Об авторах

Себастьян Рашка, автор ставшего бестселлером 1-го издания этой книги, обладает многолетним опытом написания кода на языке Python. Он проводил многочисленные семинары по практическому применению науки о данных, машинному обучению и глубокому обучению, включая руководство по машинному обучению на SciPy – ведущей конференции, посвященной научным расчетам с помощью Python.

Несмотря на то что исследовательские проекты Себастьяна сосредоточены главным образом на решении задач в области вычислительной биологии, ему нравится писать и говорить на темы науки о данных, машинного обучения и языка Python в общем, и он стремится помочь людям разрабатывать решения, управляемые данными, без обязательного знания подоплеки машинного обучения.

Недавно его работа и вклад были отмечены званием выдающегося аспиранта 2016–2017, а также наградой ACM Computing Reviews' Best of 2016. В свободное время Себастьян любит участвовать в проектах с открытым кодом, а методы, которые он реализовал, теперь успешно используются в состязаниях по машинному обучению, таких как Kaggle.

Я хотел бы воспользоваться этой возможностью, чтобы поблагодарить замечательное сообщество Python и разработчиков пакетов с открытым кодом, которые помогли мне создать идеальную среду для научных исследований и науки о данных. Также я хочу поблагодарить своих родителей, всегда поощряющих и поддерживающих меня в выборе пути и занятия, в которое я страстно влюблен.

Выражаю особую благодарность основным разработчикам scikit-learn. Как участник этого проекта, я получил удовольствие от работы с прекрасными людьми, которые не только очень хорошо осведомлены в том, что касается машинного обучения, но также являются великолепными программистами. В заключение я хотел бы поблагодарить Эли Каверка, безвозмездно просмотревшего книгу и предоставившего ценный отзыв о новых главах.

Вахид Мирджалили получил звание PhD в машиностроении, работая над новаторскими методами для крупномасштабных вычислительных эмуляций молекулярных структур. В настоящее время он сосредоточил свою научно-исследовательскую работу на приложениях машинного обучения в разнообразных проектах компьютерного зрения в отделении компьютерных наук и инженерии Университета штата Мичиган.

Вахид избрал Python в качестве главного языка программирования, и на протяжении своей научно-исследовательской карьеры накопил громадный опыт в написании кода Python. Он преподавал программирование на Python инженерной группе в Университете штата Мичиган, что дало ему возможность помочь студентам понять разные структуры данных и разрабатывать эффективный код на Python.

Наряду с тем, что обширные исследовательские интересы Вахида сконцентрированы на приложениях глубокого обучения и компьютерного зрения, он особенно интересуется использованием приемов глубокого обучения для усиления приватности в биометрических данных, таких как изображения лиц, чтобы не раскрывалась информация сверх той, что пользователи намеревались показывать. Кроме того, он также сотрудничает с командой инженеров, работающих над беспилотными автомобилями, где проектирует модели на основе нейронных сетей для слияния многоспектральных изображений с целью обнаружения пешеходов.

Я хотел бы поблагодарить моего руководителя диссертации PhD, доктора Аруна Росса, за предоставленную мне возможность работать над новаторскими задачами в его исследовательской лаборатории. Я также выражаю благодарность доктору Вишну Боддети за то, что он разжег у меня интерес к глубокому обучению и прояснил его основные концепции.

О технических рецензентах

Джаред Хаффман – предприниматель, геймер, рассказчик, фанатик машинного обучения и страстный любитель баз данных. Последние 10 лет он посвятил себя разработке программного обеспечения и анализу данных. Его предыдущая работа охватывала разнообразные темы, включая безопасность сетей, финансовые системы и бизнес-аналитику, а также веб-службы, инструменты для разработчиков и методологию ведения бизнеса. В последнее время Джаред был основателем команды, занимающейся наукой о данных в Minecraft, с концентрацией на больших данных и машинном обучении. В свободное от работы время он обычно играет или наслаждается прекрасным Тихоокеанским Северо-Западом вместе с друзьями и семьей.

Я хотел бы поблагодарить издательство Packt за предоставленную мне возможность поработать с такой великолепной книгой, свою жену за постоянное ободрение и мою дочь за то, что она спала большую часть ночей, пока я пересматривал и отлаживал код.

Хуай Энь Сунь (**Райан Сунь**) получил степень магистра по статистике в Национальном университете Цзяотун в Тайване. В настоящее время он работает экспертом по аналитическим данным для линейки продуктов в PEGATRON. Главными областями его исследований являются машинное обучение и глубокое обучение.

Предисловие

Благодаря новостям и социальным медиаресурсам вам наверняка известен тот факт, что машинное обучение стало одной из самых захватывающих технологий нашего времени. Крупные компании, такие как Google, Facebook, Apple, Amazon и IBM, вполне обоснованно делают значительные инвестиции в исследования и приложения машинного обучения. Хотя может показаться, что сейчас машинное обучение превратилось в надоедливое словечко, это определенно не так. Сфера машинного обучения открывает путь к новым возможностям и становится незаменимой в повседневной жизни. Это очевидно из разговора с голосовым помощником в наших смартфонах, рекомендации правильного товара нашим заказчикам, препятствования мошенничеству с кредитными картами, фильтрации спама из почтовых ящиков, обнаружения и диагностирования медицинских заболеваний – список можно продолжать и продолжать.

Если вы хотите стать специалистом-практиком в области машинного обучения, лучше решать задачи или, может быть, даже подумываете о том, чтобы заняться исследованиями в этой сфере, тогда настоящая книга для вас. Однако для новичка теоретические концепции, лежащие в основе машинного обучения, могут оказаться непреодолимыми. В последние годы вышло много книг, ориентированных на практику, которые помогут начать работу с машинным обучением через реализацию мощных алгоритмов обучения.

Ознакомление с практическими примерами кода и проработка образцов приложений – замечательный способ погрузиться в эту сферу. Конкретные примеры помогают проиллюстрировать более широкие концепции, применяя изложенный материал сразу на практике. Тем не менее, помните о том, что большая мощь предполагает и большую ответственность! Кроме предложения практического опыта работы с машинным обучением, используя язык программирования Python и библиотеки на Python для машинного обучения, эта книга ознакомит вас с математическими концепциями, лежащими в основе алгоритмов машинного обучения, которые жизненно важны для

успешного применения машинного обучения. Следовательно, данная книга не является чисто практической; это книга, в которой обсуждаются необходимые детали, касающиеся концепций машинного обучения, а также предлагаются интуитивно понятные и вместе с тем информативные объяснения того, как работают алгоритмы машинного обучения, как их использовать, и что самое важное, как избежать распространенных ловушек.

Если вы введете поисковый термин “machine learning” (“машинное обучение”) в системе Google Scholar, то она возвратит ошеломляюще огромное количество публикаций – 4 180 000 (по состоянию на ноябрь 2018 года). Конечно, обсудить особенности всех алгоритмов и приложений, которые появились в последние 60 лет, попросту невозможно. Однако в книге мы совершим захватывающее путешествие, которое охватит все жизненно важные темы и концепции, чтобы дать вам хороший старт в освоении данной области. На тот случай, если вы обнаружите, что ваша жажда знаний не удовлетворена, в книге приводятся многочисленные ссылки на полезные ресурсы, которыми можно воспользоваться для отслеживания за выдающимися достижениями в этой сфере.

Если вы уже хорошо знаете теорию машинного обучения, тогда книга покажет, как применить имеющиеся знания на практике. Если вы использовали приемы машинного обучения ранее и хотите лучше понять, как фактически работает машинное обучение, то эта книга для вас. Не беспокойтесь, если вы – полный новичок в области машинного обучения; у вас даже еще больше поводов для заинтересованности. Есть надежда, что машинное обучение изменит ваше представление о задачах, подлежащих решению, и покажет, как взяться за них, высвободив всю мощь данных.

Прежде чем мы глубже погрузимся в сферу машинного обучения, давайте ответим на самый важный ваш вопрос: почему выбран язык Python? Ответ прост: он мощный и вдобавок очень доступный. Python стал наиболее популярным языком программирования для науки о данных, потому что позволяет нам забыть о скучных частях программирования и предлагает среду, где можно сделать быстрый набросок своих идей и сразу же привести их в действие.

Мы как авторы искренне заявляем, что исследование машинного обучения сделало нас лучшими учеными, мыслителями и решателями задач. В этой книге мы хотим поделиться с вами этим знанием. Знание добывается изучением. Ключом является наш энтузиазм, и подлинное мастерство вла-

дения навыками может быть достигнуто только практикой. Дорога впереди временами может быть ухабистой, а некоторые темы более сложными, чем другие, но мы надеемся на то, что вы воспользуетесь возможностью и сконцентрируетесь на вознаграждении. Не забывайте, что мы путешествуем вместе, и повсюду в книге мы будем добавлять в ваш арсенал многие мощные приемы, которые помогут решить даже самые трудные задачи в управляемой данными манере.

Что рассматривается в этой книге

В главе 1, “Наделение компьютеров способностью обучения на данных”, мы предложим введение в основные подобласти машинного обучения для решения разнообразных задач. Вдобавок в главе обсуждаются важные шаги для создания типовой модели машинного обучения путем построения конвейера, который проведет через последующие главы.

В главе 2, “Обучение простых алгоритмов МО для классификации”, мы обратимся к истокам машинного обучения, представив классификаторы на основе двоичного персептрона и адаптивных линейных нейронов. Глава является кратким введением в фундаментальные основы классификации образцов и сосредоточена на взаимодействии алгоритмов оптимизации и машинного обучения.

В главе 3, “Обзор классификаторов на основе машинного обучения с использованием scikit-learn”, описаны важные алгоритмы машинного обучения, предназначенные для классификации, и приведены практические примеры применения одной из самых популярных и всеобъемлющих библиотек машинного обучения с открытым кодом – scikit-learn.

В главе 4, “Построение хороших обучающих наборов с помощью предварительной обработки данных”, показано, как иметь дело с распространенными проблемами в необработанных наборах данных, такими как недостающие данные. В ней обсуждаются подходы к идентификации наиболее информативных признаков в наборах данных, а также объясняется, каким образом подготовить переменные разных типов с целью использования в качестве надлежащего входа для алгоритмов машинного обучения.

В главе 5, “Сжатие данных с помощью понижения размерности”, описаны важные приемы сокращения количества признаков в наборе данных с целью получения меньших наборов, которые все-таки сохраняют большинство по-

лезной и отличительной информации. Здесь обсуждается стандартный подход понижения размерности посредством анализа главных компонентов, а также проводится его сравнение с приемами линейной и нелинейной трансформации с учителем.

В главе 6, “Изучение практического опыта оценки моделей и настройки гиперпараметров”, обсуждаются правила для оценки эффективности прогнозирующих моделей. Кроме того, в ней описаны различные метрики для измерения эффективности моделей и приемы для точной настройки алгоритмов машинного обучения.

В главе 7, “Объединение разных моделей для ансамблевого обучения”, представлены концепции рационального объединения нескольких алгоритмов обучения. В ней объясняется, как построить ансамбль экспертов, чтобы преодолеть слабость индивидуальных учеников, вырабатывая в результате более точные и надежные прогнозы.

В главе 8, “Применение машинного обучения для смыслового анализа”, обсуждаются важные шаги по преобразованию текстовых данных в содержательные представления для алгоритмов машинного обучения, прогнозирующих мнения людей на основе их текстов.

В главе 9, “Встраивание модели машинного обучения в веб-приложение”, продолжается работа с прогнозирующей моделью из предыдущей главы, а также демонстрируются важные шаги разработки веб-приложений со встроенными моделями машинного обучения.

В главе 10, “Прогнозирование значений непрерывных целевых переменных с помощью регрессионного анализа”, описаны приемы моделирования линейных взаимосвязей между целевыми и объясняющими переменными для прогнозирования значений с непрерывным масштабом. После представления различных линейных моделей также обсуждаются подходы с полиномиальной регрессией и на основе деревьев.

В главе 11, “Работа с непомеченными данными – кластерный анализ”, внимание переходит на другую подобласть машинного обучения – обучение без учителя. Здесь применяются алгоритмы из трех фундаментальных семейств алгоритмов кластеризации для нахождения групп объектов, которые обладают определенной степенью подобия.

В главе 12, “Реализация многослойной искусственной нейронной сети с нуля”, представленная в главе 2 концепция оптимизации, основанная на градиентах, расширяется для построения мощных многослойных нейронных

сетей на базе популярного алгоритма обратного распространения с помощью Python.

В главе 13, “Распараллеливание процесса обучения нейронных сетей с помощью TensorFlow”, опираясь на материал предыдущей главы, предоставляется практическое руководство по более эффективному обучению нейронных сетей. Основное внимание в главе сосредоточено на TensorFlow – библиотеке Python с открытым кодом, которая позволяет задействовать множество ядер современных графических процессоров.

В главе 14, “Погружаемся глубже – механика TensorFlow”, библиотека TensorFlow рассматривается более подробно с объяснением основных концепций вычислительных графов и сеансов. Вдобавок в главе раскрываются такие темы, как сохранение и визуализация графов нейронных сетей, которые очень пригодятся в оставшихся главах книги.

В главе 15, “Классификация изображений с помощью глубоких сверточных нейронных сетей”, обсуждаются архитектуры глубоких нейронных сетей, которые стали новым стандартом в областях компьютерного зрения и распознавания изображений – сверточные нейронные сети. В главе рассматриваются главные концепции сверточных слоев как средств выделения признаков, а также демонстрируется применение архитектур сверточных нейронных сетей при решении задачи классификации изображений для достижения почти совершенной правильности классификации.

В главе 16, “Моделирование последовательных данных с использованием рекуррентных нейронных сетей”, представлена еще одна популярная архитектура нейронных сетей для глубокого обучения, которая особенно хорошо подходит при работе с последовательными данными и данными временных рядов. В главе к текстовым данным применяются различные архитектуры рекуррентных нейронных сетей. Сначала рассматривается задача смыслового анализа, а затем задача порождения полностью нового текста.

Что необходимо при работе с этой книгой

Выполнение примеров кода, приводимых в книге, требует установки Python 3.6.0 или более новой версии на машине с macOS, Linux или Microsoft Windows. В книге будут повсеместно использоваться важные библиотеки Python для научных расчетов, такие как SciPy, NumPy, scikit-learn, Matplotlib и pandas.

В главе 1 будут представлены инструкции и полезные советы по настройке среды Python и указанных основных библиотек. К имеющейся совокупности мы добавим дополнительные библиотеки, предоставляя инструкции по установке в соответствующих главах: библиотеку NLTK для обработки естественного языка (глава 8), веб-инфраструктуру Flask (глава 9), библиотеку Seaborn для визуализации статистических данных (глава 10) и библиотеку TensorFlow для эффективного обучения нейронных сетей на графических процессорах (главы 13–16).

Для кого предназначена эта книга

Если вы стремитесь выяснить, как использовать язык Python для получения ответов на критически важные вопросы, связанные с данными, тогда возьмите эту книгу – хотите вы начать с нуля или же расширить свои знания науки о данных, настоящая книга будет важным и незаменимым ресурсом.

Соглашения

Для представления различных видов информации в книге используется несколько стилей текста. Ниже приведен ряд примеров таких стилей с объяснениями, что они означают.

Фрагменты кода в тексте, имена таблиц баз данных, имена каталогов, имена и расширения файлов, имена путей, фиктивные URL и пользовательский ввод представляются в следующем виде: “За счет применения настройки `out_file=None` мы напрямую присваиваем данные точек переменной `dot_data` вместо записывания промежуточного файла `tree.dot` на диск”.

А вот как представляется блок кода:

```
>>> from sklearn.neighbors import KNeighborsClassifier
>>> knn = KNeighborsClassifier(n_neighbors=5, p=2,
...                            metric='minkowski')
>>> knn.fit(X_train_std, y_train)
>>> plot_decision_regions(X_combined_std, y_combined,
...                       classifier=knn, test_idx=range(105,150))
>>> plt.xlabel('petal length [standardized]')
>>> plt.ylabel('petal width [standardized]')
>>> plt.show()
```

Любой ввод или вывод в командной строке записывается так:

```
pip3 install graphviz
```

Новые термины и **важные слова** выделяются *курсивом*. Слова, которые отображаются на экране, например, в меню или диалоговых окнах, выделяются следующим образом: “По щелчку на ссылке Dashboard (Инструментальная панель) в правом верхнем углу мы получаем доступ к панели управления”.

Здесь приводятся предостережения и важные примечания.

Здесь даются советы и трюки.

Загрузка кода примеров

Исходный код всех примеров, рассмотренных в книге, доступен для загрузки на веб-сайте издательства и по ссылке `https://github.com/PacktPublishing/Python-Machine-Learning-Second-Edition`. После загрузки распакуйте файл с помощью последней версии архиваторов:

- WinRAR / 7-Zip для Windows;
- Zipeg / iZip / UnRarX для Mac;
- 7-Zip / PeaZip для Linux.

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