Оглавление

Введение

Добро пожаловать в путешествие по динамичному миру современного машинного обучения (Machine Learning, сокращенно ML)! Эта книга укажет путь от классической роли дата-сайентиста, выросшей из бизнес-аналитики, к современным многокомпонентным системам, которые находятся на переднем крае технологий. Примеры кода из этой книги можно найти на GitHub по адресу `https://github.com/machine-learning-upgrade`.

Мы писали эту книгу так, чтобы ее можно было прочитать за раз. Это не справочник методов и не энциклопедия по машинному обучению. Наша цель — осветить проблемы, связанные с современным машинным обучением, и мы уделили особое внимание управлению версиями данных, отслеживанию экспериментов, постпродакшен-мониторингу моделей и развертыванию, чтобы предоставить код и примеры, которые позволят вам сразу же начать применять лучшие практики.

Глава 1 закладывает основы: в ней объясняется, как рабочий процесс управления машинным обучением эволюционировал от традиционных, более линейных фреймворков для науки о данных, таких как CRISP-DM, до появления приложений на основе больших языковых моделей (Large Language Model, LLM). В этой главе особый акцент делается на необходимости единого фреймворка, который открывает замечательные возможности для создания приложения на основе LLM.

Глава 2 познакомит вас с комплексным подходом к машинному обучению: в ней вы узнаете о его жизненном цикле, принципах работы продакшен системы машинного обучения и основе нашего приложения LLM.

Глава 3 подробно рассматривает дата-центричный подход, акцентируясь на роли данных в современном ML. В этой практической главе мы создадим эмбеддинги и задействуем векторные базы данных для поиска по текстовому сходству. Мы сочетаем этические принципы и стратегии управления версиями данных, чтобы обеспечить ответственный и комплексный подход.

Затем следует Глава 4, в которой мы объясняем, как правильно выбрать LLM, использовать LangChain и оптимизировать производительность LLM.

После изложения первых четырех глав мы переходим к главе 5, чтобы собрать наши компоненты и перейти от прототипа к приложению. Мы также покажем, как создать дашборды и API), чтобы предоставить вывод модели конечным пользователям.

Но это еще не все. Глава 6 завершает жизненный цикл машинного обучения, рассматривая вопросы мониторинга моделей, повторной тренировки конвейеров, а также перспективы будущих стратегий внедрения и коммуникации с заинтересованными сторонами.

Наконец, в Главе 7 мы подводим итоги лучших практик, с которыми вы познакомились в ходе этого путешествия, рассказываем о новейших тенденциях развития моделей LLM и предоставляем ресурсы для дальнейшего обучения.

Эта книга — больше, чем просто руководство. Это увлекательное путешествие по ландшафтам современного машинного обучения, которое вооружит вас инструментами и знаниями, необходимыми для навигации по этим ландшафтам. Так что пристегните ремни, друзья, и поехали!

Благодарности

Написание этой книги было совместным путешествием, вдохновленным общими взглядами и потрясающей поддержкой тех, кто сделал это путешествие возможным.

Мы безмерно благодарны редакции издательства Wiley, в особенности Джеймсу Минателю и Гасу Миклосу, чья преданность делу и профессионализм превратили нашу рукопись в книгу. Мы ценим ваше стремление к совершенству.

Наша глубокая благодарность техническому редактору Харприту Сахоте, который предоставил нам ценные замечания и убедил нас доработать наши идеи и улучшить рукопись. Его идеи и рекомендации сыграли решающую роль в подготовке окончательного варианта книги!

Мы заранее выражаем искреннюю благодарность читателям, которых заинтересовал результат нашей работы. Мы надеемся, что эта книга даст вам ценные идеи и вдохновит вас на новые исследования.

Каждому, кто прямо или косвенно участвовал в совместной работе, спасибо за то, что вы были частью этого путешествия.

С благодарностью,
Кристен Керер
Калеб Кайзер

Информация об авторах

Кристен Керер с 2010 года занимается разработкой и доработкой моделей машинного обучения для компаний электронной коммерции, организаций здравоохранения и поставщиков коммунальных услуг. В 2018 году она вошла в рейтинг в области науки о данных и аналитики с 95 тысячами подписчиков в сфере науки о данных. Кристен является создателем Data Moves Me. Ранее она была преподавателем и экспертом по предмету в Emeritus Institute of Management. Кристен получила степень магистра прикладной статистики в Worcester Polytechnic Institute и степень бакалавра математики.

Калеб Кайзер — фуллстек-разработчик из Comet. Ранее Калеб входил в команду основателей Cortex Labs. Он также работал в Scribe Media в команде авторской платформы и получил степень бакалавра изящных искусств в области писательского мастерства в Школе художественного института Чикаго.

Информация о техническом редакторе

Харприт Сахота называет себя хакером в области генеративного искусственного интеллекта. Он получил степень бакалавра и магистра в области статистики и математики, и с 2013 года работает в «мире данных» в качестве актуария (*прим. ред.*: актуарий — специалист по страховым и финансовым расчетам), биостатистика, дата-сайентиста и инженера по машинному обучению с опытом в области статистики, машинного обучения, MLOps, LLMOps и генеративного искусственного интеллекта (с фокусом на мультимодальной генерации с дополнением извлеченной информацией). Он любит экспериментировать с новыми технологиями и проводить время со своей женой Роми и детьми Джугаадом и Джиндом. Его книга «Practical Retrieval Augmented Generation» вышла в издательстве Wiley в 2025 году.

ГЛАВА 1

Введение в современное машинное обучение

В течение последних 20 лет наука о данных в основном занималась использованием данных при принятии решений в бизнесе. Типичные проекты, связанные с наукой о данных, были сосредоточены на сборе, очистке и моделировании данных или создании панели инструментов, и в конечном итоге подготавливали презентацию для обмена результатами с заинтересованными сторонами. Этот конвейер стал основой многих важных решений в бизнесе, помог бизнесу увеличить доходы, используя множество панелей инструментов.

Традиционно бизнес-аналитикой (Business Intelligence, сокращенно — BI) называли проекты, в которых для принятия обоснованных решений использовался описательный анализ. В теории BI — это одно из применений науки о данных. Технически наука о данных относится ко всей практике применения статистических методов (включая моделирование), программированию и знаниям в конкретной области данных, тогда как применение бизнес-аналитики ограничено основанным на данных подходом к принятию бизнес-решений, больше ориентированным на описательную и диагностическую аналитику, а не на прогнозирующую аналитику, которой занимаются дата-сайентисты. Однако мы считаем, что все аналитики и специалисты по BI работают в области «наук о данных».

Если вы в последнее десятилетие работали аналитиком или дата-сайентистом, то скорее всего вы в основном занимались бизнес-аналитикой.

Не все специалисты согласятся с этим предположением, указав, что бизнес-аналитика, как мы ее традиционно понимаем, относится к сфере деятельности аналитиков по BI, в то время как у дата-сайентистов, как правило, более разнообразные и ориентированные на исследования задачи. Хотя в некоторой степени такие возражения справедливы, но распределение

обязанностей и функций разных должностей в типичном аналитическом агентстве не позволяет провести четкую границу между наукой о данных и BI, и при работе с данными всегда будет использоваться как наука о данных, так и BI.

Например, в качестве аналитика в типичной компании вы, вероятно, будете готовить ответы на вопросы «Что произошло?», используя описательную аналитику для составления обзора предыдущих результатов. Вы можете использовать Excel, SQL и программное обеспечение визуализации для создания отчетов и панелей инструментов. Вы, вероятно, будете отслеживать ключевые показатели эффективности (KPI) и помогать принимать стратегические решения на основе исторических данных. Также есть вероятность, что BI или бизнес-аналитика KPI, источники данных и модели машинного обучения (если таковые имеются), используемые в этом процессе, будут настроены для вас как для аналитика еще до начала проекта, и вы будете ими управлять.

Теперь, когда у компании появляются совершенно новые данные для обработки, они часто становятся доступными для нетехнических пользователей через еще одну панель инструментов (именно здесь часто возникают бесконечные споры о преимуществах Tableau над Power BI).

В целом это полезный способ разграничить роли дата-сайентистов и специалистов по BI в большинстве компаний. Дата-сайентист обычно отвечает за технические наукоемкие проекты в отделе аналитики: изучение новых источников данных, внедрение прогнозирующей аналитики, проведение тестов для проверки гипотез, исследование новых моделей машинного обучения и т. д. Однако бóльшая часть их работы по-прежнему сосредоточена на том, что мы можем обобщенно определить как бизнес-аналитику, или ее обслуживание.

По крайней мере, так было раньше.

Наука о данных выделяется из бизнес-аналитики

Следует подчеркнуть, что работа в области науки о данных часто ориентирована на исследования, особенно когда речь идет о машинном обучении. Когда в продвинутой аналитике, которой сейчас занимаются многие дата-сайентисты, вы создаете модель для машинного обучения, то скорей всего она будет использоваться только для исследований. Еще несколько лет назад казалось, что дата-сайентисты никогда не будут работать с моделями,

которые будут использоваться на практике или внедрены в продукт. Можно создать модель сегментации клиентов в зависимости от их поведения или построить прогнозируемую модель, чтобы лучше понять своих клиентов. Новая информация о клиентах может привести к появлению новых функций или изменений в продукте в результате дальнейшей проверки гипотез. Тем не менее, сама модель обычно больше не используется после того, как результаты исследования представлены руководству компании.

Это происходит не потому, что дата-сайентисты не хотят обучать модели, которые имеют огромное влияние на бизнес, а потому, что исторически это было действительно сложно. В 2022 году аналитическое агентство Gartner опубликовало результаты опроса компаний, использующих машинное обучение (ML) в США, Германии и Великобритании, который показал, что только 54 % моделей, разработанных их дата-сайентистами, были внедрены в продакшен системы, и это после многих лет их разработки в экосистеме ML!

Основная часть пользы для бизнеса от деятельности дата-сайентистов связана с их скромной работой с системами бизнес-аналитики, а польза от их участия в громких проектах построения моделей обычно ограничивается упоминанием в личном резюме. Но наконец такая ситуация начинает меняться!

Моделирование само по себе стало более полезным для практической деятельности, и модели все чаще используются в продуктах. Новые инструменты, такие как AutoML, и усовершенствования существующих фреймворков ML значительно упростили обучение полезных моделей практически на любом типе данных. Трансферное обучение, популяризованное компьютерным зрением и еще более стимулированное взрывным ростом больших языковых моделей, во многих отношениях еще больше упростило обучение эффективных моделей. Даже развертывание стало более доступным, поскольку экосистема инфраструктуры ML в последние годы быстро развивается.

В результате дата-сайентисты все чаще занимаются моделированием сценариев использования, которые приносят пользу бизнесу, и это отвлекает их от традиционной работы в области бизнес-аналитики. Все чаще дата-сайентисты отвечают за построение и мониторинг моделей машинного обучения. Однако в этом дивном новом мире у дата-сайентистов возникают совершенно новые задачи и обязанности, и это является основной мотивацией написания данной книги — помочь вам перейти от мира науки о данных, ориентированного на BI, к новому миру, в котором дата-сайентисты разрабатывают многокомпонентные системы машинного обучения для продакшена.

Многие из принципов, рассматриваемых в этой книге, можно в целом отнести к MLOps (или LLMOps), но стоит пояснить, что наша цель не состоит в том, чтобы вы стали инженерами MLOps, поскольку управление инструментами и инфраструктурой не входит в задачи дата-сайентистов. Напротив, мы надеемся, что, понимая принципы MLOps, дата-сайентисты смогут создавать надежные, масштабируемые и воспроизводимые модели, которые найдут применение на практике.

Прежде чем углубляться в эти принципы, следует немного изучить изменения машинного обучения, которые способствовали этому переходу.

От CRISP-DM к современности. Многокомпонентные системы машинного обучения

Машинное обучение прошло долгий путь с момента своего появления. В 1990-х годах был введен межотраслевой стандартный процесс для интеллектуального анализа данных (CRoss-Industry Standard Process for Data Mining, CRISP-DM) для описания типичных этапов проекта моделирования данных. Долгое время он был основным фреймворком для управления рабочим процессом машинного обучения. CRISP-DM сыграл важную роль в продвижении идеи о том, что наука о данных не должна быть ситуативной, а должна стать организованным процессом. Фреймворк CRISP-DM включал шесть ключевых этапов:

1. **Понимание бизнеса** — определение целей и требований проекта с точки зрения бизнеса.
2. **Понимание данных** — сбор и изучение данных для понимания их качества и структуры.
3. **Подготовка данных** — очистка, преобразование и систематизация данных для их использования в машинном обучении.
4. **Моделирование** — построение и оценка моделей машинного обучения для решения задач бизнеса.
5. **Оценка** — оценка эффективности моделей в достижении целей бизнеса.
6. **Внедрение** — интеграция модели в продакшен.

Эта четкая линейная последовательность идеально подходит для некоторых из описанных ранее проектов по обработке данных. В этом контексте

модель является лишь частью конвейера, созданного для генерации определенного отчета.

Однако современные системы машинного обучения гораздо больше похожи на продукты, чем на конвейеры. Они состоят из множества взаимосвязанных компонентов и применяются для решения гораздо более обширного спектра задач. На базовом техническом уровне они ставят несколько сложных задач:

- **объем и разнообразие данных:** с появлением больших данных проекты машинного обучения часто связаны с огромными наборами данных различных типов, включая текст, изображения и структурированные данные. Для обработки этих различных источников данных необходимы новые подходы;
- **усложнение алгоритмов:** алгоритмы машинного обучения стали более сложными, особенно с появлением глубокого обучения, обучения с подкреплением и т. д. Для реализации и обучения этих алгоритмов требуются специализированные инструменты и фреймворки;
- **масштабируемость:** создание отчетов раз в месяц — это одно, но выполнение вычислений в режиме реального времени по запросам тысяч одновременно работающих пользователей — это более сложная задача;
- **совместная работа:** машинное обучение часто предполагает участие команды дата-сайентистов, инженеров по данным и экспертов в конкретных областях. Инструменты и платформы для совместной работы стали незаменимыми для управления такими межфункциональными командами;
- **этические соображения и контроль:** хотя это и не является строго технической проблемой, растущее влияние машинного обучения на общество сделало этические соображения и практики контролирования более приоритетными. Обеспечение справедливости, прозрачности и соответствия требованиям стало неотъемлемой частью систем машинного обучения.

К счастью, некоторые очень талантливые дата-сайентисты и инженеры уже много лет работают над этими проблемами, и в современной экосистеме машинного обучения есть много инструментов, которые значительно облегчают решение этих задач. У нас есть решения для управления версиями данных, отслеживания экспериментов и организации совместной работы разных команд, а также реестры моделей и инструменты для мониторинга моделей в продакшене. «Озера» и хранилища данных позволяют нам

эффективно управлять, хранить и запрашивать большие объемы данных. Фреймворки машинного обучения значительно упростили моделирование. Существуют даже библиотеки с открытым исходным кодом, обеспечивающие этичное и ответственное использование моделей машинного обучения.

В этой книге мы рассмотрим все эти инструменты и построим на их основе реальные проекты. Но прежде чем углубиться в тему, давайте уделим немного времени обсуждению, пожалуй, самого значительного изменения в экосистеме машинного обучения за последнее десятилетие: появлению больших языковых моделей.

Появление моделей LLM увеличило мощность и сложность машинного обучения

Теперь мы рассмотрим одно из самых революционных событий в области машинного обучения: появление больших языковых моделей (LLM), которые значительно увеличили как мощность, так и сложность систем машинного обучения. Такие модели, как GPT-3, BERT и их последователи, переопределили границы возможного в области понимания и генерации естественного языка.

Эти модели, отличающиеся огромным размером и предобученными весами, показали свою универсальность и способность справляться с самыми разными задачами. Например, модели LLM обеспечили беспрецедентную производительность таких функций понимания естественного языка (Natural Language Understanding, NLU), как анализ тональности, суммаризации текста, перевод на другой язык и ответы на вопросы. Они коренным образом изменили ландшафт обработки естественного языка (Natural Language Processing, NLP). Модели LLM могут генерировать связный, учитывающий контекст текст в различных стилях и областях и являются движущей силой распространения инструментов генерации контента, чатботов и написания текстов с помощью ИИ.

> Предварительно обученная модель искусственного интеллекта, как правило, тренируется на большом наборе данных для выполнения определенной задачи. Данные могут быть любого типа, в том числе изображения, текст, аудио, таблицы и т. д., в зависимости от конкретного использования. Как правило, это передовые модели глубокого обучения.

Кроме того, предварительно обученные модели LLM стали мощными инструментами для трансферного обучения. Благодаря тонкой настройке данных для конкретной области, эти модели можно адаптировать к различным приложениям, что снижает потребность в интенсивной маркировке данных для каждого нового случая использования. Их можно даже применять к задачам, выходящим за рамки текста, таким как изображения и аудио. Такая конвергенция модальностей открывает новые возможности для сложных мультимодальных приложений, таких как создание подписей к изображениям, визуальные вопросы-ответы и многое другое.

Однако расширение возможностей приводит к увеличению сложности. Рассмотрим, например, следующие типичные задачи LLM:

- **Понимание языка.** Модели LLM могут понимать нюансы языка и контекста, что позволяет создавать более сложные и контекстно-зависимые системы искусственного интеллекта, такие как чат-боты. Однако чат-боты требуют частых генераций (т. е. запуска перегенерации на обученной модели искусственного интеллекта для прогнозирования или решения задачи). Как мы можем обработать запросы 1000 одновременных пользователей, каждому из которых требуется результат каждые несколько секунд, с помощью модели с 17 миллиардами параметров?
- **Извлечение знаний.** Модели LLM могут извлекать структурированные знания из неструктурированного текста, поэтому широко применяются для интеллектуального анализа данных, поиска информации и контроля контента. Но как мы получаем и храним эти знания? Что мы делаем для того, чтобы наша система могла даже динамически находить их?
- **Генерация контента.** LLM могут генерировать высококреативный и соответствующий контексту контент, например, стихи, код и даже целые статьи. Эта сложность распространяется на создание произведений искусства, музыки и литературы с помощью ИИ. Как генерировать контент, обеспечивая при этом соблюдение авторского права? Как защититься от расизма или других экстремистских идеологий? А как быть с дезинформацией?
- **Мультимодальный ИИ**. интеграция LLM с другими моделями глубокого обучения, такими как сверхточные нейронные сети (Convolutional Neuron Networks, CNN), для выполнения задач, например обработки изображений, приводит к созданию сложных и мощных