MÔ HÌNH HỌC MÁY VÀ MỘT SỐ THÁCH THỨC BẢO MẬT

ThS Ngô Minh Phước
Trung tâm Công nghệ Thông tin, Bộ Khoa học và Công nghệ

Mô hình học máy có thể được sử dụng để xây dựng các hệ thống phòng thủ như phát hiện mã độc và tấn công mạng, tuy nhiên chúng cũng nhanh chóng trở thành đối tượng tấn công mới của các tác nhân độc hại. Vì vậy, việc xây dựng các mô hình học máy mạnh mẽ, đáp ứng miễn nhiễm với sự can thiệp từ các tác nhân bên ngoài là điều cần thiết. Bài viết chỉ ra những thách thức bảo mật đối với mô hình học máy, từ đó đề xuất một số giải pháp phòng thủ bảo mật cho các mô hình này.

Học máy và xu hướng phát triển

Học máy là một lĩnh vực của trí tuệ nhân tạo (AI) liên quan đến việc nghiên cứu và xây dựng các kỹ thuật cho phép các hệ thống học tự động từ dữ liệu để giải quyết những vấn đề cụ thể. Học máy có mối quan hệ rất mật thiết đối với thống kê, sử dụng các mô hình thống kê để ghi nhớ sự phân bố của dữ liệu. Tuy nhiên, không đơn thuần là ghi nhớ, học máy có khả năng tổng quát hóa và đưa ra những dự đoán trong tương lai.

Trong bối cảnh chuyển đổi số như hiện nay, nhiều tổ chức, doanh nghiệp đang tích cực nghiên cứu, triển khai các ứng dụng học máy vào các mô hình kinh doanh của họ như: Amazon, Google... Các tổ chức có thể sử dụng học máy để thiết kế và thực hiện quá trình chuyển đổi số, tập trung vào kế hoạch để đạt được lợi thế cạnh tranh ở hiện tại và trong tương lai (hình 1). Học máy phân tích và so sánh các mẫu trong dữ liệu lớn để cung cấp thông tin chi tiết về hành vi của khách hàng cũng như các thông tin có liên quan. Công nghệ này cũng đưa ra đề xuất về cách cải thiện các quy trình của tổ chức và các tương tác với khách hàng. Thay vì chỉ phát hiện ra những thay đổi hành vi của khách hàng sau khi doanh số đã tụt giảm, học máy có thể hỗ trợ xác định những thay đổi hành vi này ngay khi chúng xảy ra, cho phép doanh nghiệp triển khai các giải pháp để tránh sụt giảm doanh số cũng như cải thiện các cơ hội kinh doanh.

Theo xu hướng hiện tại, học máy đang trên con đường trở thành một công nghệ phổ biến trong vài năm tới. Theo báo cáo của Công ty phân tích Research and Markets (Hoa Kỳ), ngân sách đầu tư cho mô hình học máy sẽ tăng lên 8,8 tỷ USD trong năm 2022, từ mức 1,4 tỷ USD trong năm 2017, dự báo đến năm 2024, 75% các tổ chức, doanh nghiệp sẽ chuyển từ thí điểm sang vận hành mô hình AI và học máy.

Thách thức bảo mật đối với mô hình học máy

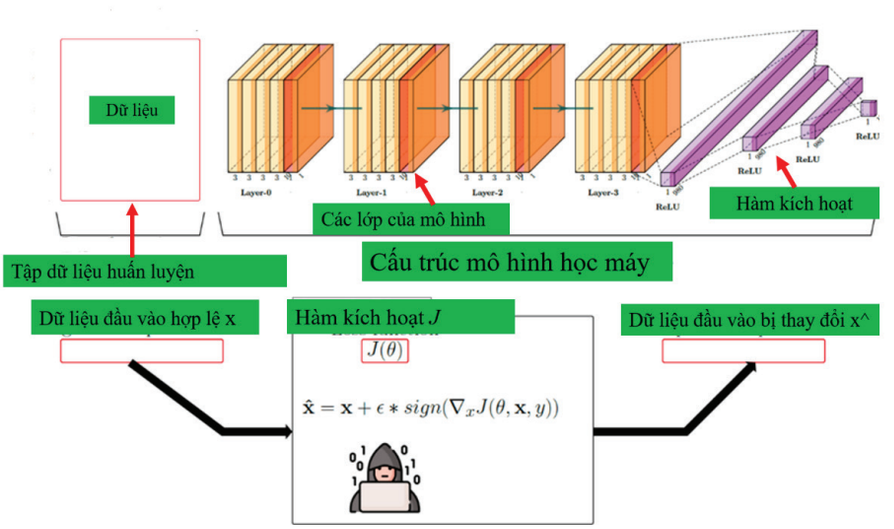
Học máy có những tiềm năng lớn nhưng đồng thời nó cũng đối mặt với những mối đe dọa nghiêm

Hình 1. Học máy trên nền tảng dữ liệu của doanh nghiệp.

trọng. Trong các lĩnh vực quan trọng được ứng dụng mô hình học máy như chăm sóc sức khỏe, giao thông và giám sát…, các cuộc tấn công vào mô hình học máy có thể dẫn đến thất thoát tài sản hoặc gây nguy hiểm cho con người. Những mối đe dọa đối với mô hình học máy có thể bao gồm những hình thức sau:

Tấn công học máy đối nghịch (adversarial attacks): Đây là tấn công sử dụng mẫu đối nghịch. Đối với kiểu tấn công này, kẻ tấn công sẽ chỉnh sửa dữ liệu đầu vào, tạo ra các mẫu giả gọi là các mẫu đối nghịch, bằng cách thêm nhiễu (thêm các trường thông tin dư thừa) vào tập dữ liệu đầu vào hợp lệ (hình 2). Các thay đổi này khó bị con người phát hiện, nhưng lại gây ảnh hưởng lớn đến đầu ra của các mô hình học máy. Tấn công FGSM (Fast Gradient Sign Method) là một dạng của tấn công học máy đối nghịch phổ biến và mạnh mẽ, có thể được sử dụng để tấn công mô hình học máy dự đoán mã chìm tín hiệu định hướng RF beamforming (kỹ thuật massive MIMO) trong mạng 6G.

Tấn công đầu độc dữ liệu (data poisoning attacks): Các mô hình học máy thường được huấn luyện lại bằng dữ liệu mới thu thập được sau khi triển khai để thích ứng với các thay đổi trong phân phối đầu vào. Ví dụ, một hệ thống phát hiện xâm nhập liên tục thu thập các mẫu trên mạng và huấn luyện lại mô hình để phát hiện các cuộc tấn công mới. Trong cuộc tấn công đầu độc dữ liệu, kẻ tấn công thực hiện xâm nhập và sửa đổi dữ liệu huấn luyện khiến cho mô hình học máy hoạt động



Hình 2. Kẻ tấn công thay đổi dữ liệu đầu vào của mô hình học máy.

không chính xác, không có khả năng phát hiện các cuộc tấn công mới.

Tấn công cửa hậu (backdoor attacks): Là một phương pháp vượt qua thủ tục xác thực người dùng thông thường hoặc thực hiện truy nhập từ xa tới một máy tính, cố gắng không bị phát hiện bởi việc giám sát thông thường. Các kỹ thuật tấn công cửa hậu có thể được nhúng vào các mô hình học máy và rất khó bị phát hiện. Thông thường backdoor là một đoạn mã nằm trong phần mềm, hoặc phần mềm nằm trong một phần cứng cho phép truy cập hệ thống từ xa nhằm lấy thông tin, hỗ trợ, phân tích hoặc dùng cho các mục đích khác. Đối với mô hình học máy, backdoor có thể được nhúng vào bằng cách thêm một số nơ-ron (các nút mạng) trong cấu trúc mạng. Hầu hết các cuộc tấn công kiểu này xảy ra trong quá trình huấn luyện mô hình.

Tấn công suy luận thành viên (membership inference attacks): Là hình thức truy vấn mô hình học máy để xác định xem một bản ghi dữ liệu cụ thể nào được sử dụng trong tập dữ liệu huấn luyện của mô hình. Các mô hình học máy được đào tạo trên số lượng hàng nghìn, hàng triệu bản ghi dữ liệu, trong nhiều trường hợp, các tập dữ liệu này có thể chứa thông tin nhạy cảm như tên, ngày sinh, địa chỉ, mật khẩu, số thẻ tín dụng, dữ liệu sức khỏe và các chi tiết cá nhân khác. Các cuộc tấn công suy luận thành viên nhằm mục đích tìm hiểu những thông tin bí mật này bằng cách thăm dò mô hình học máy với dữ liệu đầu vào khác nhau, đối chiếu với kết quả đầu ra để tiếp tục thăm dò dữ liệu đầu vào.

Tấn công trích xuất mô hình (model inversion attacks): Trích xuất mô hình học máy để tạo lại một phần hoặc toàn bộ dữ liệu đào tạo của chúng. Trong một cuộc tấn công trích xuất mô hình hoặc dữ liệu huấn luyện, kẻ tấn công sẽ phân tích đầu vào, đầu ra và thông tin bên ngoài của hệ thống để suy đoán các tham số hoặc dữ liệu huấn luyện của mô hình.

Trên cơ sở những mối đe dọa nghiêm trọng đối với mô hình học máy, có thể thấy nguyên nhân của các mối đe dọa đối với mô hình học máy xuất phát từ một số cơ sở như sau:

Thứ nhất, các mô hình học máy sẽ hoạt động tốt hơn khi khối lượng dữ liệu mà chúng được đào tạo tăng lên, khi đó các tổ chức có khả năng phải xử lý khối lượng lớn thông tin nhạy cảm, dữ liệu riêng tư..., điều đó đồng nghĩa với nguy cơ về mất an toàn dữ liệu cũng tăng lên.

Thứ hai, các mô hình học máy được đào tạo trước và chia sẻ trên internet, vốn đã trở nên rất phổ biến trong những năm gần đây. Các nhà phát triển hệ thống chưa có nhiều kinh nghiệm hoặc nguồn lực để đào tạo mô hình học máy của riêng họ mà tải xuống các mô hình được đào tạo trước này từ một trong số các nền tảng web và trực tiếp tích hợp chúng vào ứng dụng của họ. Nhưng các mô hình được đào tạo trước có thể trở thành nguồn gốc của các mối đe dọa nêu trên.

Một số giải pháp an toàn

Để hạn chế những nguy cơ về bảo mật đối với các mô hình học máy nêu trên, một số giải pháp an toàn cần được thực hiện như sau:

Một là, nâng cao tính ổn định của các mô hình học máy bằng các cơ chế như xác thực mô hình. Một số kỹ thuật xác thực có thể được sử dụng như: kỹ thuật xác thực chéo K-Fold, lấy mẫu con ngẫu nhiên, kỹ thuật bootstraping… Kỹ thuật xác thực chéo K-Fold là một phương pháp thống kê được sử dụng để ước tính hiệu suất của các mô hình học máy giúp so sánh và lựa chọn mô hình tốt nhất cho một vấn đề trong trường hợp dữ liệu không nhiều.

Hai là, xây dựng cấu trúc bảo mật an toàn cho mô hình với nhiều cơ chế, nhiều lớp. Ví dụ như cần có cơ chế quản lý dữ liệu đào tạo mô hình và áp dụng các biện pháp kiểm soát để đảm bảo rằng dữ liệu này không thể bị sửa đổi một cách độc hại. Nếu hoàn toàn phải đào tạo với dữ liệu nhạy cảm, cần xem xét các kỹ thuật bảo mật hiệu quả như ẩn danh hoặc mã hóa dữ liệu nhạy cảm. Quá trình mã hóa dữ liệu bằng phương pháp dùng thuật toán phức tạp để khóa, mã hóa một tập tin hoặc một nội dung để những hệ thống khác hoặc người khác không thể sử dụng được, không thể đọc được cho đến khi được giải mã. Trong đó, quá trình giảm thiểu tấn công được thiết kế riêng một cơ chế phòng thủ đối với các cuộc tấn công được nhận diện. Tiếp theo là sử dụng phương án thiết lập bằng bảo mật mô hình giúp nâng cao tính mạnh mẽ của mô hình bằng việc xác thực. Thêm nữa, việc bảo mật cấu trúc đã giúp định hình được một cấu trúc an toàn dựa trên nhiều cơ chế bảo mật.

Ba là, thiết lập cơ chế phòng thủ đối với các cuộc tấn công đã biết thông qua các hệ thống giám sát và phát hiện tấn công. Những hệ thống này cho phép nhanh chóng phát hiện, xác định và giải quyết các bất thường của mô hình, giúp hệ thống hạn chế được sự cố ngưng trệ và tăng thời gian hoạt động. Các cơ chế phòng thủ được nhận diện thông qua quá trình thu thập dữ liệu, huấn luyện và dự đoán mô hình. Thông qua hệ thống AI, có thể tiếp cận với các kỹ thuật phòng thủ đối với tấn công né tránh, tấn công đầu độc, tấn công cửa hậu, và tấn công trích xuất mô hình.

Trước những thách thức về tính bảo mật, mục đích của mô hình học máy là xác định các dữ liệu độc hại và ngăn chặn chúng can thiệp vào hệ thống. Không có giải pháp thực sự tối ưu để đáp ứng được mô hình học máy hoạt động với độ chính xác, bảo mật tuyệt đối, tuy nhiên đã có một số giải pháp kỹ thuật tốt giúp mô hình này ngăn chặn lại các tác nhân độc hại. Mỗi hệ thống học máy cần được đảm bảo tính bền vững trước các mối đe dọa thông qua phòng thủ bảo mật bằng nhiều cơ chế khác nhau như xác thực mô hình, bảo mật nhiều lớp, hệ thống giám sát tự động và phát hiện tấn công ✍

TÀI LIỆU THAM KHẢO

1. https://portswigger.net/daily-swig/take-threats-against-machine-learning-systems-seriously-security-firm-warns.

2. https://www.techtarget.com/searchenterpriseai/In-depth-guide-to-machine-learning-in-the-enterprise.

3. A. Moldsvor, et al. (2021), "Adversarial machine learning security problems for 6G: mmWave beam prediction use-case", *2021 IEEE International Black Sea Conference on Communications and Networking (BlackSeaCom)*, DOI:10.1109/BlackSeaCom52164.2021.9527756.