cơ sở dữ liệu huấn luyện ngoại tuyến

.

4. Tài liệu tham khảo

1. Tài liệu bài giảng.
2. Cơ chế lũy thừa, https://programming-dp.com/notebooks/ch9.html
3. Multiplicative Weights, https://dpcourse.github.io/lecnotes-web/lec-15-16-MW.pdf
4. Online mechanism, https://maxkasy.github.io/home/files/other/ML_Econ_Oxford/differential_privacy.pdf