УДК 004.056.5 URL: https://ptsj.bmstu.ru/catalog/icec/insec/1019.html

БЕЗОПАСНОСТЬ И ОБЗОР УГРОЗ МАШИННОГО ОБУЧЕНИЯ

Н.С. Лосев losevns@student.bmstu.ru
Е.В. Глинская glinskaya@bmstu.ru
SPIN-код: 5430-3023

МГТУ им. Н.Э. Баумана, Москва, Российская Федерация

Машинное обучение получило широкое распространение в преобразовании различных аспектов нашей жизни с помощью интеллектуальных цифровых решений. Использование машинного обучения в критических инфраструктурах влечет за собой желание злоумышленников получить к нему доступ, заполучив алгоритмы, методы и данные, лежащие в основе работы моделей, чтобы расширить свое влияние и извлечь выгоду из уязвимостей системы. В статье рассмотрены и проанализированы основные уязвимости и классификация различных атак машинного обучения как объекта защиты. Также представлены эффективные методы, препятствующие будущим атакам, и технологии, смягчающие возможные опасные последствия.

Ключевые слова: машинное обучение, стратегии противодействия, атаки злоумышленников, защита данных, методы защиты, модели атак, уязвимости, критические инфраструктуры

Введение. Машинное обучение — это численная оптимизация параметрических моделей для описания определенного набора данных [1]. Машинное обучение используется в различных системах в самых разных приложениях, таких как здравоохранение, обработка изображений, компьютерное зрение, классификации и т. д. Информационная безопасность также использует машинное обучение для обнаружения новых атак, не встречавшихся прежде [2].

Злоумышленники манипулируют машинным обучением, внедряя вредоносное программное обеспечение и снижая конфиденциальность жертвы. Анализ воздействия каждого типа атаки позволяет получить информацию, которая поможет сформировать стратегию и ответить на следующие исследовательские вопросы:

– необходимо определить наиболее важные типы атак на машинное обучение, которые следует изучить и проанализировать (см. рисунок);

– следует определить наиболее эффективные стратегии и решения, препятствующие злоумышленнику и смягчающие последствия атак [3].

Система машинного обучения как объект защиты. Ниже приведено краткое описание этапов обучения и применения систем машинного обучения. На вход модели поступают низкоуровневые признаки, отражающие объ-

 ISSN 2541-8009

екты реального мира. Процесс их получения подразделяют на два этапа: измерение величин и их преобразование. Признаки являются выборкой из некоторой генеральной совокупности, недоступной для обучения.

Классификация атак на системы машинного обучения

Модель имеет множество гиперпараметров, которые не меняются в процессе обучения и параметров, которые настраиваются. Гиперпараметры определяют архитектуру модели.

В целом система машинного обучения состоит из двух основных этапов:

1) этап обучения: обучающий алгоритм извлекает информацию из обучающих данных, а модель обучается;

2) этап тестирования: обученной модели предоставляются новые данные, называемые тестовыми данными, чтобы понять, соответствует ли модель ожиданиям [4].

Типы атак, в основе которых лежит обработка и разработка моделей. Атака с отравлением использует в начальной фазе обучение модели машинного обучения с использованием подготовленного набора данных. Внедряя фальсифицированные выборки или изменяя существующие, злоумышленник влияет на процесс обучения и вводит в заблуждение классификацию во время тестирования [5].

Атаку уклонения используют для ввода в заблуждение данных тестирования, чтобы снизить точность тестирования целевой модели. Конечной целью будут являться неверно интерпретированные входные данные тестиро-

вания, позволяющие нарушить целостность машинного обучения во время тестирования.

Главная цель атаки с использованием инверсии модели — украсть разработанную систему. Злоумышленник извлекает представление о базовой модели с помощью атаки для дальнейшего использования обучающих данных модели в своих целях.

Атака с логическим выводом о принадлежности — атака, при которой злоумышленник может анализировать выходные данные, имея доступ к целевой модели машинного обучения жертвы. Нападающий может повторно создать набор обучающих данных для целевой модели машинного обучения, проанализировав собранные результаты [6].

Уровни знания нарушителя об атакуемой системе машинного обучения. Эти уровни представлены на рисунке.

«Черный ящик» — атака, при которой противник ничего не знает о жертве. У противника нет конкретной цели, лишь намеренье снизить общую точность целевой модели. Данный сценарий наиболее реалистичен, поскольку злоумышленник обычно не знает целевую систему [7].

«Белый ящик» — атака, при которой противник обладает полной информацией о целевой системе. Подобные атаки предназначены главным образом для достижения конкретной цели и чаще всего применяются к атакам с отравлением и уклонением [8].

Также существует «Серый ящик», сочетающий в себе свойства перечисленных выше уровней и применяемый при наличии недостаточной информации о модели.

Типы атак, в основе которых лежат возможности и намерения противника. Целевые атаки формируются на основе определенных целей. Они основаны на глубоком понимании злоумышленником целевой модели и ее уязвимостей, которые можно использовать, а также на четких целях, которые необходимо достичь.

Нецелевая атака направлена на то, чтобы каким-либо образом нарушить модель жертвы без заранее определенных целей. Этот тип атаки предназначен для выявления уязвимостей в модели машинного обучения жертвы независимо от достижения целей.

Стратегии противодействия атакам. Для защиты моделей машинного обучения разрабатывают различные методы смягчения последствий атак. Ниже представлен анализ существующих решений безопасности по типам атак.

Для противодействия атак с отравлением применяют очистку данных — предварительную обработку обучающих наборов и удаление ошибочных или искаженных выборок. Другим методом служит обучение системы данными,

 ISSN 2541-8009

содержащими помехи выборок. Это позволяет обученной модели в будущем выявлять искаженные данные самостоятельно.

Для противостояния атакам с уклонением также применяют состязательное обучение, при котором определенная часть набора данных намеренно искажается для повышения устойчивости к атакам подобного типа вследствие того, что жертва будет в курсе враждебных образцов данных. Другой метод — упрочнение модели, при котором в процессе машинного обучения создается преграда для злоумышленника во время тестирования. При этом модель жертвы становится устойчивой и надежной.

Для отпора атакам с инверсией модели и с логическим выводом о принадлежности применяют различную конфиденциальность моделей машинного обучения. Это затрудняет злоумышленнику анализ полученных данных и извлечение конфиденциальной информации. Также применяют рандомизацию вероятности выходных данных путем добавления к ним шума [9]. Это повышает конфиденциальность системы, препятствуя действиям злоумышленников [6].

Заключение. В ходе исследований были проанализированы различные типы враждебных атак, их влияние на инфраструктуру, а также средства защиты против них. Таким образом, необходимо совершенствовать решения по смягчению последствий и разрабатывать решения, которые будут обеспечивать безопасность машинного обучения, а также автоматизируют процесс разработки моделей [10]. Основное внимание при защите систем следует уделять атакам с отравления и уклонения, при которых вредоносные данные могут быть перенесены на другие системы, а также могут быть распространены на любую систему [4].

Литература

[1] Коротеев М.В. *Основы машинного обучения на Python*. Москва, КноРус, 2024, 432 с.

[2] Ulyanikhin E. Machine learning in information security field. Язык в сфере профессиональной коммуникации. *Междунар. науч.-практ. конф. преподавателей, аспирантов и студентов: сб. тр*. Екатеринбург, Издательский Дом «Ажур», 2020, pp. 704–707.

[3] Грибунин В.Г., Гришаненко Р.Л., Лабазников А.П., Тимонов А.А. Безопасность систем машинного обучения. Защищаемые активы, уязвимости, модель нарушителя и угроз, таксономия атак. *Известия Института инженерной физики*, 2021, № 3 (61), с. 65–71.

[4] Lahe A.D., Singh G. A Survey on Security Threats to Machine Learning Systems at Different Stages of its Pipeline. *International Journal of Information Technolo-*

gy and Computer Science, 2023, vol. 15, no. 2, pp. 23–34. https://doi.org/10.5815/ijitcs.2023.02.03

[5] Gupta P., Yadav K., Gupta B.B. et al. A Novel Data Poisoning Attack in Federated Learning based on Inverted Loss Function. *Computers & Security*, 2023, vol. 130, art. 103270. https://doi.org/10.1016/j.cose.2023.103270

[6] Paracha A., Arshad Ju., Farah M.B., Ismail Kh. Machine learning security and privacy: a review of threats and countermeasures. *EURASIP Journal on Information Securit*y, 2024, vol. 2024, no. 1, art. 10. https://doi.org/10.1186/s13635-024-00158-3

[7] Bai Ya., Wang Y., Zeng Yu. et al. Query efficient black-box adversarial attack on deep neural networks. *Pattern Recognition*, 2023, vol. 133, art. 109037. https://doi.org/10.1016/j.patcog.2022.109037

[8] Wu Di., Qi S., Qi Y. et al. Understanding and defending against White-box membership inference attack in deep learning. *Knowledge-Based Systems*, 2023, vol. 259, art. 110014. https://doi.org/10.1016/j.knosys.2022.110014

[9] Попков Ю.С. Машинное обучение и рандомизированное машинное обучение: сходство и различие. *Системный анализ и информационные технологии САИТ-2019. Восьмая Междунар. конф.: сб. тр.* Иркутск, ФИЦ ИУ РАН, 2019, с. 10–25. https://doi.org/10.14357/SAIT2019001

[10] Астапов Р.Л., Мухамадеева Р.М. Автоматизация подбора параметров машинного обучения и обучение модели машинного обучения. *Актуальные научные исследования в современном мире*, 2021, № 5–2 (73), с. 34–37.

Поступила в редакцию 12.10.2024

Лосев Никита Сергеевич — студент кафедры «Информационная безопасность», МГТУ им. Н.Э. Баумана, Москва, Российская Федерация.

Глинская Елена Вячеславовна — старший преподаватель кафедры «Информационная безопасность», МГТУ им. Н.Э. Баумана, Москва, Российская Федерация.

Научный руководитель — Басараб Михаил Алексеевич, доктор физико-математических наук, заведующий кафедрой «Информационная безопасность», МГТУ им. Н.Э. Баумана, Москва, Российская Федерация.

Ссылку на эту статью просим оформлять следующим образом:
Лосев Н.С., Глинская Е.В. Безопасность и обзор угроз машинного обучения. *Политехнический молодежный журнал*, 2025, № 01 (96). URL: https://ptsj.bmstu.ru/catalog/icec/insec/1019.html

 ISSN 2541-8009

SECURITY AND THREATS OVERVIEW IN THE MACHINE LEARNING

N.S. Losev losevns@student.bmstu.ru
E.V. Glinskaya glinskaya@bmstu.ru
SPIN-code: 5430-3023

Bauman Moscow State Technical University, Moscow, Russian Federation

Machine learning is widespread in transforming various aspects of our lives with the intelligent digital solutions. The use of machine learning in the critical infrastructure entails the desire of intruders to gain access to it by obtaining the algorithms, methods, and data underlying the models in order to expand their influence and benefit from the system vulnerabilities. The paper considers and analyzes main vulnerabilities and classification of various machine learning attacks as an object of protection. It also presents efficient methods to prevent future attacks and technologies to mitigate the possible dangerous consequences.

Keywords: machine learning, countermeasure strategies, intruder attack data protection, protection methods, attack models, vulnerabilities, critical infrastructure

Received 12.10.2024

Losev N.S. — Student, Department of Information Security, Bauman Moscow State Technical University, Moscow, Russian Federation.

Glinskaya E.V. — Senior Lecturer, Department of Information Security, Bauman Moscow State Technical University, Moscow, Russian Federation.

Scientific advisor — Basarab M.A., Dr. Sci. (Phys. & Math.), Head of the Department of Information Security, Bauman Moscow State Technical University, Moscow, Russian Federation.

Please cite this article in English as:
Losev N.S., Glinskaya E.V. Security and threats overview in the machine learning. *Politekhnicheskiy molodezhnyy zhurnal*, 2025, no. 01 (96). (In Russ.). URL: https://ptsj.bmstu.ru/catalog/icec/insec/1019.html

