УДК 004.4
ББК 32.973.2-018.2
Ф42

Mark Fenner
Machine Learning with Python for Everyone

Authorized translation from the English language edition,
entitled MACHINE LEARNING WITH PYTHON FOR EVERYONE,
1st Edition; ISBN 0134845625; by FENNER, MARK;
published by Pearson Education, Inc.,
publishing as Addison-Wesley Professional.

RUSSIAN language edition published by EKSMO Publishing House.

Феннер, Марк.

Ф42 Машинное обучение с помощью Python для всех : руководство по созданию систем машинного обучения: от основ до мощных инструментов / Марк Феннер ; [перевод с английского М. А. Райтмана]. — Москва : Эксмо, 2024. — 672 с.— (Мировой компьютерный бестселлер).

ISBN 978-5-04-187899-3

Машинное обучение затронуло практически все отрасли жизни. Для него используются многие языки программирования, но наиболее популярным является Python. Несмотря на то что в основе машинного обучения обычно лежат продвинутые математические концепции, обучать сложные модели, не имея глубоких математических знаний, вполне возможно.

Эта книга представляет собой руководство, доступное для понимания любому человеку, что позволяет при любом уровне подготовки быстро улучшить знания и навыки в области машинного обучения, освоить некоторые сложные и интересные методы.

УДК 004.4
ББК 32.973.2-018.2

ISBN 978-5-04-187899-3

Посвящается моему сыну Итану —
с вечной надеждой на лучшее завтра

Предисловие

Вне зависимости от того, называется это статистикой, наукой о данных, машинным обучением или искусственным интеллектом, изучение закономерностей на основе данных меняет наш мир. Машинное обучение уже затронуло (или вот-вот затронет) практически все отрасли. Стремительное совершенствование аппаратного и программного обеспечения способствует быстрому прогрессу в этой области, однако основное внимание большинства людей сосредоточено именно на ПО.

Для машинного обучения используются многие языки программирования, включая R, C/C++, Fortran и Go, но наиболее популярным является Python. Своей известностью он во многом обязан библиотеке scikit-learn, позволяющей не только с легкостью обучать множество различных моделей, но и конструировать признаки, оценивать качество модели и проверять ее работу на новых данных. Проект scikit-learn быстро превратился в одну из важнейших и мощнейших библиотек Python.

Несмотря на то что в основе машинного обучения лежат продвинутые математические концепции, обучать сложные модели, не имея глубоких знаний в области исчисления и матричной алгебры, вполне возможно. Многим людям гораздо проще подойти к машинному обучению через программирование, чем через математику. Именно поэтому в данной книге я использую код Python в качестве основы, добавляя математические подробности по мере необходимости. Подобно книгам R for Everyone и Pandas for Everyone, книга «Машинное обучение с помощью Python для всех» представляет собой руководство, доступное для понимания любому человеку, желающему познакомиться с захватывающей областью математики и вычислений.

Марк Феннер многие годы занимается объяснением научных концепций и принципов машинного обучения людям разного уровня подготовки, оттачивая свою способность раскладывать сложные идеи на простые компоненты. Этот опыт позволяет ему объяснять непростые вещи на конкретных примерах, сводя к минимуму использование технического жаргона. Благодаря этому книгу легко читать, кроме того, она содержит множество примеров кода, с которыми читатель может работать на своем компьютере.

Учитывая стремительный рост числа желающих освоить и внедрить машинное обучение, важно разработать практические ресурсы, помогающие людям сделать это как можно быстрее и правильнее. Содержательная и увлекательная книга Марка является как раз таким ресурсом. «Машинное обучение с помощью Python для всех» полностью оправдывает свое название, позволяя людям любого уровня подготовки быстро улучшить знания и навыки в области машинного обучения и делая эту важную сферу гораздо более доступной.

Джаред Ландер,
редактор серии

Оглавление

Предисловие 6
Вступление 11
Аудитория 11
Подход к изложению материала 12
Структура книги 13
Благодарности 14
Об авторе 16

Часть I. Первые шаги

Глава 1. Поговорим о процессе обучения 19
1.1. Добро пожаловать 19
1.2. Сфера применения, терминология, прогнозирование и данные 20
1.3. Включение машины в процесс машинного обучения 25
1.4. Примеры обучающихся систем 27
1.5. Оценка обучающихся систем 30
1.6. Процесс создания обучающихся систем 33
1.7. Допущения и реальность обучения 36
1.8. Подведение итогов 38
Глава 2. Основные технические сведения 41
2.1. Настройка среды программирования 41
2.2. Потребность в математическом языке 41
2.3. Программное обеспечение для машинного обучения 43
2.4. Вероятность 43
2.5. Линейные комбинации, взвешенные суммы и скалярные произведения 52
2.6. Геометрический взгляд на вещи: точки в пространстве 61
2.7. Нотация и прием «плюс один» 71
2.8. Добавление крутизны, избавление от смирительной рубашки и нелинейность 73
2.9. NumPy против «всей этой математики» 75
2.10. Проблемы с плавающей запятой 82
2.11. Подведение итогов 84
Глава 3. Предсказание категорий: знакомство с классификацией 86
3.1. Задачи классификации 86
3.2. Простой набор данных для задач классификации 87

3.3. Тренировка и тестирование: избегайте натаскивания ... 90
3.4. Оценка: подсчет результатов экзамена ... 94
3.5. Простой классификатор № 1: ближайшие соседи, отношения на расстоянии и допущения ... 96
3.6. Простой классификатор № 2: наивный байесовский алгоритм, вероятность и нарушенные обещания ... 102
3.7. Упрощенная оценка классификаторов ... 105
3.8. Подведение итогов ... 119
Глава 4. Предсказание числовых значений: знакомство с регрессией ... 124
4.1. Простой набор данных для задач регрессии ... 124
4.2. Регрессия методом *k*-ближайших соседей и сводная статистика ... 127
4.3. Линейная регрессия и ошибки ... 132
4.4. Оптимизация: выбор лучшего ответа ... 139
4.5. Простая оценка и сравнение регрессоров ... 144
4.6. Подведение итогов ... 148

Часть II. Оценка

Глава 5. Оценка и сравнение обучающихся систем ... 153
5.1. Оценка и почему меньше значит больше ... 153
5.2. Терминология для описания этапов обучения ... 155
5.3. Майор Том, что-то не так: переобучение и недообучение ... 163
5.4. От ошибок к затратам ... 172
5.5. (Повторная) выборка: получение большего за счет меньшего ... 176
5.6. Разложение ошибки на смещение и дисперсию ... 195
5.7. Графический метод оценки и сравнения ... 204
5.8. Сравнение моделей с помощью перекрестной проверки ... 209
5.9. Подведение итогов ... 211
Глава 6. Оценка классификаторов ... 215
6.1. Базовые классификаторы ... 215
6.2. Помимо точности: показатели эффективности классификации ... 218
6.3. ROC-кривые ... 229
6.4. Еще один метод многоклассовой классификации: один против одного ... 243
6.5. Кривые прецизионность — полнота ... 248
6.6. Кумулятивный отклик и Lift-кривые ... 251
6.7. Более сложный способ оценки классификаторов: дубль два ... 254
6.8. Подведение итогов ... 266
Глава 7. Оценка регрессоров ... 270
7.1. Базовые регрессоры ... 270

7.2. Дополнительные метрики для оценки регрессоров ... 272
7.3. Графики остатков ... 282
7.4. Знакомство с понятием стандартизации ... 290
7.5. Более сложный способ оценки регрессоров: дубль два ... 295
7.6. Подведение итогов ... 302

Часть III. Дополнительные методы и основные сведения

Глава 8. Дополнительные методы классификации ... 309
8.1. Переосмысление процесса классификации ... 309
8.2. Деревья решений ... 312
8.3. Классификаторы на основе метода опорных векторов ... 323
8.4. Логистическая регрессия ... 335
8.5. Дискриминантный анализ ... 347
8.6. Допущения, смещения и классификаторы ... 366
8.7. Сравнение классификаторов: дубль три ... 369
8.8. Подведение итогов ... 372
Глава 9. Дополнительные методы регрессии ... 378
9.1. Линейная регрессия на скамейке штрафников: регуляризация ... 379
9.2. Регрессия методом опорных векторов ... 386
9.3. Кусочно-постоянная регрессия ... 394
9.4. Деревья регрессии ... 401
9.5. Сравнение регрессоров: дубль три ... 403
9.6. Подведение итогов ... 406
Глава 10. Ручное конструирование признаков:
манипулирование данными ради удовольствия и прибыли ... 409
10.1. Конструирование признаков: терминология и мотивация ... 409
10.2. Отбор признаков и сокращение объема данных:
избавление от мусора ... 414
10.3. Масштабирование признаков ... 415
10.4. Дискретизация ... 420
10.5. Кодирование категориальных переменных ... 423
10.6. Зависимости и взаимодействия ... 434
10.7. Манипуляции с целевыми переменными ... 445
10.8. Подведение итогов ... 451
Глава 11. Настройка гиперпараметров и конвейеры ... 455
11.1. Модели, параметры, гиперпараметры ... 456
11.2. Настройка гиперпараметров ... 458
11.3. Прыжок в рекурсивную кроличью нору:
вложенная перекрестная проверка ... 468

11.4. Конвейеры ... 477
11.5. Конвейеры и совместная настройка гиперпараметров ... 480
11.6. Подведение итогов ... 482

Часть IV. Добавление сложности

Глава 12. Объединение моделей ... 487
12.1. Ансамбли ... 487
12.2. Голосующие ансамбли ... 490
12.3. Бэггинг и случайные леса ... 491
12.4. Бустинг ... 500
12.5. Сравнение ансамблевых методов на основе деревьев решений ... 505
12.6. Подведение итогов ... 509
Глава 13. Модели, которые конструируют признаки за нас ... 513
13.1. Отбор признаков ... 516
13.2. Построение признаков с помощью ядер ... 537
13.3. Анализ главных компонент: обучение без учителя ... 557
13.4. Подведение итогов ... 579
Глава 14. Конструирование признаков для предметных областей: предметно-ориентированное обучение ... 587
14.1. Работа с текстом ... 588
14.2. Кластеризация ... 599
14.3. Работа с изображениями ... 601
14.4. Подведение итогов ... 616
Глава 15. Взаимосвязи, расширения и дальнейшие направления ... 620
15.1. Оптимизация ... 620
15.2. Построение модели линейной регрессии из примитивных компонентов ... 624
15.3. Построение модели логистической регрессии из примитивных компонентов ... 629
15.4. Построение SVM из примитивных компонентов ... 636
15.5. Нейронные сети ... 637
15.6. Графовые вероятностные модели ... 644
15.7. Подведение итогов ... 654

Приложение А. Листинг модуля mlwpy.py ... 657
Предметный указатель ... 665

Вступление

В 1983 году на экраны вышел фильм «Военные игры». Меня, уже подростка, абсолютно потрясли возможность ядерного апокалипсиса, почти магический способ взаимодействия главного героя с компьютерными системами, но больше всего — потенциал машин, способных *учиться*. Я потратил годы на изучение стратегических ядерных арсеналов Востока и Запада (к счастью, подойдя к этому с почти детской наивностью), но до моих первых серьезных опытов в компьютерном программировании прошло почти десять лет. Обучать компьютер выполнению конкретных задач оказалось здорово. Изучение тонкостей сложных систем и их использование в соответствии с собственными потребностями стало для меня отличным опытом. Тем не менее мне еще предстояло сделать большой шаг вперед. Несколько лет спустя я начал работать над своей первой *обучающейся* программой. Именно тогда я понял, что нашел интеллектуальное пристанище. Теперь я приглашаю вас в мир *компьютерных программ, способных учиться*.

Аудитория

Я написал книгу «Машинное обучение с помощью Python для всех» для абсолютных новичков в сфере машинного обучения. Более того, если вы не обладаете продвинутыми математическими знаниями, *я даже не попытаюсь это изменить*. Многие книги по машинному обучению переполнены математическими концепциями и уравнениями, но я сделал все возможное, чтобы свести их количество к *минимуму*. Однако, учитывая название, я ожидаю от вас обладания некоторыми базовыми навыками работы с Python. Умея *читать* код на этом языке, вы сможете извлечь из наших обсуждений гораздо больше пользы. В то время как многие авторы книг по машинному обучению полагаются на математику, я стараюсь объяснять материал с помощью историй, изображений и кода на языке Python. Кое-где в тексте *будут* встречаться уравнения, которые в большинстве случаев при желании можно пропустить. Однако, если я хорошо справился со своей задачей, вы получите достаточно контекста, чтобы понять, о чем они говорят.

По какой причине эта книга оказалась у вас в руках? Наиболее вероятный ответ на этот вопрос: вы хотите *узнать больше* о машинном обучении. Возможно, вы студент вводного курса, посвященного машинному обучению,

бизнес-аналитик, рабочие обязанности которого уже не могут ограничиваться анализом содержимого электронных таблиц, любитель, желающий расширить свои знания, или ученый, ищущий новый способ анализа данных. Машинное обучение постепенно затрагивает все больше аспектов жизни общества. В зависимости от имеющегося опыта вы найдете в этой книге что-то свое. Даже искушенный в математике читатель, желающий освоить принципы машинного обучения с помощью Python, может многое из нее почерпнуть.

Итак, моя цель — объяснить читателю *процесс* и наиболее важные *концепции* машинного обучения на конкретных примерах, используя scikit-learn и некоторые другие библиотеки Python. В ходе чтения вы познакомитесь с наиболее распространенными паттернами, стратегиями, ловушками и подводными камнями, применимыми к любой обучающейся системе, которую вам когда-либо придется исследовать, создавать или использовать.

Подход к изложению материала

Многие авторы объясняют математические темы вроде машинного обучения с помощью уравнений, представляя их так, будто рассказывают историю непосвященным. По-моему, такой подход ставит в тупик многих — даже любителей математики! Лично мне гораздо проще представить процесс машинного обучения, объединив визуальные и словесные описания с примерами *работающего кода*. Я компьютерщик в душе и по образованию. Я люблю создавать вещи — именно это позволяет мне *по-настоящему* их понять. Возможно, вы слышали фразу: «Если вы действительно хотите в чем-то разобраться, объясните это кому-нибудь другому». Существует и другая ее версия: «Если вы действительно хотите в чем-то разобраться, научите это делать компьютер!» Именно такой подход я и собираюсь использовать при объяснении концепций машинного обучения. Опишу наиболее важные и часто используемые инструменты и методы машинного обучения, обходясь минимумом математических подробностей. Более того, вы сразу же сможете применить знания на практике. Однако, вместо того чтобы создавать все программы с нуля, мы встанем на плечи гигантов и применим несколько очень мощных, экономящих время программных библиотек (о них немного позднее).

Мы не станем подробно изучать все эти библиотеки — материала слишком много. Вместо этого проявим практичность и воспользуемся самым подходящим инструментом для решения конкретной задачи. Я объясню достаточно, чтобы познакомить вас с той или иной концепцией, а затем мы сразу перейдем к ее применению. Для математически подкованных читателей я дам ссылки

на более подробные материалы, к которым они смогут обратиться. Бóльшую их часть я приведу в конце глав, чтобы остальные могли с легкостью пропустить лишнюю для себя информацию.

Если вы еще не решили, стоит ли тратить время на чтение этой книги, я дам вам некоторое представление о темах, обсуждение которых выходит за ее рамки. Мы не станем рассматривать математические доказательства и использовать математику для объяснения различных концепций. Все это содержится во множестве других книг, и в конце глав вы найдете ссылки на те из них, которые я могу рекомендовать. Я предполагаю, что вы владеете навыком программирования на языке Python хотя бы на начальном или среднем уровне. Что касается более сложных тем и задач, решение которых требует использования таких сторонних пакетов, как NumPy или Pandas, я объясню достаточно, чтобы позволить вам разобраться с каждым методом и контекстом его применения.

Структура книги

В **части I** мы заложим фундамент. В главе 1 вы найдете несколько словесных и концептуальных описаний процесса машинного обучения. В главе 2 мы рассмотрим несколько математических и вычислительных концепций, которые часто применяются в этой области. Главы 3 и 4 помогут вам сделать первые шаги в построении, тренировке и оценке обучающихся систем, которые классифицируют примеры (классификаторы) и количественно их оценивают (регрессоры).

В **части II** акцент нашего внимания сместится на наиболее важный аспект прикладных систем машинного обучения — на реалистичную оценку их эффективности. В главе 5 мы поговорим об общих методах оценки, применимых ко всем обучающимся системам. В главах 6 и 7 к этим общим методам добавятся способы оценки классификаторов и регрессоров.

В **части III** мы расширим набор методов обучения и подробнее рассмотрим компоненты обучающейся системы. В главах 8 и 9 представлены дополнительные методы классификации и регрессии. В главе 10 мы обсудим *конструирование признаков*, то есть как придать данным форму, пригодную для использования в процессе обучения. В главе 11 показано, как объединить несколько компонентов в единую обучающуюся систему и настроить ее внутренние механизмы для повышения производительности.

В **части IV** мы выйдем за рамки основ и обсудим современные методы, продвигающие область машинного обучения вперед. В главе 12 рассмотрим обучающиеся системы, состоящие из нескольких аналогичных систем меньшего размера. В главе 13 обсудим методы обучения, предполагающие автоматическое

конструирование признаков. В главе 14, представляющей собой своеобразный краеугольный камень, применим описанные в книге методы к двум особенно интересным типам данных: изображениям и тексту. В главе 15 вы найдете краткий обзор многих рассмотренных ранее методов и объяснение их соотношения с такими более сложными архитектурами, как нейронные сети и графические модели.

В этой книге основное внимание уделяется методам машинного обучения. Попутно мы рассмотрим ряд алгоритмов обучения и других методов обработки данных. Однако при этом не станем стремиться к полноте изложения. Мы обсудим наиболее распространенные методы и лишь кратко коснемся двух больших подразделов машинного обучения: графических моделей и глубоких нейронных сетей. Тем не менее мы узнаем, как рассмотренные методы соотносятся с этими более продвинутыми концепциями.

Еще одна тема, которую мы не затронем, — это реализация конкретных алгоритмов обучения. Мы поработаем на основе алгоритмов, уже доступных в scikit-learn и других библиотеках, используя их в качестве компонентов более крупных решений. Тем не менее кто-то же отвечает за механизм черного ящика, обрабатывающего наши данные. Если вас действительно интересуют аспекты его реализации, то вы в хорошей компании, потому что я их просто обожаю! Порекомендуйте друзьям купить экземпляр этой книги, чтобы я мог выпустить продолжение с описанием этих низкоуровневых деталей.

Благодарности

Хочу выразить благодарность нескольким людям, которые внесли большой вклад в реализацию данного проекта. Я очень признателен своему редактору в издательстве Pearson, Дебре Уильямс Коли, помогавшей мне на каждом этапе работы над этой книгой. Она продемонстрировала высочайший уровень профессионализма как во время наших первоначальных встреч и при поиске темы, удовлетворяющей нас обоих, так и в процессе моей работы над многими (многими!) черновиками, в ходе которой она мягко подталкивала меня, заставляя двигаться вперед и карабкаться к вершине по самым крутым склонам. За все это я сердечно ее благодарю.

Моя жена, доктор Барбара Феннер, также заслуживает большей похвалы и благодарности, чем я могу выразить ей в этом кратком разделе. В дополнение к бремени, которое приходится нести партнеру любого автора, она также являлась главным читателем моих черновиков *и* нашим бесстрашным иллюстратором. Она проделала огромную работу по составлению всех диаграмм, которые

не получалось сгенерировать на компьютере. Несмотря на то что эта книга не является нашим первым совместным академическим проектом, он оказался самым продолжительным. За время работы я пришел к выводу, что терпению жены буквально нет предела. Спасибо тебе, Барбара!

Мой главный технический корректор Мэрилин Рот с неизменной благожелательностью исправляла мои ошибки, даже самые вопиющие. Благодаря ее вкладу книга «Машинное обучение с помощью Python для всех» стала неизмеримо лучше. *Спасибо*.

Я также хочу поблагодарить нескольких сотрудников редакционной коллегии издательства Pearson: Алину Кирсанову и Дмитрия Кирсанова, Джули Нахил и многих других людей, с которыми я не имел удовольствия познакомиться. Эта книга не вышла бы в свет без вас, вашего трудолюбия и профессионализма. *Спасибо вам*.

Об авторе

Марк Феннер с 1999 года преподает информатику и математику взрослым людям, от первокурсников до ветеранов отрасли. Параллельно он проводит исследования в области машинного обучения, биоинформатики и компьютерной безопасности. В рамках своих проектов он занимался разработкой, реализацией и повышением производительности систем машинного обучения и численных алгоритмов, анализом безопасности репозиториев программного обеспечения, созданием обучающихся систем для обнаружения аномального поведения пользователей, вероятностным моделированием функции белка, а также анализом и визуализацией экологических и микроскопических данных. Он обожает информатику и математику, историю и экстремальные виды спорта. Когда он не пишет книги, не преподает и не программирует, он катается по лесу на горном велосипеде или потягивает пиво у бассейна. Марк является обладателем второго дана по дзюдо и сертифицированным специалистом по оказанию экстренной помощи в условиях дикой природы. Он и его супруга закончили Аллегейни-колледж и Питтсбургский университет. Марк имеет степень доктора философии в области компьютерных наук. Он живет на северо-востоке Пенсильвании со своей семьей и работает в собственной компании Fenner Training and Consulting, LLC.

Часть I
Первые шаги

Глава 1. Поговорим о процессе обучения
Глава 2. Основные технические сведения
Глава 3. Предсказание категорий: знакомство с классификацией
Глава 4. Предсказание числовых значений: знакомство с регрессией