Я посвящаю эту книгу Эйбрил

Содержание

Вступительное слово

Вероятностное программирование – это программная среда, которая позволяет создавать гибкие байесовские статистические модели в программном коде. После создания такой модели для обработки в ней данных могут быть использованы мощные алгоритмы логического вывода, работающие независимо. Такое сочетание гибкого определения модели и механизма автоматического логического вывода предоставляет исследователю мощный инструмент для быстрого создания, анализа и постепенного усовершенствования новых статистических моделей. Подобный итеративный подход абсолютно противоположен ранее применявшемуся способу подгонки байесовских моделей к данным: ранее используемые алгоритмы логического вывода обычно работали только с одной конкретной моделью. При этом требовались глубокие и прочные математические знания и навыки для формирования модели и разработки схемы логического вывода, что существенно замедляло итеративный цикл: изменение модели, модификация процесса логического вывода. Таким образом, вероятностное программирование делает статистическое моделирование доступным практически для всех, значительно снижая требования к уровню математической подготовки и сокращая время, требуемое для успешного создания новых моделей и нового, ранее недоступного, глубокого понимания исследуемых данных.

Сама идея вероятностного программирования не нова: BUGS, самый первый инструмент такого типа, появился в 1989 году. Количество моделей, для которых этот инструмент успешно применялся, было крайне ограниченным, а логический вывод выполнялся медленно, поэтому первое поколение языков этого типа не получило широкого распространения на практике. В наши дни существует множество специализированных языков вероятностного программирования, которые широко используются как для академических научных исследований, так и в компаниях Google, Microsoft, Amazon, Facebook и Uber для решения крупномасштабных и сложных задач. Что же изменилось? Главным фактором роста значимости вероятностного программирования и эволюции от состояния занимательной игрушки до мощного механизма, способного решать сложнейшие крупномасштабные задачи, стало появление алгоритма выборки на основе гамильтонова метода Монте-Карло, на несколько порядков более мощного, чем предыдущие алгоритмы выборки. Несмотря на то что этот алгоритм был разработан в 1987 году, только в последнее время системы вероятностного программирования Stan и PyMC3 сделали эту методику выборки широко доступной и удобной в практическом применении.

Предлагаемая книга представляет собой практический вводный курс по использованию этого чрезвычайно мощного и гибкого инструментального средства. Она, несомненно, окажет большое воздействие на ваш образ мыш-

ления и на понимание путей решения сложных аналитических задач. Лишь немногие люди подошли бы для написания такой книги лучше, чем один из основных разработчиков системы PyMC3 Освальдо Мартин (Osvaldo Martin). Освальдо обладает редким талантом подробного постепенного объяснения сложных тем, упрощая их понимание. Его глубокое знание и понимание этих тем, основанное на солидном практическом опыте, позволяет вести читателя по наиболее эффективному пути освоения этой области, которая иначе могла бы показаться недоступной. Наглядные иллюстрации и схемы, примеры программного кода делают эту книгу в высшей степени полезным практическим ресурсом, с помощью которого вы сможете в полной мере овладеть всеми необходимыми теоретическими основами.

Читатели, которые приобрели данную книгу, сделали правильный выбор. Это не простой и не быстрый путь. В наше время, когда широко рекламируется глубокое обучение, как методика решения всех текущих и будущих аналитических задач, более осмотрительный и взвешенный подход к созданию специализированных моделей для конкретной цели, возможно, не выглядит столь привлекательным. Но вы сможете решать задачи, которые трудно решаются любыми другими способами.

Это не говорит о том, что глубокое обучение не является весьма перспективной методикой. В действительности само по себе вероятностное программирование не ограничено классическими статистическими моделями. Изучая современную литературу по машинному обучению, вы наверняка обнаружите, что байесовская статистика определяется как мощный инструментальный комплекс для формирования и исследования следующего поколения глубоких нейронных сетей. Таким образом, эта книга вооружит читателя не только знаниями и навыками решения трудных аналитических задач, но также позволит создать более широкомасштабную основу для одного из самых великих достижений человечества: разработки искусственного интеллекта. Желаю успеха.

Томас Виецки (Thomas Wiecki),
д-р философии (PhD),
руководитель отдела исследований
в компании Quantopian (Бостон, США)

Об авторе

Освальдо Мартин (Osvaldo Martin) – ученый-исследователь агентства The National Scientific and Technical Research Council (CONICET) (Аргентина), занимающийся разработками в области структурной биоинформатики протеина, полисахаридов и молекул РНК. Обладает большим опытом использования цепей Маркова с применением метода Монте-Карло для имитации молекулярных систем, предпочитает пользоваться языком программирования Python для решения задач анализа данных.

Освальдо являлся преподавателем курсов по структурной биоинформатике, науке о данных и байесовскому анализу данных. Также возглавлял организационный комитет PyData в Сан-Луисе (Аргентина) в 2017 году. Освальдо – один из основных разработчиков программных систем PyMC3 и ArviZ.

«Я благодарен Ромине за ее постоянную поддержку. Также хочу поблагодарить Уолтера Лападула (Walter Lapadula), Билла Энгелса (Bill Engels), Эрика Х Ма (Eric J Ma) и Остина Рошфора (Austin Rochford) за бесценные замечания, поправки, комментарии и предложения к черновому варианту книги. Особая благодарность основным разработчикам и всем участникам проектов PyMC3 и ArviZ. Создание этой книги стало возможным благодаря поддержке, позитивному отношению и упорному труду, который все они вложили и вкладывают в эти библиотеки, а также в формирование великолепного сообщества пользователей».

О рецензентах

Эрик Х Ма (Eric J Ma) – ученый в области обработки данных в институте биомедицинских исследований корпорации Novartis. Занимается исследованиями биомедицинских данных с применением в основном байесовских статистических методов с целью улучшения качества медицинского обслуживания пациентов. До работы в корпорации Novartis был действительным членом научного общества Insight Health Data летом 2017 года, защитил докторскую диссертацию весной 2017 года.

Кроме того, Эрик является разработчиком ПО с открытым исходным кодом, ранее возглавлял разработку nxviz, пакета визуализации для NetworkX, и pujanitor, открытого API для очистки данных на языке Python. Также участвовал в разработке ряда инструментальных средств с открытым исходным кодом, включая PyMC3, Matplotlib, bokeh и CuPy.

Основной жизненный принцип (девиз) Эрика можно найти в Евангелии от Луки 12:48.

Остин Рошфор (Austin Rochford) – главный научный сотрудник в Monetate Labs, где он занимается разработкой продуктов, позволяющих розничным продавцам персонализировать свой рынок с учетом миллиардов событий, происходящих ежегодно. По образованию Остин математик, являющийся активным пропагандистом байесовских методов.

Предисловие

Методы байесовской статистики разрабатываются уже более 250 лет. Не только признание и одобрение, но не меньшее пренебрежение и даже неприятие постоянно сопровождали эту ветвь математики. На протяжении нескольких последних десятилетий байесовская статистика стала привлекать все большее внимание специалистов, занимающихся статистикой, и почти всех других ученых, инженеров и даже людей, не принадлежащих к миру науки. Подобный рост интереса стал возможным благодаря теоретическим и вычислительным разработкам, выполненным в основном во второй половине XX века. Разумеется, современная байесовская статистика является главным образом вычислительной статистикой. Необходимость в создании гибких и прозрачных моделей и более глубокая и подробная интерпретация статистических моделей и методов анализа также внесли свой вклад в тенденцию роста.

В предлагаемой книге применяется прагматический подход к изучению байесовской статистики, здесь не уделяется слишком много внимания другим статистическим парадигмам и их взаимосвязям с байесовской статистикой. Главная цель книги – научить практическому выполнению байесовского анализа данных. Философские теоретические дискуссии интересны, но на страницах этой книги вы их не обнаружите, попробуйте поискать в других, более подходящих местах.

В книге излагается методический подход моделирования в статистике, она поможет научиться мыслить в терминах вероятностных моделей и применять теорему Байеса для вывода логических следствий из используемых моделей и данных. Такой подход также является вычислительным, модели кодируются с использованием PyMC3, библиотеки поддержки байесовской статистики, которая скрывает от конечного пользователя большинство математических подробностей и вспомогательных вычислений, и ArviZ, пакета языка Python для исследовательского анализа байесовских моделей.

Байесовские методы с теоретической точки зрения основаны на теории вероятностей, поэтому неудивительно, что многие книги по байесовской статистике содержат множество математических формул, требующих определенного уровня математической подготовки. Изучение математических основ статистики может оказать немалую помощь при создании более качественных моделей и улучшить интуитивное понимание задач, моделей и результатов. Тем не менее такие библиотеки, как PyMC3, позволяют изучать и применять методы байесовской статистики даже при скромном объеме математических знаний, в чем вы сможете убедиться сами, читая эту книгу.

Для кого предназначена эта книга

Если вы студент, специалист по обработке данных, исследователь в области естественных или общественных наук или разработчик, начинающий изучать байесовский анализ данных и вероятностное программирование, то эта книга предназначена для вас. Книга представляет собой введение в байесовский анализ, поэтому не требует предварительных знаний в области статистики, хотя некоторый практический опыт использования языка Python и библиотеки NumPy был бы полезен.

Краткое содержание книги

В главе 1 «Вероятностное мышление» рассматриваются основные концепции байесовской статистики и ее практическое применение для анализа данных. Здесь содержится большинство базовых понятий и положений, которые используются в следующих главах книги.

Глава 2 «Вероятностное программирование» дает обзор концепций из предыдущей главы с точки зрения вычислительной практики. Здесь представлена ознакомительная информация о библиотеке вероятностного программирования PyMC3, а также о библиотеке ArviZ (пакет Python), предназначенной для исследовательского анализа байесовских моделей. На нескольких конкретных примерах рассматриваются иерархические модели.

В главе 3 «Моделирование с использованием линейной регрессии» рассматриваются основные элементы линейной регрессии, весьма широко применяемой модели и структурного компонента более сложных моделей.

В главе 4 «Обобщение линейных моделей» описывается, как расширить область применения линейных моделей для распределений, отличающихся от распределения Гаусса, открывая путь к решению многих задач анализа данных.

В главе 5 «Сравнение моделей» обсуждаются методы сравнения, выбора и усреднения моделей с использованием факторов Байеса, WAIC, LOO, а также основные характерные особенности и детали применения этих методов.

В главе 6 «Смешанные модели» рассматриваются способы повышения гибкости моделей с помощью объединения простых распределений для создания более сложных. Здесь представлена первая непараметрическая модель, рассматриваемая в книге: процесс Дирихле.

В главе 7 «Гауссовы процессы» описывается теоретическая концепция, лежащая в основе гауссовых процессов, а также способы ее применения для создания непараметрических моделей для решения широкого спектра задач.

В главе 8 «Механизмы логического вывода» представлено введение в методы числовой аппроксимации апостериорного распределения (вероятности), а также весьма важная с практической точки зрения тема: как точно определить надежность аппроксимированного апостериорного распределения.

В главе 9 «Что дальше» предлагается список информационных ресурсов, которыми можно воспользоваться для более глубокого изучения байесовского анализа, а также очень короткое итоговое резюме автора.

Максимально эффективное использование книги

Код в этой книге написан на языке Python версии 3.6. Для установки программной среды Python и всех необходимых библиотек рекомендуется воспользоваться Anaconda, специализированным для научных вычислений дистрибутивом. Получить более подробную информацию о дистрибутиве Anaconda и скачать его можно на сайте **https://www.anaconda.com/download/**. При этом в вашей системе будет установлено множество полезных пакетов на языке Python. После этого потребуется установить еще два пакета. Для установки библиотеки PyMC3 используйте утилиту `conda`:

```
conda install -c conda-forge pymc3
```

Чтобы установить пакет ArviZ, можно выполнить следующую команду:

```
pip install arviz
```

Другой способ установки необходимых пакетов, после того как дистрибутив Anaconda установлен в системе, – перейти по адресу **https://github.com/aloctavodia/BAP** и загрузить файл описания среды *bap.yml*. С помощью этого файла можно установить все необходимые пакеты следующей командой:

```
conda env create -f bap.yml
```

Все пакеты языка Python, использованные при написании этой книги, перечислены ниже:

- IPython 7.0;
- Jupyter 1.0 (или Jupyter-lab 0.35);
- NumPy 1.14.2;
- SciPy 1.1;
- pandas 0.23.4;
- Matplotlib 3.0.2;
- Seaborn 0.9.0;
- ArviZ 0.3.1;
- PyMC3 3.6.

При выполнении кода, приведенного в каждой главе, предполагается, что вы установили и импортировали, по крайней мере, некоторые из перечисленных выше пакетов. Вместо копирования и вставки кода из книги рекомендуется скачать файлы исходного кода из репозитория **https://github.com/aloctavodia/BAP** и запускать их с помощью Jupyter Notebook или Jupyter Lab. Автор регулярно обновляет этот репозиторий после выхода новых версий PyMC3 и ArviZ.

Если при выполнении кода из книги возникает техническая проблема или обнаружена опечатка либо какая-либо другая ошибка, то передайте информацию об этом в указанный репозиторий, и автор попытается устранить проблему как можно быстрее.

Большинство иллюстраций в книге сгенерировано при выполнении программного кода. Основная схема такова: сначала приводится блок кода, за которым сразу же следует соответствующая иллюстрация (сгенерированная при выполнении приведенного выше кода). Автор надеется, что такая схема окажется привычной для тех, кто использует Jupyter Notebook/Lab, и не вызовет никаких затруднений у других читателей.

Скачивание исходного кода примеров

Скачать файлы с дополнительной информацией для книг издательства «ДМК Пресс» можно на сайте **www.dmkpress.com** или **www.дмк.рф** на странице с описанием соответствующей книги.

Загрузка цветных иллюстраций

Мы также предоставляем файл в формате PDF, содержащий цветные изображения снимков экрана и схем из данной книги. Этот файл доступен по адресу **https://www.packtpub.com/sites/default/files/downloads/9781789341652_ColorImages.pdf**.

Типографские соглашения, принятые в книге

В этой книге используется несколько стилей выделения некоторых элементов текста.

`Фрагмент кода в тексте` – ключевые слова, операторы, имена переменных и функций непосредственно в тексте. Пример: «Большая часть приведенного выше кода предназначена для построения и вывода графической схемы, вероятностные вычисления выполняются в строке `y = stats.norm(mu, sd).pdf(x)`».

Блок кода отображается в следующем формате:

```
μ = 0.
σ = 1.
X = stats.norm(μ, σ)
x = X.rvs(3)
```

Курсив – имена файлов, каталогов и прочих объектов.

Полужирный шрифт – важные (ключевые) слова, элементы пользовательского интерфейса или слова, которые выводятся на экран.

Отзывы и пожелания

Мы всегда рады отзывам наших читателей. Расскажите нам, что вы думаете об этой книге – что понравилось или, может быть, не понравилось. Отзывы важны для нас, чтобы выпускать книги, которые будут для вас максимально полезны.

Вы можете написать отзыв на нашем сайте **www.dmkpress.com**, зайдя на страницу книги и оставив комментарий в разделе «Отзывы и рецензии». Также можно послать письмо главному редактору по адресу **dmkpress@gmail.com**; при этом укажите название книги в теме письма.

Если вы являетесь экспертом в какой-либо области и заинтересованы в написании новой книги, заполните форму на нашем сайте по адресу **http://dmkpress.com/authors/publish_book/** или напишите в издательство по адресу **dmkpress@gmail.com**.

Список опечаток

Хотя мы приняли все возможные меры для того, чтобы обеспечить высокое качество наших текстов, ошибки все равно случаются. Если вы найдете ошибку в одной из наших книг – возможно, ошибку в основном тексте или программном коде, – мы будем очень благодарны, если вы сообщите нам о ней. Сделав это, вы избавите других читателей от недопонимания и поможете нам улучшить последующие издания этой книги.

Если вы найдете какие-либо ошибки в коде, пожалуйста, сообщите о них главному редактору по адресу **dmkpress@gmail.com**, и мы исправим это в следующих тиражах.

Нарушение авторских прав

Пиратство в интернете по-прежнему остается насущной проблемой. Издательства «ДМК Пресс» и Springer очень серьезно относятся к вопросам защиты авторских прав и лицензирования. Если вы столкнетесь в интернете с незаконной публикацией какой-либо из наших книг, пожалуйста, пришлите нам ссылку на интернет-ресурс, чтобы мы могли применить санкции.

Ссылку на подозрительные материалы можно прислать по адресу электронной почты **dmkpress@gmail.com**.

Мы высоко ценим любую помощь по защите наших авторов, благодаря которой мы можем предоставлять вам качественные материалы.

Глава 1

Вероятностное мышление

«Теория вероятностей – это не что иное, как здравый смысл, сведенный к вычислениям».

– Пьер Симон Лаплас

В этой главе мы начнем изучать основные концепции байесовской статистики и познакомимся с некоторыми инструментами из байесовского арсенала. Здесь приводятся небольшие фрагменты кода на языке Python, но глава в основном теоретическая, поэтому почти все концепции, описываемые в ней, будут многократно использоваться на протяжении всей книги. Из-за обилия теоретического материала глава может показаться немного усложненной для программистов-кодеров, но я полагаю, что такой подход способствует упрощению эффективного применения байесовской статистики при решении практических задач.

В этой главе рассматриваются следующие темы:

- статистическое моделирование;
- вероятность и неопределенность;
- теорема Байеса и статистический вывод;
- статистический вывод с одним параметром и классическая задача о подбрасывании монеты;
- выбор априорного распределения вероятностей и почему людям часто не нравится эта процедура, хотя она обязательна;
- взаимодействие с байесовским анализом (представление результатов).

Статистика, модели и подход, принятый в этой книге

Статистика занимается сбором, организацией (упорядочением), анализом и интерпретацией данных, следовательно, знание статистики чрезвычайно важно для анализа данных. При анализе данных используются два основных статистических метода:

- разведочный анализ данных (РАД)[1] (Exploratory Data Analysis – EDA) – используются числовые обобщающие характеристики, такие как среднее

[1] Термин взят из русской версии Википедии, хотя более подходящим кажется вариант «исследовательский анализ данных». – *Прим. перев.*

значение, мода, стандартное отклонение и вероятные отклонения (этот раздел разведочного анализа данных также называют описательной статистикой). Кроме того, при разведочном анализе данные исследуются визуально с применением широко известных инструментальных средств, таких как гистограммы и диаграммы рассеяния;

- статистический вывод (inferential statistics) – вывод утверждений на основе текущих данных. Это может быть необходимо для понимания некоторых конкретных явлений, или для прогнозирования в будущем точек данных (которые ранее не наблюдались), либо для выбора одного из нескольких альтернативных объяснений результатов наблюдений. Статистический вывод – это набор методов и инструментов, которые помогают ответить на типы вопросов, перечисленные выше.

! В этой книге главное внимание сосредоточено на выполнении байесовского статистического вывода и последующем применении метода разведочного анализа данных для обобщения, интерпретации, проверки и предъявления результатов байесовского статистического вывода.

В большинстве вводных курсов по статистике, по крайней мере для людей, незнакомых со статистикой, материал излагается как набор готовых рецептов, более или менее похожих на следующий: войти в статистическую кладовую, выбрать одну из банок и открыть ее, добавить данные по вкусу, перемешивать до получения твердого P-значения (P-критерия), предпочтительно меньшего 0.05. Главная цель курсов с таким подходом – научить выбирать правильную банку в статистической кладовой. Мне никогда не нравился подобный подход, главным образом потому, что наиболее частым результатом такого обучения становится группа сбитых с толку людей, неспособных хорошо понимать, даже на концептуальном уровне, совокупность различных изучаемых методов. Мы будем придерживаться другой методики: также будем изучать некоторые конкретные рецепты, но при этом предпочитая домашнее приготовление, а не консервы из банок. То есть мы будем учиться смешивать свежие ингредиенты, наиболее подходящие для разнообразных блюд, и, что более важно, такой подход позволит применять изученные концепции в последующих примерах, содержащихся в книге.

Применение подобного подхода возможно по двум причинам:

- онтологическая причина – статистика является формой моделирования, унифицированной с помощью математического аппарата, основанного на теории вероятностей. Использование вероятностного подхода обеспечивает единую универсальную точку зрения на все, что может показаться весьма несопоставимыми методами, – статистические методы и методы машинного обучения выглядят намного более похожими друг на друга при взгляде с вероятностной точки зрения;
- техническая причина – современное программное обеспечение, например библиотека PyMC3, позволяет специалистам-практикам, таким как

вы и я, определять и создавать модели решений относительно простым способом. Многие из этих моделей оставались неразрешимыми (следовательно, неприменимыми на практике) буквально несколько лет назад или требовали слишком высокого уровня математической и технической подготовки.

Работа с данными

Данные – это важнейший ингредиент в статистике и науке о данных (даталогии). Данные поступают из различных источников, таких как эксперименты, компьютерные имитации, опросы и полевые наблюдения. Если мы являемся ответственными за генерацию или сбор данных, то всегда в первую очередь необходимо тщательно продумать и сформулировать вопросы, на которые нужно получить ответы, и определить используемые для этого методы, и только после этого приступать к обработке данных. В действительности существует целая область статистики, занимающаяся сбором данных, – планирование эксперимента (experimental design). В наше время, когда поток данных достиг невероятных размеров, мы иногда можем забыть о том, что сбор данных не всегда является простым и дешевым делом. Например, всем известно, что Большой адронный коллайдер генерирует сотни терабайтов данных в день, но не все помнят о том, что для его создания потребовались годы ручного и умственного труда.

В качестве обобщенного правила можно интерпретировать процесс генерации данных как случайный (стохастический), поскольку в этом процессе существует онтологическая, техническая и/или эпистемологическая неопределенность, то есть система по своей внутренней сущности является случайной, также существуют технические проблемы, добавляющие шум или ограничивающие наши возможности измерения с произвольной точностью, а кроме того, некоторые концептуальные теоретические ограничения, скрывающие от нас подробности. Из-за всех вышеперечисленных причин всегда необходимо интерпретировать данные в контексте используемых моделей, включая ментальные и формальные. Данные без моделей ни о чем не говорят.

В книге предполагается, что все необходимые данные уже собраны. Кроме того, данные считаются предварительно подготовленными и очищенными, что чрезвычайно редко встречается в реальной практике. Эти исходные предположения сделаны для того, чтобы полностью сосредоточиться на основной теме книги. Очень существенное замечание, особенно для начинающих изучать анализ данных: даже несмотря на то, что подготовка и очистка данных не рассматривается в нашей книге, это весьма важные практические навыки, которыми вы должны овладеть и развивать их для успешной работы с данными.

Очень полезной способностью при анализе данных является умение написать код на каком-либо языке программирования, например на Python. Обработка данных является неизбежной необходимостью с учетом того, что мы живем в беспорядочном мире с еще более беспорядочными данными, поэтому

умение писать программный код помогает решать эти задачи. Даже если вы настолько удачливы, что находящиеся в вашем распоряжении данные предварительно обработаны и очищены, умение писать код все равно останется полезным навыком, потому что современная байесовская статистика реализована в основном с помощью языков программирования, таких как Python или R.

Если вы хотите более подробно узнать, как использовать Python для очистки и обработки данных, рекомендуется изучить превосходную книгу «Python Data Science Handbook» Джейка Ван-дер-Пласа (Jake VanderPlas).

Байесовское моделирование

Модели – это упрощенные описания конкретной системы или процесса, которые, по тем или иным причинам, нас интересуют. Эти описания преднамеренно формулируются так, чтобы отобразить самые важные аспекты системы, но не уделять внимание каждой малозначимой подробности. Это одна из причин, по которой более сложная модель не всегда является лучшим вариантом.

Существует множество различных типов моделей, но в этой книге мы ограничимся байесовскими моделями. В общем случае процесс байесовского моделирования включает три основных этапа.

1. Выбираются некоторые данные и делаются предварительные предположения о том, как эти данные могли быть сгенерированы (извлечены), затем формируется модель с помощью объединения структурных компонентов, известных под названием распределения вероятностей. В основном эти модели представляют собой достаточно грубые приближения (аппроксимации), но в большинстве случаев это именно то, что нам нужно.
2. Используется теорема Байеса для добавления данных в сформированные модели, и выполняются логические выводы по совокупности данных и наших предварительных предположений. Обычно это называют формированием или уточнением условий работы модели по имеющимся данным.
3. Модель критически оценивается посредством проверки степени осмысленности результатов по различным критериям, включая данные, уровень наших экспертных знаний в области исследований, а иногда путем сравнения нескольких моделей.

Вообще говоря, три перечисленных выше этапа в большинстве случаев выполняются итеративно и без соблюдения строгого порядка. Мы будем возвращаться для повторения любого из этих этапов в произвольные моменты времени: возможно, была совершена ошибка при написании кода, или был найден способ изменить и усовершенствовать модель, или возникла необходимость добавления новых данных или сбора данных другого типа.

Байесовские модели также называют вероятностными моделями, потому что они создаются с использованием вероятностей. Почему сделан именно такой выбор? Потому что вероятности являются самым правильным мате-

матическим инструментом для моделирования неопределенности. Поэтому необходимо поближе познакомиться с этой «новой землей», покрытой сетью разветвляющихся дорог.

Теория вероятностей

Название этого раздела может показаться излишне претенциозным, ведь мы не собираемся изучить теорию вероятностей всего на нескольких страницах, да это и не было моим намерением. Я хотел лишь представить несколько общих и наиболее важных концепций, необходимых для лучшего понимания байесовских методов, достаточных для освоения материала данной книги. При необходимости мы будем более подробно рассматривать или вводить новые концепции, относящиеся к теории вероятности. Для более глубокого изучения теории вероятности настоятельно рекомендую книгу «Introduction to Probability» Джозефа К Блитцштайна (Joseph K Blitzstein) и Джессики Хван (Jessica Hwang). Другой весьма полезной книгой может оказаться «Mathematical Theory of Bayesian Statistics» Сумио Ватанабе (Sumio Watanabe), поскольку по названию понятно, что эта книга в большей степени ориентирована на байесовские методы, чем первая, но она более сложна с математической точки зрения.

Объяснение смысла вероятностей

Несмотря на то что теория вероятностей является вполне сформировавшейся и прочно обоснованной математической дисциплиной, существуют и другие интерпретации смысла термина «вероятность». С байесовской точки зрения вероятность – это мера, которая определяет в числовом выражении уровень неопределенности высказывания. С учетом этого определения вероятности абсолютно допустимо и даже естественно задать вопрос о вероятности существования жизни на Марсе, о вероятности того, что масса электрона составляет 9.1×10^{-31} кг, или о вероятности того, что 9 июля 1816 года в Буэнос-Айресе был солнечный день. Но при этом следует отметить, например, что жизнь на Марсе либо существует, либо не существует, то есть итоговый результат бинарный, это тип вопроса с ответом да-нет. Но, учитывая то, что мы не уверены в самом факте существования жизни на Марсе, разумным образом действий является попытка определить, насколько вероятна жизнь на Марсе. Поскольку приведенное выше определение вероятности связано с эпистемологическими, то есть познавательными, функциями нашего мышления, его часто называют субъективным определением вероятности. Однако отметим, что любой человек с научным складом ума не будет использовать свои личные верования или «информацию, полученную от ангела», для ответа на вопросы такого рода, а воспользуется всеми доступными геофизическими данными о Марсе, обратится к своим знаниям в области биомеханики, чтобы определить необходимые условия для жизни, и т. д. Таким образом, байесовские вероятности, следовательно, и вся байесовская статистика, так же субъективны (или объективны),