

Цунами машинного обучения

В 2006 году Джеффри Хинтон и др. опубликовали статью (<https://hom1.info/136>)¹, в которой было показано, как обучать глубокую нейронную сеть, способную распознавать рукописные цифры с передовой точностью (>98%). Они назвали такой прием “глубоким обучением” (*Deep Learning*). Глубокая нейронная сеть является (весьма) упрощенной моделью коры нашего головного мозга, состоящей из стопки слоев искусственных нейронов. Обучение глубокой нейронной сети в то время считалось невозможным² и с 1990-х годов большинство исследователей отказалось от этой затеи. Указанная статья возродила интерес научной общественности и вскоре многочисленные новые статьи продемонстрировали, что глубокое обучение было не только возможным, но способным к умопомрачительным достижениям. Никакой другой прием машинного обучения (*Machine Learning*) не мог даже приблизиться к таким достижениям (при помощи гигантской вычислительной мощности и огромных объемов данных). Энтузиазм быстро распространился на многие другие области машинного обучения.

Спустя 10 лет или около того машинное обучение завоевало отрасль: оно лежит в основе большей части магии сегодняшних высокотехнологичных продуктов, упорядочивая результаты ваших поисковых запросов, приводя в действие распознавание речи в вашем смартфоне, рекомендуя к просмотру видеоролики и одержав победу над чемпионом мира по игре го. Вы не успеете оглянуться, как машинное обучение начнет вести ваш автомобиль.

¹ Джеффри Хинтон и др., *A Fast Learning Algorithm for Deep Belief Nets* (Быстрый алгоритм обучения для глубоких сетей доверия), *Neural Computation* 18 (2006 г.): с. 1527–1554.

² Несмотря на тот факт, что глубокие сверточные нейронные сети Яна Лекуна хорошо работали при распознавании изображений с 1990-х годов, они не были до такой степени универсальными.

Машинное обучение в ваших проектах

Итак, естественно вы потрясены машинным обучением и желали бы присоединиться к компании!

Возможно, вы хотите дать своему домашнему роботу собственный мозг? Добиться, чтобы он распознавал лица? Научить его ходить?

А может быть, у вашей компании имеется масса данных (журналы пользователей, финансовые данные, производственные данные, данные машинных датчиков, статистические данные от линии оперативной поддержки, отчеты по персоналу и т.д.) и вероятнее всего вы сумели бы найти там несколько скрытых жемчужин, просто зная, где искать. С помощью машинного обучения вы могли бы решить перечисленные ниже задачи, а также многие другие (<https://homl.info/usecases>):

- сегментировать заказчиков и установить наилучшую маркетинговую стратегию для каждой группы;
- рекомендовать товары каждому клиенту на основе того, что покупают похожие клиенты;
- определять, какие транзакции, возможно, являются мошенническими;
- прогнозировать доход в следующем году.

Какой бы ни была причина, вы решили освоить машинное обучение и внедрить его в свои проекты. Великолепная идея!

Цель и подход

В книге предполагается, что вы почти ничего не знаете о машинном обучении. Ее цель — предоставить вам концепции, инструменты и идеи, которые необходимы для реализации программ, способных *обучаться на основе данных*.

Мы рассмотрим многочисленные методики, начиная с простейших и самых часто используемых (таких как линейная регрессия) и заканчивая методиками глубокого обучения, которые регулярно побеждают в состязаниях.

Вместо того чтобы реализовывать собственную миниатюрную версию каждого алгоритма, мы будем применять фреймворки Python производственного уровня.

- Библиотека Scikit-Learn (<http://scikit-learn.org/>) очень проста в использовании, но эффективно реализует многие алгоритмы машинного обучения, что делает ее великолепной отправной точкой при изучении машинного обучения.
- TensorFlow (<http://tensorflow.org/>) является более сложной библиотекой для распределенных численных расчетов. Она позволяет эффективно обучать и запускать очень большие нейронные сети, потенциально распределяя вычисления между сотнями серверов с множеством графических процессоров. Библиотека TensorFlow (TF) была создана в компании Google и поддерживает многие из их крупномасштабных приложений машинного обучения. В ноябре 2015 года она стала продуктом с открытым кодом.
- Библиотека Keras (<https://keras.io/>) представляет собой высокоуровневый API-интерфейс для глубокого обучения, который делает очень простым обучение и запуск нейронных сетей. Она может функционировать поверх библиотек TensorFlow, Theano либо инструментального комплекта Microsoft Cognitive Toolkit (ранее известного как CNTK). Библиотека TensorFlow поставляется с собственной реализацией такого API-интерфейса, называемой `tf.keras`, которая предлагает поддержку для ряда расширенных возможностей TensorFlow (скажем, способности эффективно загружать данные).

Предпочтение в книге отдается практическому подходу, стимулируя интуитивное понимание машинного обучения через конкретные работающие примеры, поэтому теории здесь совсем немного. Хотя вы можете читать книгу, не прибегая к ноутбуку, я настоятельно рекомендую экспериментировать с кодом примеров, который доступен в виде тетрадей Jupyter по адресу <https://github.com/ageron/handson-ml2>.

Предварительные требования

В книге предполагается, что вы обладаете некоторым опытом программирования на языке Python и знакомы с главными библиотеками Python для научных расчетов, в частности NumPy (<http://numpy.org/>), pandas (<http://pandas.pydata.org/>) и Matplotlib (<http://matplotlib.org/>).

Вдобавок если вас интересует, что происходит внутри, тогда вы должны также понимать математику на уровне колледжа (исчисление, линейную алгебру, теорию вероятностей и статистику).

Если вы пока еще не знаете язык Python, то веб-сайт <http://learnpython.org/> станет прекрасным местом, чтобы приступить к его изучению. Также неплохим ресурсом будет официальное руководство на Python.org (<https://docs.python.org/3/tutorial/>).

Если вы никогда не работали с Jupyter, то в главе 2 будет описан процесс установки и основы: это замечательный инструмент, который полезно иметь в своем арсенале.

Если вы не знакомы с библиотеками Python для научных расчетов, то предоставленные тетради Jupyter включают несколько подходящих руководств. Доступно также краткое руководство по линейной алгебре.

Дорожная карта

Книга содержит две части. В части I раскрываются перечисленные ниже темы.

- Что собой представляет машинное обучение, задачи, которые оно пытается решать, а также основные категории и фундаментальные концепции его систем.
- Шаги в типовом проекте машинного обучения.
- Обучение путем подгонки модели к данным.
- Оптимизация функции издержек.
- Обработка, очистка и подготовка данных.
- Выбор и конструирование признаков.
- Выбор модели и подстройка гиперпараметров с использованием перекрестной проверки.
- Проблемы машинного обучения, в особенности недообучение и переобучение (компромисс между смещением и дисперсией).
- Наиболее распространенные алгоритмы обучения: линейная и полиномиальная регрессия, логистическая регрессия, метод k ближайших соседей, метод опорных векторов, деревья принятия решений, случайные леса и ансамблевые методы.

- Понижение размерности обучающих данных в целях борьбы с “проклятием размерности”.
- Другие методики обучения без учителя, включая кластеризацию, оценку плотности и обнаружение аномалий.

В части II рассматриваются следующие темы.

- Что собой представляют нейронные сети и для чего они пригодны.
- Построение и обучение нейронных сетей с применением TensorFlow и Keras.
- Самые важные архитектуры нейронных сетей: нейронные сети прямого распространения для табличных данных, сверточные сети для компьютерного зрения, рекуррентные сети и сети с долгой краткосрочной памятью (LSTM) для обработки последовательностей, кодировщики/декодировщики и преобразователи для обработки естественного языка. автокодировщики и порождающие состязательные сети для генеративного обучения.
- Методики обучения глубоких нейронных сетей.
- Построение агента (например, бота в игре), который способен узнавать хорошие стратегии методом проб и ошибок с использованием обучения с подкреплением.
- Эффективная загрузка и предварительная обработка крупных объемов данных.
- Широкомасштабное обучение и развертывание моделей TensorFlow.

Первая часть основана главным образом на применении Scikit-Learn, в то время как вторая — на использовании TensorFlow и Keras.



Не спешите с погружением: несмотря на то, что глубокое обучение, без всякого сомнения, является одной из самых захватывающих областей в машинном обучении, вы должны сначала овладеть основами. Кроме того, большинство задач могут довольно хорошо решаться с применением простых приемов, таких как случайные леса и ансамблевые методы (обсуждаются в части I). Глубокое обучение лучше всего подходит для решения сложных задач, подобных распознаванию изображений, распознаванию речи или обработке естественного языка, при условии, что имеется достаточный объем данных, вычислительная мощность и терпение.

Изменения во втором издании

Второе издание книги преследует шесть основных целей.

1. Раскрытие дополнительных тем о машинном обучении: больше методик для обучения без учителя (в числе которых кластеризация, обнаружение аномалий, оценка плотности и модели со смесями распределений); больше методик для обучения глубоких сетей (включая самонормализующиеся сети); дополнительные методики компьютерного зрения (в том числе Xception, SENet, выявление объектов с помощью YOLO и семантическая сегментация с использованием R-CNN); обработка последовательностей с применением сверточных нейронных сетей (включая WaveNet); обработка естественного языка с использованием рекуррентных нейронных сетей, сверточных нейронных сетей и преобразователей; порождающие состязательные сети.
2. Раскрытие дополнительных библиотек и API-интерфейсов (Keras, Data API, TF-Agents для обучения с подкреплением), а также широкомасштабное обучение и развертывание моделей TF с применением API-интерфейса Distribution Strategies, TF-Serving и платформы AI Platform инфраструктуры Google Cloud. Вдобавок предлагается краткое введение в TF Transform, TFLite, TF Addons/Seq2Seq и TensorFlow.js.
3. Обсуждение ряда последних важных результатов, полученных при исследовании глубокого обучения.
4. Переход с TensorFlow на TensorFlow 2 и повсеместное использование реализации API-интерфейса Keras библиотеки TensorFlow (`tf.keras`).
5. Обновление кода примеров для применения последних версий Scikit-Learn, NumPy, pandas, Matplotlib и других библиотек.
6. Прояснение некоторых разделов и исправление ошибок благодаря многочисленным отзывам читателей.

В книгу добавлены новые главы, некоторые главы были переписаны и несколько глав переупорядочено. Дополнительные сведения об изменениях во втором издании доступны по ссылке <https://homl.info/changes2>.

Другие ресурсы

Для освоения машинного обучения доступно много превосходных ресурсов. Учебные курсы по машинному обучению Эндрю Ына на Coursera (<https://homl.info/ngcourse>) изумительны, хотя требуют значительных затрат времени (вероятно, нескольких месяцев).

Кроме того, существует много интересных веб-сайтов о машинном обучении, в числе которых, конечно же, веб-сайт с выдающимся руководством пользователя Scikit-Learn (<https://homl.info/skdoc>). Вам также может понравиться веб-сайт Dataquest (<https://www.dataquest.io/>), который предлагает очень полезные интерактивные руководства, и блоги по машинному обучению вроде перечисленных на веб-сайте Quora (<https://homl.info/1>). Наконец, веб-сайт по глубокому обучению (<http://deeplearning.net/>) содержит хороший список ресурсов для дальнейшего изучения.

Доступно множество других вводных книг по машинному обучению, включая перечисленные ниже.

- В книге Джоэла Грусса *Data Science from Scratch* (O'Reilly) представлены основы машинного обучения и реализации ряда основных алгоритмов с помощью чистого кода на Python (с нуля).
- Книга Стивена Марсленда *Machine Learning: An Algorithmic Perspective* (Chapman and Hall) представляет собой замечательное введение в машинное обучение и раскрывает широкий спектр вопросов с примерами кода на Python (также с нуля, но с использованием NumPy).
- Книга Себастьяна Рашки *Python Machine Learning* (Packt Publishing; *Python и машинное обучение: машинное и глубокое обучение с использованием Python, Scikit-learn и TensorFlow*, пер. с англ., Диалектика) также является великолепным введением в машинное обучение и задействует библиотеки Python с открытым кодом (Pylearn 2 и Theano).
- Книга Франсуа Шолле *Deep Learning with Python* (Manning) ориентирована на практику и раскрывает большой диапазон тем в ясной и лаконичной манере, чего и можно было ожидать от автора великолепной библиотеки Keras. Предпочтение в ней отдается примерам кода, а не математической теории.

- Книга Андрея Буркова *The Hundred-Page Machine Learning Book* отличается краткостью и посвящена впечатляющему спектру тем, представляя их доступными терминами и не избегая математических уравнений.
- Книга Ясера С. Абу-Мустафы, Малика Магдона-Исмаила и Сюань-Тянь Линя *Learning from Data* (AMLBook) демонстрирует довольно теоретический подход к машинному обучению, который обеспечивает глубокое понимание многих аспектов, в частности компромисса между смещением и дисперсией (см. главу 4).
- Прекрасная (и большая) книга Стюарта Рассела и Питера Норвига *Artificial Intelligence: A Modern Approach, 3rd Edition* (Pearson) охватывает невероятный объем тем, в числе которых и машинное обучение. Она помогает уловить общую картину машинного обучения.

Наконец, присоединение к веб-сайтам состязаний по машинному обучению, таким как [Kaggle.com](https://www.kaggle.com/), даст вам возможность приложить свои навыки к решению реальных задач, получая помощь от ряда выдающихся профессионалов в области машинного обучения.

Типографские соглашения, используемые в книге

В книге применяются следующие типографские соглашения.

Курсив

Используется для новых терминов.

Моноширинный

Применяется для URL, адресов электронной почты, имен и расширений файлов, листингов программ, а также внутри абзацев для ссылки на программные элементы наподобие имен переменных и функций, баз данных, типов данных, переменных среды и ключевых слов.

Моноширинный полужирный

Используется для представления команд или другого текста, который должен набираться пользователем буквально.

Моноширинный курсив

Применяется для текста, который должен быть заменен значениями, предоставленными пользователем, или значениями, определяемыми контекстом.



Этот элемент содержит совет или указание.



Этот элемент содержит замечание общего характера.



Этот элемент содержит предупреждение или предостережение.

Код примеров

Имеется множество тетрадей Jupyter, заполненных дополнительными материалами, такими как код примеров и упражнения, которые доступны для загрузки по адресу <https://github.com/ageron/handson-ml2>.

Код ряда примеров в книге не содержит повторяющихся разделов или деталей, которые очевидны или не имеют отношения к машинному обучению. Такой подход позволяет сосредоточить внимание на важных частях кода и экономит пространство для раскрытия большего числа тем. Полный код всех примеров находится в тетрадях Jupyter.

Имейте в виду, что когда код примеров отображает вывод, то он показан с приглашениями на ввод Python (`>>>` и `...`), как в командной оболочке Python, чтобы отделить код от вывода. Скажем, в следующем коде определяется функция `square()`, после чего она вызывается и отображается результат возведения 3 в квадрат:

```
>>> def square(x):  
...     return x ** 2  
...  
>>> result = square(3)  
>>> result  
9
```

Если код ничего не отображает, тогда приглашения на ввод не используются. Однако временами результат может быть показан в виде комментария, например:

```
def square(x):  
    return x ** 2  
  
result = square(3)    # значение result равно 9
```

Использование кода примеров

Настоящая книга призвана помочь вам выполнять свою работу. Обычно, если в книге предлагается код примеров, то вы можете применять его в собственных программах и документации. Вы не обязаны обращаться к нам за разрешением, если только не используете значительную долю кода. Скажем, написание программы, в которой задействовано несколько фрагментов кода из этой книги, разрешения не требует. Для продажи или распространения компакт-диска с примерами из книг O'Reilly разрешение обязательно. Ответ на вопрос путем цитирования данной книги и ссылки на код примера разрешения не требует. Для встраивания значительного объема кода примеров, рассмотренных в этой книге, в документацию по вашему продукту разрешение обязательно.

Мы высоко ценим указание авторства, хотя и не требуем этого. Установление авторства обычно включает название книги, фамилию и имя автора, издательство и номер ISBN. Например: “*Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow*, 2nd Edition, by Aurélien Géron (O'Reilly). Copyright 2019 Aurélien Géron, 978-1-492-03264-9”. Если вам кажется, что способ использования вами кода примеров выходит за законные рамки или не соответствует упомянутым выше разрешениям, тогда свяжитесь с нами по следующему адресу электронной почты: permissions@oreilly.com.

Об иллюстрации на обложке

Животное, изображенное на обложке книги — это огненная, или пятнистая, или обыкновенная саламандра (*Salamandra salamandra*), амфибия, встречающаяся в большей части Европы. Ее черная лоснящаяся кожа отличается большими желтыми пятнами на голове и спине, которые сигнализируют о наличии алкалоидных токсинов. Они являются вероятным источником общего названия этой амфибии: контакт с алкалоидными токсинами (которые она также умеет разбрызгивать на короткие расстояния) вызывает конвульсии и гипервентиляцию. Болезненные яды или влажность кожи саламандры (либо то и другое) привели к ошибочному убеждению, что эти существа не только способны выжить, будучи помещенными в огонь, но также могут потушить его.

Огненные саламандры живут в прохладных лесах, прячась в сырых расщелинах и под упавшими стволами деревьев вблизи заводей или других пресноводных водоемов, которые облегчают их размножение. Хотя большую часть своей жизни огненные саламандры проводят на земле, они рожают своих детенышей в воде. Огненные саламандры питаются в основном насекомыми, пауками, слизнями и червями. Они могут вырастать до 30 см в длину, а в неволе способны жить до 50 лет.

Численность огненной саламандры была уменьшена по причине уничтожения их лесного ареала и отлова особей с целью продажи в качестве домашних животных, но самая большая угроза, с которой они сталкиваются — это восприимчивость их влагопроницаемой кожи к загрязнителям и микробам. С 2014 г. из-за привнесенного грибка они исчезли в некоторых частях Нидерландов и Бельгии.

Многие животные, изображенные на обложках книг O'Reilly, находятся под угрозой уничтожения; все они важны для нашего мира. Чтобы узнать больше о том, чем вы можете помочь, посетите веб-сайт animals.oreilly.com.

Изображение на обложке взято из книги *Wood's Illustrated Natural History* (Иллюстрированное естествознание Вуда).

Ждем ваших отзывов!

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 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>