Оглавление

Предисловие

Введение

Машинное обучение применяет математические модели к данным, чтобы делать аналитические выводы или предсказания. Эти модели принимают признаки на вход. *Признак* — это числовое представление некоторого аспекта сырых данных. Он находится между данными и моделью в процессе машинного обучения. *Конструирование признаков* — это извлечение признаков из необработанных данных и приведение их к формату, пригодному для обработки моделью машинного обучения. Это один из важнейших шагов во всем процессе, так как правильно подобранные признаки облегчают сложное моделирование и, как следствие, способствуют выводу более качественных результатов. Но, несмотря на всю важность, отдельно данная тема обсуждается достаточно редко. Возможно, это происходит потому, что правильные признаки можно определить только в контексте модели и данных, а так как данные и модели могут быть очень разнообразными, сложно выделить общую тактику конструирования признаков для различных проектов.

Тем не менее конструирование признаков — это не ситуативная практика. Есть более глубокие принципы работы, но лучше всего их можно проиллюстрировать на конкретном примере. Каждая глава решает определенную проблему: как представить текстовые данные или изображение, как понизить размерность автоматически сгенерированных признаков, когда и как их нормализовать, и др. Взгляните на эту книгу как на серию взаимосвязанных историй, а не как на единый роман. В этих историях представлены краткие сцены из большого разнообразия всевозможных техник конструирования признаков. А все вместе они образуют единую концепцию.

Знание предмета подразумевает не только изучение определений и умение выводить формулы. Этого недостаточно, чтобы понимать, как работает механизм и что конкретно он делает, — нужно также разобраться, почему он спроектирован именно таким образом, как соотносится с другими техниками и каковы преимущества и недостатки каждого из подходов. Знание предмета заключается в точной осведомленности в работе процессов, интуитивном понимании

внутренних механизмов и связывании этого в единую сеть знания. Никто не становится экспертом в чем-либо, просто читая книгу, хотя хорошая книга и открывает новые двери. Чтобы найти применение идеям, необходима практика, а это цикличный процесс. На каждом круге мы глубже изучаем их и становимся все более искусными. Цель этой книги в том, чтобы ускорить процесс реализации идей.

В первую очередь мы попытаемся научить вас рассуждать, а уже потом мыслить математически. Вместо того чтобы просто описывать, *как* что-то работает, мы пробуем объяснить *почему*. Наша цель — приоткрыть завесу того, что скрывается за идеями, чтобы вы могли понять, когда и как их применять. Вы столкнетесь с множеством описаний и картинок, и это связано с тем, что каждый воспринимает информацию по-своему. Математические формулы приведены для того, чтобы сделать представление более точным, а также чтобы связать эту книгу с другими потрясающими теориями.

Примеры кода в книге приведены на языке Python, при этом использованы различные бесплатные модули с открытым кодом. Библиотека NumPy (**www.numpy.org**) предоставляет числовые векторы и операции над матрицами. Pandas (**pandas.pydata.org**) предоставляет DataFrame (кадр данных) — строительный блок анализа данных в Python. Scikit-learn (**scikit-learn.org/stable/**) — это пакет общего назначения для машинного обучения, в который включен широкий набор математических моделей и преобразования признаков. Matplotlib (**matplotlib.org**) и библиотека стилей Seaborn (**seaborn.pydata.org**) предоставляют поддержку схем и визуализации. Вы можете найти примеры в нашем репозитории на GitHub (**github.com/alicezheng/feature-engineering-book**).

Первые несколько глав могут показаться вам не слишком насыщенными, так как это лишь вступление для тех, кто только начал свой путь в анализ данных и машинном обучении. Первая глава раскроет фундаментальные основы процесса машинного обучения (данные, модели, признаки и т. д.). Во второй главе мы рассмотрим базовое конструирование признаков для числовых данных: фильтрацию, связывание, масштабирование, логарифмические преобразования и степенные, а также признаки взаимодействия. Глава третья погрузит нас в конструирование признаков для текста на естественном языке, открывая такие техники, как мультимножество слов, *n*-граммы и обнаружение устойчивых фраз. В главе четвертой мы изучим TF-IDF (частота встречаемости термина) в качестве примера масштабирования признаков и обсудим, почему это работает. В пятой главе темп начнет нарастать — мы рассмотрим эффективные техники кодирования для качественных переменных, включая хеширование признаков и двоичный счет. К моменту, когда мы доберемся до анализа основных

компонентов (PCA — principal component analysis) в главе шестой, мы уже глубоко внедримся в мир машинного обучения. В седьмой главе метод *k*-средних рассматривается как техника извлечения признаков, которая демонстрирует пользу идеи объединения моделей. Восьмая глава посвящена изображениям, которые с точки зрения извлечения признаков намного сложнее, чем текстовые данные. Мы изучим две техники ручного извлечения — SIFT и HOG, — прежде чем перейти к объяснению глубокого обучения — новейшего метода извлечения признаков для изображений. Завершающим аккордом станет демонстрация нескольких различных техник в девятой главе, а также мы создадим рекомендации для набора данных научных работ.

В цвете
Иллюстрации в этой книге лучше всего изучать в цвете. Советуем распечатать швейцарский рулет из седьмой главы в цветном виде и вложить его в книгу. Чувство прекрасного будет вам очень благодарно.

Конструирование признаков — обширная тема, и всё новые методы появляются каждый день, особенно в области автоматического изучения признаков. Чтобы не превышать разумные объемы книги, нам пришлось опустить некоторые вещи. Мы обойдем стороной частотный анализ для аудио, хотя это великолепная тема, которая тесно связана с характеристическим анализом в линейной алгебре (а ее мы затронем в четвертой и шестой главах). Мы также опустим обсуждение случайных признаков, которые не менее близки к частотному анализу. Зато займемся изучением признаков с помощью глубокого обучения для изображений, но не будем слишком погружаться в его бесчисленные модели, которые сейчас активно развиваются. В книгу не входят продвинутые идеи, вроде случайной проекции, моделей извлечения признаков сложных текстов (таких как word2vec и кластеризация Брауна), а также моделей скрытого пространства, латентного размещения Дирихле или факторизации матрицы. Если эти слова ничего для вас не значат, то вы счастливчик!

В книге много информации о базовых идеях машинного обучения, например понятиях модели и вектора, хотя памятка все равно приводится. Опыт в линейной алгебре или оптимизации будет полезным, но он не обязателен.

Условные обозначения

В этой книге использовались следующие типографические обозначения:

Курсивный шрифт
: Обозначает новые термины, имена файлов и их расширения.

Полужирный шрифт
: Обозначает URL-адреса и адреса электронной почты.

`Моноширинный шрифт`
: Используется в листингах программ, а также в тексте, который относится к элементам программы — именах переменных или функций, баз данных, типов данных, переменных среды, утверждениях или ключевых словах.

Полужирный моноширинный шрифт
: Указывает на команды или любой другой текст, который пользователь должен ввести сам.

`Курсивный моноширинный шрифт`
: Указывает на текст, который следует заменить значениями пользователя или значениями, следующими из контекста.

Вы встретите множество уравнений линейной алгебры. Мы используем следующие обозначения в соответствии с нотациями: скаляры написаны в строчном регистре курсивом (например, a), векторы — строчным жирным (например, $\mathbf{v}$), а матрицы обозначены заглавными буквами жирным и курсивом (например, $\boldsymbol{U}$).

Этим элементом обозначена подсказка или предложение.

Этот элемент обозначает общие указания.

Этот элемент означает предупреждение или предостережение.

Примеры кода

Дополнительные материалы (примеры кода, упражнения и т. д.) можно скачать здесь: **github.com/alicezheng/feature-engineering-book**.

Благодарности

Прежде всего мы хотели бы поблагодарить наших редакторов Шеннон Кат и Джеффа Блаэля, которые заботливо вели двух начинающих авторов по длинному и неизвестному пути издания книги (мы делали это впервые). Без ваших правок эта книга не увидела бы свет. Спасибо также Бену Лорика, идейному вдохновителю O'Reilly, чьи поддержка и одобрение превратили одну сумасшедшую идею в настоящую книгу. Спасибо Кристен Браун и всему производственному отделу O'Reilly за вашу чрезвычайную внимательность к деталям и невероятное терпение в ожидании нашего ответа.

Говорят, чтобы вырастить ребенка, нужно семь нянек, а чтобы издать книгу, понадобится целое государство аналитиков. Мы очень ценим все замечания и предложения по улучшению. Андрэас Мюллер, Сефу Раман и Энтони Атала нашли время в своем плотном графике, чтобы проанализировать книгу с технической точки зрения. Энтони не только делал это практически со скоростью звука, но еще и предоставил нам вычислительные мощности, чтобы мы могли экспериментировать. Знания статистики и опыт в применении машинного обучения Тома Даннинга просто легендарны. Он не скупился на время и идеи и буквально подарил нам метод *k*-средних и пример к нему, который мы описываем в одной из глав. Оуэн Чжен поделился приемами работы с показателем отклика на платформе Kaggle. Они были добавлены к алгоритму машинного обучения, связанному со счетчиком товаров в корзине, подобранному Мишей Биленко. Спасибо также Алексу Отту, Франциско Мартину и Дэвиду Гаррисону за ваши отзывы и обратную связь.

Особая благодарность Элис Чжен

Я бы хотела поблагодарить команды GraphLab/Dato/Turi за поддержку на первых этапах проекта. Идея родилась из общения с нашими пользователями. В процессе разработки принципиально новой платформы для специалистов по анализу данных мы выяснили, что миру нужно более систематическое и глубокое понимание конструирования признаков. Спасибо Карлосу Гестрину за то, что

позволил мне выйти из напряженного развития начинающего проекта и полностью погрузиться в написание книги.

Спасибо тебе, Аманда, начинавшая как технический консультант, а в итоге ставшая тем, кто помог этой книге увидеть свет. Ты лучше всех! И сейчас, когда книга завершена, нам нужен новый проект. Но только при условии, что мы точно также будем вносить правки за чашечкой чая или кофе и заказанной из ресторана едой.

Особая благодарность моей подруге и помощнице Дейзи Томпсон за ее непрекращающуюся поддержку на протяжении всего пути. Без твоей помощи мне понадобилось бы намного больше времени, чтобы на это решиться, и я бы просто возненавидела это все. Ты привнесла в книгу свет и легкость, собственно, что ты делаешь со всеми своими проектами.

Особая благодарность Аманды Казари

Так как это всего лишь книга, а не достижение всей жизни, я постараюсь поблагодарить за работу над ней как можно короче.

Большое спасибо Элис, которая сначала пригласила меня как технического консультанта, а потом сделала соавтором. Я продолжаю учиться у тебя многому, в том числе придумывать смешные шутки о математике и простым языком объяснять сложные идеи.

И второму в списке, но не по значимости, огромное спасибо моему мужу Мэтью, который взял на себя почти невыполнимую роль моего вдохновителя и заставлял идти к цели и никогда не позволял от нее отступить. Ты лучший партнер и мой любимый соучастник преступлений. В горе и радости ты вдохновляешь меня, и я хочу, чтобы ты мной гордился.

ГЛАВА 1

Процесс машинного обучения

Прежде чем погрузиться в конструирование признаков, давайте рассмотрим общий процесс машинного обучения. Это поможет увидеть более общую картину того, как можно его применить. Начнем с небольшого размышления над базовыми понятиями — *данные* и *модели*.

Данные

То, что мы называем *данными*, — это наблюдение за феноменами реального мира. Например, данные фондовой биржи включают ежедневное отслеживание стоимости ценных бумаг, обнародование информации о заработке отдельных компаний и даже статьи аналитиков, где они делятся своим мнением. Персональные биометрические данные — это ежеминутное измерение сердечного ритма, уровень сахара в крови, давление и т. д. Данными о наблюдении за клиентами могут быть такие замечания: «Элис купила две книги в воскресенье», «Боб посетил эти три страницы на веб-сайте» и «Чарли кликнул на ссылку со специальным предложением на прошлой неделе». Можно привести множество примеров данных из различных областей.

Каждый фрагмент данных предоставляет узкое окно в ограниченную область реальности. А набор всех этих наблюдений дает нам общую картину. Но эта картина беспорядочна, так как состоит из тысячи мелких частиц, в каждой из которых существует погрешность и неполнота данных.

Задачи

Зачем мы собираем данные? Есть вопросы, на которые они могут дать нам ответы. Вопросы вроде «В какие ценные бумаги стоит инвестировать?», или «Как вести более здоровый образ жизни?», или «Как понять клиентов, чьи вкусы постоянно меняются, чтобы я мог предложить им лучший сервис?»

Путь от данных к ответам полон неверных стартов и тупиков (рис. 1.1). То, что выглядит как многообещающий подход, не всегда таким является. И то,

что изначально кажется просто догадкой, в итоге может привести к наилучшему решению. Работа с потоками данных обычно многоэтапный и цикличный процесс. Например, за стоимостью ценных бумаг наблюдают на бирже, затем эту информацию собирает воедино брокерское агентство вроде Thomson Reuters, заносит в базу данных, их покупает другая компания и преобразует в хранилище Hive на кластере серверов Hadoop. Из хранилища их извлекают одним скриптом, разбивают на подгруппы, обрабатывают и очищают другим скриптом, помещают в файл и снова преобразуют в формат, с которым может работать ваша любимая библиотека R, Python или Scala. Прогнозы цен снова помещаются в CSV-файл и разбираются алгоритмом оценки. Модель при этом изменяется несколько раз, переписывается на C++ или Java вашими разработчиками, и через нее вновь пропускают все данные, прежде чем предсказания попадут в новую базу данных.

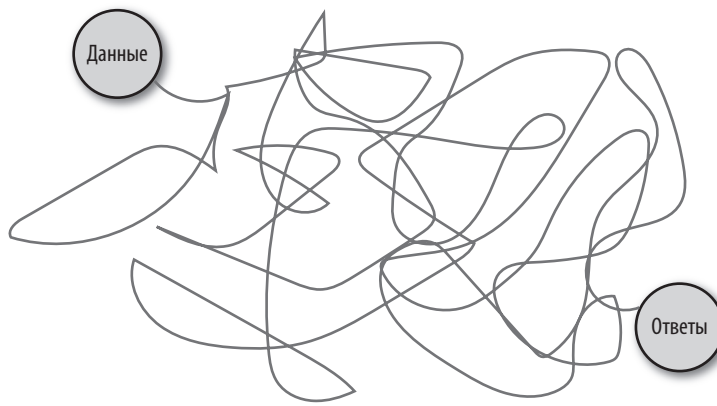

Рис. 1.1. *Сад разветвляющихся дорог на пути от данных к ответам*

Тем не менее, если мы отбросим множество инструментов и задействованных систем, то увидим, что в процессе участвуют две математические сущности — хлеб и масло машинного обучения — *модели* и *признаки*.

Модели

Пытаться понять мир через данные все равно что пытаться собрать реальность из пазлов, где не хватает нужных деталей, зато есть лишние. И тут вмешивается математическое моделирование, в частности статистическое моделирование. Набор статистических данных, содержит концепцию таких характеристик данных как *неверные*, *избыточные* или *недостающие*. Неверные данные появляются

в результате ошибки в измерениях. Избыточные данные содержат множество элементов, которые передают одну и ту же информацию. Например, день недели можно представить как переменную категории, которая принимает значения «Понедельник», «Вторник»… «Воскресенье» или выражается целочисленным значением от 0 до 6. Если информация о дне недели не представлена в некоторых фрагментах данных, то вы столкнулись с недостающими данными.

Математическая модель данных описывает отношения различных аспектов данных. Например, модель, предсказывающая стоимость акций, может быть представлена формулой, которая связывает историю прибыли компании, предшествующие цены и особенности отрасли, для которой делается прогноз. Модель, рекомендующая музыку, может измерять степень сходства пользователей на основе их музыкальных привычек и рекомендовать одинаковых артистов тем, кто слушает одни и те же песни.

Математические формулы связывают численные величины друг с другом. Но сырые данные редко бывают числовыми. (Действие «Элис купила трилогию „*Властелин колец*“ в среду» не представлено числами, точно так же, как отзыв, который она написала после прочтения книги.) Должно быть что-то, что объединит эти данные. Вот тут и нужны признаки.

Признаки

Признак — это числовое представление сырых данных. Существует много способов превратить сырые данные в числовые измерения, поэтому в итоге признаки могут выглядеть совершенно по-разному. Очевидно, признаки должны получаться из доступных данных. Возможно, менее очевидный факт, что они также привязаны к модели. Некоторые модели больше подходят для определенных типов признаков, и наоборот. Хороший признак соответствует решаемой задаче и должен легко встраиваться в модель. *Конструирование признаков* — процесс формулирования наиболее подходящих признаков на основе данных информации, модели и задачи.

Количество признаков тоже имеет значение. Если информативных признаков недостаточно, то модель не сможет выполнить конечную задачу. Если же их слишком много или большинство из них окажутся не имеющими отношения к делу, то модель будет более дорогой и сложной в обучении. А обучение не всегда проходит без сучка, без задоринки, что в итоге повлияет на конечную производительность модели.

Оценка модели

Признаки и модели находятся между сырыми данными и желаемыми результатами (рис. 1.2). В процессе машинного обучения мы выбираем не только модель, но и признаки. Это феноменально гибкий механизм, где выбор одного влияет на выбор другого. Хорошие признаки облегчают последующее моделирование, а итоговая модель благодаря им выполняет задачу наиболее точно. Плохо извлеченные признаки требуют намного более сложной модели, чтобы достичь того же уровня производительности. Далее мы расскажем о различных типах признаков и обсудим их преимущества и недостатки для разных типов данных и моделей. Впрочем, хватит слов, приступаем!

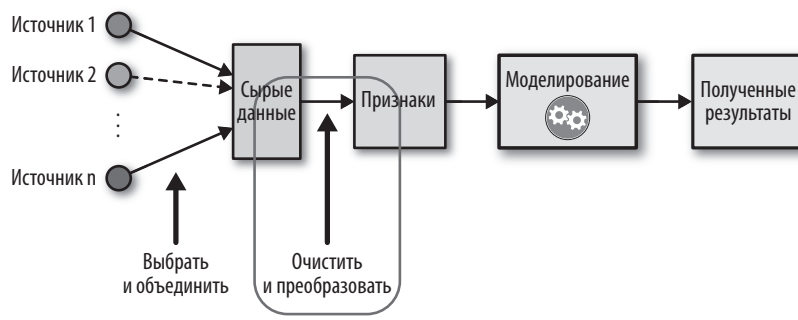

Рис. 1.2. *Место конструирования признаков в потоке машинного обучения*