Как защитить машинное обучение в системах безопасности

Атаки на искусственный интеллект

Александр Чистяков, Алексей Андреев
«Лаборатория Касперского», Департамент исследования угроз

kaspersky

Введение

Согласно исследованию компании Cisco[1], доля руководителей служб безопасности, уверенных в успехе систем машинного обучения, сократилась с 77% опрошенных в 2018 году до 67% в 2019 году; число заинтересованных в решениях на основе искусственного интеллекта также уменьшилось с 74% до 66%.

В 2010-е годы искусственный интеллект (AI) и машинное обучение (ML) стали активно применяться в информационной безопасности. Предполагалось, что эти технологии дадут ответ на перегретые ожидания предыдущего бума Больших Данных. Теперь, когда индустрия научилась собирать огромное количество данных, следовало извлечь из них что-нибудь полезное. Производители систем безопасности «следующего поколения» (Next-Gen) использовали модный тренд по полной программе – они стали утверждать, что традиционный антивирус умер, и его с успехом заменят более продвинутые антивирусы на основе искусственного интеллекта (хотя практика сравнительных тестов не подтвердила этот тезис).

За прошедшие годы AI-лихорадка несколько поутихла. Согласно исследованию компании Cisco[1], доля руководителей служб безопасности, уверенных в успехе систем машинного обучения, сократилась с 77% опрошенных в 2018 году до 67% в 2019 году; число заинтересованных в решениях на основе искусственного интеллекта также уменьшилось с 74% до 66%. Даже аналитики Gartner, ранее восхищавшиеся системами безопасности «следующего поколения», в 2019 году стали более аккуратными в формулировках. «Искусственный интеллект не означает автоматически «лучший продукт» в современной безопасности», – пишут они.

Одна из главных причин охлаждения энтузиазма (хотя мы пока не говорим о «зиме искусственного интеллекта», а только о некоторых завышенных ожиданиях) состоит в том, что алгоритмы машинного обучения, выпущенные из стерильных лабораторий в реальный мир, могут давать большое количество ошибок и сбоев. Причём эти ошибки можно вызвать намеренно с помощью специальных атак, от которых большинство ML-систем защищены очень плохо или не защищены вовсе.

Детектор людей обманут специальной картинкой. Иллюстрация из исследования KU Leuven[2]

Такие выводы особенно тревожны для индустрии безопасности. Неудивительно, что за последние несколько лет наиболее популярной темой докладов на ИБ-конференциях вместо «искусственного интеллекта» стали «атаки на искусственный интеллект». Самые наглядные примеры обхода систем машинного обучения демонстрируются в сфере распознавания образов – дорожный знак с наклеенной точкой идентифицируется как совершенно другой знак, бумажные очки обманывают распознавание лиц, а яркая картинка на одежде приводит систему к выводу, что вы вообще не человек.

Конечно, системы идентификации не полагаются на одно только автоматическое распознавание лиц. Однако ошибки в этой сфере уже приводят к тому, что невинных людей называют преступниками. В ноябре 2018 года полицейская система Шанхая обвинила популярную китайскую предпринимательницу в нарушении правил перехода улицы – камера безопасности распознала её лицо в рекламном постере на борту автобуса. Примерно в то же время студент из Нью-Йорка подал в суд на компанию Apple за то, что система безопасности этой компании ошибочно идентифицировала его как вора. В мае нынешнего года власти Сан-Франциско запретили использование систем распознавания лиц в

[1] The CISO Benchmark Report 2019

[2] Simen Thys, Wiebe Van Ranst, Toon Goedeme. Fooling automated surveillance cameras: adversarial patches to attack person detection. - arXiv, 2019

полиции и других городских службах, чтобы предотвратить подобные ошибки и злоупотребления.

Современные системы обнаружения вредоносных программ используют алгоритмы машинного обучения, очень схожие с теми, что применяются в системах распознавания образов. При этом последствия атаки на ML-антивирус могут быть весьма серьёзными: не пойманный вовремя троян означает тысячи заражённых машин и убытки в миллионы долларов.

Однако подобных проблем можно избежать, если заранее позаботиться о безопасности систем машинного обучения. В этой статье мы расскажем о том, как организованы наиболее популярные атаки на ML-алгоритмы, и как защититься от таких атак.

Общие понятия

Машинное обучение (ML) – подмножество технологий искусственного интеллекта (AI), которое можно описать как «набор алгоритмов, позволяющих компьютерам обучаться тому, на что они изначально не запрограммированы». Иными словами, ML-алгоритм – это такая программа, которая сама может создавать программы для решения различных задач.

Алгоритм машинного обучения может обучаться на выборке уже решённых примеров задачи (это называется **обучение с учителем**), либо он должен самостоятельно выявить общие черты и связи в заданном множестве объектов **(обучение без учителя)**.

В случае обучения с учителем, ML-алгоритм может работать **в режиме обучения** либо в **боевом режиме**. В режиме обучения алгоритм получает **обучающую выборку**, в которой **объекты** представлены своими **признаками** и **метками классов**. В частности, если ML-алгоритм используется для выявления вредоносного ПО, объектом может быть файл, признаками – различные мета-данные файла и данные о его поведении (логи статистики, используемые API-функции и т.д.), а метки могут быть двух видов – «вредоносный» или «чистый».

На основе обучающей выборки, состоящей из известных вредоносов и известных чистых файлов, ML-алгоритм должен создать такую **прогностическую модель**, которая сможет корректно классифицировать (то есть размечать как вредоносные или чистые) другие объекты, ранее не известные системе (новые файлы). После обучения построенная ML-модель переходит в **боевой режим** и используется для выявления вредоносного ПО.

При обучении без учителя нас интересует выявление скрытых **паттернов** или **кластеров** – групп схожих объектов или связанных событий. В этом случае данные передаются в программу без предварительной разметки, и ML-алгоритм сам должен выявить **корреляции**. В индустрии кибербезопасности обучение без учителя часто применяется для **поведенческого анализа** и **выявления аномалий**.

Виды атак и защиты от них

Искажение разметки

Чтобы провести такую атаку на алгоритм машинного обучения, злоумышленник должен получить доступ к обучающей выборке и добавить в выборку объекты с неправильной разметкой. Обученная на искажённой выборке, ML-модель будет ошибаться и на других схожих объектах.

Но так ли легко получить доступ к обучающим выборкам? На самом деле, да. Многие вендоры обмениваются информацией об угрозах через специальные потоки данных об угрозах (threat intelligence feeds), и уже известны случаи, когда злоумышленники манипулировали такими механизмами. Например, агрегатор данных об угрозах VirusTotal атаковали специально созданными чистыми файлами с признаками вредоносов. После того, как один из антивирусных сканеров определяет такой файл как вредоносный, эту ошибочную классификацию начинают использовать другие системы безопасности, что вызывает цепную реакцию ложных срабатываний по всему миру – схожие чистые файлы детектируются как вредоносы.

Защита:
Двойная проверка и ансамбли методов

Все размеченные файлы, которые «Лаборатория Касперского» получает от третьих сторон, проходят дополнительную проверку классификации по нашим собственным базам. Ошибки классификации на основе искажённой разметки также можно сократить, используя ансамбли – совместную работу разных обучающих алгоритмов под контролем экспертов-людей.

Искажение обучающей выборки

Имея доступ к обучающей выборке, злоумышленник может добавить в неё специальные объекты, которые ухудшают качество работы прогностической модели. При этом разметка может быть корректной, однако сами объекты являются нестандартными. Например, это может быть файл, который по каким-то параметрам значительно отличается от типичных файлов для данной задачи («чёрный лебедь»).

Это угроза особенно серьёзна в связи с тем, что многие разработчики ML-решений, включая производителей систем безопасности «следующего поколения», используют публичные обучающие выборки, которые могут быть легко «отравлены» третьими сторонами.

Защита:
Берегите выборки, выявляйте странности

Если атакующий не знает, на каких примерах вы тренировали свою модель, ему будет трудно создать аномальный объект, выпадающий из нормального распределения. В «Лаборатории Касперского» мы сами собираем обучающие выборки, и они не являются публичными: лишь некоторые примеры проанализированных вредоносов экспортируются во внешние потоки данных об угрозах. А логи наших поведенческих ML-моделей основаны на специальном внутреннем протоколе, который недоступен для посторонних.

Мы также разработали систему оценки уровня доверия для прогноза на заданном объекте. Это позволяет нашей ML-системе поведенческого анализа отличать «странные» файлы и не использовать их для классификации, чтобы они не испортили модель.

Кроме того, хорошим средством от подобных атак является многоуровневая защита. Даже если вредоносу удастся обмануть статический анализ на основе машинного обучения, угрозу распознают другие технологии – например, динамический анализ в эмуляторе или в песочнице.

Атаки «белого ящика» и «черного ящика»

Злоумышленник, не имеющий доступа к обучающей выборке, всё равно может активно взаимодействовать с системой машинного обучения. В частности, если атакующий получил доступ к самой модели в локальном пользовательском продукте, он может скрытно исследовать модель, сколько пожелает, и даже воссоздать её исходный код. Так он узнает архитектуру модели и выявит, какие признаки объектов используются. После этого злоумышленник может создать вредоносное ПО, которое будет маскировать или усиливать нужные признаки. Это называют атакой «белого ящика» или «кражей модели».

Пример подобной атаки – обход искусственного интеллекта Cylance, обнаруженный в июле 2019 г. Восстановив исходный код системы безопасности, исследователи нашли название популярной игры внутри некоего списка исключений. Тогда они взяли несколько строк кода из главного исполняемого файла данной игры, добавили эти строки во вредоносный файл и обнаружили, что система безопасности перестала детектировать его как вредоносный.

Если исходный код недоступен, злоумышленник может атаковать ML-модель «грубым перебором», то есть делать небольшие изменения во вредоносных файлах и раз за разом тестировать модель на этих изменённых файлах – до тех пор, пока не будет обнаружено её слабое место. Подобная атака «чёрного ящика» является довольно трудоёмкой, однако атакующий может её автоматизировать, используя «враждебный» искусственный интеллект (adversarial AI) для генерации большого количества вредоносных образцов.

Защита:
Используйте машинное обучение в облаке

С облачной ML-моделью злоумышленник не сможет поиграть у себя дома. Так работает, например, наша система машинного обучения для детектирования угроз на Android-устройствах. Программа-клиент на мобильном устройстве пользователя собирает признаки нового приложения и посылает эти мета-данные в облачную ML-модель, обученную на миллионах вредоносов. Ответ облачной системы детектирования тут же пересылается обратно на мобильное устройство.

Теоретически, злоумышленник может атаковать грубым перебором и облачную модель. Однако такая атака «чёрного ящика» требует очень много времени и может быть легко отслежена.

Дискретные правила обнаружения на основе ML

Облачные и лабораторные ML-модели можно использовать для создания статичных детектирующих записей, которые добавляются в базы продукта при обновлении. Даже если атакующему удастся восстановить логику отдельной записи, он сможет обмануть только одну запись. Но он не сломает ML-модель, которая автоматически сгенерирует новые правила обнаружения с учётом новой угрозы.

Так, в случае нашей системы обнаружения на основе гибких свёрток (Similarity Hash Detection System) внутренняя система машинного обучения «Лаборатории Касперского» используется для выявления признаков целых групп вредоносных файлов. На основе этих признаков система генерирует гибкие свёртки, которые отправляются в базы локальных продуктов через облачную систему KSN. Продукт на конечном устройстве вычисляет гибкую свёртку для анализируемого файла, а затем сравнивает её с базой гибких свёрток, полученных через KSN. Такой подход позволяет продукту выявлять целые семейства полиморфных вредоносов – безо всякого риска атаки на ML-модель.

Доказуемо устойчивая ML-модель

Ещё один способ защитить машинное обучение от подобных атак – построить такую модель, которую нельзя сломать с помощью изменений во вредоносных образцах. Для нашей ML-системы поведенческого анализа мы разработали концепцию монотонных классификационных моделей, которая обеспечивает доказуемую стабильность прогнозов в случае попыток внедрения какого-либо шума или «чистой» активности в поведение программы. Прогнозы такой модели меняются монотонно: добавление новых строк в лог исполнения файла может только увеличить вероятность того, что файл будет детектироваться как вредоносный. Такие ML-модели можно использовать даже на конечных устройствах для распознавания вредоносов в режиме реального времени.

Атаки на предобученные и аутсорсинговые ML-модели

Недостаток ресурсов зачастую вынуждает разработчиков использовать чужие архитектуры машинного обучения, созданные сторонними разработчиками для решения стандартных задач обработки данных. Эти готовые и всем доступные решения могут быть хорошо знакомы злоумышленникам, что облегчает проведение атаки «белого ящика».

На одной из недавних конференций по безопасности представитель крупного производителя ПО гордо объявил: «Мы открыто публикуем все наши ML-модели на GitHub». На вопрос о риске атак выступающий ответил, что обычные хакеры вряд ли способны на такое, поскольку подобная атака требует серьёзных навыков. Это был очень смелый ответ, ведь буквально в соседнем зале той же конференции демонстрировались исследования по взлому систем машинного обучения.

Ещё один вектор атаки связан с аутсорсингом: некоторые ML-модели обучаются сторонними командами экспертов или публичными ML-сервисами. Такие модели могут содержать бэкдоры.

Защита:

Не полагайтесь на решения третьих сторон

В «Лаборатории Касперского» мы обучаем ML-модели самостоятельно на собственной аппаратной базе. Для такого «домашнего» подхода есть ещё одна важная причина: интерпретируемость. При создании и настройке системы машинного обучения разработчик имеет дело с тонкой архитектурой из тысяч узлов и весовых коэффициентов. Чтобы удостовериться в адекватности прогнозов, а также понимать причины возможных ошибок и вовремя обнаруживать атаки, нужно уметь интерпретировать результаты работы системы на основе машинного обучения. Поэтому мы регулярно занимаемся анализом интерпретируемости наших моделей. Провести подобный анализ ML-системы, разработанной на стороне, гораздо труднее или вообще невозможно.

Утечки через обученные модели

В некоторых ML-системах злоумышленник, отправляя в систему специально подобранные объекты, способен получить информацию о том, какие объекты использовались в обучающей выборке. Это может представлять угрозу, если объекты содержат конфиденциальную информацию (например, персональные медицинские данные) либо если сам факт использования объекта в обучении не должен разглашаться (например, преступник может обнаружить, что его фотография использовалась для обучения полицейской системы распознавания).

Защита:
Ограничивайте доступ, анонимизируйте данные

Для такой атаки взломщику обычно нужен полный доступ к ML-модели, чтобы многократно тестировать её, прежде чем он сможет добыть нужную информацию. Поэтому один из методов защиты – использовать многоуровневые облачные ML-модели вместо моделей в пользовательских продуктах. Другая хорошая идея – анонимизировать данные, которые применяются в обучении. Также можно использовать ML-модели, работающие с зашифрованными данными.

Атаки на уровне железа

Некоторые методы машинного обучения требуют значительного объёма вычислений, результаты которых могут отличаться на разных процессорах. К примеру, модель, обученная на мощном компьютере во внутренней инфраструктуре компании-разработчика, затем должна работать на пользовательском смартфоне. Злоумышленник может создать специальный файл, который будет неверно классифицироваться на некоторых моделях телефонов, что в свою очередь приведёт к неправильной разметке в других системах безопасности (аналогично атакам, искажающим обучающую выборку).

Защита:
Методы, независимые от локальной архитектуры

Данная угроза – ещё один повод отказаться от ML-моделей на конечных устройствах. Мы предпочитаем использовать облачные ML-модели либо дискретные детектирующие записи, созданные внутренними системами машинного обучения (см. выше).

Стоит также отметить, что некоторые методы машинного обучение (например, решающие деревья) меньше зависят от разницы в аппаратуре, чем другие методы (нейронные сети).

Заключение

Невзирая на явную перегретость бума машинного обучения и искусственного интеллекта в наши дни, эти технологии играют большую роль в современной информационной безопасности. «Лаборатория Касперского» начала применять машинное обучение задолго до прихода вендоров «нового поколения», и сейчас такие алгоритмы используются на многих стадиях нашего детектирующего конвейера. Это и методы кластеризации при обработке входящего потока файлов во внутренней инфраструктуре, и модели на основе глубокого обучения для облачных систем обнаружения, и созданные на основе машинного обучения детектирующие записи в базах продуктов на конечных устройствах.

Тем не менее, как показывают наши исследования, ML-алгоритмы могут быть уязвимы для разнообразных атак. Поэтому при использовании машинного обучения в безопасности рекомендуется применять следующие полезные принципы:

- Производитель решений в области безопасности должен понимать основные требования, предъявляемые к ML-системам в реальном, потенциально враждебном мире. К таким требованиям относятся: минимальный процент ложных срабатываний, интерпретируемость модели и устойчивость к действиям потенциального противника. Аудиты безопасности и пентесты систем машинного обучения (ML red-teaming) должны стать необходимой практикой при разработке ML-систем.
- Оценивая защищённость ML-решений, необходимо выяснить, насколько эти решения зависят от чужих данных и сторонних разработок (импорт данных об угрозах, публичные обучающие выборки, аутсорсинговые ML-модели), поскольку с этим связаны многие атаки на системы машинного обучения.
- Нельзя считать методы искусственного интеллекта «универсальным лекарством». Такие методы должны быть частью многоуровневых систем безопасности, где альтернативные технологии защиты и человеческая экспертиза работают вместе, помогая друг другу.

Дополнительная информация

- Kaspersky TechoWiki – хороший источник информации о продвинутых технологиях безопасности, включая искусственный интеллект, машинное обучение и защиту на основе анализа поведения.
- Securelist.ru предоставит вам самые свежие и подробные данные о современном вредоносном ПО, таргетированных атаках и других трендах киберкриминального мира.

Cyber Threats News: www.securelist.com
IT Security News: business.kaspersky.com
IT Security for SMB: kaspersky.com/business
IT Security for Enterprise: kaspersky.com/enterprise

www.kaspersky.com