|  |
| --- |
| **ĐẠI HỌC KHOA HỌC TỰ NHIÊN - ĐẠI HỌC QUỐC GIA TP HCM**  **KHOA CÔNG NGHỆ THÔNG TIN** |
| Background pattern  Description automatically generated with low confidenceLogo  Description automatically generated  **ĐỒ ÁN**  **PHÂN TÍCH DỮ LIỆU**  **BẢO TOÀN TÌNH RIÊNG TƯ** |
| **Giảng viên phụ trách:**  PGS.TS Nguyễn Đình Thúc  **Sinh viên thực hiện:**  18120423 - Trịnh Tấn Khoa  18120341 – Hoàng Văn Duy  18120426 – Lê Anh Khương  *Học Kỳ 1 – Năm Học 2021-2022*  *Thành phố Hồ Chí Minh, tháng 01 năm 2022* |

MỤC LỤC

[1. TẠO CƠ SỞ DỮ LIỆU HUẤN LUYỆN OFFLINE 3](#_Toc93586124)

[a) Giới thiệu. 3](#_Toc93586125)

[b) Mục tiêu. 4](#_Toc93586126)

[c) Phân tích thật toán. 4](#_Toc93586127)

[d) Kiểm tra. 6](#_Toc93586128)

[e) Kết luận 7](#_Toc93586129)

[2. XÂY DỰNG MÔ HÌNH HUẤN LUYỆN TRỰC TUYẾN TRÊN THUẬT GIẢI HỒI QUY TUYẾN TÍNH VỚI 2 CHUYÊN GIA 7](#_Toc93586130)

[a) Giới thiệu. 7](#_Toc93586131)

[b) Mục tiêu. 7](#_Toc93586132)

[c) Phân tích thuật toán. 8](#_Toc93586133)

[d) Kiểm tra. 9](#_Toc93586134)

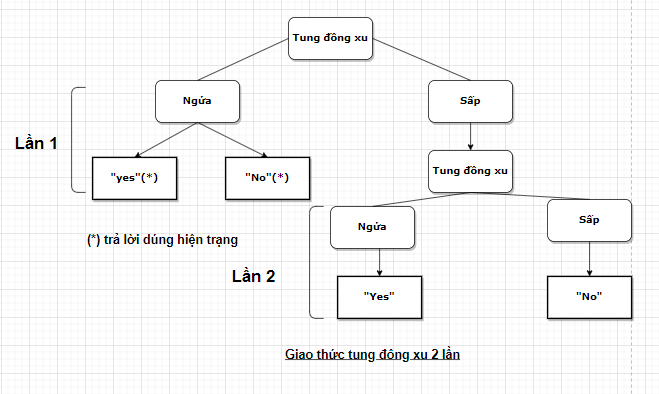
[e) Kết luận. 11](#_Toc93586135)

[3. Tài liệu tham khảo. 12](#_Toc93586136)

# TẠO CƠ SỞ DỮ LIỆU HUẤN LUYỆN OFFLINE

## Giới thiệu.

* Thuật toán tung đồng xu:
* Tung đồng xu lần 1:
* Nếu mặt ngửa thì sẽ trả lời đúng kết qua. Nghĩa là có nói yes, không nói no.
* Nếu mặt sấp, tung đồng xu lần thứ 2:
* Trả lời “yes” nếu lần thứ 2 được ngửa,.
* Trả lời “no” nếu lần thứ 2 được sấp.
* Phân tích kết quả:
* Nếu nhận câu trả lời là “Yes” thì khả năng kết quả đúng là 75%.
* Gọi f là xác suất f trả lời đúng thực sự, f = Pr(“Yes”). Thì trong tập dữ liệu ngẫu nhiên, giá trị của f là: f = (1 + 2p)/2, với **p** là xác suất trả lời “yes”.
* **Độ chính xác của ước lượng** sẽ tăng khi số người tham gia khảo sát lớn
* Mô hình của giao thức tung đồng xu



## Mục tiêu.

* Tạo cơ sở dữ liệu cho mô hình huấn luyện ngoại tuyến với cơ sở dữ liệu đã được thả nhiễu.

## Phân tích thật toán.

* Input: Data mảng chứa dữ liệu cần được thả nhiễu.
* Output: result: mảng chứa dữ liệu đã được thả nhiễu.

Ảnh có chứa văn bản

Mô tả được tạo tự động

* Input:
* N - int: số lượng mẫu đã được thả nhiễu.
* SE - float: sai số ước lượng của dữ liệu.
* Output: q độ bảo mật với cớ mẫu n

Ảnh có chứa văn bản

Mô tả được tạo tự động

* Tạo dữ liệu ngẫu nhiên với 500 mẫu
* Dữ liệu bao gồm giá trị {0, 1} tương ứng với {“No”, “Yes”}

Ảnh có chứa văn bản

Mô tả được tạo tự động

## Kiểm tra.

* Với cớ mẫu là 500 thì độ chính xác của thuật toán là 0.8944.

Ảnh có chứa văn bản, thiết bị điện tử, bàn phím

Mô tả được tạo tự động

## Kết luận

* Có thể áp dụng giao thức tung đồng xu để làm nhiễu dữ liệu như: làm nhiễu dữ liệu bệnh nhân mắc bênh X, làm nhiễu dữ liệu những nhân viên lương trên 3000$...

# XÂY DỰNG MÔ HÌNH HUẤN LUYỆN TRỰC TUYẾN TRÊN THUẬT GIẢI HỒI QUY TUYẾN TÍNH VỚI 2 CHUYÊN GIA

## Giới thiệu.

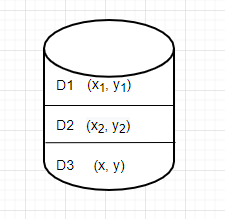
* Mô hình huấn luyện trực tuyến: là huấn luyện dựa vào mô hình có trọng số lớn trong các mô hình đã được huấn luyện trước đó. Với mỗi mô hình có trọng số biểu thị độ tin tưởng dựa độ sai lệnh kết quả đầu ra. Với mỗi vòng lặp trọng số của mô hình đó luôn cập nhật. Kết quả trả về là một mô hình huấn luyện mới.
* Hồi quy tuyến tính đơn giản: "Hồi quy tuyến tính" là một phương pháp thống kê để hồi quy dữ liệu với biến phụ thuộc có giá trị liên tục trong khi các biến độc lập có thể có một trong hai giá trị liên tục hoặc là giá trị phân loại. Nói cách khác "Hồi quy tuyến tính" là một phương pháp để dự đoán biến phụ thuộc (Y) dựa trên giá trị của biến độc lập (X). Nó có thể được sử dụng cho các trường hợp chúng ta muốn dự đoán một số lượng liên tục. Ví dụ, dự đoán giao thông ở một cửa hàng bán lẻ, dự đoán thời gian người dùng dừng lại một trang nào đó hoặc số trang đã truy cập vào một website nào đó v.v...
* Mô hình huấn luyện trực tiếp trên thuật toán hồi quy tuyến tính: là huấn luyện mô hình dự đoán kết quả theo mô hình tuyến tinh() dựa vào nhiều mô hình tuyến tính được huấn luyện từ nhiều tập dữ liệu khác nhau.

## Mục tiêu.

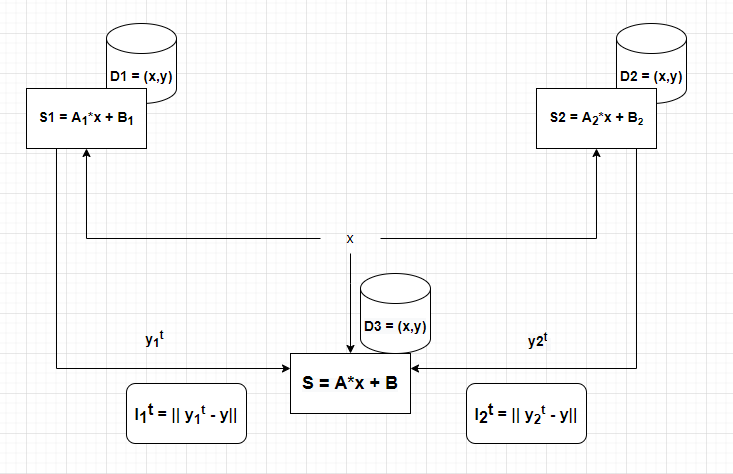
* Huấn luyện mô hình dự đoán kết quả theo mô hình tuyến tinh() dựa 2 mô hình tuyến tính được huấn luyện từ 2 tập dữ liệu khác nhau ( gọi là 2 chuyên gia khác nhau).

## Phân tích thuật toán.

* Chia dữ liệu để huấn luyện riêng cho từng chuyên gia: dữ liệu được chia làm 3 phần
* D1: bao gồm {x,y} là dữ liệu huấn luyện ngoại tuyến cho chuyên gia 1.
* D2: bao gồm {x,y} là dữ liệu huấn luyện ngoại tuyến cho chuyên gia 2
* D2: bao gồm {x,y} là dữ liệu huấn luyện trực tuyến dựa trên 2 chuyên gia.



* Mô hình huấn luyện trực tuyến:
* Chọn mô hình để huấn luyện bằng trọng số, trọng số cao thì được chọn.
* Lấy x từ D3 để huấn luyện trên mô hình được chọn.
* Kết quả của quá trình huấn luyện là ( t là vong lặp, i là mô hình để huấn luyện)
* Mô hình S tính độ mất mát của mô hình ( = || - y||) và cập nhật lại trọng số cho S1 hoặc S2.



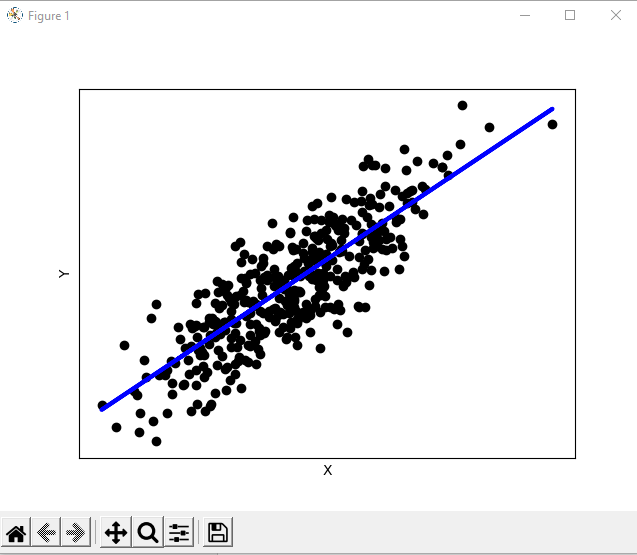
## Kiểm tra.

* Kết quả huấn luyện trên tập dữ liệu của D3
* Trọng số của 2 mô hình ban đầu là [1 , 1]
* Với i = 0 là mô hình S1, i = 1 là mô hình S2
* Với T = 20 thì trọng số của các mô hình được cập nhật liên tục.
* L\_t: là mất mát của 2 mô hình.
* Độ tiết nuối của thuật giải là Regert.

Ảnh có chứa văn bản

Mô tả được tạo tự động

* Đồ thị của mô hình trên tập D3



## Kết luận.

* Mô hình hồi quy tuyến tính dựa trên huấn luyện Online tạo mô hình có thể thích ứng với dữ liệu rời rạc ( khác dữ liệu tính trong học Offline). Khi dữ liệu bị thay đổi mô hình đươc cập nhật và điều chỉnh nhanh chóng.

# Tài liệu tham khảo.

<https://courses.fit.hcmus.edu.vn/pluginfile.php/170551/mod_resource/content/6/DP_PTDLDBTRT.pdf>

<https://www.cs.princeton.edu/courses/archive/spring18/cos511/scribe_notes/0411.pdf>

<https://viblo.asia/p/linear-regression-hoi-quy-tuyen-tinh-trong-machine-learning-4P856akRlY3>

link source code: <https://github.com/hvanduy-us/PTDL_DBTRT>