Machine Learning Security: Inference/Membership Attack

LÃ NGỌC ÁNH - 20521065
TRẦN ĐẠI DƯƠNG - 20521226
TRƯƠNG ĐÌNH TRỌNG THANH - 20520766

1 Giới thiệu

1.1 Tổng quan vấn đề

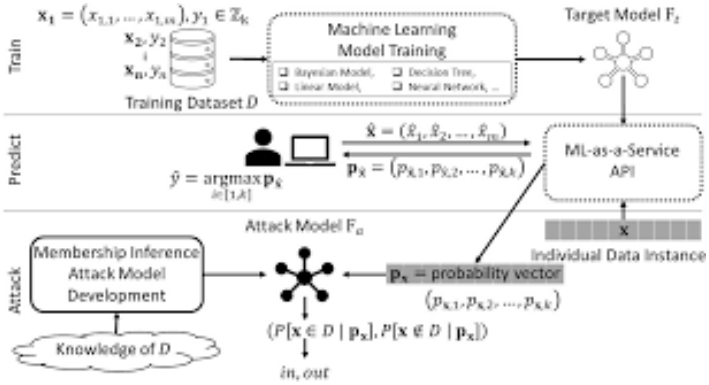
Dịch vụ học máy đã chứng kiến một sự bùng nổ quan tâm đến sự phát triển đến từ nền tảng đám mây dịch vụ. Amazon, Microsoft, IBM và Google tất cả đã đưa ra các hình thức máy học máy học như một dịch vụ. Các dịch vụ này cho phép các công ty tận dụng công nghệ máy học hiệu quả và trí tuệ nhân tạo mà không yêu cầu kiến thức chuyên môn. Các nền tảng máy học như một dịch vụ cho phép người dùng tải lên dữ liệu của họ, chạy các phân tích dữ liệu khác nhau hoặc các chuyên gia xây dựng mô hình và triển khai các mô hình được đào tạo cho các dịch vụ của riêng họ.
Với bối cảnh này, mối quan tâm mới đã được trao cho các lỗ hổng tiềm ẩn của các dịch vụ học máy như vậy. Một trong những lỗ hổng đó là Membership inference.

1.2 Trường hợp cụ thể

Một trung tâm điều trị ung thư tận dụng nền tảng máy học như một dịch vụ để phát triển một mô hình dự đoán, khi bệnh nhân đưa ra dữ liệu làm đầu vào, có thể dự đoán kết quả sức khỏe liên quan đến ung thư. Sau đó, trung tâm điều trị sử dụng triển khai đám mây tùy chọn để tạo một dịch vụ của riêng họ, trong đó người dùng có thể đăng nhập cung cấp thông tin sức khỏe của chính họ và đổi lại nhận được thông báo trước. Cuộc tấn công Membership inference xem xét một tình huống trong đó người dùng dự đoán hộp đen như vậy dịch vụ là một đối thủ. Đối thủ này có thể cung cấp thông tin sức khỏe của một cá nhân X khác và dựa trên đầu ra của mô hình, cố gắng suy luận xem X có phải là bệnh nhân ung thư ở Trung tâm điều trị. Có hai bên chính quan tâm đến Bản thảo đã nhận được X; sửa đổi Y. bảo vệ chống lại các cuộc tấn công suy luận thành viên như vậy: Bệnh nhân X và trung tâm điều trị ung thư. Các bệnh nhân trước đây của trung tâm điều trị ung thư, chẳng hạn như bệnh nhân X, coi tư cách thành viên của họ là tư nhân và không muốn thông tin của họ là thông tin công khai. Ví dụ, hãy xem xét trường hợp của một bệnh nhân ở trung tâm điều trị, Alice. Hãy để Alice đang được xem xét cho một công việc tại công ty của Bob. Bob có thể tận dụng dịch vụ của trung tâm điều trị ung thư để suy ra Alice có phải là bệnh nhân hay không. Khi biết Alice được đưa vào cơ sở dữ liệu của trung tâm điều trị ung thư, Bob quyết định không thuê Alice mà ủng hộ một ứng viên mà anh tin rằng sẽ có chi phí chăm sóc sức khỏe thấp hơn cho công ty.

2 Membership Inference attacks

2.1 Mô hình chung



Hình 1: Minh họa quá trình hình thành M.I

Dữ liệu bao gồm:

- 1 tập dữ liệu huấn luyện D
- Mô hình phân loại Ft được huấn luyện dựa trên D

Phương thức hoạt động của Membership Interface Cos0=1

- Nhà cung cấp dịch vụ máy học có thể cung cấp dịch vụ phân loại thông qua một API dự đoán. API này cung cấp cho người dùng hộp đen truy cập vào Ft

- Người dùng có thể gửi các truy vấn dự đoán với dữ liệu của riêng họ tới dịch vụ và nhận các dự đoán phân loại. Một đối thủ sử dụng một dịch vụ như vậy để thu thập thông tin về tập D riêng tư mà trên đó mô hình chuyển hướng trước Ft đã được đào tạo 1 cách riêng tư.

- Bằng cách tận dụng mọi kiến thức chung hoặc kiến thức nền tảng về tập dữ liệu đào tạo D hoặc mô hình mục tiêu Ft, đối thủ xây dựng mô hình tấn công suy luận thành viên Fa để triển khai cho việc phát động các cuộc tấn công suy luận thành viên trong thời gian thực.

2.2 Shadow Model

- Mục tiêu của kẻ tấn công là xây dựng một mô hình tấn công có thể nhận ra những khác biệt như vậy trong hành vi của mô hình mục tiêu và sử dụng chúng để phân biệt các thành viên với những thành viên không phải là thành viên của tập dữ liệu đào tạo của mô hình mục tiêu chỉ dựa trên đầu ra của mô hình mục tiêu.

- Nhiều mô hình "shadow"nhằm hoạt động tương tự như mô hình mục tiêu. Ngược lại với mô hình đích, chúng tôi biết sự thật cơ bản cho mỗi shadow model, tức là, liệu một bản ghi nhất định có nằm trong tập dữ liệu huấn luyện của nó hay không. Do đó có thể sử dụng đào tạo có giám sát về đầu vào và đầu ra tương ứng (mỗi đầu ra được gắn nhãn “In” hoặc “Out”) của shadow model để dạy mô hình tấn công cách phân biệt đầu ra của mô hình bóng trên các thành viên của bộ dữ liệu đào tạo với kết quả đầu ra trên không phải là thành viên.

2.3 Phương thức khởi tạo shadow model training data

1 số phương pháp để tạo ra các data:

- Model-based synthesis: Nếu kẻ tấn công không có thực dữ liệu đào tạo cũng như bất kỳ số liệu thống kê nào về phân phối của nó, có thể tạo dữ liệu đào tạo tổng hợp cho các mô hình shadow bằng cách sử dụng chính mô hình mục tiêu. Mục đích là các bản ghi được được phân loại theo mô hình mục tiêu với độ tin cậy cao.

  - Tạo nên 1 thuật toán tìm kiếm để kiếm được dữ liệu có độ tin cậy cao nhất, bắt đầu bằng 1 dữ liệu bất kì -> mỗi vòng lặp ta thay đổi 1 vài feature của dữ liệu -> để có thể đến tới vùng có độ tin cậy cao hơn -> và giữ dữ liệu với dộ tin cậy cao.

  - Bao gồm 2 phần:
  1) Tìm kiếm sử dụng thuật toán hill-climbing, khoảng số possible data record -> kiếm input được classified bởi mô hình mục tiêu với độ tin cậy cao.
  2) Sample systhesis data từ những record này.

- Statistics-based synthesis: Kẻ tấn công có được thông tin về nơi training data được lấy. Ví dụ kẻ tấn công có thể có kiến thức trước về phân phối cận biên của các tính năng khác nhau. Trong các thí nghiệm của chúng tôi, tạo ra tổng hợp hồ sơ đào tạo cho các mô hình shadow một cách độc lập lấy mẫu giá trị của từng tính năng từ biên của chính nó phân bổ.

- Noisy real data: Kẻ tấn công có thể có khả năng truy cập vào dữ liệu tương tự với training data của mô hình mục tiêu.

2.4 Huấn luyện mô hình tấn công

- Ý tưởng chính đằng sau kỹ thuật đào tạo bóng tối của chúng tôi là các mô hình tương tự được đào tạo trên các bản ghi dữ liệu tương đối giống nhau bằng cách sử dụng cùng một dịch vụ hoạt động theo một cách tương tự.

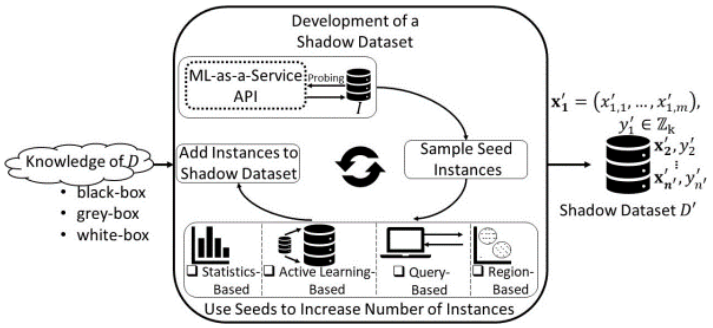
- Truy vấn từng mô hình bóng với tập dữ liệu đào tạo của riêng nó và với một bộ kiểm tra rời rạc có cùng kích thước. Các kết quả đầu ra trên tập dữ liệu huấn luyện được gắn nhãn “in”, phần còn lại được gắn nhãn “out”. Bây giờ, kẻ tấn công có một tập dữ liệu các bản ghi, tương ứng đầu ra của các mô hình bóng tối và các nhãn in/out. Các mục tiêu của mô hình tấn công là suy ra các nhãn từ bản ghi và đầu ra tương ứng.

- Nếu chúng tôi sử dụng model-based synthesis dựa trên mô hình từ tất cả dữ liệu đào tạo thô cho mô hình tấn công được rút ra từ các bản ghi được phân loại theo mô hình mục tiêu với cao lòng tin. Tuy nhiên, điều này đúng đối với cả các bản ghi

được sử dụng trong tập dữ liệu đào tạo của mô hình shadow và cho các bản ghi test tách khỏi các bộ dữ liệu này. Vì vậy, nó không phải là trường hợp mà mô hình tấn công chỉ đơn giản là học cách nhận ra các đầu vào phân loại với độ tin cậy cao. Thay vào đó, nó học cách thực hiện một nhiệm vụ phức tạp hơn nhiều: làm thế nào để phân biệt giữa đào tạo đầu vào được phân loại với độ tin cậy cao và khác, không đào tạo đầu vào cũng được phân loại với độ tin cậy cao. Trên thực tế, ta chuyển đổi vấn đề nhận ra mối quan hệ phức tạp giữa các thành viên của tập dữ liệu huấn luyện và đầu ra của mô hình thành một vấn đề binary classification.

3 Thực nghiệm

3.1 Phát triển bộ Shadow DataSet



Hình 2: Phát triển bộ Shadow DataSet

Cho một mô hình mục tiêu Ft, tập dữ liệu huấn luyện D của nó, và kiến thức đối nghịch hộp đen, sự phát triển của một bộ dữ liệu bóng D' là bước đầu tiên để tạo mô hình tấn công suy luận thành viên. D' bao gồm n' đào tạo trường hợp (x'1; y1'); (x'2; y2'); ... (x'n'; y'n') trong đó mỗi x' i bao gồm m đặc trưng tương đương với các đặc trưng trong D và mỗi yi' là một nhãn lớp dự đoán trong Zk. Lưu ý rằng k và m đã biết thông qua API dịch vụ và do đó nhất quán trên D và D'. Các cardinality của D, tuy nhiên, vẫn chưa được biết và do đó n và n' có khả năng khác nhau. Tạo tập dữ liệu bóng tối quy trình tận dụng API dịch vụ dự đoán để quản lý tạo và kiểm soát chất lượng của D', như trong Hình 2.

Kết quả của việc thăm dò API này, đối thủ có thể xây dựng tập dữ liệu khung D', tương tự như D về cấu trúc và lý tưởng nhất là bất kỳ x' > D' nào cũng phải là một phiên bản khả thi có thể được đưa vào D.

Các mô hình bóng tối phải được đào tạo theo cách tương tự như mô hình mục tiêu. Điều này thật dễ dàng nếu thuật toán đào tạo của mục tiêu (ví dụ: mạng thần kinh, SVM, hồi quy logistic) và cấu trúc mô hình (ví dụ: hệ thống dây điện của mạng thần kinh) đã được biết. Học máy như một dịch vụ khó khăn hơn. Ở đây, loại và cấu trúc của mô hình đích không được biết, nhưng kẻ tấn công có thể sử dụng chính xác cùng một dịch vụ (ví dụ: Google Prediction API) để huấn luyện mô hình bóng như đã được sử dụng để huấn luyện mô hình đích.

Algorithm 1 Data synthesis using the target model

procedure SYNTHESIZE(class : c)
  $\mathbf{x} \leftarrow$ RANDRECORD( ) ▷ *initialize a record randomly*
  $y_c^* \leftarrow 0$
  $j \leftarrow 0$
  $k \leftarrow k_{max}$
  for $iteration = 1 \cdots iter_{max}$ **do**
    $\mathbf{y} \leftarrow f_{\text{target}}(\mathbf{x})$ ▷ *query the target model*
    if $y_c \geq y_c^*$ **then** ▷ *accept the record*
      if $y_c > \text{conf}_{min}$ and $c = \arg\max(\mathbf{y})$ **then**
        if $\text{rand}() < y_c$ **then** ▷ *sample*
          return $\mathbf{x}$ ▷ *synthetic data*
        end if
      end if
      $\mathbf{x}^* \leftarrow \mathbf{x}$
      $y_c^* \leftarrow y_c$
      $j \leftarrow 0$
    else
      $j \leftarrow j + 1$
      if $j > rej_{max}$ **then** ▷ *many consecutive rejects*
        $k \leftarrow \max(k_{min}, \lceil k/2 \rceil)$
        $j \leftarrow 0$
      end if
    end if
    $\mathbf{x} \leftarrow$ RANDRECORD($\mathbf{x}^*$, k) ▷ *randomize k features*
  end for
  return $\perp$ ▷ *failed to synthesize*
end procedure

Hình 3: Statistics-based synthesis Model

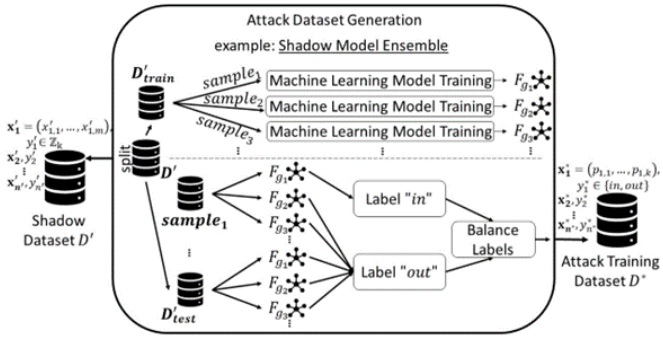
3.2 Tạo dữ liệu đào tạo cho các Shadow Model

Để huấn luyện các mô hình bóng tối, kẻ tấn công cần dữ liệu huấn luyện được phân phối tương tự như dữ liệu huấn luyện của mô hình đích. chúng tôi đã phát

triển một số phương pháp để tạo dữ liệu đó.

Chúng tôi sẽ chọn Model-based synthesis. Bởi Nếu kẻ tấn công không có dữ liệu đào tạo thực cũng như bất kỳ số liệu thống kê nào về phân phối của nó, thì anh ta có thể tạo dữ liệu đào tạo tổng hợp cho các mô hình bóng tối bằng chính mô hình đích. Trực giác là các bản ghi được phân loại theo mô hình mục tiêu với độ tin cậy cao sẽ tương tự về mặt thống kê với tập dữ liệu đào tạo của mục tiêu và do đó cung cấp nguồn thức ăn tốt cho các mô hình bóng tối.
Quá trình tổng hợp diễn ra theo hai giai đoạn: (1) tìm kiếm, sử dụng một thuật toán leo đồi, không gian của các bản ghi dữ liệu có thể để tìm đầu vào được phân loại theo mô hình mục tiêu với lòng tin cao; (2) mẫu dữ liệu tổng hợp từ các hồ sơ này. Sau đó quá trình này tổng hợp một bản ghi, kẻ tấn công có thể lặp lại nó cho đến khi tập dữ liệu đào tạo cho các mô hình bóng tối đã đầy.
Xem Hình 3 để biết mã giả tổng hợp của chúng tôi thủ tục. Đầu tiên, sửa lớp c mà kẻ tấn công muốn tạo dữ liệu tổng hợp. Giai đoạn đầu tiên là một quá trình lặp đi lặp lại. Bắt đầu bằng cách khởi tạo ngẫu nhiên một bản ghi dữ liệu x. Gia sư rằng kẻ tấn công chỉ biết định dạng cú pháp của bản ghi dữ liệu, lấy mẫu giá trị cho từng tính năng một cách ngẫu nhiên từ trong số tất cả các giá trị có thể có của tính năng đó. Trong mỗi lần lặp, đề xuất một kỷ lục mới. Hồ sơ đề xuất chỉ được chấp nhận nếu nó làm tăng mục tiêu leo đồi: xác suất của được mô hình mục tiêu phân loại là lớp c.
Mỗi lần lặp liên quan đến việc đề xuất một bản ghi ứng cử viên mới bằng cách thay đổi k tính năng được chọn ngẫu nhiên của tính năng được chấp nhận mới nhất ghi x. Điều này được thực hiện bằng cách lật các đối tượng nhị phân hoặc lấy mẫu lại các giá trị mới cho các đối tượng thuộc các loại khác. chúng tôi khởi tạo k thành kmax và chia cho 2 khi rejmax đề xuất tiếp theo bị từ chối. Điều này kiểm soát đường kính tìm kiếm xung quanh bản ghi được chấp nhận để đề xuất một bản ghi mới. chúng tôi đặt giá trị nhỏ nhất của k để kmin. Điều này kiểm soát tốc độ của tìm kiếm các bản ghi mới với khả năng phân loại cao hơn xác suất yc.
Giai đoạn thứ hai, giai đoạn lấy mẫu bắt đầu khi mô hình mục tiêu xác suất yc rằng bản ghi dữ liệu được đề xuất được phân loại là thuộc về lớp c lớn hơn xác suất cho tất cả các lớp khác và cũng lớn hơn một ngưỡng. Cái này đảm bảo rằng nhãn dự đoán cho bản ghi là c và rằng mô hình mục tiêu đủ tự tin trong dự đoán nhãn của nó. chúng tôi chọn bản ghi như vậy cho tập dữ liệu tổng hợp với xác suất yc và, nếu lựa chọn không thành công, hãy lặp lại cho đến khi một bản ghi được chọn

3.3 Quy trình tạo nên ATTACK model training set



Hình 4: Thuật toán đào tạo

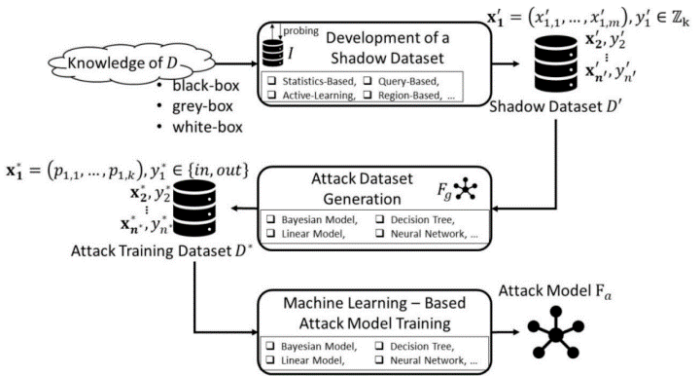
Sau khi hoàn thành sự phát triển của bóng tối tập dữ liệu D', đối thủ sẽ tiếp tục sử dụng bóng tập dữ liệu D' để phát triển tập dữ liệu tấn công thành viên cho đào tạo một bộ phân loại nhị phân làm mô hình tấn công cuối cùng, như được hiển thị trong Hình 5. Cho rằng mỗi trường hợp trong D' bao gồm một vectơ đặc trưng và lớp đã biết của nó, ký hiệu là (x'; y'), đối thủ có thể xác định chức năng tạo tấn công được biểu thị bởi Fg : (Rm; Zk) -> (Rk; Z2). Fg lấy một cặp lớp vector đặc trưng (x'; y') làm đầu vào và xuất ra một cuộc tấn công ví dụ đào tạo, bao gồm hai phần thông tin: một vectơ xác suất p = (p1; p2;... pk) và nhãn lớp nhị phân, biểu thị "In" hoặc "Out". Có một số cách tiếp cận để tạo ra Fg sử dụng D'. Ví dụ, đối phương có thể đào tạo một mô hình mới trên D' mô phỏng riêng tư mô hình mục tiêu Ft. Trong trường hợp này, chúng tôi gọi Fg là một mô hình bóng tối của Ft. Cho rằng đối thủ không biết tập huấn luyện ban đầu D cũng như kích thước của D, đối thủ có thể tận dụng các kỹ thuật học tập đồng bộ, chẳng hạn như dữ liệu nhóm dựa trên phân vùng, nhóm dựa trên mô hình hoặc mô hình nhóm hỗn hợp, để cải thiện chất lượng của bóng mô hình với mục tiêu mô phỏng chính xác mục tiêu người mẫu Ft. Chức năng tạo tập huấn luyện mô hình tấn công Fg : (Rm; Zk) -> (Rk; Z2) do đó có thể được xem như một tập hợp của một tập hợp các mô hình

bóng tối. Những mô hình bóng tối này tìm cách mô tả ranh giới quyết định của mô hình mục tiêu.
Cụ thể hơn, các mô hình bóng nhằm mục đích phản ánh độ nhạy của ranh giới quyết định mục tiêu đối với cá nhân trường hợp. Xem xét cách tiếp cận tập hợp dựa trên phân vùng dữ liệu. Kẻ thù phân vùng tập dữ liệu bóng D' thành Dtrain ' và Dtest'. Dtrain ' sau đó được chia thành q phân vùng (q > 1), một phân vùng cho mỗi mô hình bóng tối. Mỗi phân vùng của D' sau đó huấn luyện sẽ được sử dụng để huấn luyện một mô hình bóng đơn Fgi. Ở đây chúng tôi cố ý không chỉ định máy học loại mô hình của Fgi vì đây là một lựa chọn thiết kế khác được thực hiện bởi một kẻ thù. Quyết định có thể được thông báo nếu đối thủ biết loại mô hình của Ft hoặc được chọn bằng cách sử dụng một số tiêu chí khác, chúng tôi để lại quyết định này không xác định loại bỏ các ràng buộc rằng một đối thủ phải biết loại mô hình của Ft. Tiếp theo, Dtest ' sẽ được đánh giá theo Fgi. Các đầu ra tương ứng sau đó sẽ được dán nhãn là "Out". Ngoài ra, một mẫu có kích thước |Dtest '| được lấy từ D' phân vùng đào tạo được sử dụng để đào tạo Fgi và được đánh giá dựa trên Fgi với các đầu ra tương ứng được gắn nhãn là "In". Qua kết hợp các cặp nhãn đầu ra này, chúng tôi có được cuộc tấn công huấn luyện dữ liệu D.
Hình 4 làm nổi bật quy trình tạo ra một cuộc tấn công tập huấn luyện mô hình. Một đối thủ có thể chọn một mô hình duy nhất cho hiệu quả hoặc một tập hợp các mô hình để tăng kích thước hoặc tính tổng quát của D. Đối thủ có thể đa dạng hóa mô hình loại trong một bản hòa tấu khi loại của Ft' không xác định. Nếu một toàn bộ được sử dụng, có các lựa chọn về kích thước, lấy mẫu và tổng hợp phải được thực hiện và có thể được thông báo bởi kiến thức của kẻ thù về mục tiêu theo nhiều cách khác nhau. chúng tôi nhấn mạnh rằng trong khi chúng tôi xây dựng suy luận thành viên tấn công sử dụng một mô hình chung, tồn tại nhiều biến thể triển khai.
Tác dụng của các phương pháp tập hợp. Kết hợp nhiều mô hình khác nhau làm giảm nguy cơ chọn giả thuyết sai trong không gian giả thuyết của một vấn đề cụ thể. Ngoài ra, nhiều mô hình cho phép tìm kiếm cục bộ hiệu quả hơn, điều mà nhiều thuật toán học máy thực hiện theo nhiều cách khác nhau và hạn chế tác động của vấn đề tối ưu cục bộ. Cuối cùng, sự kết hợp của các giả thuyết đã chọn cho phép mở rộng không gian giả thuyết. Hai cách phổ biến để thực hiện sự đa dạng này là đóng gói và tăng cường.

3.4 Phát triển Mô hình Membership Attack

Chúng tôi rút ra được công thức chung để thực thi mô hình này:

- Tạo nên data để đưa vào shadow dataset D'
- Sử dụng shadow model dataset to develop shadow model bằng chính nền tảng ML được sự dủng trên target model để giả dạng hành vi của target model, từ đó tạo nên data từ prediction của shadow model cho ATTACK model
- Sử dụng binary classifier để tạo nên dự đoán



Hình 5: Mô hình Attack

4 Đánh giá

CIFAR 10	2500	5000	10 000	15 000
Precision	0.7054	0.7044	0.7113	0.7062
Recall	0.7036	0.7026	0.7064	0.7044
F1-score	0.7030	0.7019	0.7222	0.7038
Accuracy	0.7366	0.7026	0.7046	0.7044

Hình 6: Độ chính xác của cuộc tấn công suy luận thành viên chống lại các mạng thần kinh được đào tạo trên bộ dữ liệu CIFAR 10. Các biểu đồ hiển thị độ chính xác cho các lớp khác nhau trong khi thay đổi kích thước của tập dữ liệu huấn luyện. Các giá trị trung bình được kết nối trên các kích thước tập huấn luyện khác nhau. Độ chính xác trung bình (từ kích thước tập dữ liệu nhỏ nhất đến lớn nhất) là 0,70, 0,71 đối với CIFAR-10

4.1 DataSet

CIFAR-10 là bộ dữ liệu điểm chuẩn dùng để đánh giá các thuật toán nhận dạng ảnh. CIFAR-10 bao gồm các hình ảnh màu 32×32 trong 10 lớp, với 6.000 hình ảnh mỗi lớp. Tổng cộng, có 50.000 hình ảnh đào tạo và 10, 000 hình ảnh thử nghiệm. CIFAR-100 có cùng định dạng với CIFAR-10, nhưng nó có 100 lớp chứa 600 hình ảnh mỗi lớp. Có 500 hình ảnh đào tạo và 100 hình ảnh thử nghiệm cho mỗi lớp. Sử dụng các phần khác nhau của tập dữ liệu này trong cuộc tấn công thử nghiệm để cho thấy ảnh hưởng của kích thước tập dữ liệu đào tạo đối với độ chính xác của cuộc tấn công.

4.2 Mô hình mục tiêu

Chúng tôi đã đánh giá các cuộc tấn công suy luận của mình trên một nền tảng chúng tôi đã triển khai tại local. Trong tất cả các cuộc tấn công của chúng tôi coi các mô hình là hộp đen.

Chúng tôi đã sử dụng nền tảng theo hai cấu hình: cài đặt mặc định (10, 1e 6) và (100, 1e 4).

Neural network đã trở thành một cách tiếp cận phổ biến để học máy quy mô lớn. Chúng tôi sử dụng Torch7 và các gói nn của nó, một thư viện học sâu có đã được sử dụng và mở rộng bởi các công ty Internet lớn như Facebook. Trên bộ dữ liệu CIFAR, chúng tôi đào tạo một mạng thần kinh tích chập tiêu chuẩn mạng (CNN) với hai lớp tích chập và tổng hợp tối đa cộng với một lớp được kết nối đầy đủ có kích thước 128 và một lớp SoftMax. Chúng tôi sử dụng Tanh làm chức năng kích hoạt. Chúng tôi thiết lập việc học tỷ lệ thành 0,001, tỷ lệ học tập phân rã thành 1e - 07 và đào tạo epoch tối đa đến 100

4.3 Thiết lập thử nghiệm

Chúng tôi thay đổi kích thước của tập huấn luyện cho CIFAR bộ dữ liệu, để đo lường sự khác biệt về độ chính xác của cuộc tấn công. Đối với tập dữ liệu CIFAR-10, chúng tôi chọn 2, 500; 5, 000; 10.000; và 15, 000.
Các thử nghiệm trên bộ dữ liệu CIFAR được chạy cục bộ, dựa trên các mô hình của chúng tôi, vì vậy chúng tôi có thể thay đổi mô hình cấu hình và đo lường tác động đến độ chính xác của cuộc tấn công.

Chúng tôi đặt số lượng mô hình shadow là 20 cho bộ dữ liệu CIFAR 10.

4.4 Độ chính xác của cuộc tấn công

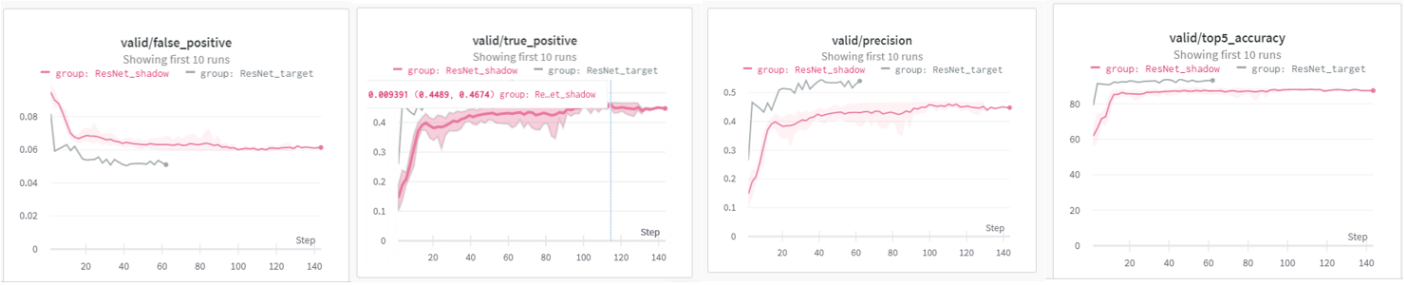
Mục tiêu của kẻ tấn công là xác định xem một bản ghi đã cho là một phần của tập dữ liệu đào tạo của mô hình mục tiêu. chúng tôi đánh giá cuộc tấn công này bằng cách thực hiện nó trên các bản ghi được xáo trộn ngẫu nhiên từ tập dữ liệu kiểm tra và huấn luyện của mục tiêu.

Trong đánh giá cuộc tấn công của chúng tôi, chúng tôi sử dụng các bộ có cùng kích thước (nghĩa là số thành viên bằng nhau và không phải là thành viên) để tối đa hóa sự không chắc chắn của suy luận, do đó độ chính xác cơ sở là 0,5 và được mô tả qua Hình 6 và Hình 7. Hàm Top5-accuracy có tác dụng để hiển thị chỉ số accuracy của mỗi model. Dựa vào hình có thể thấy đường của target ngắn hơn shadow là vì target accuracy cao nên sẽ dừng sớm. Còn shadow model thì vẫn còn thấp nên phải chạy đến khi nào accuracy đạt cao nhất thì dừng. Và vì chỉ có 20 shadow nên precision, TP, FP còn thấp.

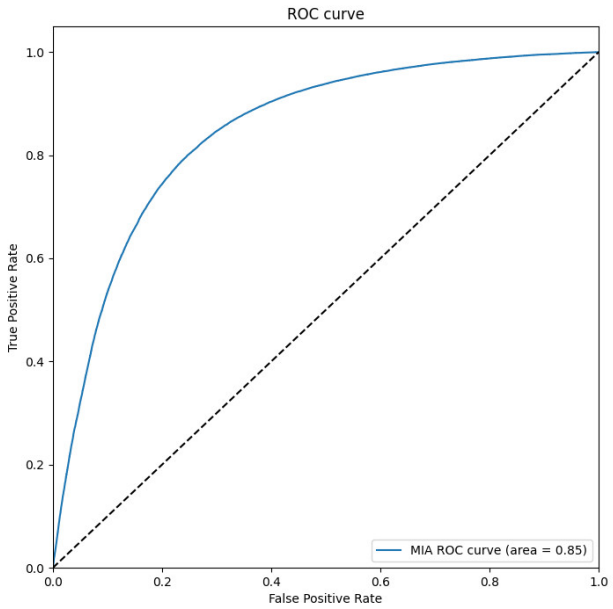
Hình 8 biểu thị việc đánh giá trên tập dữ liệu nội bộ với dữ liệu có sẵn và kiến thức toàn vẹn, chúng tôi nhận thấy tỉ lệ dự đoán nhận biết dương tính thật khá cao đối với việc giải quyết sự khác biệt giữa tập dữ liệu nằm trong mô hình mục tiêu và tập dữ liệu nằm ngoài.

4.5 Tóm tắt quá trình thực hiện Model

- Đầu tiên khai báo số mô hình shadow, tỷ lệ học, tên dataset, số lần train,. . . . vào file .yaml và tạo các file .ipynb trong đó có seed dùng để tạo các bản ghi ngẫu nhiên , file metrics dùng để tính các độ chính xác, trung bình,v.v. Sau đó chạy hết các file đó.
- Tạo 1 file để thêm dataset về member và non-member
- Train target model và sau đó lưu các danh mục thành khung dữ liệu (dataframe), từ đó tổng hợp ra được các dữ liệu từ mô hình target và lưu vào dataset
- Train nhiều shadow models, trong lúc train thì tác giả có điều chỉnh mô hình để từ đó tạo ra các dataset về member và non-member
- Sau khi train các shadow sẽ có dataset attack, thì ta sẽ dùng dataset về train tấn công. Cuối cùng sẽ có biểu đồ ROC như hình



Hình 7: Đánh giá M.I.A thông qua tập CIFAR-10 chạy trên 2500 tập dữ liệu mô phỏng



Hình 8: Kết quả biểu đồ

5 Kết luận

Chúng tôi đã trình bày khuôn khổ tổng quát đầu tiên để phát triển mô hình tấn công suy luận thành viên. Công thức chung này cho phép mô tả chuyên sâu các cuộc tấn công suy luận thành viên chống lại các loại mô hình học máy khác nhau. Thông qua thử nghiệm rộng rãi và bằng chứng thực nghiệm, chúng tôi chỉ ra thời điểm và lý do các mô hình máy học có thể dễ bị tấn công suy luận thành viên. Bằng cách khám phá nhiều loại mô hình học máy và mối tương quan của chúng đối với ba giai đoạn của quá trình tạo cuộc tấn công, chúng tôi trình bày năm đặc điểm thú vị của các cuộc tấn công suy luận thành viên:

- Chúng là các cuộc tấn công dựa trên dữ liệu
- Các mô hình tấn công có thể chuyển nhượng được
- Loại mô hình mục tiêu là một chỉ báo rõ ràng về lỗ hổng của mô hình
- Các kỹ thuật tạo dữ liệu tấn công không cần phải phản ánh rõ ràng mô hình mục tiêu và các cuộc tấn công suy luận thành viên có thể tồn tại như các cuộc tấn công nội bộ trong các hệ thống liên kết. chúng tôi cũng bao gồm một cuộc thảo luận về các biện pháp đối phó và phương pháp giảm thiểu chống lại các cuộc tấn công suy luận thành viên.

Nghiên cứu của chúng tôi về các cuộc tấn công suy diễn tư cách thành viên và quyền riêng tư của thành viên vẫn tiếp tục theo nhiều khía cạnh. Đầu tiên, chúng tôi đang tham gia vào việc phát triển các biện pháp đối phó và phương pháp phòng thủ. Thứ hai, chúng tôi hiện đang nghiên cứu quy mô và sự đa dạng của các cuộc tấn công suy luận thành viên trong các hệ thống học tập hợp tác và liên kết. Thứ ba, chúng tôi đang điều tra các mối quan hệ phức tạp giữa các cuộc tấn công suy luận thành viên, quyền riêng tư của thành viên và quyền riêng tư khác biệt

6 Tài liệu tham khảo

[1] Truex, Stacey, et al. "Demystifying membership inference attacks in machine learning as a service."IEEE Transactions on Services Computing (2019).

[2] Shokri, Reza, et al. "Membership inference attacks against machine learning models."2017 IEEE symposium on security and privacy (SP). IEEE, 2017.