Предисловие

Цунами машинного обучения

В 2006 году Джеффри Хинтон и др. опубликовали статью[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"). Обучение глубокой нейронной сети в то время считалось невозможным[2] 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 — ML) не мог даже приблизиться к таким достижениям (при помощи гигантской вычислительной мощности и огромных объемов данных). Этот энтузиазм быстро распространился на многие другие области машинного обучения.

Промелькнуло 10 лет и машинное обучение завоевало отрасль: теперь оно лежит в основе большей части магии сегодняшних высокотехнологичных продуктов, упорядочивая результаты ваших поисковых запросов, приводя в действие распознавание речи в вашем смартфоне и рекомендуя видеоролики. Вдобавок оно еще и одержало победу над чемпионом мира по игре го. Не успеете оглянуться, и машинное обучение начнет вести ваш автомобиль.

---

[1] Статья доступна на домашней странице Хинтона, находящейся по адресу http://www.cs.toronto.edu/~hinton/.

[2] Несмотря на тот факт, что глубокие сверточные нейронные сети Яна Лекуна хорошо работали при распознавании изображений с 1990-годов, они были не настолько универсальны.

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, скорее всего, вы сумели бы найти там несколько скрытых жемчужин, просто зная, где искать. Ниже перечислены примеры того, что можно было бы делать:

- сегментировать заказчиков и установить лучшую 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;
- многое другое (`https://www.kaggle.com/wiki/DataScienceUseCases`).

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, идеи и инструменты, которые необходимы для фактической реализации программ, способных *обучаться на основе данных*.

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

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

- Библиотека Scikit-Learn (http://scikit-learn.org/) очень проста в использовании, при этом она эффективно реализует многие алгоритмы машинного обучения, что делает ее великолепной отправной точкой для изучения машинного обучения.
- TensorFlow (http://tensorflow.org/) является более сложной библиотекой для распределенных численных расчетов с применением графов потоков данных. Это позволяет эффективно обучать и запускать очень большие нейронные сети, потенциально распределяя вычисления между тысячами серверов с множеством графических процессоров. Библиотека TensorFlow была создана в Google и поддерживает много крупномасштабных приложений машинного обучения. В ноябре 2015 года она стала продуктом с открытым кодом.

В книге отдается предпочтение 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, настоятельно рекомендуется экспериментировать с примерами кода, которые доступны в виде тетрадей Jupyter по адресу https://github.com/ageron/handson-ml.

Предварительные требования

В книге предполагается, что вы обладаете некоторым опытом программирования на 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/).

К тому же, 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 его 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 математическое руководство по линейной алгебре.

Дорожная карта

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

- Что такое машинное обучение? Какие задачи оно пытается решать? Каковы основные категории и фундаментальные концепции систем машинного обучения?
- Главные 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.
- Главные проблемы машинного обучения, в частности недообучение и переобучение (компромисс между смещением и дисперсией).
- Понижение размерности обучающих данных в целях борьбы с "проклятием размерности".
- 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.

В **части II** рассматриваются следующие темы.

- Что собой представляют нейронные сети? Для чего они пригодны?
- Построение и обучение нейронных сетей с применением TensorFlow.
- Самые важные архитектуры нейронных сетей: нейронные сети прямого распространения, сверточные сети, рекуррентные сети, сети с долгой краткосрочной памятью (LSTM) и автокодировщики.

- Приемы обучения глубоких нейронных сетей.
- Масштабирование нейронных сетей для гигантских наборов данных.
- Обучение с подкреплением.

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

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 использованием 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.

Другие ресурсы

Для освоения машинного обучения доступно много ресурсов. Учебные курсы по машинному обучению Эндрю Ына на Coursera (https://www.coursera.org/learn/machine-learning/) и учебные курсы по нейронным сетям и глубокому обучению Джеффри Хинтона (https://www.coursera.org/course/neuralnets) 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 (http://scikit-learn.org/stable/user_guide.html). Вам также может понравиться веб-сайт Dataquest (https://www.dataquest.io/), предлагающий очень полезные интерактивные руководства, и блоги по машинному обучению вроде перечисленных на веб-сайте Quora (http://goo.gl/GwtU3A). Наконец, веб-сайт по глубокому обучению (http://deeplearning.net/) содержит хороший список ресурсов для дальнейшего изучения.

Разумеется, доступны многие другие вводные книги по машинному обучению, включая перечисленные ниже.

- Джоэл Грус, *Data Science from Scratch* (http://shop.oreilly.com/product/0636920033400.do) (O'Reilly). В этой книге представлены основы машинного обучения и реализации ряда основных алгоритмов с помощью только кода Python (с нуля).
- Стивен Марсленд, *Machine Learning: An Algorithmic Perspective* (Chapman and Hall). Эта книга является замечательным введением в машинное обучение. В ней подробно раскрыт широкий спектр вопросов и приведены примеры кода на Python (также с нуля, но с применением NumPy).
- Себастьян Рашка, *Python Machine Learning* (Packt Publishing). Также великолепное введение в машинное обучение, в котором задействованы библиотеки Python с открытым кодом (Pylearn 2 и Theano).
- Ясер С. Абу-Мустафа, Малик Магдон-Исмаил и Сюань-Тянь Линь, *Learning from Data* (AMLBook). За счет достаточно объемного теоретического подхода к машинному обучению эта книга обеспечивает глубокое понимание многих аспектов, в частности компромисса между смещением и дисперсией (глава 4).
- Стюарт Рассел и Питер Норвиг, *Artificial Intelligence: A Modern Approach, 3rd Edition* (Pearson) (*Искусственный интеллект. Современный подход,* 2-е изд., Диалектика). Прекрасная (и большая) книга, в которой раскрыт невероятный объем вопросов, включая машинное обучение. Она помогает уловить общую картину машинного обучения.

Наконец, замечательный способ изучения предусматривает присоединение к веб-сайтам состязаний по машинному обучению, таким как Kaggle.com (https://www.kaggle.com/). Это 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, определенными контекстом.

Этот элемент содержит совет или указание.

Этот элемент содержит замечание общего характера.

Этот элемент содержит предупреждение или предостережение.

Использование примеров кода

Добавочные материалы (примеры кода, упражнения и т.д.) доступны для загрузки по адресу `https://github.com/ageron/handson-ml`.

Настоящая книга призвана помочь вам выполнять свою работу. Обычно, если в книге предлагается пример кода, 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 пример кода разрешения не требует. Для встраивания значительного объема примеров кода, 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 and TensorFlow* by Aurélien Géron (O'Reilly). Copyright 2017 Aurélien Géron, 978-1-491-96229-9".

Если вам кажется, что способ использования вами примеров кода выходит за законные рамки или упомянутые выше разрешения, тогда свяжитесь с нами по следующему адресу электронной почты: permissions@oreilly.com.

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 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