Предисловие

Машинное обучение стало неотъемлемой частью различных коммерческих и исследовательских проектов, начиная от постановки медицинского диагноза с последующим лечением и заканчивая поиском друзей в социальных сетях. Многие полагают, что технологию машинного обучения могут использовать только крупные компании, обладающие мощными командами аналитиков. В этой книге мы хотим показать вам, с какой легкостью можно самостоятельно построить модели машинного обучения, и рассказать, как это можно сделать на практике. Прочитав эту книгу, вы сможете построить собственную систему машинного обучения, которая позволит вам выяснить настроение пользователей Твиттера или получить прогнозы по поводу глобального потепления. Область применения машинного обучения безгранична и, учитывая все многообразие данных, имеющихся на сегодняшний день, в действительности ограничивается лишь вашим воображением.

Кому стоит прочитать эту книгу

Данная книга адресована действующим и начинающим специалистам по машинному обучению, решающим реальные задачи. Эта книга является вводной и не требует предварительных знаний в области машинного обучения или искусственного интеллекта. Речь здесь пойдет об использовании языка Python и библиотеки scikit-learn, мы рассмотрим все этапы создания *успешного* проекта по машинному обучению. Методы, которые мы обсуждаем в этой книге, пригодятся ученым и исследователям, а также специалистам по анализу данных, работающим в различных коммерческих сферах. Максимальную отдачу от книги вы сможете получить, если хотя бы немного знакомы с языком Python и библиотеками NumPy и matplotlib.

Мы сознательно приложили немалые усилия к тому, чтобы в большей степени сосредоточиться на практических аспектах использования алгоритмов машинного обучения вместо детального изложения их математического

обоснования. А поскольку той основой, на которой строится машинное обучение, является именно математика (в частности, теория вероятностей), мы в этой книге решили не вдаваться в детальное описание и обсуждение используемых алгоритмов. Если же вас интересует именно математический аппарат алгоритмов машинного обучения, мы рекомендуем вам обратиться к книге издательства Springer *The Elements of Statistical Learning*, авторами которой являются Тревор Хасти (*Trevor Hastie*), Роберт Тибширани (*Robert Tibshirani*) и Джером Фридман (*Jerome Friedman*).[2] Эта книга свободно доступна на сайте авторов: **`http://statweb.stanford.edu/~tibs/ElemStatLearn/`**. Кроме того, в нашей книге мы не будем рассказывать о том, как написать тот или иной алгоритм машинного обучения "с нуля". Вместо этого изложение будет сосредоточено исключительно на практическом применении большого спектра моделей, уже реализованных в библиотеке scikit-learn и других подобных ей библиотеках.

Почему мы написали эту книгу

Существует масса книг по машинному обучению и искусственному интеллекту. Однако все они предназначены для студентов старших курсов и аспирантов, обучающихся по специальности "компьютерные науки", и полны математических подробностей. Это резко контрастирует с тем фактом, что машинное обучение в настоящее время широко используется в качестве важного прикладного инструмента в научных и коммерческих проектах. Сегодня применение машинного обучения не требует наличия научной степени. Однако существует очень мало ресурсов, в которых все важные аспекты применения машинного обучения на практике освещались бы доступно, без необходимости предварительного освоения читателем сложных математических курсов. Мы надеемся, что эта книга окажет реальную помощь тем людям, которые хотят использовать машинное обучение здесь и сейчас, не тратя годы на изучение математики, линейной алгебры и теории вероятностей.

Структура книги

В целом эта книга организована следующим образом.

- В главе 1 кратко рассказывается об основных понятиях теории машинного обучения и сферах применения этой технологии. Здесь же

[2] Эта книга выйдет в издательстве "Диалектика" в 2017 году. — *Примеч. ред.*

описывается процедура установки основных библиотек, которые мы будем использовать на протяжении всей книги.

- В главах 2 и 3 освещаются актуальные алгоритмы машинного обучения, которые в настоящее время широко используются на практике, а также анализируются их преимущества и недостатки.
- В главе 4 обсуждается важность определенного представления данных, которое можно получить с помощью алгоритмов машинного обучения, а также рассказывается о том, какие аспекты данных требуют особого внимания.
- В главе 5 освещаются передовые методы, предназначенные для оценки качества модели и настройки параметров, при этом особое внимание уделено перекрестной проверке и решетчатому поиску.
- В главе 6 излагаются принципы построения конвейеров для связывания моделей в единую цепочку и инкапсуляции рабочего потока.
- В главе 7 рассказывается о том, как применять методы, описанные в предыдущих главах, к текстовым данным, а также кратко освещаются некоторые методы обработки текста.
- В главе 8 дается общий обзор различных аспектов машинного обучения.

Несмотря на то что в главах 2 и 3 дается описание достаточно большого количества наиболее популярных алгоритмов, вполне возможно, что начинающему специалисту будет совсем необязательно знать их все. Если вам необходимо в сжатые сроки построить систему машинного обучения, мы предлагаем начать чтение книги с главы 1 и начальных разделов главы 2, в которых кратко рассказывается об основных принципах машинного обучения. Затем вам следует перейти к разделу "Выводы и перспективы" в главе 2, который включает в себя обзор всех моделей машинного обучения с учителем, освещаемых в этой книге. Здесь вы сможете выбрать ту модель, которая будет наилучшим образом соответствовать вашим задачам, после чего обратиться к разделу, посвященному этой модели, чтобы ознакомиться с деталями ее использования. Далее вы можете воспользоваться методами, описанными в главе 5, чтобы оценить качество полученной модели и требуемым образом настроить ее параметры.

Онлайн-ресурсы

Изучая материал этой книги, обязательно воспользуйтесь сайтом библиотеки scikit-learn: ***http://scikit-learn.org/stable/***. Здесь вы найдете подробную документацию об используемых в ней классах и функциях Python, а также массу полезных примеров. Кроме того, существует видеокурс Андреаса Мюллера *Advanced Machine Learning with scikit-learn*, дополняющий эту книгу. Вы можете найти его по адресу ***http://bit.ly/advanced_machine_learning_scikit-learn***.

Условные обозначения, принятые в этой книге

В этой книге используются определенные соглашения, направленные на облегчение восприятия материала.

- *Выделение курсивом* используется для обозначения новых терминов, названий книг и других печатных изданий.
- `Моноширинный шрифт` используется для листингов программ и внутри абзацев для выделения элементов программ (названий переменных или функций, баз данных, типов данных, переменных среды, операторов и ключевых слов), а также имен файлов и расширений файлов.
- **`Полужирный моноширинный шрифт`** употребляется для выделения команд или другого текста, который должен вводиться самим пользователем.
- *`Курсивный моноширинный шрифт`* применяется для выделения параметров, которые при выполнении расчетов должны быть замещены фактическими значениями, вводимыми пользователем или определяемыми из контекста.
- ***`Полужирный курсивный моноширинный шрифт`*** применяется для выделения URL-адресов и адресов электронной почты.
- И наконец рубленый полужирный курсивный шрифт используется в этой книге для выделения названий различных программных и других продуктов — отдельных программ, библиотек, комплексных приложений, пакетов данных и т.д.

Пиктограммы, используемые в этой книге

В том случае, если требовалось подчеркнуть что-нибудь действительно важное или нечто особенное, на поля в левой части страницы авторы помещали следующие пиктограммы.

Эта пиктограмма отмечает полезный совет или подсказку.

Такой пиктограммой выделяются общие замечания.

Данная пиктограмма обозначает некоторое важное предупреждение или предостережение.

Использование примеров программного кода

Все примеры программного кода и упражнения, которые приводятся в этой книге, доступны для скачивания по адресу ***https://github.com/amueller/introduction_to_ml_with_python***.

Данная книга призвана оказать вам помощь в решении задач, связанных с машинным обучением. Вы можете свободно использовать примеры программного кода из этой книги в своих программах и документации. Вам не нужно обращаться в издательство за разрешением, если вы не собираетесь воспроизводить существенные части программного кода. Например, если вы разрабатываете программу и используете в ней несколько фрагментов программного кода из книги, вам не нужно обращаться за разрешением. Однако в случае продажи или распространения компакт-дисков с примерами из этой книги вам необходимо получить разрешение от издательства O'Reilly. Если вы отвечаете на вопросы, цитируя данную книгу или примеры из нее, разрешение не требуется. Но при включении значительного объема программного кода из этой книги в свою документацию необходимо будет получить разрешение от издательства.

Мы приветствуем, но не требуем добавлять ссылку на первоисточник при цитировании. Под ссылкой на первоисточник мы подразумеваем

указание авторов, издательства и ISBN книги, например "*An Introduction to Machine Learning with Python* (O'Reilly) by Andreas C. Mueller and Sarah Guido. Copyright 2017 Sarah Guido and Andreas Mueller, 978-1-449-36941-5".

Если вы считаете, что использование вами примеров программного кода выходит за разрешенные рамки, присылайте свои вопросы на нашу электронную почту: ***permissions@oreilly.com***.

От издательства "Диалектика"

Вы, читатель этой книги, и есть главный ее критик. Мы ценим ваше мнение и хотим знать, что было сделано нами правильно, что можно было сделать лучше и что еще вы хотели бы увидеть изданным нами. Нам интересны любые ваши замечания в наш адрес.

Мы ждем ваших комментариев и надеемся на них. Вы можете прислать нам бумажное или электронное письмо либо просто посетить наш веб-сайт и оставить свои замечания там. Одним словом, любым удобным для вас способом дайте нам знать, нравится ли вам эта книга, а также выскажите свое мнение о том, как сделать наши книги более интересными для вас.

Отправляя письмо или сообщение, не забудьте указать название книги и ее авторов, а также свой обратный адрес. Мы внимательно ознакомимся с вашим мнением и обязательно учтем его при отборе и подготовке к изданию новых книг.

Наши электронные адреса:

E-mail: `info@dialektika.com`
WWW: `http://www.dialektika.com`

Наши почтовые адреса:

в России: 195027, Санкт-Петербург, Магнитогорская ул., д. 30, ящик 116
в Украине: 03150, Киев, а/я 152