Моей жене и семье, благодарю вас
за вашу любовь, поддержку и терпение.
Джон

Моей семье.
Брайан

Дедушке Д'Арси за вдохновение.
Аоифе

Предисловие

При написании этой книги наша цель состояла в том, чтобы создать доступный, вводный учебник по основам машинного обучения и описать способы использования методов машинного обучения на практике для решения задач аналитического прогнозирования в бизнесе, науке и других областях. Таким образом, книга выходит за рамки стандартных тем, охватываемых книгами по машинному обучению, и описывает жизненный цикл проекта по аналитическому прогнозированию — подготовку данных, выбор признаков и развертывание модели.

Книга предназначена для преподавания методов машинного обучения, интеллектуального анализа данных, аналитического прогнозирования и искусственного интеллекта студентам и аспирантам естественных, технических, экономических и гуманитарных наук. Тот факт, что в книге приводятся тематические исследования, иллюстрирующие применение машинного обучения в отраслевом контексте анализа данных, также делает ее подходящей для практиков, которые ищут введение в эту область, и студентов соответствующих специальностей.

Структура книги основана на нашем многолетнем опыте преподавания методов машинного обучения, а подход и материалы были разработаны и проверены на практике. Для того чтобы сделать материал доступным, при написании книги мы придерживались следующих руководящих принципов.

1. Необходимо объяснять только наиболее важные и популярные алгоритмы, а не делать полный обзор методов машинного обучения. Как преподаватели, мы считаем, что предоставление студентам глубоких знаний о ключевых концепциях дает им прочную основу, опираясь на которую они могут самостоятельно исследовать эту область. Этот более четкий фокус позволяет больше времени уделять объяснению, демонстрации, конкретизации и использованию фундаментальных алгоритмов.
2. Следует неформально объяснять, что делает алгоритм, прежде чем давать формальное описание того, как он это делает. Такое неформальное введение в каждую тему дает студентам прочную основу для перехода к темам технического характера. Наш опыт обучения этому материалу смешанным аудиториям студентов, аспирантов и специалистов показал, что такие неформальные введения позволяют им легко войти в тему.

3. Необходимо приводить полностью проработанные примеры. В книге мы полностью проработали все примеры, чтобы читатель мог тщательно проверить, насколько он понял содержание книги.

Структура книги

При изложении технических вопросов важно показать применение концепций к решению реальных проблем. По этой причине мы представляем машинное обучение в контексте аналитического прогнозирования (predictive data analytics)[1] — важной и быстро развивающейся области машинного обучения. Связь между машинным обучением и анализом данных проходит через каждую главу книги. В главе 1 мы делаем введение в машинное обучение и объясняем роль, которую оно играет в рамках стандартного жизненного цикла проекта по анализу данных. В главе 2 мы закладываем основы для решения задач аналитического прогнозирования на основе машинного обучения, которое отвечает потребностям бизнеса. Все алгоритмы машинного обучения предполагают, что набор данных доступен для обучения, поэтому в главе 3 мы объясняем, как проектировать, создавать и проверять качество набора данных перед его использованием в модели аналитического прогнозирования.

Основной материал по машинному обучению изложен в главах 4–7. Каждая из этих глав представляет отдельный подход к машинному обучению: глава 4 — обучение через сбор информации, глава 5 — обучение на аналогиях, глава 6 — обучение путем прогнозирования вероятных результатов, глава 7 — обучение с помощью поиска решений, минимизирующих ошибки. Все эти главы имеют одну и ту же структуру.

- В первой части каждой главы дается неофициальное введение в материал, представленный в главе, за которым следует подробное объяснение основных технических понятий, необходимых для понимания материала. Затем описывается стандартный алгоритм машинного обучения, используемый в этом подходе, и приводится подробно проработанный пример.
- Во второй части каждой главы описываются различные способы обобщения стандартного алгоритма и его известные варианты.

---

[1] *В русскоязычной литературе используются также термины* ***интеллектуальный анализ данных****, а также* ***предикативная, прогнозная или предсказательная аналитика****.* — Примеч. ред.

Такая структура, состоящая из двух частей, обеспечивает естественное разделение материала, изложенного в главе. В результате тема может быть включена в курс просто в виде первой части (основная идея, основы, стандартный алгоритм и пример работы), а затем — если есть время — ее можно расширить за счет второй части. В главе 8 объясняется, как оценивать эффективность моделей прогнозирования, и представлен ряд оценочных показателей. Эта глава также состоит из двух частей с последующими обобщениями и вариантами. Во всех этих главах связь с более широким контекстом аналитического прогнозирования обеспечивается подробными и полными примерами из реального мира, а также ссылками на наборы данных и/или документы, на которых основаны эти примеры.

Связь между более широким бизнес-контекстом и машинным обучением наиболее ярко проявляется в тематических исследованиях, представленных в главе 9 (предсказание оттока клиентов) и в главе 10 (классификация галактик). В частности, в этих тематических исследованиях подчеркивается важность ряда аспектов, помимо построения модели, таких как понимание бизнеса, определение проблемы, сбор и подготовка данных, а также представление выводов, имеющих решающее значение для успеха проекта по аналитическому прогнозированию. Наконец, в главе 11 обсуждается ряд фундаментальных тем машинного обучения, а также подчеркивается, что выбор подходящего подхода к компьютерному обучению для данной задачи включает в себя факторы, не зависящие от модели, т.е. необходимо согласовывать характеристики модели с потребностями бизнеса.

Как использовать книгу

В ходе нашей преподавательской карьеры мы выработали понимание того, какое количество материала является разумным для вводного курса, который читается в течение одного семестра, и более полного курса, занимающего два семестра. Для того чтобы облегчить использование книги в этих разных контекстах, ее структура сделана модульной — главы мало зависят друг от друга. В результате лектор, использующий книгу, может планировать свой курс, выбирая разделы книги, материал которых он хотел бы изложить, и не беспокоиться о зависимости между разделами. При изложении в рамках курса материал глав 1, 2, 9–11 обычно занимает от двух до трех академических часов, а материал глав 3–8 — от четырех до шести часов.

В табл. 1 мы перечислили ряд предлагаемых планов курсов, ориентированных на различные контексты. Все эти курсы включают в себя главу 1 (“Методы

машинного обучения для аналитического прогнозирования") и главу 11 ("Искусство машинного обучения для аналитического прогнозирования"). Первый курс (*Краткое введение в машинное обучение — КВМО*) предназначен для преподавания в одном семестре. В нем основное внимание уделяется тому, чтобы дать студентам глубокое понимание двух подходов к машинному обучению, а также правильной методологии, используемой при оценке моделей машинного обучения. В предложенном нами курсе мы решили охватить весь материал из главы 4 ("Информационное обучение") и главы 7 ("Обучение на основе ошибок"). Тем не менее вместо этого можно использовать главу 5 ("Обучение на основе сходства") и главу 6 ("Вероятностное обучение"). Введение в машинное обучение — также идеальный план курса для короткого (одна неделя) профессионального курса. Второй курс (*Расширенное введение в машинное обучение— РВМО*) — это еще один курс машинного обучения на один семестр. Однако здесь основное внимание уделяется охвату ряда подходов к машинному обучению и, опять же, подробно рассматриваются оценки. Для более продолжительного курса машинного обучения (*Полный курс машинного обучения — ПКМО*) на протяжении двух семестров мы предлагаем охватить подготовку данных (раздел 3.6), все главы о машинном обучении и главу об оценках.

Однако есть контексты, в которых основное внимание в курсе уделяется не только машинному обучению. Мы также предоставляем материал для курсов, которые посвящены аналитическому прогнозированию. *Краткое введение в аналитическое прогнозирование* (КВАП) определяет курс на один семестр. Этот курс представляет собой введение в аналитическое прогнозирование и дает студентам глубокое понимание того, как разрабатываются модели машинного обучения для удовлетворения потребностей бизнеса, как работают и оцениваются модели прогнозирования, а также описывает тематический пример. Краткое введение в аналитическое прогнозирование также является идеальным коротким (одна неделя) курсом профессиональной подготовки. Если есть больше времени, то этот курс можно дополнить до *расширенного введения в аналитическое прогнозирование* (РВАП), чтобы студенты получили более глубокое и широкое понимание машинного обучения, а также изучили второй тематический пример.

Таблица 1. Описание предлагаемых курсов

Глава	Раздел	КМВО	РВМО	ПКМО	КВАП	РВАП
1		●	●	●	●	●
2					●	●

Окончание табл. 1

Глава	Раздел	КМВО	РВМО	ПКМО	КВАП	РВАП
3	3.1, 3.2				●	●
	3.3, 3.4				●	●
	3.5				●	●
	3.6	●	●	●		●
4	4.1, 4.2, 4.3	●	●	●	●	●
	4.4.1	●		●		
	4.4.2	●		●		
	4.4.3	●		●		
	4.4.4	●	●	●		
	4.4.5	●	●	●		●
5	5.1, 5.2, 5.3		●	●		●
	5.4.1		●	●		●
	5.4.2			●		
	5.4.3		●	●		●
	5.4.4			●		
	5.4.5			●		
	5.4.6		●	●		●
6	6.1, 6.2, 6.3		●	●		●
	6.4.1		●	●		●
	6.4.2			●		
	6.4.3			●		
	6.4.4			●		
7	7.1, 7.2, 7.3	●	●	●	●	●
	7.4.1	●		●		●
	7.4.2	●		●		●
	7.4.3	●		●		●
	7.4.4	●	●	●		●
	7.4.5	●	●	●		
	7.4.6	●	●	●		
	7.4.7	●	●	●		
8	8.1, 8.2, 8.3	●	●	●	●	●
	8.4.1	●	●	●		●
	8.4.2	●	●	●		●
	8.4.3	●	●	●		●
	8.4.4	●	●	●		●
	8.4.5	●	●	●		●
	8.4.6					●
9					●	●
10						●
11		●	●	●	●	●

Интернет-ресурсы

Веб-сайт

`www.machinelearningbook.com`

предоставляет доступ к широкому спектру материалов, дополняющих книгу. Этот материал включает в себя слайды лекций, полный набор рисунков, используемых в книге, видеолекции на основе книги, образцы кода и список ошибок (надеемся, короткий), а также решения для всех упражнений, приведенных в конце каждой главы. Для задач, которые не отмечены звездочкой, на веб-сайте книги приводятся указания. Решения задач, которые отмечены звездочкой, содержатся в руководстве для преподавателей, которое можно получить в издательстве MIT Press по запросу.

Благодарности

Начиная писать книгу, мы знали, что нам предстоит большая работа. Однако мы недооценили объем поддержки, которая нам понадобится, и получили ее от других людей. Мы рады воспользоваться возможностью выразить им признательность за вклад в книгу. Мы хотели бы поблагодарить наших коллег и студентов за их помощь и терпение, а также сотрудников MIT Press, в частности Мари Лафкин Ли (Marie Lufkin Lee), и нашего технического редактора Мелани Мэллон (Melanie Mallon). Хотим поблагодарить и обоих анонимных рецензентов, которые предоставили полезные комментарии на первый вариант рукописи. Каждому из нас также повезло иметь поддержку близких друзей и семьи, что было неоценимо при работе над книгой.

Джон хотел бы поблагодарить Роберта Росса (Robert Ross), Саймона Добника (Simon Dobnik), Йозефа ван Генэбита (Josef van Genabith), Алана Макдоннелла (Alan Mc Donnell), Лорейн Бирн (Lorraine Byrne) и всех его друзей по баскетболу. Он также хотел бы поблагодарить своих родителей (Джона и Бетти) и сестер — без их поддержки он не научился бы читать и писать. Наконец, он хотел выразить признательность Афре. Работа над этой книгой не началась бы без ее вдохновения и не была бы завершена без ее терпения.

Брайан благодарит своих родителей (Лиама и Ройшэн) и семью за их поддержку, а также выражает признательность Падрейгу Каннингему (Pádreig Cunningham) и Саре Джейн Делани (Sarah Jane Delany), которые открыли ему глаза на машинное обучение.

Аоифе хотела бы поблагодарить своих родителей (Майкла и Мейрид) и семью, а также всех людей, которые поддерживали ее, особенно очень ценных клиентов компании The Analytics Store, которые предоставили ей свои данные для обработки!

Обозначения

В этом разделе мы приводим краткий обзор формальных обозначений, используемых в этой книге.

Условные обозначения

В этой книге мы обсудим использование алгоритмов машинного обучения для настройки моделей прогнозирования на основе наборов данных. В следующем списке приведены обозначения, используемые для разных элементов в наборе данных. На рис. 1 показаны основные обозначения на основе простого образца данных.

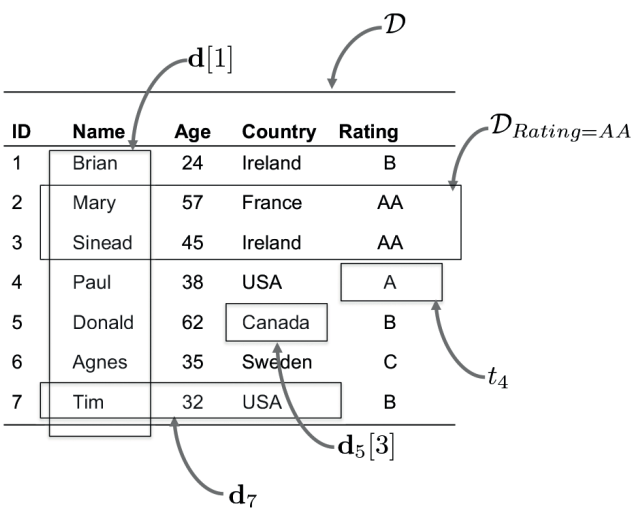

Рис. 1. Обозначения элементов наборов данных, принятые в книге

Наборы данных

- Символ $\mathcal{D}$ обозначает набор данных.
- Набор данных состоит из n экземпляров, ($\mathbf{d}_1$, t_1) — ($\mathbf{d}_n$, t_n), где $\mathbf{d}$ — набор m описательных признаков, t — целевой признак.
- Подмножество набора данных обозначается символом $\mathcal{D}$ с индексом, указывающим определение подмножества. Например, $\mathcal{D}_{f=l}$ представляет подмножество экземпляров из набора данных $\mathcal{D}$, где признак f имеет значение l.

Векторы признаков

- Вектор признаков обозначается строчными полужирными буквами. Например, **d** обозначает вектор описательных признаков экземпляра в наборе данных, а **q** — вектор описательных признаков в тестовом экземпляре.

Экземпляры

- Для индексации списка экземпляров используются нижние индексы.
- $\mathbf{x}_i$ обозначает i-й экземпляр в наборе данных.
- $\mathbf{d}_i$ обозначает описательные признаки i-го экземпляра в наборе данных.

Индивидуальные признаки

- Отдельные признаки обозначаются строчными буквами (например, f, a, b, c ...).
- Для индексации элементов в векторе признаков используются квадратные скобки [] (например, $\mathbf{d}[j]$ — это значение j-го признака в векторе **d**).
- t — целевой признак.

Индивидуальные признаки в конкретном экземпляре

- $\mathbf{d}_i[j]$ обозначает значение j-го описательного признака i-го экземпляра в наборе данных.
- a_i — это значение признака a в i-м экземпляре в наборе данных.
- t_i — это значение целевого признака i-го экземпляра в наборе данных

Индексы

- Обычно буква i используется для индексирования экземпляров в наборе данных, а буква j — для индексации признаков в векторе.

Модели

- Для обозначения модели используется буква $\mathbb{M}$.
- $\mathbb{M}_{\mathbf{w}}$ — это модель $\mathbb{M}$, параметризованная вектором параметров **w**.
- $\mathbb{M}_{\mathbf{w}}[\mathbf{d}]$ — это выход модели, параметризованной параметрами **w** для описательных признаков **d**.

Размер набора

- Две вертикальные черты (| |) обозначают количество вхождений (например, $|a = l|$ — это число, обозначающее, сколько раз в наборе данных выполняется условие $a = l$).

Названия и значения признаков

- Для ссылок на признаки в тексте используются малые прописные буквы (например, Позиция, Размер претензии и Кредитоспособность).
- Для категориальных признаков используется курсив, чтобы указать уровни в области значений при ссылке на признак по имени (например, *центровой*, *аа* и *мягкие ткани*).

Условные обозначения для вероятностей

Для ясности в главе 6 мы используем дополнительные условные обозначения для вероятностей.

Обычные события

- Обычные события, в которых неопределенному признаку (или набору признаков) присваивается значение (или набор значений), обозначаются прописными буквами. Обычно для этой цели используются последние буквы алфавита, например X, Y, Z.
- Для перебора событий используются индексы и прописные буквы. Так, запись $\sum_i P(X_i)$ следует интерпретировать как суммирование по множеству событий, которые являются полным присваиванием признаков при событии X (т.е. все возможные комбинации значений признаков при событии X).

Именованные признаки

- Признаки, явно названные в тексте, обозначаются прописными начальными буквами их имен. Например, признак, названный Менингит, обозначается как M.

События, связанные с бинарными признаками

- Если именованный признак является бинарным, то для обозначения события, в котором признак является истинным, используется строчная начальная буква имени признака, а строчная начальная буква, которой предшествует символ ¬, обозначает событие, в котором он является ложным. Итак, *m* означает событие Менингит = *true*, а ¬*m* — событие Менингит = *false*.

События, связанные с небинарными признаками

- Для перебора значений в области значений признака используются строчные буквы с индексами:

  $$\sum_i P(X_i) = P(m) + P(\neg m) .$$

- В ситуациях, когда буква, например *X*, обозначает совместное событие, запись $\sum_i P(X_i)$ следует интерпретировать как суммирование по всем возможным комбинациям значений признаков при событии *X*.

Вероятность события

- Вероятность того, что признак *f* равен значению *v*, записывается как $P(f = v)$.

Распределение вероятностей

- Для того чтобы отличить распределение вероятностей **P**() от функции вероятностной массы *P*(), используется полужирный шрифт.
- Мы придерживаемся соглашения о том, что первым элементом в векторе распределения вероятностей является вероятность истинного значения. Например, распределение вероятности для бинарного признака *A* с вероятностью истины, равной 0,4, будет записано так: **P**(*A*) = <0,4; 0,6>.