Оглавление

От переводчика

В последние несколько лет наблюдается небывалый рост объема, скорости получения и сложности данных в самых разных областях жизнедеятельности человека. Неудивительно, что и спрос на специалистов, способных извлечь полезную информацию из этих потоков данных, сегодня высок, как никогда раньше. Важную роль в подготовке таких специалистов играет учебная литература по современным методам статистического анализа. Написать хороший учебник — это титанический труд, однако авторы книги, которую Вы сейчас держите в руках, справились с этой задачей блестяще. Простота изложения материала, многочисленные практические примеры и хорошо продуманные лабораторные работы и упражнения сделали книгу «An Introduction to Statistical Learning with Applications in R» чрезвычайно популярной в академических кругах и среди аналитиков коммерческих организаций во всем мире. Для меня было честью выполнить перевод этой работы, и я рад, что теперь она стала доступной и для русскоязычных читателей.

Я сделал все, что было в моих силах, чтобы максимально точно передать текст оригинала, однако возможность повстречать неточности в использованной терминологии и банальные опечатки, к сожалению, полностью не исключена. О любых замечаниях сообщайте, пожалуйста, по адресу rtutorialsbook@gmail.com — это позволит внести соответствующие правки при публикации последующих изданий.

Я благодарен Дмитрию Мовчану и всей команде «ДМК Пресс» за помощь с подготовкой и изданием этой книги, а также Артему Груздеву, Дмитрию Дерябину и Александру Вишератину за оказанные ими консультации и внимательное прочтение рукописи. Наконец, я хотел бы поблагодарить свою жену Светлану за ее бесконечное терпение и понимание в отношении всех моих начинаний, одним из которых стала работа над этим переводом.

Сергей Мастицкий *Лондон, март 2015 г.*

Предисловие

К статистическому обучению относят набор инструментов, предназначенных для моделирования и понимания сложно организованных данных. Это недавно разработанная область статистики, которая развилась параллельно с достижениями в компьютерных науках и особенно машинном обучении. Данная область охватывает многие методы, включая лассо и разреженную регрессию, классификационные и регрессионные деревья, бустинг и метод опорных векторов.

Одновременно со взрывообразным ростом круга задач, связанных с «большими данными», статистическое обучение стало очень популярным во многих научных областях, а также в маркетинге, финансах и других бизнес–дисциплинах. Люди с навыками статистического обучения очень востребованны.

Одна из первых книг в этой области — «*Основы статистического обучения*» (ОСО)[1] — была опубликована в 2001 г., а в 2009 г. вышло ее второе издание. ОСО стала очень популярной книгой среди не только статистиков, но и специалистов из смежных областей. Одна из причин такой популярности заключается в относительно легкодоступном стиле изложения. Однако ОСО предназначена для людей с основательной математической подготовкой. Новая книга «*Введение в статистическое обучение*» возникла в связи с ощутимой необходимостью в более широком и не таком техническом изложении материала. В этой новой книге мы освещаем многие из тех же тем, которые присутствуют в ОСО, но уделяем основное внимание практическому применению соответствующих методов, а не их математическим деталям. Мы разработали лабораторные работы, иллюстрирующие реализацию каждого метода с использованием статистического пакета R. Эти лабораторные работы позволяют читателю получить ценный практический опыт.

Эта книга подойдет для студентов и магистрантов, углубленно изучающих статистику или родственные дисциплины, а также для представителей других наук, которые желают применять инструменты статистического обучения для анализа своих данных. Ее можно использовать в качестве учебника для курса, длящегося один или два семестра.

Мы благодарим за ценные комментарии следующих читателей черновых вариантов этой книги: Паллави Басу, Александру Чулдечову, Патрика Данахера, Уилла Фитиана, Луэллу Фу, Сэма Гросса, Макса Гразьера Г'Селла, Кортни Паулсон, Ксингао Кьяо, Элизу Шенг, Ноа Симон, Кена Минга Тана и Ксина Лу Тана.

[1] Hastie T., Tibshirani R., Friedman J. (2001) The Elements of Statistical Learning. Springer, 745 p.

«Делать предсказания трудно, особенно в отношении будущего».

Йоги Берра

Джеймс Гарет (Лос–Анджелес, США)

Даниела Уиттен (Сиэттл, США)

Тревор Хасти (Пало Альто, США)

Роберт Тибширани (Пало Альто, США)

Глава 1

Введение

Обзор задач статистического обучения

Под *статистическим обучением* понимают огромный набор инструментов, предназначенных для *понимания данных*. Эти инструменты можно разделить на две группы: *обучение с учителем* и *обучение без учителя*. В общих чертах статистическое обучение подразумевает построение статистической модели для предсказания, или оценивания, некоторой *выходной переменной* на основе одной или нескольких *входных переменных*. Подобные проблемы встречаются в настолько разнящихся областях, как бизнес, медицина, астрофизика и государственное управление. При обучении без учителя имеются входные переменные, но нет предсказываемой переменной; тем не менее мы можем выявить закономерности и структуру в таких данных. В качестве иллюстрации некоторых практических приложений статистического обучения ниже мы кратко обсудим три реальных набора данных, рассматриваемых в этой книге.

Данные по заработной плате

В этом примере мы исследуем связь нескольких факторов с уровнем заработной платы у группы мужчин из центрально–атлантического региона США (в этой книге мы будем ссылаться на соответствующие данные как «набор данных `Wage`»). В частности, мы хотим выяснить зависимость между заработной платой работника (переменная `wage`) и его возрастом (`age`), уровнем образования (`education`), а также календарным годом (`year`). Посмотрите, например, на график, представленный слева на рис. 1.1, где показана связь между заработной платой и возрастом работников из этого набора данных. Имеется свидетельство в пользу того, что `wage` увеличивается по мере возрастания `age`, а затем снова снижается примерно после 60 лет. Синяя линия, которая соответствует оценке среднего уровня `wage` для заданного значения `age`, позволяет увидеть этот тренд более четко.

Зная возраст работника, мы можем *предсказать* его заработную плату по этой кривой. Однако на рис. 1.1 также хорошо виден значительный разброс относительно этого среднего значения, из чего следует, что сама по себе переменная `age` вряд ли позволит с большой точностью предсказать `wage` для конкретного человека.

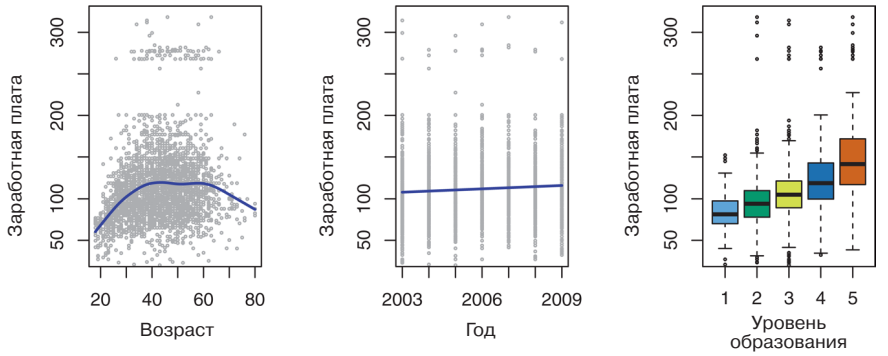

РИСУНОК 1.1. *Таблица* `Wage` *с данными по заработной плате мужчин из центрально–атлантического региона США.* Слева: `wage` *как функция от* `age`*. В среднем* `wage` *увеличивается одновременно с* `age` *до возраста около 60 лет, после чего начинает снижаться.* В центре: `wage` *как функция от* `year`*. В период с 2003 по 2009 г. имеет место медленный, но устойчивый рост* `wage` *в среднем на 10 000$ в год.* Справа: *диаграмма размахов* `wage` *как функции от* `education`*, где 1 соответствует самому низкому уровню образования (неоконченная средняя школа), а 5 – самому высокому уровню (ученая степень). В среднем* `wage` *возрастает с уровнем образования*

У нас имеется также информация по уровню образования каждого работника и его заработной плате `wage` за каждый год `year`. Графики, представленные в центре и справа на рис. 1.1, показывают `wage` в зависимости от `year` и `education` и свидетельствуют о том, что каждый из этих факторов связан с `wage`. С 2003 по 2009 г. значения зарплаты с каждым годом линейно возрастают примерно на 10 000$, хотя этот рост очень слабый, по сравнению с разбросом в данных. Зарплаты также выше у людей с более высоким уровнем образования: работники с наименьшим уровнем образования (1) в целом зарабатывают гораздо меньше, чем работники с самым высоким уровнем (5). Очевидно, что наиболее точное предсказание `wage` для конкретного человека будет получено при объединении информации по его возрасту `age`, уровню образования `education` и году `year`. В главе 3 мы обсудим линейную регрессию, которую можно применить для предсказания `wage` по этим данным. В идеале мы должны предсказывать `wage` с учетом нелинейного характера связи этой переменной с `age`. В главе 7 мы рассмотрим класс методов, предназначенных для решения данной проблемы.

Данные по рынку акций

В случае с набором данных `Wage` предсказывается *непрерывное*, или *количественное*, выходное значение. Часто такую ситуацию называют проблемой *восстановления регрессии*. Однако в некоторых случаях мы можем столкнуться с необходимостью предсказать нечисловое значение, т. е. *категориальную*, или *качественную*, выходную переменную. Так, в главе 4 мы рассмотрим набор данных по рынку акций, который описывает днев-

ные изменения индекса Standard & Poor's 500 (S&P) в течение 5–летнего периода (с 2001 по 2005 г.). Мы будем ссылаться на него как на «набор данных `Smarket`». Задача заключается в предсказании *возрастания* или *снижения* индекса на основе его удельного изменения за последние 5 дней. Здесь проблема статистического обучения не подразумевает предсказания числового значения. Вместо этого предсказывается рост (`Up`) или снижение (`Down`) рынка акций для того или иного дня. Это известно как проблема *классификации*. Модель, способная с высокой точностью предсказывать направление движения рынка, была бы очень полезной!

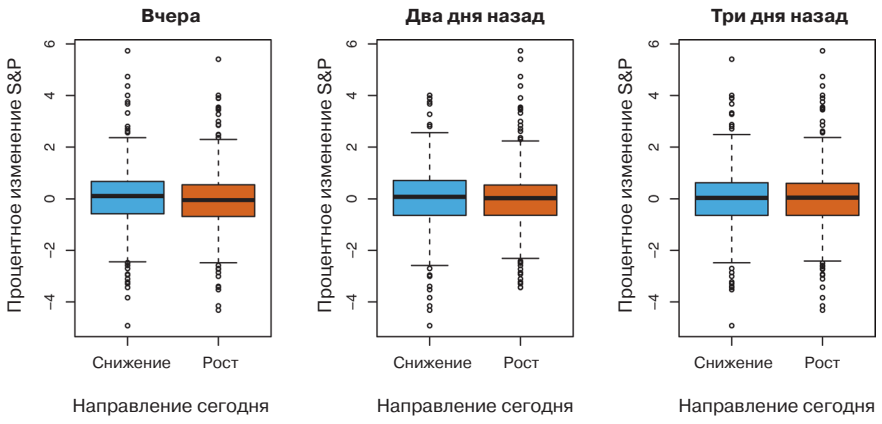

РИСУНОК 1.2. Слева: *диаграмма размахов, отражающая процентное изменение индекса S&P по сравнению со вчерашним значением для дней, когда происходили рост или снижение рынка (по данным* `Smarket`). В центре и справа: *то же, но показаны процентные изменения по сравнению с двумя и тремя предыдущими днями соответственно*

На рис. 1.2 слева представлена диаграмма размахов, отражающая процентные изменения индекса акций по сравнению с предыдущим днем: для 648 дней, когда в следующие за ними дни рынок вырос, и для 602 дней, когда рынок ушел вниз. Эти две диаграммы почти идентичны, что указывает на невозможность простой стратегии по использованию вчерашнего состояния индекса S&P для предсказания его сегодняшнего положения. Остальные графики, на которых приведены диаграммы размахов для процентных изменений в сравнении с двумя и тремя предыдущими днями, также указывают на отсутствие выраженной связи между прошлым и текущим состояниями индекса. Безусловно, отсутствие связи здесь ожидаемо, иначе при наличии тесных корреляций между следующими друг за другом днями мы могли бы использовать простую торговую стратегию для получения прибыли. Тем не менее в главе 4 мы подробно исследуем эти данные при помощи нескольких методов статистического обучения. Интересно, что есть некоторые указания на наличие слабых закономерностей в этих данных, предполагающие возможность правильного предсказания направления движения рынка примерно в 60% случаев (по крайней мере, для этого 5–летнего периода; рис. 1.3).

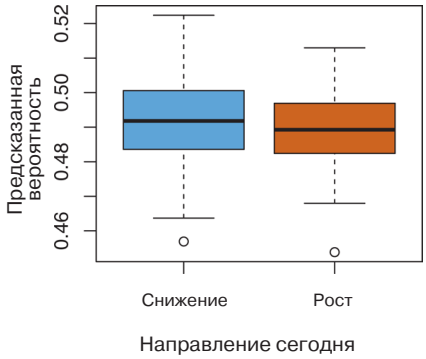

РИСУНОК 1.3. *Мы подогнали квадратичную дискриминантную модель для части данных* `Smarket`*, соответствующей периоду с 2001 по 2004 г., и предсказали вероятность снижения рынка акций для данных за 2005 г. В среднем предсказанная вероятность снижения рынка выше для дней, когда снижение в действительности имело место. На основе этих результатов мы можем правильно предсказать направление движения рынка в 60% случаев*

Данные по экспрессии генов

Два предыдущих примера иллюстрируют данные, в которых есть как входные, так и выходные переменные. Однако еще один важный класс проблем охватывает ситуации, в которых мы наблюдаем только входные переменные, без соответствующей зависимой переменной. Например, выполняя маркетинговые исследования, мы могли бы располагать демографической информацией для ряда уже имеющихся или потенциальных клиентов. У нас может возникнуть желание понять, какие клиенты похожи друг на друга, для чего мы объединили бы отдельных людей в группы в соответствии с их наблюдаемыми характеристиками. Такая ситуация известна как проблема *кластеризации*. В отличие от предыдущих примеров, здесь мы не пытаемся предсказать какую–либо выходную переменную.

Мы посвящаем главу 10 обсуждению методов статистического обучения, предназначенных для решения проблем, в которых нет естественной выходной переменной. Мы рассматриваем набор данных `NCI60`, состоящий из 6830 значений уровня экспрессии генов в 64 линиях раковых клеток. Вместо предсказания какой–то конкретной выходной переменной нам интересно выяснить наличие групп, или кластеров, среди этих клеточных линий на основе измерений генной экспрессии. Решить этот вопрос нелегко, отчасти из–за наличия тысяч значений уровня экспрессии для каждой линии, которое затрудняет визуализацию данных.

График, показанный на рис. 1.4 слева, решает эту проблему путем представления каждой из 64 клеточных линий при помощи всего лишь двух чисел — Z_1 и Z_2. Это первые две *главные компоненты* данных, которые сводят 6830 значений уровня экспрессии по каждой линии до двух чисел, или *измерений*. Несмотря на вероятность потери некоторой части информации в результате такого снижения размерности, теперь появля-

ется возможность визуально исследовать данные на наличие кластеров. Выбор числа кластеров часто бывает трудной проблемой. Однако график, приведенный на рис. 1.4 слева, указывает на наличие не менее четырех групп клеточных линий, которые мы пометили разными цветами. Теперь мы можем подробнее изучить клеточные линии из каждого кластера на предмет их сходства по типу рака и тем самым лучше понять взаимосвязь между уровнями генной экспрессии и раком.

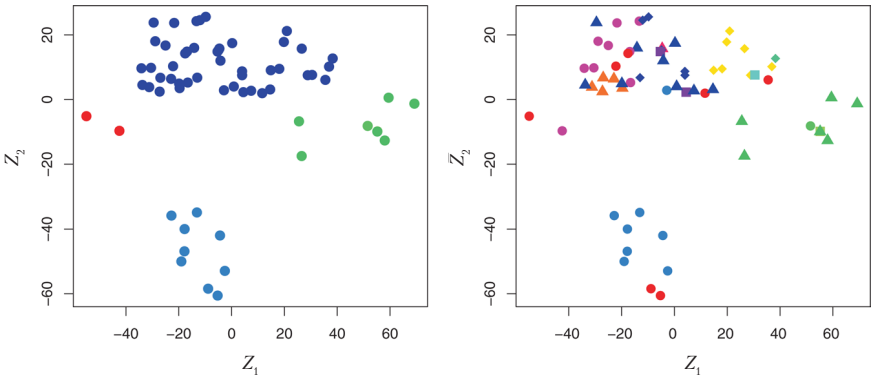

РИСУНОК 1.4. Слева: *представление данных по уровню экспрессии генов* `NCI60` *в двумерном пространстве, образованном переменными* Z_1 *и* Z_2. *Каждая точка соответствует одной из 64 клеточных линий. Клеточные линии образуют примерно четыре группы, которые мы представили разными цветами.* Справа: *то же, что и слева, за тем исключением, что мы выделили каждый из 14 типов рака при помощи символов разной формы и цвета. Клеточные линии, соответствующие одному типу рака, стремятся располагаться в этом двумерном пространстве рядом*

Оказывается, что в случае с этим конкретным набором данных клеточные линии соответствуют 14 различным типам рака. (Эта информация, однако, не была использована при создании диаграммы на рис. 1.4 слева.) Справа на рис. 1.4 показаны те же данные, однако 14 типов рака отмечены символами разной формы и цвета. Хорошо видно, что клеточные линии с одинаковым типом рака стремятся располагаться близко друг к другу в этом двумерном представлении. Кроме того, несмотря на то что информация по раку не была использована для создания первого графика, полученная кластеризация в значительной мере соответствует действительным типам рака, наблюдаемым на втором графике. Это в определенной мере является независимым подтверждением верности нашего кластерного анализа.

Краткая история развития статистического обучения

Несмотря на то что термин «*статистическое обучение*» достаточно новый, многие из основополагающих концепций этой дисциплины были раз-

работаны очень давно. В начале XIX века Лежандр и Гаусс опубликовали статьи по *методу наименьших квадратов*. Этот подход был впервые успешно применен для решения проблем астрономии. Линейная регрессия используется для предсказания значений количественных переменных, таких, например, как заработная плата. Для предсказания значений качественных переменных (например, выживет пациент или нет, пойдет рынок акций вверх или вниз) Фишер в 1936 г. предложил *линейный дискриминантный анализ*. В 1940–х г. разные авторы предложили альтернативный подход — *логистическую регрессию*. В начале 1970–х г. Нельдер и Веддербурн ввели термин «*обобщенные линейные модели*» для целого класса методов статистического обучения, которые включают как линейную, так и логистическую регрессию в качестве частных случаев.

К концу 1970–х г. стали доступными многие другие методы обучения на основе данных. Однако почти всегда это были линейные методы, поскольку с вычислительной точки зрения подгонка *нелинейных* зависимостей в то время была неосуществимой. К началу 1980–х гг. вычислительные технологии, наконец, были усовершенствованы до уровня, который больше не ограничивал работу с нелинейными методами. В середине 1980–х г. Брейман, Фридман, Ольшен и Стоун ввели *деревья регрессии и классификации* и стали одними из первых, кто детально продемонстрировал большой потенциал для практической реализации этого метода, включая перекрестную проверку для выбора модели. В 1986 г. Хасти и Тибширани ввели термин «*обобщенные аддитивные модели*» для класса нелинейных дополнений обобщенных линейных моделей, а также разработали соответствующее программное обеспечение.

С тех пор благодаря появлению *машинного обучения* и других дисциплин, статистическое обучение развилось в новую ветвь статистики, уделяющую основное внимание обучению с учителем и без учителя, а также прогнозированию. В последние годы прогресс в статистическом обучении был связан с ростом доступности мощного и относительно удобного программного обеспечения, каковым является популярная и бесплатная система `R`. Потенциально это может привести к дальнейшей трансформации дисциплины из набора методов, используемых и разрабатываемых статистиками и специалистами в области компьютерных наук, в неотъемлемый набор инструментов для гораздо более широкого сообщества.

Об этой книге

Книга «Основы статистического обучения» (ОСО), написанная Хасти, Тибширани и Фридманом, была впервые опубликована в 2001 г. С тех пор она превратилась в важную справочную работу по фундаментальным основам статистического обучения. Ее успех обусловлен широким и детальным рассмотрением многих тем статистического обучения, а также тем фактом, что (в сравнении со многими специализированными учебниками по статистике) она доступна для широкой аудитории. Однако больше всего успех ОСО связан с тематикой этой книги. На момент публикации интерес к области статистического обучения начинал свой взрывообразный рост. ОСО стала одной из первых доступных и всеобъемлющих вводных работ по этой теме.

С момента публикации ОСО статистическое обучение продолжило свой расцвет. Развитие этой дисциплины приняло две формы. Наиболее заметный рост был связан с разработкой новых и усовершенствованных подходов статистического обучения, предназначенных для получения ответов на широкий круг вопросов в ряде научных областей. Однако статистическое обучение расширило также и свою аудиторию. В 1990–х г. рост доступности вычислительных ресурсов вызвал волну интереса к этой области со стороны неспециалистов по статистике, которым не терпелось начать использовать современные статистические инструменты для анализа своих данных. К сожалению, высокотехничная природа этих методов означала, что сообщество их пользователей оставалось ограниченным преимущественно экспертами по статистике, компьютерным и смежным областям, имеющими необходимую подготовку (и время) для освоения и реализации соответствующих методов.

В последние годы новое и усовершенствованное программное обеспечение значительно облегчило практическое применение многих методов статистического обучения. В то же время во многих областях, таких как бизнес, здравоохранение, генетика, социальные науки и т. д., произошло осознание того, что статистическое обучение является мощным инструментом для решения важных практических задач. Как следствие оно перестало быть чем–то, что представляет преимущественно академический интерес, и превратилось в популярную дисциплину с огромной потенциальной аудиторией. Несомненно, этот тренд продолжится по мере роста доступности огромных объемов данных и программного обеспечения, предназначенного для их анализа.

Цель книги «*Введение в статистическое обучение*» (ВСО) состоит в том, чтобы содействовать превращению статистического обучения из академической в популярную практическую дисциплину. ВСО не предназначена для замены ОСО, которая является гораздо более обстоятельной работой как по числу рассматриваемых в ней методов, так и по глубине их описания. Мы рассматриваем ОСО в качестве справочника для профессионалов (имеющих ученые степени по статистике, машинному обучению или сходным направлениям), которым необходимо понимать технические детали, лежащие в основе подходов статистического обучения. Однако сообщество пользователей методов машинного обучения расширилось и включает людей с более широким кругом интересов и с разным образованием. Поэтому мы убеждены, что сейчас появилось место для менее технической и более доступной версии ОСО.

В ходе преподавания этих тем на протяжении многих лет мы обнаружили, что они представляют интерес для магистрантов и аспирантов из настолько далеких друг от друга дисциплин, как бизнес–администрирование, биология и компьютерные науки, а также для ориентированных на количественные дисциплины студентов старших курсов. Для этой разнородной аудитории важно иметь возможность понимать модели, их предпосылки, а также сильные и слабые стороны различных методов. Однако многие технические детали методов статистического обучения, такие как алгоритмы оптимизации и теоретические свойства методов, для этой аудитории не представляют большого интереса. Мы убеждены, что таким студентам не нужно иметь глубокого понимания этих аспектов, для того чтобы начать осознанно применять различные методы и сделать

вклад в соответствующие научные дисциплины с помощью инструментария статистического обучения.

Книга ВСО основана на следующих четырех предпосылках.

1. *Многие методы статистического обучения применимы и полезны для широкого круга академических и практических дисциплин, выходящих далеко за рамки статистической науки.* Мы убеждены, что многие современные процедуры статистического обучения должны стать (и станут) настолько же широко доступными и используемыми, как классические методы наподобие линейной регрессии. В связи с этим вместо попытки охватить все возможные подходы (а это невыполнимая задача) мы сконцентрировались на представлении методов, которые считаем наиболее широко применимыми.

2. *Статистическое обучение не следует рассматривать как набор «черных ящиков».* Не существует метода, который одинаково хорошо сработает во всех возможных ситуациях. Без понимания всех «винтиков» внутри «ящика»» и без взаимодействия с этими «винтиками» невозможно выбрать наилучший «ящик». Поэтому мы предприняли попытку тщательно описать модель, идею, допущения и компромиссы, лежащие в основе каждого рассматриваемого нами метода.

3. *Несмотря на важность понимания функции, выполняемой каждым «винтиком», нет необходимости уметь конструировать саму машину, находящуюся внутри «ящика».* Поэтому мы минимизировали обсуждение технических деталей, имеющих отношение к процедурам подгонки моделей и теоретическим свойствам методов. Мы предполагаем, что читатель чувствует себя комфортно с простейшими математическими концепциями, но мы не ожидаем от него ученой степени в области математических наук. Например, мы почти полностью исключили использование матричной алгебры, и всю книгу можно понять без знания матриц и векторов.

4. *Мы предполагаем, что читатель интересуется применением методов статистического обучения для решения практических проблем.* Чтобы удовлетворить этот интерес и мотивировать к применению обсуждаемых методов, после каждой главы мы приводим раздел с лабораторными работами. В каждой лабораторной работе мы знакомим читателя с реалистичным практическим приложением методов, рассмотренных в соответствующей главе. Когда мы преподавали этот материал в наших курсах, мы отводили на лабораторные работы примерно треть всего времени и нашли их чрезвычайно полезными. Многие студенты, которые поначалу испытывали затруднения при работе с командным интерфейсом R, усвоили необходимые навыки в течение семестра. Мы использовали R потому, что эта система является бесплатной и достаточно мощной для реализации всех рассмотренных в книге методов. Кроме того, она имеет расширения, которые можно загрузить для реализации буквально тысяч дополнительных методов. Но важнее всего то, что R предпочитают академические статистики, и новые методы часто становятся доступными в R на несколько лет раньше того, как они появляются в платных

программах. Тем не менее лабораторные работы в ВСО автономны, и их можно пропускать, если читатель желает использовать другое программное обеспечение или не намерен применять обсуждаемые методы к реальным проблемам.

Кому следует прочесть эту книгу?

Эта книга предназначена для всех, кто интересуется применением современных статистических методов для моделирования и прогнозирования на основе данных. Эта группа читателей включает ученых, инженеров, финансовых аналитиков, а также людей с меньшей технической и математической подготовкой, которые имеют образование в таких областях, как социальные науки или бизнес. Мы ожидаем, что читатель прослушал как минимум один вводный курс по статистике. Знание линейной регрессии также полезно, но не обязательно, поскольку в главе 3 мы даем обзор ключевых концепций, лежащих в основе этого метода. Уровень математики в этой книге умеренный, и детальное знание матричных операций не требуется. Книга содержит введение в язык статистического программирования `R`. Предыдущий опыт программирования на другом языке, вроде `MATLAB` или `Python`, полезен, но не обязателен.

Мы успешно преподавали материал на этом уровне магистрантам и аспирантам, изучающим бизнес, компьютерные науки, биологию, науки о Земле, психологию и многие другие направления естественных и гуманитарных наук. Эта книга также могла бы оказаться подходящей для студентов последних курсов, которые уже прослушали курс по линейной регрессии. В контексте математически более строгого курса, где основным учебником является ОСО, ВСО можно было бы использовать в качестве дополнительного источника для преподавания вычислительных аспектов различных методов.

Обозначения и простая матричная алгебра

Выбор системы обозначений для учебника — это всегда сложная задача. В большинстве случаев мы применяем те же условные обозначения, что и в ОСО.

Мы будем использовать n для обозначения числа отдельных значений, или наблюдений, в нашей выборке. При помощи p мы будем обозначать число имеющихся переменных, на основе которых можно делать предсказания. Например, набор данных `Wage` состоит из 12 переменных для 3000 людей, так что у нас есть $n = 3000$ наблюдений и $p = 12$ переменных, таких как `year`, `age`, `wage` и др. Заметьте, что на протяжении всей этой книги для обозначения имен переменных мы используем цветной шрифт: `Имя Переменной`.

В некоторых примерах p может быть довольно большим, порядка нескольких тысяч или даже миллионов; подобная ситуация достаточно часто возникает, например, при анализе современных биологических данных или данных по интернет–рекламе.

Обычно при помощи x_{ij} мы будем обозначать i–е значение j–й переменной, где $i = 1, 2, \ldots, n$, а $j = 1, 2, \ldots, p$. На протяжении этой книги

i будет использоваться для индексирования выборок или отдельных наблюдений (от 1 до n), а j — для индексирования переменных (от 1 до p). С помощью $\mathbf{X}$ мы обозначаем матрицу размером $n \times p$, чей (i, j)–й элемент — это x_{ij}. Другими словами,

$$\mathbf{X} = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{pmatrix}.$$

Читателям, не знакомым с матрицами, полезно будет мысленно представлять $\mathbf{X}$ в виде таблицы чисел с n строками и p столбцами.

В ряде случаев нам будут интересны строки матрицы $\mathbf{X}$, которые мы записываем как $x_1, x_2, \ldots, x_n$. Здесь x_i представляет собой вектор длиной p, содержащий значения p переменных для i-го наблюдения. Другими словами,

$$x_i = \begin{pmatrix} x_{i1} \\ x_{i2} \\ \vdots \\ x_{ip} \end{pmatrix}. \tag{1.1}$$

(По определению, векторы представлены в виде столбцов.) Например, для набора данных `Wage` x_i — это вектор длиной 12, состоящий из значений `year`, `age`, `wage` и других переменных для i–го человека. В других случаях вместо строк нам будут интересны столбцы матрицы $\mathbf{X}$, которые мы записываем как $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_p$. Каждый из этих столбцов является вектором длиной n, т. е.

$$\mathbf{x}_j = \begin{pmatrix} x_{1j} \\ x_{2j} \\ \vdots \\ x_{nj} \end{pmatrix}.$$

Например, в случае с данными `Wage` $\mathbf{x}_1$ содержит $n = 3000$ значений `year`.

Используя эту нотацию, матрицу $\mathbf{X}$ можно записать как

$$\mathbf{X} = (\mathbf{x}_1 \quad \mathbf{x}_2 \quad \cdots \quad \mathbf{x}_p),$$

или

$$\mathbf{X} = \begin{pmatrix} x_1^T \\ x_2^T \\ \vdots \\ x_n^T \end{pmatrix}.$$

Символ ^{T} обозначает *транспозицию* матрицы или вектора. Так, например,

$$\mathbf{X}^T = \begin{pmatrix} x_{11} & x_{21} & \cdots & x_{n1} \\ x_{12} & x_{22} & \cdots & x_{n2} \\ \vdots & \vdots & & \vdots \\ x_{1n} & x_{2n} & \cdots & x_{np} \end{pmatrix},$$

тогда как

$$x_i^T = (x_{i1} \quad x_{i2} \quad \cdots \quad x_{ip}).$$

Мы используем y_i для обозначения i–го наблюдения переменной, которую мы хотим предсказать (например, `wage`). Следовательно, в векторной форме мы записываем набор всех n наблюдений как

$$\mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}.$$

Тогда наши наблюдаемые данные состоят из пар $\{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$, где каждый элемент x_i — это вектор длиной p. Если $p = 1$, то x_i является просто скаляром.

В этой книге вектор длиной n всегда будет обозначаться при помощи *прописной буквы, выделенной жирным шрифтом*, т. е.

$$\mathbf{a} = \begin{pmatrix} a_1 \\ a_2 \\ \vdots \\ a_n \end{pmatrix}.$$

Однако векторы, чья длина отличается от n (например, векторы признаков длиной p, как в (1.1)), будут обозначаться при помощи прописных букв, выполненных обычным шрифтом (например, a). Скаляры также будут обозначаться при помощи таких *прописных букв*, т. е. a. В редких случаях, когда использование прописных букв с обычным шрифтом может привести к двусмысленности, мы будем пояснять, что имеется в виду. Матрицы будут обозначаться с использованием *заглавных букв, выполненных жирным шрифтом* (например, $\mathbf{A}$). Случайные переменные будут обозначаться *заглавными буквами, выполненными обычным шрифтом* (например, A), вне зависимости от их размерности.

Иногда у нас будет возникать необходимость указать размерность конкретного объекта. Чтобы показать, что объект является вектором, мы будем использовать нотацию $a \in \mathbb{R}$. Чтобы показать, что это вектор длиной k, мы будем использовать обозначение $a \in \mathbb{R}^k$ (или $a \in \mathbb{R}^n$, если он имеет длину n). Объекты, которые являются матрицами размером $r \times s$, мы будем обозначать как $\mathbf{A} \in \mathbb{R}^{r \times s}$.

Мы избегали использования матричной алгебры везде, где это было возможно. Однако в нескольких случаях полностью избегать ее становилось слишком обременительно. В этих редких случаях важно понимать концепцию умножения двух матриц. Предположим, что $\mathbf{A} \in \mathbb{R}^{r \times d}$,

а $\mathbf{B} \in \mathbb{R}^{d \times s}$. Результат умножения $\mathbf{A}$ на $\mathbf{B}$ обозначается как $\mathbf{AB}$. Элемент (i, j) матрицы $\mathbf{AB}$ вычисляется путем умножения каждой i–й строки $\mathbf{A}$ на соответствующий элемент j–го столбца $\mathbf{B}$. Иначе говоря, $(\mathbf{AB})_{ij} = \sum_{k=1}^{d} a_{ik} b_{kj}$. В качестве примера представьте, что

$$\mathbf{A} = \begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix} \quad \text{и} \quad \mathbf{B} = \begin{pmatrix} 5 & 6 \\ 7 & 8 \end{pmatrix}.$$

Тогда

$$\mathbf{AB} = \begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix} \begin{pmatrix} 5 & 6 \\ 7 & 8 \end{pmatrix} = \begin{pmatrix} 1 \times 5 + 2 \times 7 & 1 \times 6 + 2 \times 8 \\ 3 \times 5 + 4 \times 7 & 3 \times 6 + 4 \times 8 \end{pmatrix} = \begin{pmatrix} 19 & 22 \\ 43 & 50 \end{pmatrix}.$$

Заметьте, что результатом этой операции является матрица размером $r \times s$. $\mathbf{AB}$ можно вычислить только, если число столбцов в $\mathbf{A}$ равно числу строк в $\mathbf{B}$.

Структура этой книги

Глава 2 вводит основные термины и концепции, лежащие в основе статистического обучения. Эта глава описывает также классификатор на основе *K ближайших соседей* — очень простой метод, который удивительно хорошо работает для многих проблем. Главы 3 и 4 охватывают классические линейные методы регрессии и классификации. В частности, глава 3 дает обзор *линейной регрессии* — фундаментальной отправной точки всех регрессионных методов. В главе 4 мы обсуждаем два наиболее важных классических метода классификации — *логистическую регрессию* и *линейный дискриминантный анализ*.

Центральной проблемой всех случаев использования статистического обучения является выбор наилучшего метода для решения конкретной задачи. Поэтому в главе 5 мы знакомим читателя с *перекрестной проверкой* и *бутстрепом*, которые можно использовать для оценивания точности нескольких методов и выбора самого подходящего из них.

Многие недавние исследования в области статистического обучения были посвящены нелинейным методам. Однако линейные методы часто имеют преимущества по сравнению со своими нелинейными конкурентами в смысле интерпретируемости, а иногда и точности предсказаний. Поэтому в главе 6 мы рассматриваем целый ряд линейных методов, как классических, так и более современных, которые предлагают потенциальные улучшения по сравнению со стандартной линейной регрессией. Речь идет о *пошаговом отборе*, *гребневой регрессии*, *регрессии на главные компоненты*, методе *частных наименьших квадратов* и *лассо–регрессии*.

Остальные главы посвящены миру нелинейного статистического обучения. Сначала в главе 7 мы вводим несколько нелинейных методов, которые хорошо работают для проблем с одной входной переменной. Затем мы показываем, как эти методы можно использовать для построения нелинейных *аддитивных* моделей с несколькими входными переменными. В главе 8 мы исследуем методы, основанные на *решающих деревьях*, включая *бэггинг*, *бустинг* и *случайные леса*. *Метод опорных векторов*

— совокупность подходов для решения задач как линейной, так и нелинейной классификации — обсуждается в главе 9. Наконец, в главе 10 мы рассматриваем ситуацию, когда у нас имеются входные переменные, но нет выходной переменной. В частности, мы описываем *анализ главных компонент*, кластеризацию с помощью *метода K средних*, а также *иерархическую кластеризацию*.

В конце каждой главы мы приводим одну или несколько лабораторных работ с использованием R, в которых подробно разбираем примеры практического применения рассмотренных в этой главе методов. Эти лабораторные работы демонстрируют сильные и слабые стороны разных подходов, а также предоставляют полезный справочный материал по синтаксису, необходимому для реализации разных методов. Возможно, читатель предпочтет работать над этими лабораторными в своем собственном темпе, но это могут быть также и групповые сессии, являющиеся частью классной работы. В каждой лабораторной работе мы представляем результаты, которые получили на момент написания книги. Однако постоянно выходят новые версии R, и со временем используемые в лабораторных работах пакеты будут обновлены. Поэтому возможно, что в будущем представленные в соответствующих разделах результаты больше не будут в точности соответствовать результатам, полученным читателем. По мере необходимости мы будем обновлять лабораторные работы на сайте книги.

Мы используем значок  для обозначения разделов и упражнений повышенной сложности. Эти разделы могут быть легко пропущены читателями, которые не желают так глубоко погружаться в соответствующий материал или не имеют необходимой математической подготовки.

Данные, использованные в лабораторных работах и упражнениях

В этом учебнике мы иллюстрируем методы статистического обучения на практических примерах из маркетинга, финансов, биологии и других областей. Пакет ISLR, доступный на сайте книги, содержит несколько наборов данных, которые необходимы для выполнения соответствующих лабораторных работ и упражнений. Один из таких наборов данных входит в состав библиотеки MASS, а еще один является частью базового дистрибутива R. Таблица 1.1 содержит сводную информацию по данным, необходимым для выполнения лабораторных работ и упражнений. Несколько таких наборов данных, используемых в главе 2, доступно также в виде текстовых файлов на сайте книги.

Веб–сайт книги

Веб–сайт этой книги находится по адресу `www.StatLearning.com`. Он содержит целый ряд ресурсов, включая связанный с книгой R–пакет и некоторые дополнительные наборы данных.

ТАБЛИЦА 1.1. *Список наборов данных, необходимых для выполнения лабораторных работ и упражнений в этом учебнике. Все данные доступны в пакете* `ISLR`*, за исключением* `Boston` *и* `USArrests` *(они входят в состав базового дистрибутива* `R`*)*

Название	Описание
`Auto`	Расход топлива, мощность двигателя и другая информация по автомобилям
`Boston`	Стоимость жилья и другая информация по пригородам Бостона
`Caravan`	Информация по людям, которым была предложена страховка для их жилых прицепов
`Carseats`	Информация по продажам автомобильных сидений в 400 магазинах
`College`	Демографические характеристики, стоимость обучения и другие данные по колледжам США
`Default`	Данные по должникам компании, выпускающей кредитные карты
`Hitters`	Данные по заработной плате бейсбольных игроков
`Khan`	Измерения генной экспрессии для четырех типов рака
`NCI60`	Измерения генной экспрессии для 64 раковых клеточных линий
`OJ`	Информация по продажам апельсинового сока марок Citrus Hill и Minute Made
`Portfolio`	Исторические значения финансовых активов, используемые для распределения инвестиционных средств
`Smarket`	Дневные удельные изменения доходности индекса S&P за 5–летний период
`USArrests`	Криминальная статистика в расчете на 100 000 жителей в 50 штатах США
`Wage`	Данные исследования доходов мужчин в центрально–атлантическом регионе США
`Weekly`	1098 значений недельной доходности рынка акций за 21 год

Благодарности

Несколько графиков в этой книге было заимствовано из OCO: рис. 6.7, 8.3 и 10.12. Все остальные графики в книге новые.

Глава 2

Статистическое обучение

2.1 Что такое статистическое обучение?

Рассмотрим простой пример, который поможет нам приступить к изучению статистического обучения. Представьте себе, что мы являемся консультантами–статистиками, нанятыми некоторым клиентом для разработки рекомендаций по повышению продаж определенного продукта. Набор данных `Advertising` включает сведения по продажам (`sales`) этого продукта в 200 различных регионах, а также по величине регионального бюджета на рекламу продукта в средствах массовой информации (СМИ) трех видов: телевидение (`TV`), радио (`radio`) и газеты (`newspaper`). Эти данные представлены на рис. 2.1. Наш клиент не имеет прямой возможности увеличить продажи. С другой стороны, он может контролировать затраты на рекламу в каждом из трех типов СМИ. Следовательно, если мы обнаружим связь между затратами на рекламу и продажами, то сможем дать рекомендации нашему клиенту по корректированию бюджета на рекламу, что опосредованно приведет к увеличению продаж. Другими словами, наша цель состоит в разработке модели, которую можно будет использовать для верных предсказаний продаж на основе данных по бюджету для трех типов СМИ.

входная и выходная переменные

При таком сценарии уровни бюджета на рекламу в разных СМИ являются *входными переменными*, тогда как `sales` — это *выходной переменной*. Входные переменные обычно обозначаются при помощи символа X с нижним индексом, позволяющим различать отдельные переменные. Так, X_1 может обозначать бюджет `TV`, X_2 — бюджет `radio`, а X_3 — бюджет `newspaper`. Входные переменные известны под разными названиями — *предикторы*, *независимые переменные*, *признаки*, или иногда просто *переменные*. Выходную переменную — в данном случае `sales` — часто называют *откликом*, или *зависимой переменной*, и обычно обозначают при помощи символа Y. В этой книге мы будем использовать указанные термины попеременно.

независимая и зависимая переменные

предиктор

признак

отклик

В качестве более общего случая представьте, что мы наблюдаем некоторый количественный отклик Y и p отдельных предикторов $X_1, X_2, \ldots, X_p$. Мы делаем предположение о том, что существует определенная связь между Y и $X = (X_1, X_2, \ldots, X_p)$, которую в очень общей форме можно записать как

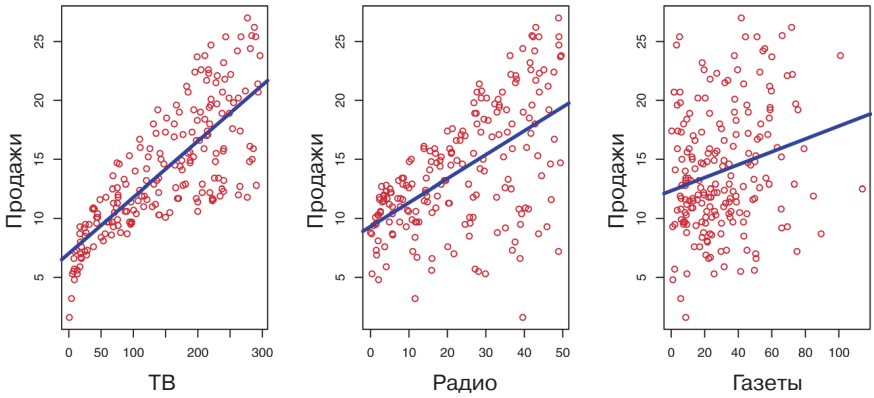

РИСУНОК 2.1. *Набор данных* `Advertising`. *На рисунке показан объем продаж* (`sales`, *тыс. единиц) в зависимости от величины бюджета на рекламу (тыс. долларов) на телевидении (*`TV`*), радио (*`radio`*) и в газетах (*`newspaper`*) в 200 регионах.*

$$Y = f(X) + \epsilon. \tag{2.1}$$

ошибка

Здесь f — это некоторая фиксированная, но не известная функция от $X_1, \ldots, X_p$, а ϵ — *ошибка*, которая не зависит от X и имеет нулевое среднее значение. В таком представлении f выражает *систематическую* информацию о Y, содержащуюся в X.

В качестве другого примера рассмотрим данные для 30 человек из таблицы `Income`, отражающие зависимость дохода (`income`) от количества лет, затраченных этими людьми на образование (`years of education`; см. рис. 2.2 слева). Этот график предполагает, что мы могли бы предсказать доход на основе длительности обучения. Однако функция, связывающая входную переменную с выходной переменной, в общем случае неизвестна. В такой ситуации мы должны оценить f на основе имеющихся наблюдений. Поскольку таблица `Income` содержит имитированные данные, то f известна и показана на рис. 2.2 справа в виде кривой голубого цвета. Вертикальные отрезки соответствуют ошибкам ϵ. Заметьте, что некоторые из 30 наблюдений лежат выше голубой линии, а некоторые — ниже, но в целом среднее значение ошибок примерно равно нулю.

В общем случае функция f может включать более одной входной переменной. На рис. 2.3 мы изображаем `income` как функцию от длительности обучения и стажа (`seniority`). Здесь f представляет собой двумерную плоскость, которую мы должны оценить на основе имеющихся данных.

В сущности, под статистическим обучением понимают совокупность методов для оценивания f. В этой главе мы рассматриваем некоторые из ключевых теоретических концепций, применяемых при нахождении f, а также инструменты для определения качества полученных оценок.

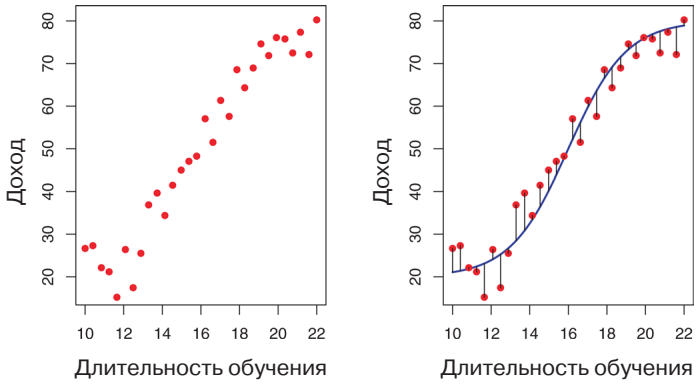

РИСУНОК 2.2. *Набор данных* `Income`. Слева: *красные точки соответствуют значениям переменных* `income` *(доход, десятки тыс. долларов) и* `years` *(длительность обучения, лет) у 30 человек.* Справа: *голубая кривая соответствует истинной функции связи между* `income` *и* `years`*, которая обычно неизвестна (в этом случае она известна, поскольку данные были имитированы). Черные вертикальные линии показывают ошибки соответствующих наблюдений. Заметьте, что некоторые ошибки положительны (когда наблюдение лежит выше голубой кривой), а некоторые — отрицательны (когда наблюдение находится ниже кривой). В целом среднее значение ошибок примерно равно нулю*

2.1.1 Зачем оценивать f?

Существуют две основные причины, по которым мы хотели бы оценить f: *предсказание* и *статистический вывод*. Обсудим каждую из этих причин по порядку.

Предсказание

Во многих ситуациях набор входных переменных X легко доступен, однако получить выходную переменную Y не так просто. Благодаря тому, что ошибки имеют нулевое среднее значение, при таком сценарии мы можем предсказать Y с помощью

$$\hat{Y} = \hat{f}(X), \tag{2.2}$$

где $\hat{f}$ представляет собой нашу оценку f, а $\hat{Y}$ — предсказанное значение Y. В подобной ситуации $\hat{f}$ часто рассматривают как *черный ящик*, в том смысле, что если $\hat{f}$ обеспечивает верные предсказания Y, то точная форма этой функции для нас не важна.

В качестве примера предположим, что $X_1, \ldots, X_p$ являются характеристиками образца крови пациента, которые легко измерить в лаборатории, а Y — это переменная, отражающая риск того, что пациент проявит резко негативную реакцию на определенный лекарственный препарат. Естественным будет желание предсказать Y по X, поскольку таким образом

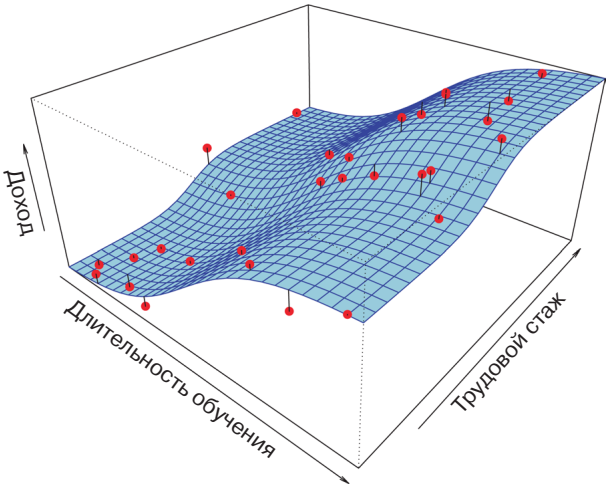

РИСУНОК 2.3. *График показывает* `income` *как функцию от* `years` *и* `seniority` *(переменные из набора данных* `Income`*). Голубая плоскость отражает истинную зависимость* `income` *от* `years` *и* `seniority`*, которая известна, поскольку эти данные были имитированы. Красные точки показывают наблюденные значения для 30 человек*

мы сможем избежать назначения данного препарата пациентам с высоким риском негативной реакции, т. е. пациентам с высокими значениями оценок Y.

устранимая и неустранимая ошибки

Точность $\hat{Y}$ в качестве предсказанного значения Y зависит от двух величин, которые мы будем называть *устранимой ошибкой* и *неустранимой ошибкой*. Обычно $\hat{f}$ не будет идеальной оценкой f, и эта неточность приведет к возникновению некоторой ошибки. Такая ошибка является *устранимой*, поскольку потенциально мы можем улучшить точность $\hat{f}$, используя более подходящий статистический метод для оценивания f. Но даже если бы имелась возможность достичь настолько идеальной оценки f, что $\hat{Y} = f(X)$, наше предсказанное значение все равно содержало бы в себе некоторую ошибку! Подобная ошибка известна как *неустранимая*, поскольку как бы хорошо мы ни оценили f, мы не сможем снизить ошибку, привнесенную за счет ϵ.

Почему же неустранимая ошибка превышает нулевое значение? Величина ϵ может включать неучтенные переменные, которые полезны для предсказания Y, но поскольку мы их не измеряем, то f не может использовать их для предсказания. Например, риск негативной реакции может варьировать у того или иного пациента в определенный день в зависимости условий производства самого лекарственного препарата или от общего состояния здоровья пациента в тот день.

Представьте себе некоторую конкретную оценку $\hat{f}$ и набор предикторов X, которые дают предсказание $\hat{Y} = \hat{f}(X)$. Допустите на мгновение, что и $\hat{f}$, и X являются фиксированными величинами. Тогда можно легко показать, что

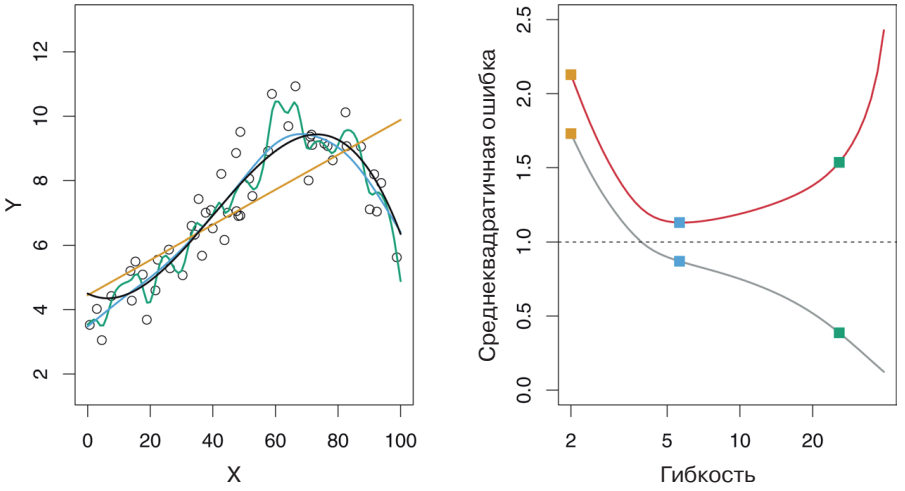

РИСУНОК 2.9. Слева: *данные, имитированные на основе f, показаны полыми точками черного цвета. Представлены три способа подгонки f: линейная регрессия (оранжевая линия) и две модели гладких сплайнов (голубая и зеленая линии).* Справа: *среднеквадратичная ошибка на обучающих (серая линия) и контрольных данных (красная линия), а также минимально возможное значение ошибки для контрольных данных (пунктирная линия). Квадратные символы соответствуют ошибкам на обучающих и контрольных выборках, которые были получены для трех моделей, изображенных на графике слева*

монотонно снижается по мере возрастания гибкости. В этом примере истинная функция f нелинейна, в связи с чем линейная модель (оранжевый цвет) недостаточно гибка для получения удовлетворительной оценки f. Из всех трех методов зеленая кривая имеет наименьшую MSE на обучающих данных, поскольку она соответствует наиболее гибкой из трех моделей, представленных на графике слева.

В этом примере истинная функция f нам известна, и поэтому мы также можем вычислить MSE для очень большой контрольной совокупности в зависимости от гибкости модели. (Конечно, как правило, f неизвестна, и такие вычисления будут невозможны.) MSE на контрольной выборке показана на рис. 2.9 справа в виде красной кривой. Как и в случае с MSE на обучающих данных, MSE на контрольной выборке сначала снижается по мере увеличения гибкости. Однако в какой–то момент MSE на контрольной выборке снова начинает возрастать. Как следствие оранжевая и зеленая кривые имеют самые высокие ошибки на контрольных данных. Голубая кривая имеет минимальную MSE на контрольной выборке, что не должно удивлять, поскольку визуально эта модель выглядит как наилучшая оценка f (см. рис. 2.9 слева). Горизонтальная прерывистая линия показывает $Var(\epsilon)$ — неустранимую ошибку из (2.3), которая соответствует наименьшей достижимой MSE на контрольных данных для всех возможных методов. Следовательно, сглаживающий сплайн, показанный в виде голубой линии, близок к оптимуму.

Как видно на рис. 2.9 (справа), по мере возрастания гибкости метода статистического обучения происходит монотонное снижение MSE на обучающих данных и *U–образное* изменение MSE на контрольных данных. Это является фундаментальным свойством статистического обучения, которое остается справедливым для любого имеющегося набора данных и для любого применяемого статистического метода. По мере увеличения гибкости модели MSE на обучающей выборке будет снижаться, но MSE на контрольной выборке необязательно будет вести себя тем же образом. Ситуацию, когда некоторый метод обеспечивает небольшую MSE на обучающих данных и высокую MSE на проверочных данных, называют *переобучением* модели. Это происходит потому, что наша процедура статистического обучения слишком усердно пытается найти закономерности в обучающих данных и в результате может обнаружить некоторые закономерности, которые просто случайны и никак не связаны с истинными свойствами неизвестной функции f. При переобучении модели MSE на контрольной выборке будет большой потому, что предполагаемые закономерности, найденные нашим методом в обучающих данных, в проверочных данных просто не существуют. Заметьте, что вне зависимости от того, случилось ли переобучение, мы почти всегда ожидаем, что MSE на обучающих данных будет ниже, чем MSE на контрольных данных, поскольку большинство методов статистического обучения так или иначе пытаются минимизировать ошибку обучения. К переобучению относят также частный случай, когда менее гибкая модель обеспечивает меньшую MSE на контрольных данных.

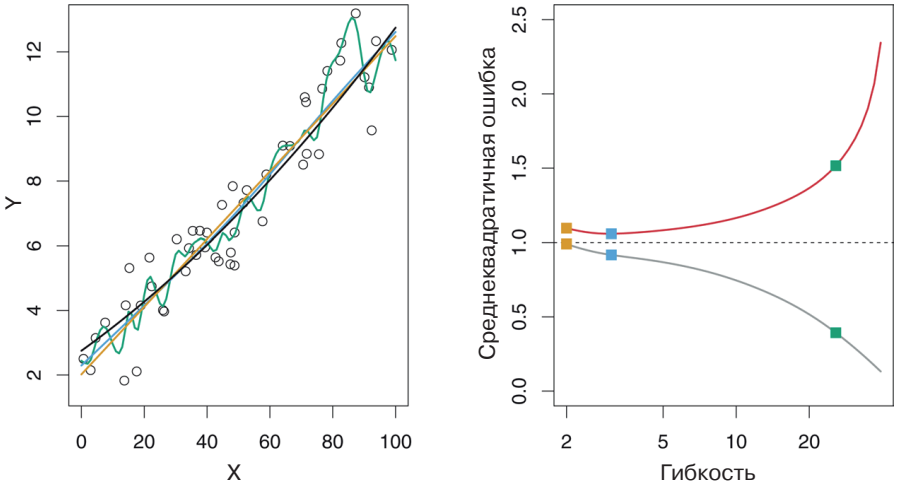

РИСУНОК 2.10. *То же, что на рис. 2.9, но с истинной функцией f, которая намного ближе к линейной. В этой ситуации линейная регрессия очень хорошо описывает данные*

На рис. 2.10 приведен пример, в котором истинная функция f приблизительно линейна. Как и раньше, мы наблюдаем монотонное снижение MSE на обучающих данных по мере возрастания гибкости модели, а кривая MSE на контрольных данных имеет U–образную форму. Одна-

ко в связи с тем, что истинная зависимость близка к линейной, MSE на контрольных данных лишь незначительно снижается перед последующим возрастанием, и поэтому оранжевая линия, подогнанная по методу наименьших квадратов, подходит для описания данных гораздо лучше, чем очень гибкая зеленая кривая. Наконец, рис. 2.11 иллюстрирует пример, в котором функция f в значительной мере нелинейна. Кривые MSE для обучающих и контрольных выборок по–прежнему демонстрируют те же общие закономерности, но теперь имеет место быстрое снижение обеих этих кривых, перед тем как MSE на контрольных данных начинает медленно возрастать.

На практике вычислить MSE по обучающим данным относительно легко, однако оценить MSE на контрольных данных гораздо труднее, поскольку они обычно отсутствуют. Как показывают предыдущие три примера, уровень гибкости, соответствующий модели с наименьшей контрольной MSE, может значительно варьировать в зависимости от свойств данных. В этой книге мы обсуждаем целый ряд подходов, которые можно использовать на практике получения оценки этого минимального значения. Одним из важных методов, предназначенных для оценивания MSE на контрольных данных, является *перекрестная проверка*[6] (глава 5).

перекрестная проверка

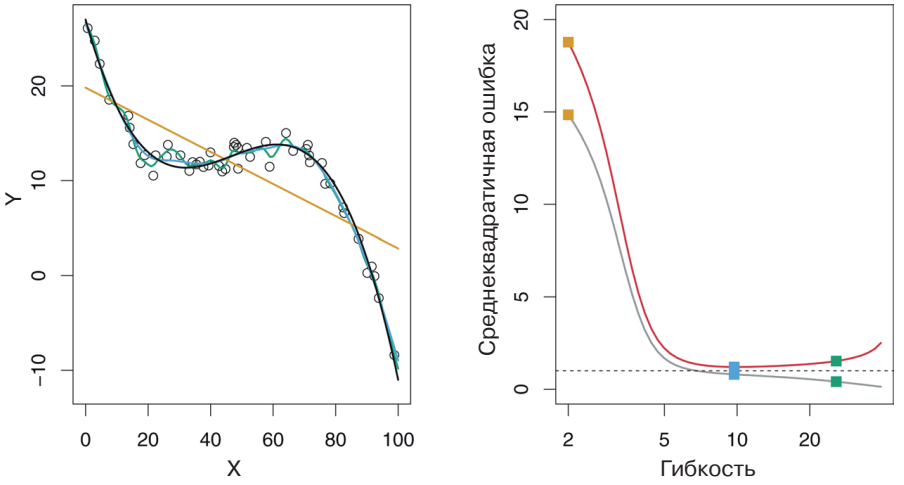

РИСУНОК 2.11. *То же, что на рис. 2.9, но с истинной функцией f, существенно отличной от линейной. В этой ситуации линейная регрессия описывает данные очень плохо*

2.2.2 Компромисс между смещением и дисперсией

Оказывается, что U–образная форма кривых, описывающих MSE на контрольных данных (рис. 2.9–2.11), является результатом двух конкурирующих свойств методов статистического обучения. Хотя математическое доказательство выходит за рамки этой книги, можно показать, что для

[6] Используются также термины «кросс–проверка», «кросс–валидация» и «скользящий контроль». — *Прим. пер.*