Классификация атак

А.С. Казимиров

Атаки на нейронные сети

# Классификация атак

## Классификация атак по трем основным признакам

Ранее мы в общих чертах рассмотрели, как модели машинного обучения могут себя вести в состязательных условиях. Далее нам нужно не просто понять, как машинное обучение можно использовать в состязательных условиях (например, для обнаружения вредоносного ПО), но и каким образом такие параметры вносят уязвимости в традиционные подходы к обучению. Принципиальное обсуждение таких уязвимостей сосредоточено вокруг точных моделей угроз. Рассмотрим общую классификацию моделей угроз или атак в контексте машинного обучения. Было предпринято несколько попыток классифицировать атаки на алгоритмы машинного обучения. Предлагаемая ниже классификация связана с некоторыми из них и направлена ​​на выявление наиболее важных особенностей атак, которые мы обсуждаем. В частности, мы классифицируем атаки по трем параметрам: время, информация и цели.

1. Время. Первое, что необходимо учитывать при моделировании атак, — это когда происходит атака. Это соображение приводит к следующей общей дихотомии, которая является центральной для атак на машинное обучение: атаки на модели (из которых атаки уклонения — evasion attacks — являются наиболее типичными случаями) и атаки на алгоритмы (широко известные как атаки с отравлением — poisoning attacks). Атаки на модели или, точнее, на решения, принимаемые обученными моделями, предполагают, что модель уже изучена, и злоумышленник теперь либо меняет ее поведение, либо вносит изменения в наблюдаемую среду, чтобы заставить модель делать ошибочные прогнозы. Атаки отравления, напротив, происходят до обучения моделей, изменяя часть данных, используемых для обучения.

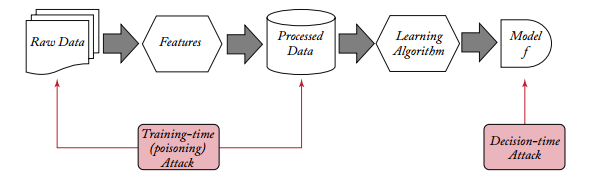


Рис. 1. Схематическое изображение различия между атаками времени принятия решения (атаками на модели) и атаками отравления (атаками на алгоритмы).

2. Информация. Вторым важным вопросом при моделировании атак является то, какая информация имеется у злоумышленника о модели обучения или алгоритме, различие, которое обычно сводится к атакам «белого ящика» (white-box attack) и «черного ящика» (black-box attack). В частности, атаки «белого ящика» предполагают, что либо модель (в случае атак на решения), либо алгоритм (в атаках с отравлением) полностью известны противнику, тогда как в атаках «черного ящика» злоумышленник имеет ограниченную информацию или не имеет никакой информации об этих атаках, хотя часть информации может получить косвенно, например, через запросы.

3. Цели. У злоумышленников могут быть разные причины для атаки, например уклонение от обнаружения или снижение доверия к алгоритму. Мы различаем два широких класса целей атаки: целевые атаки и атаки на надежность метода обучения (или просто атаки на надежность). При целенаправленной атаке цель злоумышленника состоит в том, чтобы вызвать ошибку в конкретных экземплярах определенного характера (например, заставить обученную функцию f предсказать конкретную ошибочную метку l для экземпляра x). Напротив, атака на надежность направлена ​​на снижение воспринимаемой надежности системы обучения за счет максимизации ошибки прогноза.

## Классификация по времени совершения атаки

Атаки во время принятия решения. Из всех классов атак, которые мы будем рассматривать, атаки с уклонением — основной подкласс атак, происходящих во время принятия решения, — пожалуй, исторически наиболее заметны. Известным примером уклонения является эволюция спам-трафика электронной почты, например, когда спамеры заменяют букву «о» на цифру «0» в слове «Лотерея» (которое становится «Л0терея»).

В общем, атака уклонения от классификатора на двоичные классификаторы принимает в качестве входных данных классификатор f(x) и «идеальный» экземпляр в пространстве признаков, xideal (т. е. это то, что противник хотел бы отправить, если бы не было классификатора для идентификации этих данных как вредоносных). Затем атака выводит другой экземпляр, соответствующий вектору признаков x\*. Если f(x\*)=-1, уклонение успешно, но возможно, что противнику не удастся найти адекватное уклонение (фактически, для любой значимой меры устойчивости алгоритма к уклонению необходимо, чтобы противник не мог найти уклонение, какими бы ни были f и xideal).

В качестве иллюстрации предположим, что кто-то хочет обнаружить спам в электронной почте и обучает для этой цели классификатор f (x) (где x — вектор, представляющий характеристики электронной почты). Теперь рассмотрим спамера, который ранее использовал шаблон, соответствующий вектору признаков xspam, и предположим, что f(xpam) помечает его как «спам» (+1), так что спамер не получает реакции на свои письма. Спамер будет вносить изменения в электронное письмо, чтобы получить экземпляр, который в функциональном пространстве выглядит как x\* со свойством f(x\*)=1 (т. е. он классифицируется как не спам и может передаваться в почтовые ящики пользователей). Но x\* не может быть произвольным: злоумышленник несет затраты на изменение исходного экземпляра xspam для достижения x\*, что может измерять стоимость усилий (для поддержания функциональности) или эффективность (например, ему, возможно, придется внести орфографические ошибки, которые позволят злоумышленнику избежать обнаружения, но также уменьшат вероятность того, что люди будут переходить по ссылкам в письме).

Обобщая идею атак уклонения, мы можем рассмотреть атаки времени принятия решения на многоклассовую классификацию. Пусть Y будет конечным набором меток, и предположим, что для некоторого экземпляра xideal предсказанная метка равна f(xideal)=y. Злоумышленник может захотеть изменить этот экземпляр на другой, x\*, либо для получения неверного предсказания (f(x\*) ≠ y), либо для того, чтобы классификатор предсказал целевую метку t=f(x\*). В последнее время такие атаки привлекли большое внимание под термином «состязательные примеры», в основном в приложениях машинного зрения и глубоких нейронных сетях. Потенциальная проблема заключается в том, что злоумышленник может попытаться вызвать аварию автономного транспортного средства, полагающегося на зрение, путем манипулирования воспринимаемым изображением дорожного знака, такого как знак остановки (например, путем размещения специально созданных наклеек, которые кажутся прохожим граффити).

Атаки на обучающие данные. Проблема обучения с поврежденными или зашумленными обучающими данными была предметом серьезных исследований в сообществах машинного обучения и статистики в течение нескольких десятилетий. Однако в последнее время злонамеренное повреждение обучающих данных стало рассматриваться более систематически, особенно если мы допускаем искажение значительной части данных. Природа отравляющих атак заключается в том, что злоумышленник преднамеренно манипулирует обучающими данными до обучения, чтобы алгоритм обучения делал неверный выбор. Важная концептуальная проблема с отравляющими атаками заключается в определении объема злоумышленником манипулирования обучающими данными и целей злоумышленника при этом. Один из распространенных способов обойти эти проблемы — предположить, что злоумышленник может вносить произвольные изменения в небольшое подмножество точек обучающих данных. Тогда целью будет разработка алгоритмов, устойчивых к такому произвольному повреждению обучающих данных, если количество поврежденных данных достаточно мало.

Можно также рассмотреть более конкретные модели порчи данных, которые накладывают дополнительные ограничения на действия злоумышленника. Одним из распространенных классов таких атак являются атаки с переворачиванием меток, когда злоумышленнику разрешено изменять метки не более чем C объектов обучающих данных. Как правило, такие атаки рассматриваются в контексте классификации, хотя можно также изменить метки регрессии. В большинстве случаев предполагается, что алгоритм и пространство признаков известны противнику (т. е. это white-box-атака). Отравление данных также можно рассматривать в условиях неконтролируемого обучения, например, когда оно используется для обнаружения аномалий. В этом случае злоумышленник может внести небольшие изменения в наблюдаемое нормальное поведение, которое теперь загрязняет модель, используемую для обнаружения аномалий, с целью гарантировать, что будущая атака цели будет помечена как безвредная.

## Классификация по информации, доступной злоумышленнику

Одним из наиболее важных факторов при моделировании атаки является информация, которой обладает злоумышленник о системе, которую он атакует. Мы проводим различие между атаками «белого ящика», когда злоумышленник знает все, что нужно знать, и атаками «черного ящика», когда злоумышленник имеет ограниченную информацию.

Атаки «белого ящика» предполагают, что противник точно знает либо изученную модель (например, фактический классификатор) в случае атак на время принятия решения, либо алгоритм обучения в случае атаки отравления. Это означает, например, что злоумышленнику известны все параметры модели, в том числе признаки, а в случае атаки с отравлением — гиперпараметры алгоритма обучения. Предположение о том, что у злоумышленника есть такая информация об обучающей системе, может показаться подозрительным. Тем не менее, есть важные причины, чтобы рассмотреть атаки белого ящика.

Во-первых, они предлагают естественную отправную точку с точки зрения модели: если модель устойчива к атакам белого ящика, она, безусловно, устойчива и к атакам, которые ограничены в информации.

Во-вторых, с точки зрения злоумышленника, может быть несколько способов косвенно получить достаточную информацию об изученной модели для развертывания успешной атаки. Возьмем, к примеру, атаку уклонения при детекции вредоносного ПО. Предположим, что набор используемых функций является общедоступной информацией (например, из опубликованных работ), а наборы данных, используемые для обучения детектора вредоносных программ, являются общедоступными (или, альтернативно, существуют общедоступные наборы данных, которые достаточно похожи на данные, фактически используемые для обучения). Наконец, предположим, что модель использует стандартный алгоритм обучения для изучения модели, такой как случайный лес, глубокая нейронная сеть или машина опорных векторов, и стандартные методы настройки гиперпараметров, такие как перекрестная проверка. В этом случае злоумышленник может получить идентичную или почти идентичную версию детектора, используемую в действительности!

При атаках методом «черного ящика», в отличие от атак «белого ящика», злоумышленник не имеет точной информации ни о модели, ни об алгоритме, используемом обучаемым. Важной задачей моделирования атак методом «черного ящика» является точное моделирование того, какая информация имеется у злоумышленника либо об изученной модели, либо об алгоритме.

В контексте атак методом «черного ящика» во время принятия решения один из подходов заключается в рассмотрении иерархии информации об изученной модели, доступной злоумышленнику. С точки зрения противоположного подхода, противнику вообще недоступна никакая информация. У более информированного противника могут быть некоторые данные для обучения, которые отличаются от данных, на которых была обучена реальная модель, но он не может иметь информации о конкретном классе модели или об используемых функциях. Более информированный злоумышленник может знать класс и функции модели и, возможно, алгоритм обучения, но не иметь обучающих данных, а еще более информированный противник может также иметь обучающие данные, отобранные из того же распределения, что и данные, используемые для обучения. Наконец, когда у этого же противника есть фактические обучающие данные, используемые алгоритмом обучения, результирующая атака эквивалентна атаке белого ящика, как обсуждалось выше, поскольку злоумышленник может узнать точную модель из заданных обучающих данных. Можно заметить, что, в отличие от атак «белого ящика», существует множество способов моделирования атак «черного ящика». Действительно, существует термин атака серого ящика, указывающий, что злоумышленник имеет некоторую, хотя и неполную, информацию об атакуемой системе.

Приведенная выше информационная иерархия не отвечает на естественный вопрос: откуда злоумышленник получает информацию о модели, которую он атакует (в случае атак времени принятия решения)? Важный класс моделей атаки «черный ящик» решает этот вопрос, позволяя злоумышленнику получить доступ к изученной модели с запросом. В частности, для произвольного объекта, представленного в виде вектора признаков x, злоумышленник может получить (запросить) фактическую предсказанную метку y=f(x) для неизвестной модели черного ящика f. Обычно и неявно такие модели запросов также предполагают, что злоумышленник знает как модельное пространство (например, алгоритм обучения), так и функциональное пространство. Кроме того, еще одним практическим ограничением этой модели является то, что f(x) часто наблюдается неточно или с шумом при заданном x. Например, предположим, что спамер отправляет спам-сообщение по электронной почте. Отсутствие ответа не обязательно означает, что оно было отфильтровано — вероятнее может быть так, что пользователи просто проигнорировали электронное письмо. Тем не менее, такая структура, основанная на запросах, позволяет провести элегантное теоретическое исследование того, что может сделать злоумышленник, имея очень ограниченную информацию об обучающемся.

Следуя тем же принципам, что и типичные модели атак методом «черного ящика» во время принятия решения, атаки «черного ящика» с отравлением данных позволят получить ряд знаний об алгоритме, используемом защитником. Например, в одном крайнем случае злоумышленник может вообще не иметь информации об алгоритме. Более информированный злоумышленник может знать алгоритм, но не знать гиперпараметры (такие как вес регуляризации или количество скрытых слоев в нейронной сети) или функции. Более информированный злоумышленник может знать алгоритм, функции и гиперпараметры, но не обучающие данные, которые злоумышленник пытается отравить.

## Классификация по цели атакующего

Хотя у злоумышленников может быть множество возможных целей для совершения атаки на системы машинного обучения, мы делим атаки на две основные категории с точки зрения целей злоумышленника: целевые атаки и атаки на надежность.

Целевые атаки характеризуются конкретной целью злоумышленника в отношении модельных решений. Например, рассмотрим атаку времени принятия решения на мультиклассовый классификатор с набором возможных меток L, и пусть x будет конкретным экземпляром, представляющим интерес для злоумышленника с истинной меткой y. Целью направленной атаки в этом случае будет изменение метки для x на конкретную целевую метку t≠y . В более общем смысле целевая атака характеризуется подмножеством S пространства объектов и меток X×Y, для которых злоумышленник хотел бы изменить решение, а также целевой функцией принятия решения D(x ). В наиболее распространенных условиях целевых атак на обучение с учителем злоумышленник будет стремиться получить предсказания для каждого (x, y) из S, чтобы они соответствовали целевой функции метки l(x).

Атаки на надежность, с другой стороны, пытаются максимизировать ошибки в решениях, сделанных путем обучения относительно истины. Например, при обучении с учителем злоумышленник будет стремиться максимизировать ошибку предсказания. В приложениях машинного зрения такие атаки, которые стали обычно называть нецелевыми, будут изменять изображение таким образом, чтобы вызвать ошибочное предсказание (например, распознавание объекта, отсутствующего на изображении, например, ошибочное принятие изображения знака остановки за любой другой дорожный знак).

Различие между целенаправленными атаками и атаками на надежность стирается, когда мы рассматриваем бинарную классификацию: в частности, атаки на надежность теперь становятся частным случаем, в котором метки цели l(x) являются просто альтернативными метками. В более общем плане мы отмечаем, что даже разделение между целенаправленными атаками и атаками на надежность является неполным: например, можно рассматривать атаки, целью которых является избежание предсказаний определенного класса, отличного от правильной метки.