Оливер Теобальд

МАШИННОЕ ОБУЧЕНИЕ
ДЛЯ АБСОЛЮТНЫХ НОВИЧКОВ

ВВОДНЫЙ КУРС,
ИЗЛОЖЕННЫЙ ПРОСТЫМ ЯЗЫКОМ

БОМБОРА
ИЗДАТЕЛЬСТВО
Москва 2024

УДК 004.8
ББК 32.813
Т33

Oliver Theobald
MACHINE LEARNING FOR ABSOLUTE BEGINNERS

Теобальд, Оливер.

Т33 Машинное обучение для абсолютных новичков. Вводный курс, изложенный простым языком / Оливер Теобальд; [перевод с английского М. А. Райтмана]. — Москва : Эксмо, 2026. — 208 с. — (Мировой компьютерный бестселлер).

ISBN 978-5-04-190305-3

«Машинное обучение для абсолютных новичков» Оливера Теобальда — это идеальная книга для тех, кто хочет изучить основы машинного обучения (ML) без опыта программирования. Книга содержит основные алгоритмы ML, наглядные примеры, практические работы и обучение классической статистике. Руководство включает в себя материалы по загрузке бесплатных наборов данных, методы очистки и подготовки данных для анализа, основы работы нейронных сетей и многое другое.

УДК 004.8
ББК 32.813

ISBN 978-5-04-190305-3

ИЩИТЕ НАС НА СЛЕДУЮЩИХ РЕСУРСАХ:

Ежемесячный информационный бюллетень
http://eepurl.com/gKjQij
Получайте рекомендации книг, выигрывайте бесплатные экземпляры новых изданий от авторов и знакомьтесь со статьями и новостями, посвященными машинному обучению и науке о данных.

Teachable
http://scatterplotpress.teachable.com/
Здесь вы найдете вводные видеокурсы по машинному обучению, а также бонусные видеоуроки, сопровождающие эту книгу.

Skillshare
www.skillshare.com/user/machinelearning_beginners
На этом сайте представлены вводные видеокурсы по машинному обучению и видеоуроки от других инструкторов.

Instagram[1]
machinelearning_beginners
Здесь вы найдете краткие уроки, цитаты из книг и многое другое!

[1] Соцсеть признана экстремистской и запрещена на территории РФ.

ОГЛАВЛЕНИЕ

1

ПРЕДИСЛОВИЕ

Машины претерпели большие изменения с начала промышленной революции. Они по-прежнему заполняют цеха заводов и фабрик, однако их возможности уже выходят за рамки ручного труда и позволяют решать когнитивные задачи, которые до недавнего времени были под силу только человеку. Судейство песенных конкурсов, вождение автомобилей и выявление мошеннических операций — вот лишь три примера сложных задач, которые способны решать современные машины.

Однако эти замечательные достижения вселяют страх в некоторых людей. Отчасти он связан с опасениями выживальщиков и вечным вопросом: «А что, если?» Что, если разумные машины восстанут против нас в борьбе за существование? Что, если разумные машины произведут потомство, обладающее возможностями, которыми люди никогда не собирались их наделять? Что, если легенда о сингулярности окажется правдой?

Еще один существенный страх связан с угрозой потери рабочего места, и если вы работаете таксистом или бухгалтером, то у вас есть вполне веские основания для беспокойства. Согласно совместному исследованию Национальной статистической службы и компании Deloitte UK, опубликованному BBC в 2015 году, такие рабочие профессии, как работник бара (77%), официант (90%), дипломированный бухгалтер (95%), администратор (96%) и таксист (57%), имеют высокие шансы быть автоматизированными к 2035 году[2]. Однако к результатам исследований, касающихся планируемой автоматизации рабочих мест, и предсказаниям относительно развития машин и искусственного интеллекта (ИИ) следует относиться с некоторой долей скептицизма. В своей книге «Искусственный интеллект. Этапы. Угрозы. Стратегии» Ник Бостром рассуждает о постоянном пересмотре прогнозов относительно развития ИИ и о том, что «два десятилетия — это тот оптимальный срок, который является достаточно близким, чтобы привлекать

[2] "Will A Robot Take My Job?", *The BBC*, дата обращения: 30 декабря 2017 года, **http://www.bbc.com/news/technology-34066941**

внимание и оставаться актуальным, но достаточно далеким для того, чтобы допустить возможность ряда прорывов, которые могли бы к тому времени произойти»[3, 4].

Несмотря на то что ИИ развивается довольно быстро, его широкое внедрение по-прежнему остается неизведанным путем, на котором нам предстоит столкнуться с непредвиденными проблемами, задержками и другими препятствиями. Машинное обучение не сводится к простому щелчку выключателем, позволяющим приказать машине спрогнозировать результаты «Суперкубка» и подать вам вкусный мартини.

Машинное обучение не имеет ничего общего с готовыми аналитическими решениями. Оно опирается на статистические алгоритмы, которые контролируются такими квалифицированными специалистами, как дата-сайентисты и инженеры машинного обучения. Это один из быстро растущих рынков труда, на котором предложение пока не в состоянии удовлетворить спрос.

На самом деле нехватка профессионалов, обладающих необходимым опытом и уровнем подготовки, — одно из основных препятствий, задерживающих развитие сферы ИИ. По словам Чарльза Грина, директора по идейному лидерству компании Belatrix Software:

> «Во-первых, это огромная проблема — найти дата-сайентистов, *людей, обладающих* опытом работы в сфере машинного обучения или навыками анализа и использования данных, а также тех, кто способен создавать необходимые алгоритмы. Во-вторых, несмотря на то что технология все еще находится на стадии становления, существует множество текущих разработок. Очевидно, что сфера ИИ еще очень далека от того, как мы ее себе представляем»[5].

Возможно, с чтения этой книги начнется ваш путь к получению работы в области машинного обучения, а может быть, она просто даст вам о ней

---

[3] Ник Бостром, «Искусственный интеллект. Этапы. Угрозы. Стратегии». Издательство: Манн, Иванов и Фербер, 2016.

[4] Бостром также шутливо замечает, что два десятилетия примерно соответствуют оставшейся продолжительности карьеры типичного прогнозиста.

[5] Matt Kendall, "Machine Learning Adoption Thwarted by Lack of Skills and Understanding," Nearshore Americas, дата обращения: 14 мая 2017 года, **http://www.nearshoreamericas.com/machine-learning-adoption-understanding**

базовое представление, которого будет достаточно для удовлетворения вашего любопытства.

Эта книга представляет собой высокоуровневый обзор, включающий ключевые термины, общее описание рабочего процесса и статистические основы базовых алгоритмов, которые позволят вам начать свой путь. Для разработки и программирования интеллектуальных машин вам в первую очередь необходимо хорошо усвоить классическую статистику, так как алгоритмы, реализованные на ее основе, — это сердце машинного обучения. Они представляют собой метафорические нейроны, отвечающие за искусственные когнитивные способности. Написание кода — это еще одна неотъемлемая часть машинного обучения, которая предусматривает управление и манипулирование большими объемами данных. В отличие от создания целевых веб-страниц с помощью таких конструкторов, как Wix и WordPress, машинное обучение требует использования Python, C++, R или другого языка программирования. Если вы еще не изучили соответствующий язык, вам придется это сделать, чтобы развиваться в этой области. Однако материал, который вы здесь найдете, можно освоить даже без опыта программирования.

Хотя эта книга служит вводным курсом по машинному обучению, ознакомление читателей с основами математики, компьютерного программирования и статистики — не ее цель. Для облегчения понимания материала следующих глав вам могут потребоваться базовые знания в этих областях или доступ к Интернету.

Для тех, кто хочет погрузиться в аспект машинного обучения, связанный с программированием, в главах 17 и 18 представлен весь процесс создания модели машинного обучения с помощью языка Python. Небольшое введение в программирование на языке Python вы найдете в приложении, а информацию о дополнительных обучающих ресурсах — в заключительном разделе книги.

Наконец, видеоуроки и другие онлайн-материалы, сопровождающие эту книгу, вы можете найти на сайте:

https://scatterplotpress.teachable.com/p/ml-code-exercises.

2

ЧТО ТАКОЕ МАШИННОЕ ОБУЧЕНИЕ?

В 1959 году компания IBM опубликовала в журнале *IBM Journal of Research and Development* статью с интригующим и туманным названием. Ее автор, сотрудник IBM Артур Самуэль, исследовал вопрос о применении машинного обучения в контексте игры в шашки «для проверки того факта, что компьютер можно запрограммировать таким образом, чтобы он научился играть в шашки лучше, чем человек, написавший программу»[6].

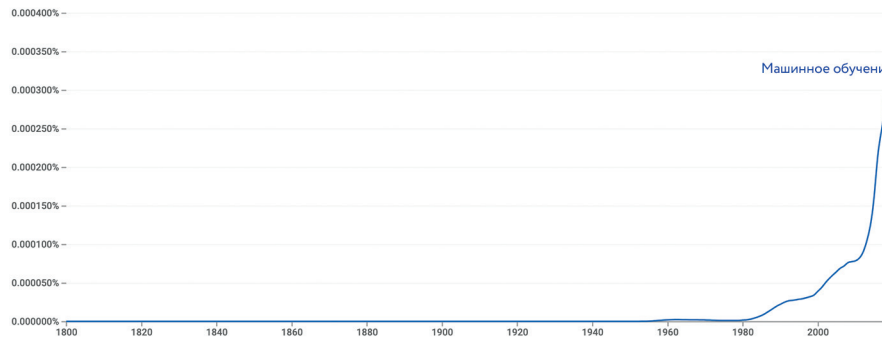

Рис. 1. История упоминания термина «машинное обучение» в опубликованных книгах. Источник: *Google Ngram Viewer*, 2017.

Несмотря на то что это была не первая опубликованная работа, в которой использовался термин «машинное обучение», Артур Самуэль считается тем человеком, который ввел понятие и дал определение машинному обучению как концепции и специализированной области, известной нам сегодня (рис. 1). В знаковой статье Самуэля под названием *Some Studies in Machine Learning Using the Game of Checkers* («Некоторые исследования в области машинного обучения на примере игры в шашки») машинное обучение

[6] Arthur Samuel, "Some Studies in Machine Learning Using the Game of Checkers," *IBM Journal of Research and Development*, Vol. 3, Issue. 3, 1959.

было представлено в качестве области информатики, которая изучает возможность компьютеров обучаться без их явного программирования.

Хотя в первоначальном определении Артура Самуэля это понятие не было рассмотрено, ключевая характеристика машинного обучения — это концепция самообучения. Она подразумевает применение статистического моделирования для выявления закономерностей и повышения производительности на основе данных и эмпирической информации без использования прямых команд. Артур Самуэль назвал это способностью обучаться без явного программирования. Он не имел в виду, что машины могут формулировать решения без какого-либо предварительного программирования. Напротив, машинное обучение в значительной степени зависит от вводимого кода. Но он заметил, что машины способны выполнять поставленную задачу, используя входные данные, вместо того чтобы полагаться на входную команду (рис. 2).

Рис. 2. Сравнение входной команды и входных данных

В качестве примера использования входной команды можно рассмотреть ввод выражения «2+2» на языке программирования Python и нажатие кнопки запуска или клавиши ввода для просмотра результата.

```
>>> 2+2
4
>>>
```

Это прямая команда с заранее запрограммированным ответом, характерная для большинства компьютерных приложений. Однако в отличие

от традиционного компьютерного программирования, при котором результаты или решения заранее определяются программистом, машинное обучение предполагает использование входных данных для построения модели принятия решений. Эти решения генерируются путем выявления существующих в данных взаимосвязей и закономерностей с помощью вероятностных рассуждений, проб и ошибок и применения других вычислительно-интенсивных методов. Это означает, что выход модели принятия решений определяется содержанием входных данных, а не правилами, заранее определенными программистом. При этом программист по-прежнему отвечает за подачу данных на вход модели, выбор подходящего алгоритма и настройку его параметров (называемых гиперпараметрами) для уменьшения ошибки прогнозирования. Однако в отличие от традиционного программирования в данном случае машина и разработчик работают на разных уровнях.

В качестве примера предположим, что в результате анализа привычек пользователей YouTube[7] модель принятия решений выявляет значимую взаимосвязь, говорящую о том, что специалисты в области науки о данных любят смотреть видео с котиками. В то же время другая модель выявляет закономерности между физическими показателями бейсболистов и вероятностью получения ими награды «Самый ценный игрок» (MVP) в текущем сезоне.

В первом сценарии машина анализирует предпочтения пользователей YouTube*, опираясь на такие показатели их вовлеченности, как лайки, подписки и повторные просмотры. Во втором сценарии машина оценивает физические показатели бейсболистов, ранее получивших награду MVP, наряду с другими их характеристиками, такими как возраст и уровень образования. Однако ни на одном из этапов модель принятия решений не получает каких-либо указаний, заставляющих ее выдать именно эти два результата. Модель использует методы машинного обучения для выявления сложных закономерностей, существующих во входных данных, без помощи со стороны человека. Это также означает, что при использовании аналогичного набора данных, собранного за другой период времени и включающего другое количество точек данных, модель может выдать несколько иной результат.

Еще одна отличительная особенность машинного обучения — способность модели улучшать качество прогнозов на основе опыта. Подобно тому

[7] Соцсеть признана экстремистской и запрещена на территории РФ.