Лекция 12
Тема: Применение машинного обучения

Простыми словами: как работает машинное обучение

В последнее время все технологические компании твердят о машинном обучении, но как оно работает, никто не рассказывает. А мы расскажем — максимально простыми словами.

В последнее время все технологические компании твердят о машинном обучении. Мол, столько задач оно решает, которые раньше только люди и могли решить. Но как конкретно оно работает, никто не рассказывает. А кто-то даже для красного словца машинное обучение называет искусственным интеллектом.

Как обычно, никакой магии тут нет, все одни технологии. А раз технологии, то несложно все это объяснить человеческим языком, чем мы сейчас и займемся. Задачу мы будем решать самую настоящую. И алгоритм будем описывать настоящий, подпадающий под определение машинного обучения. Сложность этого алгоритма игрушечная — а вот выводы он позволяет сделать самые настоящие.

Задача: отличить осмысленный текст от белиберды

Текст, который пишут настоящие люди, выглядит так:

Могу творить, могу и натворить! У меня два недостатка: плохая память и что-то еще. Никто не знает столько, сколько не знаю я.

Белиберда выглядит так:

ОРПорыав аоырОрпаыор ОрОРАыдцуцзущгкгеуб ыватьыивдцулвдлоадузцщ Йцхяь длваополц ыадолцлопиолым бамдлотдламда.

Наша задача — разработать алгоритм машинного обучения, который бы отличал одно от другого. А поскольку мы говорим об этом применительно к антивирусной тематике, то будем называть осмысленный текст «чистым», а белиберду — «зловредной». Это не просто какой-то мысленный эксперимент, похожая задача на самом деле решается при анализе реальных файлов в реальном антивирусе.

Для человека задача кажется тривиальной, ведь сразу видно, где чистое, а где зловредное, но вот формализовать разницу или тем более объяснить ее компьютеру — уже сложнее. Мы используем машинное обучение: сначала дадим алгоритму примеры, он на них «обучится», а потом будет сам правильно отвечать, где что.

Алгоритм

Наш алгоритм будет считать, как часто в нормальном тексте одна конкретная буква следует за другой конкретной буквой. И так для каждой пары букв. Например, для первой чистой фразы — «Могу творить, могу и натворить!» — распределение получится такое:

ат — 1
во — 2
гу — 2
ит — 2
мо — 2
на — 1
ог — 2
ор — 2
ри — 2
тв — 2
ть — 2

Что получилось: за буквой **«в»** следует буква **«о»** — два раза, а за буквой **«а»** следует буква **«т»** — один раз. Для простоты мы не учитываем знаки препинания и пробелы.

На этом этапе мы понимаем, что для обучения нашей модели одной фразы мало: и сочетаний недостаточное количество, и разница между частотой появления разных сочетаний не так велика. Поэтому надо взять какой-то существенно больший объем данных.

Например, давайте посчитаем, какие сочетания букв встречаются в первом томе «Войны и мира»:

то — 8411
ст — 6591
на — 6236
го — 5639
ал — 5637
ра — 5273
не — 5199
по — 5174
ен — 4211
оу — 31
мб — 2
тж — 1

Разумеется, это не вся таблица сочетаний, а лишь ее малая часть. Оказывается, вероятность встретить «**то**» в два раза выше, чем «**ен**«. А чтобы за буквой **«т»** следовала **«ж»** — такое встречается лишь один раз, в слове «отжившим».

Отлично, «модель» русского языка у нас теперь есть, как же ее использовать? Чтобы определить, насколько вероятно исследуемая нами строка чистая или зловредная, посчитаем ее «правдоподобность». Мы будем брать каждую пару букв из этой строки, определять по «модели» ее частоту (по сути, реалистичность сочетания букв) и перемножать эти числа:

F(**мо**) * F(**ог**) * F(**гу**) * F(**тв**) * ... = 2131 * 2943 * 474 * 1344 * ... = правдоподобность

Также в финальном значении правдоподобности следует учесть количество символов в исследуемой строке — ведь чем она была длиннее, тем больше чисел мы перемножили. Поэтому из произведения извлечем корень нужной степени (длина строки минус один).

Использование модели

Теперь мы можем делать выводы: чем больше полученное число — тем правдоподобнее исследуемая строка ложится в нашу модель. Стало быть, тем больше вероятность, что ее писал человек, то есть она *чистая*.

Если же исследуемая строка содержит подозрительно большое количество крайне редких сочетаний букв (например, **«ёё»**, **«тж»**, **«ъь»** и так далее), то, скорее всего, она искусственная — *зловредная*.

Для строчек выше правдоподобность получилась следующая:

Могу творить, могу и натворить! — 1805 баллов
У меня два недостатка: плохая память и что-то еще. — 1535 баллов
Никто не знает столько, сколько не знаю я. — 2274 балла
ОРПорыав аоырОрпаыор ОрОРАыдцуцзущгкгеуб ыватьыивдцулвдлоадузцщ — 44 балла
Йцхяь длваополц ыадолцлопиолым бамдлотдламда — 149 баллов

Как видите, *чистые* строки правдоподобны на 1000–2000 баллов, а *зловредные* недотягивают и до 150. То есть все работает так, как задумано.

Чтобы не гадать, что такое «много», а что — «мало», лучше доверить определение порогового значения самой машине (пусть обучается). Для этого скормим ей некоторое количество чистых строк и посчитаем их правдоподобность, а потом скормим немного зловредных строк — и тоже посчитаем. И вычислим некоторое значение посередине, которое будет лучше всего отделять одни от других. В нашем случае получится что-то в районе 500.

В реальной жизни

Давайте осмыслим, что же у нас получилось.

1. Мы выделили признаки чистых строк, а именно пары символов.

В реальной жизни — при разработке настоящего антивируса — тоже выделяют признаки из файлов или других объектов. И это, кстати, самый важный шаг: от уровня экспертизы и опыта исследователей напрямую зависит качество выделяемых признаков. Понять, что же на самом деле важно, — это все еще задача человека. Например, кто сказал, что надо использовать пары символов, а не тройки? Такие гипотезы как раз и проверяют в антивирусной лаборатории. Отмечу, что у нас для отбора наилучших и взаимодополняющих признаков тоже используется машинное обучение.

2. На основании выделенных признаков мы построили математическую модель и обучили ее на примерах.

Само собой, в реальной жизни мы используем модели чуть посложнее. Сейчас наилучшие результаты показывает ансамбль решающих деревьев, построенный методом Gradient boosting, но стремление к совершенству не позволяет нам успокоиться.

3. На основе математической модели мы посчитали рейтинг «правдоподобности».

В реальной жизни мы обычно считаем противоположный рейтинг — рейтинг вредоносности. Разница, казалось бы, несущественная, но угадайте, насколько неправдоподобной для нашей математической модели покажется строка на другом языке или с другим алфавитом. Антивирус не имеет права допустить ложное срабатывание на целом классе файлов только по той причине, что «мы его не проходили».

Альтернатива машинному обучению

20 лет назад, когда вредоносов было мало, каждую «белиберду» можно было просто задетектить с помощью сигнатур — характерных отрывков. Для примеров выше «сигнатуры» могли бы быть такими:

ОРПорыав аоырОрпаыор ОрОРАыдцуцз**зущгкгеу**б ыватьыивдцулвдлоадузцщ

Йцхяь длваополц ыадолцлопиолым бам**длотдламд**а

Антивирус сканирует файл, если встретил «**зущгкгеу**«, говорит: «Ну понятно, это белиберда номер 17». А если найдет «**длотдламд**» — то «белиберда номер 139».

15 лет назад, когда вредоносов стало много, преобладать стало «дженерик»-детектирование. Вирусный аналитик пишет правила, что для осмысленных строк характерно:

1. Длина слов от 1 до 20 символов.
2. Заглавные буквы очень редко встречаются посередине слова, цифры тоже.
3. Гласные обычно более-менее равномерно перемежаются с согласными.
4. И так далее. Если много критериев нарушено — детектируем эту строку как зловредную.

По существу, это примерно то же самое, только вместо математической модели в этом случае набор правил, которые аналитик должен вручную написать. Это хорошо работает, но требует времени.

И вот 10 лет назад, когда вредоносов стало ну просто очень много, начали робко внедряться алгоритмы машинного обучения. Поначалу по сложности они были сопоставимы с описанным нами простейшим примером, но мы активно нанимали специалистов и наращивали экспертизу. Как итог — у нас лучший уровень детектирования.

Сейчас без машинного обучения не работает ни один нормальный антивирус. Если оценивать вклад в защиту пользователей, то с методами на основе машинного обучения по статическим признакам могут посоперничать разве что методы на основе анализа

поведения. Но только при анализе поведения тоже используется машинное обучение. В общем, без него уже никуда.

Недостатки

Преимущества понятны, но неужели это серебряная пуля, спросите вы. Не совсем. Этот метод хорошо справляется, если описанный выше алгоритм будет работать в облаке или в инфраструктуре, постоянно обучаясь на огромных количествах как *чистых*, так и *вредоносных* объектов.

Также очень хорошо, если за результатами обучения присматривает команда экспертов, вмешивающихся в тех случаях, когда без опытного человека не обойтись.

В этом случае недостатков действительно немного, а по большому счету только один — нужна эта дорогостоящая инфраструктура и не менее дорогостоящая команда специалистов.

Другое дело, когда кто-то пытается радикально сэкономить и использовать только математическую модель и только на стороне продукта, прямо у клиента. Тогда могут начаться трудности.

1. Ложные срабатывания.

Детектирование на базе машинного обучения — это всегда поиск баланса между уровнем детектирования и уровнем ложных срабатываний. И если нам захочется детектировать побольше, то ложные срабатывания будут. В случае машинного обучения они будут возникать в непредсказуемых и зачастую труднообъяснимых местах. Например, эта чистая строка — «Мцыри и Мкртчян» — распознается как неправдоподобная: 145 баллов в модели из нашего примера. Поэтому очень важно, чтобы антивирусная лаборатория имела обширную коллекцию чистых файлов для обучения и тестирования модели.

2. Обход модели.

Злоумышленник может разобрать такой продукт и посмотреть, как работает модель. Он человек и пока если не умнее, то хотя бы креативнее машины — поэтому он подстроится. Например, следующая строка считается чистой (1200 баллов), хотя ее первая половина явно вредоносная: «лоыралоыврачигшуралорыловарДобавляем в конец много осмысленного текста, чтобы обмануть машину». Какой бы умный алгоритм ни использовался, его всегда может обойти человек (достаточно умный). Поэтому антивирусная лаборатория обязана иметь продвинутую инфраструктуру для быстрой реакции на новые угрозы.

Один из примеров обхода описанного нами выше метода: все слова выглядят правдоподобно, но на самом деле это бессмыслица.

3. Обновление модели.

На примере описанного выше алгоритма мы упоминали, что модель, обученная на русских текстах, будет непригодна для анализа текстов с другим алфавитом. А вредоносные файлы, с учетом креативности злоумышленников (смотри предыдущий пункт), — это как будто постепенно эволюционирующий алфавит. Ландшафт угроз меняется довольно быстро. Мы за долгие годы исследований выработали оптимальный подход к постепенному

обновлению модели прямо в антивирусных базах. Это позволяет дообучать и даже полностью переобучать модель «без отрыва от производства».

Заключение

Итак.

1. Мы рассмотрели реальную задачу.
2. Разработали реальный алгоритм машинного обучения для ее решения.
3. Провели параллели с антивирусной индустрией.
4. Рассмотрели с примерами достоинства и недостатки такого подхода.

Несмотря на огромную важность машинного обучения в сфере кибербезопасности лучшую в мире киберзащиту обеспечивает именно многоуровневый подход.

Машинное обучение в IT

Машинное обучение в IT — отрасль компьютерных технологий, в которой разработчики занимаются созданием интеллектуальных машинных процессов. Идея того, что компьютер может анализировать и систематизировать введенные значения с минимальным участием человека нашла яркое отражение во всех сферах человеческой деятельности.

Грубо говоря, компьютерная программа, на основе опыта, может делать определенные выводы или действия. Классическим примером машинного обучения или как его еще называют искусственным интеллектом, можно назвать поисковую систему Google. Она сама на основании имеющихся знаний, выдает пользователю ответ на тот, или иной запрос. Также машинное обучение Поисковой системы влияет и на успешность **ПРОДВИЖЕНИЯ САЙТА.** Ведь от принятия решений поисковика, зависит ранжирование веб-сайтов.

Существует множество методов машинного обучения, которые основываются на изучении данных. Основной способ — это использование глубокого обучения, которое основано на нейронных сетях. Мы решили обсудить тематику, потому что это может напрямую влиять на нашу индустрию **РАЗРАБОТКИ САЙТОВ**.

Понятие нейронных сетей в машинном обучении

Нейронные сети неразрывно связаны с машинным обучением в IT. Эти два понятия существуют много лет и собрали вокруг себя много полезной литературы и обучающих программ. Нейронные сети используются только в сочетании с другими методами, такими как классификация или регрессия. Их используют для обучения на примере данных, однако, есть технологии, которые могут быть описаны только с помощью другого алгоритма. В качестве такового можно использовать глубокое машинное обучение.

Глубокое обучение имеет несколько преимуществ перед традиционным машинным обучением. Оно основано на нейронных сетях, которые являются общей моделью программирования и легко обучаются и настраиваются для различных проблемных областей.

Под глубоким обучением понимают метод создания искусственных нейронных сетей, которые эффективно обучаются на огромных массивах данных. Большим преимуществом является то, что глубокие нейронные сети можно обучать в течение длительного времени, подавая определенное количество информации, что помогает сети сформировать правильные параметры и алгоритм.

Где применяется машинное обучение?

Machine Learning может быть использоваться для моделирования сложных ситуаций, где нужно делать прогнозы по многим вещам одновременно. Например, для видеоигры, в которой много вещей происходит одномоментно, и нужно предсказать исход того или иного сюжетного хода.

Инновационная технология широко используется в сфере data science для решения маркетинговых задач. Так, гигант Amazon использует искусственный интеллект для предложения пользователям наиболее востребованного товара. Для этого программа

применяет накопленный ранее пользовательский опыт. Обработка релевантных запросов в поисковиках также происходит с использованием ИИ.

Здесь важно убедиться, что данные, которые входят в созданную модель, максимально хорошо сформированы. Когда начинаете создавать модель, нужно учесть все параметры, так считают многие программисты. Машинное обучение основано на искусственном интеллекте, но учить программу обрабатывать информацию нужным образом должен именно человек.

Алгоритм машинного обучения строится на множестве методов и подходов, к ним можно отнести и метод опорных векторов. Если говорить кратко, то метод опорных векторов (согласно источнику: habr) - это стандартный линейный алгоритм, который используется в таких задачах как классификация и регрессия. Часто, метод опорных векторов называют называют SVM (англ. support vector machines).

Наивный байесовский классификатор - метод обработки больших данных, строящийся классификаторе с применением теоремы Байеса, также является одним из наиболее популярных направлений изучения алгоритмов машинного обучения и больших данных. Обучение с учителем, согласно Википедии, также является одним из основных способов машинного обучения.

Интернет вещей и технология M2M

Следующей важной отраслью где используется машинное обучение это технология M2M или как его еще называют “интернет вещей”. Данная технология решения задач, позволяет взаимодействовать разным системам между собой без, либо с минимальным присутствием человека. Например, система “умный дом”. Данная технология, также активно используется на сложных производствах, например, в нефтяной сфере. Есть специальные датчики, которые срабатывают только в момент неисправности оборудования и передают сигнал на станцию, где уже принимается решение об исправлении поломки.

Данная технология открывает огромные возможности для будущего человечества. Например, в недалеком будущем компании по производству продуктов могут быть полностью автоматизированы, экономя ценные человеческие ресурсы, используя уже наработанные и генерируемые алгоритмы. Мы убеждены, что связь систем между собой - это важнейший шаг для развития всех сфер деятельности человека.