Введение в состязательное машинное обучение

А.С. Казимиров

Атаки на нейронные сети

# введение в состязательное машинное обучение

## Задачи, в которых возникает состязательное машинное обучение

Растущее количество больших высококачественных наборов данных в сочетании со значительными техническими достижениями последних нескольких десятилетий превратили машинное обучение в основной инструмент, используемый в широком спектре задач, включая компьютерное зрение, обработку естественно языка, финансы и безопасность. Однако успех моделей машинного обучения сопровождался новыми важными вызовами: многие приложения машинного обучения носят состязательный характер. Некоторые из них являются состязательными, потому что они критически важны для безопасности, например, беспилотные автомобили. Противником в этих приложениях может быть злоумышленник, стремящийся вызвать сбои или аварии, или даже может моделировать необычные ситуации, которые выявляют уязвимости в прогнозе. Другие приложения являются состязательными, потому что их задача и/или данные, которые они используют, являются таковыми. Например, важный класс проблем в области безопасности включает в себя обнаружение вредоносного ПО, спама или вторжений. Использование машинного обучения для обнаружения вредоносных объектов побуждает противников уклоняться от обнаружения, изменив свое поведение или содержание вредоносных объектов, которые они создают.

Область состязательного машинного обучения возникла для изучения уязвимостей подходов машинного обучения в состязательных условиях и разработки методов, позволяющих сделать обучение устойчивым к состязательным манипуляциям.

По мере того, как методы машинного обучения становились массовыми в вычислительной технике, их использование увеличивалось. Интернет-реклама и алгоритмическая торговля в настоящее время немыслимы без машинного обучения, а методы машинного обучения все чаще находят свое применение в медицине, обнаружении мошенничества, компьютерном зрении, машинном переводе и понимании естественного языка. Однако наиболее важным является то, что методы машинного обучения находят все более широкое применение в области безопасности в целом и кибербезопасности в частности. Причина в том, что проблемы безопасности по определению являются состязательными. Есть защитники, например, сетевые администраторы, антивирусные компании, производители брандмауэров, пользователи компьютеров и т. п., пытающиеся сохранить производительность, несмотря на внешние угрозы, и злоумышленники, которые распространяют вредоносное ПО, рассылают спам и фишинговые письма, взламывают вычислительные устройства, крадут данные или выполнять DoS-атаки.

Естественной ролью методов машинного обучения в приложениях безопасности является детектирование, примерами которого являются обнаружение спама, вредоносных программ, вторжений и аномалий. В качестве типичного примера возьмем обнаружение вредоносной электронной почты (спама или фишинга). Мы можем начать с построения размеченного набора данных безопасных и вредоносных (например, спама) электронных писем, содержащего тексты электронных писем и любую другую соответствующую информацию (например, метаданные, такие как регистрационная информация DNS для IP-адреса отправителя). Давайте сосредоточимся на тексте электронной почты как на единственной информации о природе (вредоносной или полезной) электронного письма. Набор данных преобразуется в векторы признаков, которые фиксируют текстовое содержимое и числовые метки, соответствующие двум классам (вредоносные и безопасные). Распространенным способом численного представления документа является использование набора слов. В представлении «мешок слов» мы строим словарь слов, которые встречаются в электронных письмах, а затем создаем вектор признаков для данного электронного письма, учитывая, как часто каждое слово в словаре появлялось в тексте электронного письма. В более простом двоичном представлении набора слов каждый признак просто указывает, появилось ли соответствующее слово в тексте электронной почты; альтернативное вещественное представление учитывает количество раз, которые слово появляется в электронном письме, или частоту терминов (tf-idf). После того, как набор данных преобразован в числовой формат, мы обучаем классификатор предсказывать, является ли новое электронное письмо спамом или фишинговым на основе текста электронного письма в его представлении признаками мешка слов.

## Пример атаки на классификатор

Для достаточно большого набора данных использование современных моделей машинного обучения может обеспечить чрезвычайно эффективное обнаружение спама или фишинговой электронной почты в отношении оценки с использованием прошлых данных. Что делает этот параметр состязательным, так это то, что спам и фишинговые электронные письма генерируются намеренно, с конкретными целями, злоумышленниками. Указанные субъекты явно недовольны, если их электронные письма обнаруживаются и в результате не достигают предполагаемого места назначения (т. е. почтовых ящиков пользователей). Вариантов у спамеров два: либо прекратить заниматься рассылкой спама, либо изменить способ построения спам-шаблонов, чтобы избежать детекторов спама. Такое уклонение спамеров от враждебного классификатора является прототипом использования состязательного машинного обучения.

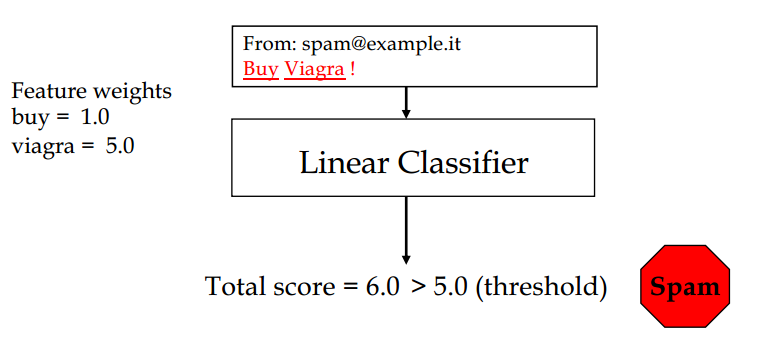


Рис. 1. Пример успешного обнаружения спама

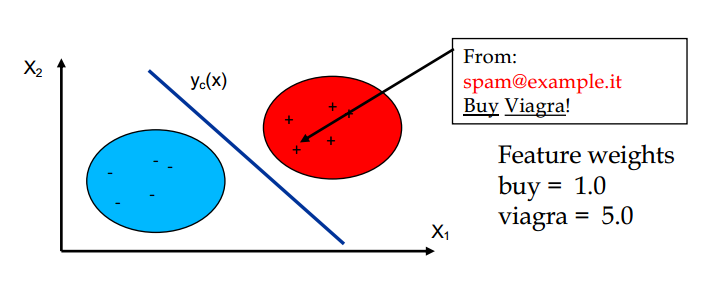


Рис. 2. Объект в пространстве признаков

Обычно это достигается путем удаления «спамных» слов (то есть слов, которые имеют тенденцию повышать показатель вредоносности детектора) и, возможно, добавления слов, которые детектор считает безопасными (последняя стратегия часто называется атакой с использованием хороших слов).

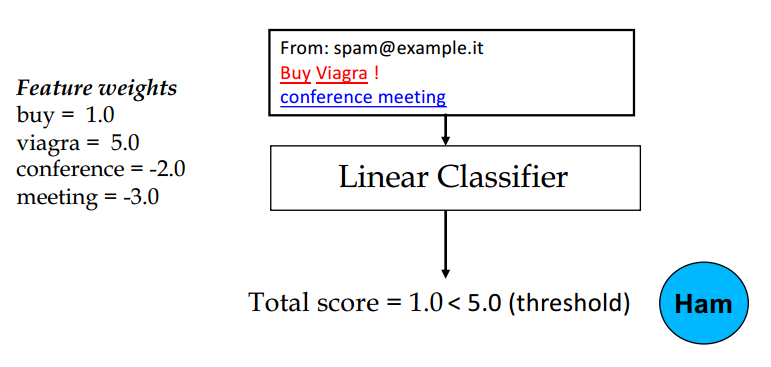


Рис. 3. Замаскированный спам.

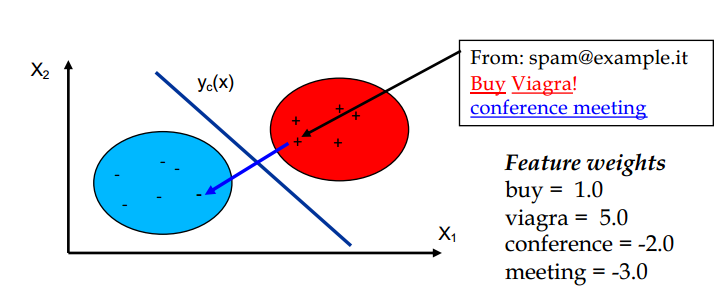


Рис. 4. Замаскированный спам в пространстве признаков.

## Состязательное машинное обучение

Эти две противоречащие друг другу цели — уклонение от обнаружения и одновременное достижение первоначальных целей атаки — являются центральными для атак уклонения в целом. При обнаружении вредоносного ПО, как и при обнаружении спама, злоумышленник может изменить код вредоносного ПО, чтобы оно выглядело безобидным для детектора, сохраняя при этом исходную или эквивалентную вредоносную функциональность. В системах обнаружения вторжений хакер хотел бы пересмотреть автоматизированные инструменты и ручные процедуры, чтобы они выглядели безобидными, но в то же время успешно выполняли эксплойт.

В то время как атаки с уклонением кажутся наиболее естественными при состязательном использовании машинного обучения, многочисленные другие примеры показывают, что масштаб проблемы состязательного машинного обучения гораздо шире. Например, использование размеченных обучающих данных с целью обучения детекторов подвергает алгоритмы обучения атакам с отравлением, когда злоумышленники манипулируют данными, которые впоследствии используются для обучения алгоритмов обучения. Анализ вредоносных программ или попыток взлома может потребовать использования кластеризации в попытке классифицировать природу атаки, но это может быть подвержено преднамеренным атакам, которые манипулируют назначением кластера, слегка изменяя характер атак, вызывая ошибочную категоризацию. Более того, состязательное обучение также имеет более широкий охват, чем приложения для кибербезопасности. В сфере физической безопасности, например, проблема обнаружения вредоносной активности с помощью видеонаблюдения носит состязательный характер: хитрые злоумышленники могут манипулировать своим внешним видом или другими факторами, такими как то, как они осуществляют свою злонамеренную деятельность, чтобы избежать обнаружения. Точно так же обнаружение мошенничества с кредитными картами, в котором используются методы обнаружения аномалий, чтобы определить, является ли конкретное действие подозрительным из-за его крайне неожиданного характера, может быть подвержено атакам, которые делают транзакции типичными для большинства пользователей кредитных карт. В качестве еще одного примера можно привести методы алгоритмической торговли, использующие машинное обучение, которые могут быть использованы конкурентами, которые совершают рыночные сделки с единственной целью манипулирования прогнозируемыми ценами и использования возникающих в результате арбитражных возможностей для получения прибыли.

Систематическое изучение состязательного машинного обучения направлено на формальное исследование проблем, связанных с использованием методов машинного обучения в состязательной среде, в которой интеллектуальный противник пытается использовать слабые места в таких методах. Два основных аспекта этого исследования: (1) моделирование и исследование атак на машинное обучение и (2) разработка методов обучения, устойчивых к манипулированию злоумышленниками.