«Искусственный интеллект – это новое электричество».

– *Эндрю Ын* (Andrew Ng), сооснователь Coursera, профессор и внештатный преподаватель Стэнфордского университета

Содержание

От издательства

Отзывы и пожелания

Мы всегда рады отзывам наших читателей. Расскажите нам, что вы думаете об этой книге – что понравилось или, может быть, не понравилось. Отзывы важны для нас, чтобы выпускать книги, которые будут для вас максимально полезны.

Вы можете написать отзыв на нашем сайте www.dmkpress.com, зайдя на страницу книги и оставив комментарий в разделе «Отзывы и рецензии». Также можно послать письмо главному редактору по адресу dmkpress@gmail.com; при этом укажите название книги в теме письма.

Если вы являетесь экспертом в какой-либо области и заинтересованы в написании новой книги, заполните форму на нашем сайте по адресу http://dmkpress.com/authors/publish_book/ или напишите в издательство по адресу dmkpress@gmail.com.

Список опечаток

Хотя мы приняли все возможные меры для того, чтобы обеспечить высокое качество наших текстов, ошибки все равно случаются. Если вы найдете ошибку в одной из наших книг, мы будем очень благодарны, если вы сообщите о ней главному редактору по адресу dmkpress@gmail.com. Сделав это, вы избавите других читателей от недопонимания и поможете нам улучшить последующие издания этой книги.

Нарушение авторских прав

Пиратство в интернете по-прежнему остается насущной проблемой. Издательство «ДМК Пресс» очень серьезно относится к вопросам защиты авторских прав и лицензирования. Если вы столкнетесь в интернете с незаконной публикацией какой-либо из наших книг, пожалуйста, пришлите нам ссылку на интернет-ресурс, чтобы мы могли применить санкции.

Ссылку на подозрительные материалы можно прислать по адресу электронной почты dmkpress@gmail.com.

Мы высоко ценим любую помощь по защите наших авторов, благодаря которой мы можем предоставлять вам качественные материалы.

Кто мы?

Вы держите в руках книгу, ставшую плодом совместных усилий разработчиков из компании Cuantum Technologies. Мы – команда людей, преданных идее создания программного обеспечения, наделенного творческим началом и помогающего решать реальные задачи. Мы стремимся создавать высококачественные и дружественные веб-приложения, отвечающие всем требованиям наших заказчиков.

Мы в нашей команде верим, что программирование не ограничивается исключительно написанием кода. Задача программирования состоит в решении насущных проблем и облегчении жизни людей. С завидным постоянством исследуем новые технологии и науки, чтобы оставаться в авангарде индустрии, и с удовольствием делимся накопленными знаниями с вами, наши читатели.

Наш подход к созданию приложений базируется на совместной работе и творческом начале. Мы работаем в тесном контакте с заказчиками, чтобы досконально понять их нужды и создать продукт, отвечающий всем их требованиям. Верим в то, что выходящие из-под пера программиста приложения должны быть интуитивно понятными, легкими в обращении и привлекательными визуально, и делаем все от нас зависящее для воплощения этой веры.

В этой книге мы постарались изложить практические рекомендации по подготовке данных для их интеллектуального анализа. Делаете ли вы свои первые шаги в программировании или уже имеете определенный опыт в написании приложений, эта книга поможет вам освоить приемы работы с данными, которые позволят вам в будущем погрузиться в мир машинного и глубокого обучения.

Наша философия

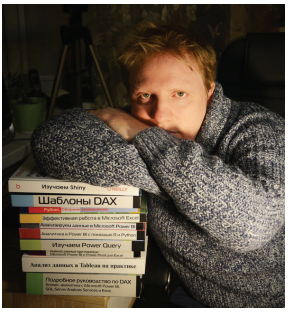
В Cuantum Technologies мы свято верим, что учиться программировать и развивать свои навыки в написании приложений можно на протяжении всей жизни. В связи с этим мы всячески поощряем стремление наших разработчиков осваивать новые технологии и техники и обеспечиваем их всеми необходимыми инструментами, для того чтобы не терять позиций в быстро развивающейся отрасли. Кроме того, мы полагаем, что программирование должно приносить радость и удовлетворение, и стараемся создавать рабочее окружение, поощряющее творческий подход и внедрение инноваций.

Наш опыт

Компания Cuantum Technologies специализируется на создании веб-приложений, помогающих решать заказчикам их насущные проблемы творчески и эффективно. Наши разработчики обладают богатым опытом написания приложений на разных языках программирования и использования различных фреймворков и технологий, включая Python, AI, ChatGPT, Django, React, Three.js, Vue.js и пр. Мы постоянно осваиваем новые технологии и стараемся реализовывать на практике все пожелания клиентов.

Также наши разработчики имеют все необходимые навыки в области анализа данных, визуализации, машинного обучения и искусственного интеллекта. Мы убеждены, что именно за этими технологиями будущее индустрии, и делаем все возможное, чтобы оставаться на переднем крае этой революции.

О переводчике



Александр Гинько, обладающий богатым опытом работы в сфере ИТ и более десяти лет посвятивший переводам книг и статей на самые разные темы, в последние годы специализируется на переводе книг в области бизнес-аналитики и программирования для издательства «ДМК Пресс» по направлениям Python, SQL, Power BI, DAX, Excel, Power Query, Tableau, R... На данный момент в активе Александра уже более 25 книг, включая одну авторскую, и он продолжает плодотворно работать над переводом и написанием новых книг.

Возможно, вам также будут интересны книги «Сверхбыстрый Python» (https://dmkpress.com/catalog/computer/programming/python/978-5-93700-226-6), «Python: практическое руководство по Pandas (200 упражнений)» (https://dmkpress.com/catalog/computer/programming/python/978-5-93700-227-3) и «Введение в статистическое обучение с примерами на Python» (https://dmkpress.com/catalog/computer/statistics/978-5-93700-217-4) в переводе Александра.

Помимо перевода книг, Александр ведет свой канал в Telegram (https://t.me/alexanderginko_books), на котором вы можете из первых уст получить ответы на все интересующие вас вопросы об уже переведенных книгах, находящихся в работе и запланированных на будущее. Также на канале можно найти промокоды на все книги Александра для покупки книг на сайте издательства «ДМК Пресс» с большими скидками. Купить книги Александра и следить за переводом новых книг в режиме реального времени можно также с помощью его бота в Telegram по адресу https://t.me/alexanderginko_books_bot.

Введение

Данные – это один из наиболее ценных активов в современном цифровом мире, и они лежат в основе всего: от принятия бизнес-решений до осуществления технического прогресса. В то же время сырые необработанные данные зачастую содержатся в разрозненном, беспорядочном и неструктурированном виде. Истинная ценность данных кроется в их сути, добраться до которой можно только путем их преобразований. Но для выполнения этих преобразований недостаточно просто знать нужные алгоритмы. Нужно хорошо понимать приемы и способы эффективной очистки, подготовки и манипулирования данными. В этой книге мы поговорим о **ключевых концепциях и техниках**, связанных с анализом данных, а также с конструированием и отбором признаков, лежащих в основе машинного и глубокого обучения.

Цель этой книги – научить вас подготавливать и преобразовывать сырые данные, конструировать новые признаки и в целом придавать исходным данным форму, пригодную для будущего интеллектуального анализа при помощи методов машинного и глубокого обучения. Работаете ли вы с небольшими объемами данных или оперируете сложными наборами данных высокой размерности, эта книга познакомит вас с эффективными техниками и приемами предварительной обработки и подготовки данных к дальнейшему анализу. По большей части мы будем пользоваться богатыми средствами и инструментами наиболее популярных библиотек Python для работы с данными, таких как Pandas, NumPy и Scikit-learn.

Почему так важно подготавливать данные и заниматься построением признаков?

В машинном обучении часто можно услышать фразу *«Данные – наше всё»*. Хотя выбор модели и тщательная настройка используемых алгоритмов играют важную роль, качество исходных данных оказывает гораздо большее влияние на эффективность итоговой модели. Этап подготовки данных и конструирования признаков зачастую занимает наиболее продолжительное время в проекте, и на то есть веские причины. Искусно подготовленные данные позволяют модели выявлять значимые шаблоны и делать точные предсказания, а также улучшают ее обобщающую способность при применении к новым данным.

Конструирование признаков (feature engineering) относится к области создания новых информативных признаков на основе сырых данных и играет важнейшую роль в процессе интеллектуального анализа данных. Именно *признакам* (feature) – преобразованным, комбинированным или созданным

на основе существующих данных – модели обязаны своим высоким предсказательным потенциалом. В процессе чтения книги вы увидите, что хорошо сконструированные признаки способны определять зависимости и шаблоны, которые невозможно выявить на основе одних только исходных данных. Такие тщательно отобранные признаки лежат в основе моделей, характеризующихся высоким качеством предсказаний, устойчивостью и интерпретируемостью результатов.

Инструменты, которые мы будем использовать

В последнее время практическим стандартом для специалистов по работе с данными стал язык программирования Python, и в этой книге мы будем в основном использовать три наиболее популярные библиотеки этого языка для преобразования данных: Pandas, NumPy и Scikit-learn.

Pandas – мощная библиотека для анализа и манипулирования данными. Pandas представляет собой интуитивно понятный фреймворк для управления исходными данными, представленными в виде строк и столбцов. Он особенно полезен в задачах первичной обработки данных, таких как очистка, фильтрация, агрегация и объединение наборов данных. Pandas поможет вам значительно облегчить процедуру обработки сырых данных и извлечения из них важных сведений.

NumPy – высокоэффективная библиотека для быстрой работы с массивами и применения математических операций к большим наборам данных. Преимущества NumPy в работе с данными делают эту библиотеку незаменимой при выполнении ресурсоемких операций с большими массивами информации, таких как масштабирование, нормализация и математические преобразования.

Scikit-learn – будучи одной из наиболее популярных библиотек в области машинного обучения, Scikit-learn не ограничивается одними лишь инструментами для построения моделей, но также предлагает богатый арсенал средств для манипулирования данными. В модулях фреймворка, связанных с предварительной обработкой данных, присутствуют средства для кодирования категориальных переменных, масштабирования непрерывных переменных, создания конвейеров обработки данных и многого другого. Библиотека Scikit-learn играет ключевую роль на этапе предварительной обработки данных, обеспечивая вашему анализу согласованность и воспроизводство.

В совокупности перечисленные библиотеки дают вам полный набор средств для управления процессом подготовки данных и конструирования признаков, включая очистку и преобразование данных, а также отбор и кодирование переменных.

Чему вы научитесь

Главы этой книги разбиты на три части, каждая из которых посвящена отдельному этапу подготовки данных и конструирования признаков.

Часть I. Подготовка к расширенному анализу данных. В первой части книги мы познакомимся с базовыми принципами промежуточного анализа данных и закроем пробелы в знаниях, касающихся обработки данных средствами Python. Мы научимся подходить к обработке данных системно и обеспечивать полную пригодность и структурированность данных перед их дальнейшим анализом. В процессе мы пройдемся по основным возможностям библиотек Pandas и NumPy и научимся выполнять требуемые операции над данными максимально эффективно. Таким образом, в этой части мы заложим основы для знакомства с более сложными техниками обработки данных, о которых будем говорить далее.

Часть II. Конструирование и отбор признаков для повышения качества моделей. Во второй части книги мы с головой погрузимся в область конструирования и отбора признаков. Мы познакомимся с эффективными способами управления пропущенными значениями, масштабирования и преобразования признаков, кодирования категориальных переменных и создания новых признаков. Область конструирования признаков требует творческого подхода и глубокого понимания поставленной задачи, и в этой части книги мы научимся выстраивать мыслительный процесс и использовать наиболее подходящие техники для повышения предсказательной способности будущей модели. Мы обсудим множество полезных методов для создания полиномиальных признаков, комбинирования существующих переменных с целью отслеживания эффектов их взаимодействий и кодирования категориальных переменных с использованием разных стратегий. Завершив чтение этой части книги, вы будете вооружены полным спектром техник и приемов для конструирования и отбора признаков, которые сможете полноценно использовать в собственных проектах.

Часть III. Очистка и предварительная обработка данных. В заключительной части книги мы сосредоточимся на критически важных задачах, связанных с очисткой данных и их предварительной обработкой. Здесь вы узнаете о передовых техниках обработки выбросов в данных, корректировки аномалий и подготовки данных для анализа временных рядов. Мы также обсудим способы снижения размерности данных, такие как метод главных компонент, незаменимые при работе с многомерными данными. Вы узнаете, как можно уменьшить сложность пространства признаков без существенного вреда для качества модели. Это позволит повысить эффективность итоговой модели и ее интерпретируемость. Освоив эти техники предварительной обработки данных, вы сможете строить хорошо структурированные наборы данных, что существенно повысит качество создаваемых моделей.

Практические примеры и реальные задачи

Каждая глава книги сопровождается примерами и упражнениями, которые позволят вам проверить полученные знания на практике в самых разных областях – от финансов до здравоохранения и розничных продаж. Столь разные примеры помогут вам понять, какие преобразования, способы кодирования

и масштабирования могут успешно применяться в тех или иных областях. Кроме того, эти примеры помогут вам критически взглянуть на конструирование признаков и научиться применять только нужные преобразования в зависимости от набора данных и итоговой модели.

Также в конце каждой главы вы встретите раздел с названием «Что может пойти не так?», в котором мы будем обсуждать распространенные ловушки и трудности, поджидающие вас в процессе конструирования и отбора признаков. Эти разделы призваны помочь вам развить критическое мышление и всегда находиться на шаг впереди, предвосхищая возможные трудности.

Важность воспроизводимости

В науке о данных воспроизводимость является одним их ключевых аспектов в построении надежных и качественных моделей. В книге мы часто будем обращаться к теме воспроизводимости, в частности при обсуждении использования конвейеров в Scikit-learn. Автоматизация шагов по преобразованию данных внутри конвейеров позволит вам применять одни и те же действия к обучающему и тестовому наборам данных и тем самым избежать утечки информации и повысить надежность результатов.

Заключение

Книга, которую вы держите в руках, является подробным руководством по освоению ключевых навыков, необходимых для подготовки данных и конструирования и отбора признаков. Эти аспекты составляют основу любого успешного проекта с применением машинного и глубокого обучения, и у вас есть прекрасная возможность овладеть ими в полной мере.

Делаете ли вы свои первые шаги в освоении науки о данных или являетесь практикующим специалистом в этой области, желающим улучшить свои навыки, эта книга поможет вам обрести уверенность при работе с большими объемами данных.

Что ж, давайте вместе отправимся в это увлекательное путешествие и научимся из сырых данных готовить восхитительные полуфабрикаты для построения успешных моделей машинного обучения!