26 ноября 2020

Задачи и инструменты ML и их практическое применение

Машинное обучение – распространившийся термин, но не все понимают его верно. В этом материале эксперты направления аналитических решений ГК «КОРУС Консалтинг» Алена Гайбатова и Екатерина Степанова расскажут, что же на самом деле такое machine learning (ML), в каких случаях эту технологию стоит использовать в проектах, а также где машинное обучение активно применяется на практике.

Как работают с данными

Уже давно на встречах с заказчиками мы стали замечать, что все путают машинное обучение, искусственный интеллект (ИИ), большие данные и другие термины из этой области.

Итак, общее название технологии – искусственный интеллект. Он бывает двух типов – сильный (он же общий) и слабый. Мы не будем особенно обсуждать сильный ИИ, так как это решения уровня Терминатора. Мы к нему потихонечку приближаемся, но до сих пор он существует только в виде собранных вместе фрагментов слабого ИИ (как, например, в «умных» колонках).

Намного интереснее поговорить о слабом искусственном интеллекте. Он тоже делится на два типа. Первый – экспертные системы, алгоритмы,

запрограммированные вручную (например, запрограммированный группой лингвистом алгоритм перевода слов из одного языка в другой).

Второй – так называемые data-driven системы, которые извлекают логику работы из каких-то исторических данных. У этого типа есть много терминов-синонимов, которые возникали с течением времени:

- модные в 90-е и нулевые data mining и knowledge discovery from database (KDD),
- data science, вошедший в обиход ближе к 2010-м,
- big data популярная ныне. Единственное исключение, точнее дополнение, которое привносит именно этот термин – наличие огромного количества сложноструктурированных данных.

Для разных задач – разные алгоритмы

В соответствии с двумя типами слабого ИИ выводы из данных мы можем сделать вручную (при экспертных системах) и с помощью машинного обучения. Оно же в свою очередь подразделяется на два типа: классический ML и deep learning (с использованием глубоких нейронных сетей с большим количеством слоев).

В проектах на базе ML используются модели. Прежде всего, прогнозные, которые отвечают на базовые вопросы: к какой группе относится объект, какое числовое значение у него будет и так далее. В зависимости от того, на какой вопрос мы отвечаем, это могут быть модель классификатора или регрессии.

Классификаторы

Классификатор – это процесс, позволяющий сказать, к какой группе будет относиться тот или иной объект. Например, у кошек есть разные характеристики: длина хвоста, цвет шерсти, масса тела и другие параметры. По ним мы можем определить, к какой породе относится кошка. Если мы решаем эту задачу с помощью алгоритма, то этот алгоритм будет называться классификатором.

Алгоритм, часто применяемый для классификации – дерево принятия решений. Если мы хотим построить дерево условий для распределения котов по породам, на моменте обучения алгоритм строит дерево условий, задавая первый вопрос.

Рыжая ли у кота шерсть? Да: мы относим его сразу к классу персидских котов, все персидские коты оказываются в одной ветке. Нет: у нас возникает следующее условие – весит ли кот меньше 3 кг. Дерево условий создается в момент обучения алгоритма, а все новые элементы проходят по нему и оказываться в той или иной группе.

Этот алгоритм удобен с точки зрения бизнес-интерпретации результатов, так как мы не всегда можем сразу определить, по каким свойствам у нас разделились группы.

Регрессоры

Регрессор – это алгоритм, который не относит предмет исследования к определенному классу, а присваивает ему определенное число. Пример

– алгоритм кредитного скоринга: у нас есть возраст заемщика, трудовой стаж, зарплата – и требуется рассчитать, через какое время клиент сможет выплатить кредит.

Самый простой такой алгоритм – линейная регрессия. Представим себе, что наши объекты - это точки на плоскости. Наша задача – сделать так, чтобы прямая, которая будет проходить на плоскости, лежала как можно ближе ко всем точкам. Тем самым мы зададим линейные коэффициенты между входными данными и выходным значением. Подобный алгоритм прост и не требует особых затрат. Им удобно пользоваться, если у нас много признаков и мало объектов.

Кластеризация

Кластеризация отвечает на вопросы о том, как разбить исследуемые объекты на группы и чем объекты внутри одной группы похожи.

Самый популярный алгоритм кластеризации – метод ближайших соседей. Снова к кошкам. Мы хотим разбить наших зверей на 4 группы. Наши объекты – снова точки на плоскости. Мы выбираем случайным образом центры наших групп, затем смотрим расстояние от центра группы до точек, ближайших к этому центру группы. После мы смещаем центры таким образом, чтобы расстояние до точек своей группы оказывалось меньше, чем до точек другой группы. Через нескольких итераций у нас получатся хорошо разделенные группы.

Сложность этого алгоритма заключается в том, что объекты не всегда хорошо делятся на группы – в связи с этим трудно оценить корректность результата

даже с помощью специальной оценки.

Нейронные сети

Первая нейронная сеть появилась еще в 1950-х гг. Сейчас при помощи нейронных сетей можно ответить на любой вопрос, но лишь с одной оговоркой: ответ не всегда можно интерпретировать.

При работе с нейросетью на вход подается большой объем данных в виде числовых значений, у каждого из которых есть определенный вес. Мы суммируем эти значения и к этой сумме применяем операцию активации, после этого получаем некий прогноз. Так как нейросети используют большое количество скрытых слоев, операции активаций и сумм может быть много. В связи с тем, что этим алгоритмом можно обрабатывать большие объемы данных, модель хорошо работает с текстом, изображением и звуком.

Дополнительно в проектах ML используются оптимизационные методы для минимизации ошибок. В условиях ограничений они стараются найти лучшее решение задачи с помощью нахождения экстремумов функции и применения статистических методов.

Обучение с подкреплением

Это и есть тот самый сильный искусственный интеллект, о котором мы уже говорили выше. К примеру, по этому принципу работают беспилотные автомобили.

Система состоит из агента и среды. Для агента задано конечное число операций (на примере машины – максимальная скорость, торможение, поворот направо или налево и так далее). После совершения действия агент получает либо вознаграждение, если его действие приводит к правильному выполнению задачи, либо наказание, если действие, наоборот, отдаляет его от выполнения задания.

Мы также пользуемся алгоритмами Uplift, нейролингвистического программирования и рекомендательными моделями. Uplift позволяет понять, нужно ли коммуницировать с объектом, НЛП использует алгоритмы для анализа текста (к примеру, на этом принципе работает функция подсказки слов в смартфоне), а рекмодели могут быть персонализированными и не персонализированными.

Теория – на практике

Посмотрим, как эти модели используются на для решения реальных задач. Мы сформулировали предпосылки для использования ML в проектах. Безусловно, они не гарантируют стопроцентного успеха, но на старте могут значительно снизить риски.

- Экономический эффект, который может принести оптимизация бизнес-процесса в несколько процентов;
- Регулярный технический или бизнесовый процесс, при оптимизации которого регулярное принятие решений на среднем уровне и/или действия по заданному алгоритму могут значительно улучшиться;

- Наличие данных, при которых может быть осуществлена оптимизация, за счет их анализа и обработки.

Одна из самых успешных отраслей в плане применения машинного обучения – это розничная торговля. Связано это с тем, что в ней много регулярных процессов

Например, категорийные менеджеры ежедневно занимаются управлением ассортиментом, промоакциями, ценообразованием, прогнозированием спроса, управлением логистикой. Оптимизация на доли процентов даже одного такого бизнес-процесса в масштабе торговой сети приобретает существенный эффект.

Задачи, которые решает ML в ритейле, включают в себя предсказание оттока клиентов, анализ продуктовых корзин, прогнозирование товаров в следующем чеке, распознавание ценников и товаров, прогноз закупок и спроса, оптимизация закупок и логистики, планирование промо, цен и ассортимента – или это лишь малая часть.

Ритейл не испытывает недостатка как в наличия разных данных, так и в их глубине истории. У ритейлеров есть история продаж, статистика поведения клиентов, история промоакций, исторический ассортимент, параметры товаров и магазинов, изображения ценников и товаров, история доставок и поступления товаров и многое другое. Оцифровка всего этого, чаще всего, не требуется.

Похуже с данными в сфере промышленности – хотя и там они есть. Это и исторические данные с датчиков о производительности, поломках, работе

бригад, данные по расходу и поставкам сырья, отгрузкам и доставкам. Для производств каждый процент простоя – это существенные потери, поэтому именно способы его сокращения, как и сокращение запасов, становятся основными задачами для оптимизации. Поэтому в числе главных задач для ML здесь – предсказание поломок оборудования, маркировка похожих поломок, выявление закономерностей поломок, выявление факторов на снижения производительности, оптимизация расхода сырья в производстве, оптимизация заказов и времени поставок сырья, прогноз скорости доставки.

Еще две отрасли, в которых распространены проекты на базе искусственного интеллекта, это банки и телекоммуникации. Это и управление клиентскими рисками (кредитный скоринг), и оптимизация регулярных рассылок клиентам. Задачи, стоящие в этих проектах, разношерстны – от предсказания оттока клиентов до маркировки клиентов, от кросс-сейл кредитов и депозитов до предсказания крупных транзакций.

Среди данных, которыми обладают подобные компании, статистика по поведению клиентов, их реакция на прошлую коммуникацию, история получения и возвратов кредитов, анкеты клиентов, параметры сотрудников, история эффективности работы персонала и другое.

Количество примеров проектов, реализуемых на базе машинного обучения, множество, и успешные кейсы будут появляться все чаще. Но главное усвоить базовые знания о том, что в действительности используют специалисты по машинному обучению, и заранее просчитать, будет ли от вашего будущего ML-проекта бизнес-эффект.

В настоящее время крупные компании вкладывают большие средства в машинное обучение, потому что данная технология не только окупается, но и помогает найти новые подходы к реализации рутинных задач. Действительно, ИИ занимает все более значимое место на рынке, но это не значит, что машины нас заменят. Мы успешно расширяем наши способности за счет машин, именно для этого и существует машинное обучение.

Источник: Habr