Введение

Понимание мира вокруг нас требует сбора и анализа данных об окружающей среде. Объединение последних технологических тенденций предоставляет новые возможности для применения анализа данных к более сложным задачам, чем когда-либо прежде.

Емкость компьютерных хранилищ увеличивается экспоненциально; хранение данных сейчас стало настолько дешевым, что компьютерным системам почти невозможно ничего забыть. Сенсорные устройства все шире и шире контролируют все, за чем только можно наблюдать: потоки видео, действия в социальных сетях и местоположение всего, что перемещается. Сетевая вычислительная среда позволяет использовать огромные количества машин для манипулирования этими данными. Каждый раз, когда вы осуществляете поиск в Google, задействуются сотни компьютеров, тщательно исследующие все ваши предыдущие действия, только для того, чтобы решить, какая реклама является наилучшей для демонстрации именно вам.

Результатом всего этого стало рождение *науки о данных* (data science) — новой области, посвященной максимизации значения обширных коллекций информации. Как дисциплина наука о данных находится где-то на пересечении статистики, информатики и машинного обучения, но стоит она отдельно, как самостоятельный персонаж. Эта книга служит введением в науку о данных, сосредоточиваясь на навыках и принципах, необходимых для построения систем, предназначенных для анализа и интерпретации данных.

Моя профессиональная практика как исследователя и преподавателя убедила меня в том, что одной из главных сложностей науки о данных является то, что она значительно сложней, чем выглядит. Любой студент, когда-либо вычислявший свой *средний балл успеваемости* (grade point average — GPA), может сказать, что выполнял элементарный статистический расчет, а рисование простого графика разброса позволит вам добавить в свое резюме упоминание о наличии опыта в визуализации данных. Однако реальный анализ и интерпретация данных требуют и технических знаний, и мудрости. Основами обладает очень много людей, но не техническими знаниями, что и вдохновило меня на написание этой книги.

Для кого написана эта книга

Меня порадовал теплый прием моей предыдущей книги, *The Algorithm Design Manual* [1], впервые опубликованной в 1997 году. Она была признана уникальным руководством по использованию алгоритмических методик для решения задач, которые нередко возникают на практике. Книга, которую вы держите в руках, — это совершенно другой материал, но предназначенный для решения тех же задач.

В частности, здесь я подчеркиваю следующие принципы, необходимые, чтобы стать хорошим аналитиком данных.

- *Главное — делать простые вещи правильно*. Наука о данных — это не ракетостроение. Студенты и практики зачастую теряются в технологических вопросах, пытаясь применить наиболее передовые методы машинного обучения, новейшие библиотеки программного обеспечения, реализации с открытым исходным кодом или шикарные методики визуализации. Однако наука о данных призвана делать правильно простые вещи: понимать область применения, очищать и интегрировать корректные источники данных, а также доходчиво представлять ваши результаты другим.

  Как бы то ни было, просто — не значит легко. Действительно, нужна существенная проницательность и опыт, чтобы задавать правильные вопросы, и ум, чтобы двигаться к правильным ответам и действенным решениям. Я противостою искушению излишне углубиться в чисто технический материал только потому, что он и так доступен. Есть много других книг, которые рассматривают сложности алгоритмов машинного обучения или статистической проверки гипотез. Моя цель — заложить основы того, что действительно имеет значение в анализе данных.

- *Думайте как программист, но действуйте как статистик*. Наука о данных подобна зонтику, объединяющему программистов, статистиков и специалистов в некой области. Однако у каждого сообщества есть его собственные специфические стили мышления и действия, которые становятся шаблонами для его членов.

  В этой книге я излагаю подходы, которые наиболее естественны для программистов, в частности алгоритмы манипулирования данными, использование машинного обучения и мастерство масштабирования. Но я также пытаюсь передавать основы статистического рассуждения: потребность понимать область применения, надлежащая оценка малого, поиск значения и жажда исследования.

> Ни у какой дисциплины нет монополии на правду. Лучшие аналитики данных объединяют инструментальные средства из нескольких областей, и эта книга представляет относительно нейтральную территорию, где конкурирующие философии вполне могут обсуждаться вместе.

Не менее важным является то, чего вы не найдете в этой книге. Я не рассматриваю специфические языки или комплекты инструментальных средств анализа данных. Вместо этого здесь обсуждаются важные принципы проектирования. Я постараюсь остаться скорее на концептуальном уровне, чем на техническом. Задача этого руководства в том, чтобы направить вас по правильному пути как можно быстрее, вне зависимости от конкретных программных инструментальных средств, которые вы находите наиболее подходящими для себя.

Для преподавателей

Эта книга содержит достаточно материала курса *Введение в науку о данных* для студентов младших или даже старших курсов. Я надеюсь, что читатель закончил по крайней мере один курс программирования и имеет хоть немного предварительного опыта работы с вероятностью и статистикой, но всегда лучше больше, чем меньше.

Я подготовил полный набор слайдов для лекций, чтобы сделать этот курс более доступным для дистанционного обучения, см. `http://www.data-manual.com`. Там же приводятся ресурсы для домашних заданий и проектов. Кроме того, в сети доступны мои видеолекции курса науки о данных для полного семестра.

Эта книга включает следующие педагогические средства.

- *Случай из жизни*. Чтобы продемонстрировать, как методы науки о данных применяются в реальном мире, я включил в книгу разделы “Случай из жизни”, содержащие рассказы о нашей практике с реальными проблемами. Мораль этих историй в том, что рассматриваемые методы — не просто теория, а полезные инструменты.
- *Фальстарты*. Большинство методов в этой книге представлены в готовом виде, за кадром остались идеи, возникавшие в ходе их разработки, и причины, по которым другие подходы потерпели неудачу. Правдивые истории иллюстрируют процесс рассуждений при решении некоторых практических задач и описывают материал, лежащий в их основе.
- *На заметку*. Фрагменты выделенного этим заголовком текста встречаются в каждой главе и подчеркивают наиболее общие концепции, которые стоит усвоить при изучении данной главы.

- *Домашняя работа*. Я предоставляю широкий диапазон различных упражнений для домашней работы и самообучения. Многие задачи представлены в традиционном стиле, но есть также задания на крупномасштабную реализацию и вопросы меньшего масштаба, как на интервью при поиске работы студентами. Всем задачам присвоена степень сложности.

  Вместо ключей к ответам, были установлены Solution Wiki, где решения всех пронумерованных задач будут востребованы краудсорсингом. Подобная система в моей книге *Algorithm Design Manual* позволила выработать сбалансированные решения, по крайне мере мне так говорят. Я отказываюсь смотреть их из принципа, так что покупателю стоит остерегаться.
- *Конкурсы Kaggle*. Kaggle (`www.kaggle.com`) поддерживает форум для аналитиков данных, позволяя им соревноваться в решении сложных реальных задач на великолепных наборах данных и проверять, насколько хороша ваша модель относительно других. Упражнения каждой главы включают три подходящих конкурса Kaggle, способных послужить источником вдохновения, материалом для самообучения, а также данными для других проектов и исследований.
- *Телевидение науки о данных*. Для широкой общественности наука о данных остается таинственной и даже зловещей. Любительское телевизионное шоу *The Quant Shop* демонстрирует то, чем наука о данных должна быть в действительности. Студенческие группы занимаются разнообразными задачами прогнозирования реальных проблем и пытаются предсказывать результат будущих событий. Ознакомьтесь с этим по адресу `http://www.quant-shop.com`.

  Была подготовлена серия из восьми 30-минутных эпизодов, каждый из которых посвящен конкретной реальной проблеме прогнозирования. Рассматриваются цены произведений искусства на аукционе, выбор победителя на конкурсе Мисс Вселенная и предсказание продолжительности жизни знаменитостей. В каждом случае мы наблюдаем, как группа студентов справляется с задачей, и вместе с ними учимся строить модель прогноза. Они делают свои прогнозы, а мы, глядя на них, видим, оказываются они правы или нет.

  В этой книге я использую передачу *The Quant Shop* для предоставления конкретных примеров сложностей прогнозирования и обсуждения науки о данных, моделируя последовательность от сбора данных до вычисления оценки. Надеюсь, что вы найдете эти примеры интересными и они сподвигнут вас на создание своих собственных конкурсов по моделированию.

- *Дополнительная информация*. Каждая глава книги завершается разделом кратких заметок, рекомендующих читателям основные первоисточники и дополнительные ссылки.

Посвящение

Моим умным и любящим дочерям Бонни и Эбби, теперь уже настоящим подросткам, в том смысле, что они не всегда рассматривают статистическое доказательство с такой живостью, как мне бы хотелось. Я посвящаю эту книгу им в надежде, что их аналитические навыки улучшатся до такой степени, что они всегда будут только соглашаться со мной.

Кроме того, я посвящаю эту книгу своей прекрасной жене Рене, которая соглашается со мной, даже когда она не согласна, и любит меня без всяких заслуживающих доверия доказательств.

Благодарности

Мой список заслуживающих благодарности людей достаточно велик, хотя некоторых я, вероятно, пропустил. Я попытаюсь перечислить их систематически, чтобы минимизировать пропуски, но заранее прошу прощения у тех, кого не упомянул по невнимательности.

В первую очередь, я благодарю тех, кто оказал мне конкретную помощь в компоновке этой книги. *Есыль Ли* (Yeseul Lee) была на этом проекте практикантом, она помогала мне с рисунками, упражнениями и другим на протяжении лета 2016 года. Вы будете видеть доказательство работы ее рук почти на каждой странице, и я очень ценю ее помощь. Аакрити Миттал (Aakriti Mittal) и Джек Зхенг (Jack Zheng) также поспособствовали созданию нескольких рисунков.

Студенты моего курса *Introduction to Data Science* (CSE 519) помогли откорректировать эту рукопись, а также нашли множество проблемных мест. Я особенно благодарю Ребекку Сифорд (Rebecca Siford), которая предложила более ста исправлений. Мои друзья и коллеги по науке о данных, Аншул Ганди (Anshul Gandhi), Ю Фан Ху (Yifan Hu), Клаус Мюллер (Klaus Mueller), Франческо Орабона (Francesco Orabona), Энди Шварц (Andy Schwartz) и Чарльз Уорд (Charles Ward), отрецензировали несколько моих глав, и я благодарю их за усилия.

Я признателен всем студентам *The Quant Shop* из Fall 2015, видео и усилия которых по моделированию и так видны на экране. Большое спасибо Жанне (Дине) Дискин-Циммерман (Jan (Dini) Diskin-Zimmerman), чьи редакторские усилия не

входили в ее служебные обязанности, я чувствовал себя просто преступником за то, что позволил ей делать это.

Как обычно, работать с моими редакторами из Springer — Уэйном Виллером (Wayne Wheeler) и Саймоном Риисом (Simon Rees) — было просто удовольствием. Я также благодарю весь производственный и маркетинговый персонал, который помог предоставить эту книгу вам, особенно Адриана Пьерона (Adrian Pieron) и Аннетт Анлауф (Annette Anlauf).

Некоторые упражнения были предложены коллегами или созданы по мотивам других источников. Восстановить первоначальные источники через несколько лет очень сложно, но благодарности за каждую задачу (по моей памяти) приведены на веб-сайте.

Большая часть моих знаний о науке о данных была получена в ходе работы с другими людьми. К ним относятся мои аспиранты, в частности Рами аль-Рфоу (Rami al-Rfou), Михаил Баутин (Mikhail Bautin), Чи-Хао Чен (Haochen Chen), Яньцин Чен (Yanqing Chen), Вивек Кулкарни (Vivek Kulkarni), Левон Ллойд (Levon Lloyd), Андрей Мехлер (Andrew Mehler), Брайан Пероззи (Bryan Perozzi), Йингтао Тьян (Yingtao Tian), Юн Тинг Йе (Junting Ye), Венбин Чжан (Wenbin Zhang) и докторант Чарльз Уорд. Я с нежностью вспоминаю всех студентов моего проекта Lydia за эти годы и напоминаю, что предложенный мною приз первому, кто назовет свою дочь Лидией, остается невостребованным. Я благодарю других своих сотрудников за рассказанные истории, в частности Брюса Фатчера (Bruce Futcher), Джастина Гардина (Justin Gardin), Арно ван де Рийта (Arnout van de Rijt) и Алексея Старова (Oleksii Starov).

Я помню всех членов мира General Sentiment/Canrock, особенно Марка Фасциано (Mark Fasciano), с которым я разделил мечту о новом и испытал то, что случается, когда данные попадают в реальный мир. Я благодарю коллег по Yahoo Labs/Research, работавших со мной в 2015-2016 годах, когда была задумана б*о*льшая часть этой книги. Я вспоминаю Аманду Стент (Amanda Stent), благодаря которой мне довелось быть в компании Yahoo на протяжении того особенно трудного года в ее истории. Я узнал очень ценные вещи от других людей, которые вели курсы, связанные с наукой о данных, включая Эндрю Ына (Andrew Ng) и Ханса-Петера Фистера (Hans-Peter Pfister). Я благодарю их всех за помощь.

Если у вас есть процедура с десятью параметрами, то некоторые вы, вероятно, пропустили.

— Алан Перлис (Alan Perlis)

Предупреждения

Для автора традиционно великодушно принимать на себя вину за любые оставшиеся неточности. Я не таков. Любые ошибки, неточности или проблемы в этой книге — это чья-то ошибка, и я оценил бы информацию о них, чтобы определить, кто виноват.

Стивен С. Скиена
Факультет информатики
Университет штата Нью-Йорк в Стоуни-Брук
http://www.cs.stonybrook.edu/~skiena
skiena@data-manual.com
Май 2017

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.

Актуальность ссылок не гарантируется.

Наши электронные адреса:

E-mail: info@dialektika.com
WWW: http://www.dialektika.com