

СОДЕРЖАНИЕ

Предисловие.....	7
Благодарности	11
Глава 1. ЧТО ТАКОЕ НАУКА О ДАННЫХ?	13
Глава 2. ЧТО ТАКОЕ ДАННЫЕ И ЧТО ТАКОЕ НАБОР ДАННЫХ?	47
Глава 3. ЭКОСИСТЕМА НАУКИ О ДАННЫХ.....	71
Глава 4. ОСНОВЫ МАШИННОГО ОБУЧЕНИЯ	93
Глава 5. СТАНДАРТНЫЕ ЗАДАЧИ НАУКИ О ДАННЫХ	139
Глава 6. КОНФИДЕНЦИАЛЬНОСТЬ И ЭТИКА	163
Глава 7. БУДУЩИЕ ТЕНДЕНЦИИ И ПРИНЦИПЫ УСПЕШНОСТИ	199
Глоссарий.....	215

ПРЕДИСЛОВИЕ

Цель науки о данных — улучшить процесс принятия решений, основывая их на более глубоком понимании ситуации с помощью анализа больших наборов данных. Как область деятельности наука о данных включает в себя ряд принципов, методов постановки задач, алгоритмов и процессов для выявления скрытых полезных закономерностей в больших наборах данных. Она тесно связана с глубинным анализом данных и машинным обучением, но имеет более широкий охват. Сегодня наука о данных управляет принятием решений практически во всех сферах современного общества. В повседневной жизни вы ощущаете на себе воздействие науки о данных, когда видите отобранные специально для вас рекламные объявления, рекомендованные фильмы и книги, ссылки на предполагаемых друзей, отфильтрованные письма в папке со спамом, персональные предложения от мобильных операторов и страховых компаний. Она влияет на порядок переключения и длительность сигналов светофоров в вашем районе, на то, как были созданы новые лекарства, продающиеся в аптеке, и то, как полиция вычисляет, где может потребоваться ее присутствие.

Рост использования науки о данных в обществе обусловлен появлением больших данных и социальных сетей, увеличением вычислительной мощности, уменьшением размеров носителей компьютерной памяти и разработкой более эффективных методов анализа и моделирования данных, таких как глубокое обучение. Вместе эти факторы означают, что сейчас процесс сбора, хранения и обработки данных стал как никогда ранее доступен для организаций. В то же время эти технические новшества и растущее применение науки

о данных означают, что этические проблемы, связанные с использованием данных и личной конфиденциальностью, тоже вышли на первый план. Цель этой книги — познакомить с наукой о данных на уровне ее основных элементов и с той степенью погружения, которая обеспечит принципиальное понимание вопроса.

Глава 1 очерчивает область науки о данных и дает краткую историю ее становления и эволюции. В ней мы также рассмотрим, почему наука о данных стала такой востребованной сегодня, и перечислим факторы, стимулирующие ее внедрение. В конце главы мы развенчаем несколько мифов, связанных с темой книги. Глава 2 вводит фундаментальные понятия, относящиеся к данным. В ней также описаны стандартные этапы проекта: понимание бизнес-целей, начальное изучение данных, подготовка данных, моделирование, оценка и внедрение. Глава 3 посвящена инфраструктуре данных и проблемам, связанным с большими данными и их интеграцией из нескольких источников. Одна из таких типичных проблем заключается в том, что данные в базах и хранилищах находятся на одних серверах, а анализируются на других. Поэтому колоссальное время тратится на перемещение больших наборов данных между этими серверами. Глава 3 начинается с описания типичной инфраструктуры науки о данных для организации и некоторых свежих решений проблемы перемещения больших наборов данных, а именно: метода машинного обучения в базе данных, использования Hadoop для хранения и обработки данных, а также разработки гибридных систем, в которых органично сочетаются традиционное программное обеспечение баз данных и решения, подобные Hadoop. Глава завершается описанием проблем, связанных с интеграцией данных в единое представление для последующего машинного обучения. Глава 4 знакомит читателя с машинным обучением и объясняет некоторые

из наиболее популярных алгоритмов и моделей, включая нейронные сети, глубокое обучение и деревья решений. В главе 5 основное внимание уделяется использованию опыта в области машинного обучения для решения реальных задач, приводятся примеры анализа стандартных бизнес-проблем и того, как они могут быть решены с помощью машинного обучения. В главе 6 рассматриваются этические вопросы науки о данных, последние разработки в области регулирования и некоторые из новых вычислительных методов защиты конфиденциальности в процессе обработки данных. Наконец, в главе 7 описаны сферы, на которые наука о данных окажет наибольшее влияние в ближайшем будущем, изложены принципы, позволяющие определить, будет ли данный конкретный проект успешным.

БЛАГОДАРНОСТИ

Джон хотел бы поблагодарить свою семью и друзей за их содействие и поддержку в процессе подготовки этой книги и посвящает ее своему отцу Джону Бернарду Келлехеру в знак признания его любви и дружбы.

Брендан хотел бы поблагодарить Грейс, Дэниела и Элеонору за их постоянную поддержку при написании всех его книг (эта уже четвертая), что позволило совмещать работу и путешествия.

ЧТО ТАКОЕ НАУКА О ДАННЫХ?

Наука о данных включает в себя набор принципов, методов постановки задач, алгоритмов и процессов для выявления скрытых полезных закономерностей в больших данных. Многие элементы этой науки были разработаны в смежных областях, таких как машинное обучение и глубинный анализ данных. Фактически термины «наука о данных», «машинное обучение» и «глубинный анализ данных» часто используются взаимозаменяемо. Эти дисциплины объединяет то, что все они направлены на улучшение процесса принятия решений посредством анализа данных. Однако, хотя наука о данных заимствует методы перечисленных областей, она имеет более широкий охват. Машинное обучение фокусируется на разработке и оценке алгоритмов выявления закономерностей в данных. Глубинный анализ данных, как правило, предполагает анализ структурированных данных и часто подразумевает акцент на коммерческих приложениях. Наука о данных учитывает и то и другое, при этом охватывает и другие проблемы: очистку и преобразование неструктурированных веб-данных и информации из социальных сетей, хранение и обработку больших неструктурированных наборов данных и вопросы, связанные с этикой и регулированием.

Используя науку о данных, мы можем выявлять различные типы закономерностей. Например, нам понадобилось выявить

закономерности, которые помогут идентифицировать группы клиентов, демонстрирующих сходное поведение и вкусы. На языке бизнеса эта задача известна как *сегментация клиентов*, а в терминологии науки о данных выявление такого типа закономерностей называется *кластеризацией*. Или, допустим, нам потребовалось выявить закономерность, которая обнаруживает продукты, которые часто покупают вместе. Опять же, в терминах науки о данных выявление такого типа закономерностей называется *поиском ассоциативных правил*. Или же нам нужны закономерности, которые выявляют странные или подозрительные события, например мошенничество со страховкой. Идентификация таких типов закономерностей известна как *обнаружение аномалий или выбросов*. Наконец, мы можем выявлять закономерности, которые помогают классифицировать что угодно. Например, закономерность классификации, выявленная в наборе данных электронной почты, могла бы выглядеть следующим образом: *если письмо содержит фразу «легкий заработок» — это, скорее всего, спам*. Поиск подобных правил классификации называется *прогнозированием*. Выбор слова «прогнозирование» может показаться странным, потому что правило не предсказывает, что произойдет в будущем: электронное письмо уже либо является, либо не является спамом. Поэтому правильнее говорить о закономерностях прогнозирования как о прогнозировании недостающего значения атрибута, а не о предсказании будущего. В этом примере мы прогнозируем, должен ли атрибут классификации электронной почты иметь значение «Спам» или нет.

Хотя науку о данных можно использовать для выявления различных типов закономерностей, мы всегда хотим, чтобы они были нетривиальными и полезными. Приведенный выше пример с электронной почтой настолько прост и очевиден, что, если бы это было единственное правило, извлеченное

Если человек может
с такой же легкостью
создать шаблон,
то, как правило,
не стоит тратить
время и усилия
на использование
науки о данных
для «обнаружения»
закономерности.

в процессе обработки данных, нас ждало бы разочарование. Этим правилом проверяется только один атрибут электронного письма: содержит ли оно фразу «легкий заработок». Если человек может с такой же легкостью создать шаблон, то, как правило, не стоит тратить время и усилия на использование науки о данных для «обнаружения» закономерности. Как правило, наука о данных становится полезной, когда у нас есть большое количество примеров и когда выявляемые закономерности слишком сложны, чтобы человек мог обнаружить их самостоятельно. В качестве нижней границы мы можем взять такое число примеров, обработка которых становится слишком трудоемкой для человека. Что касается сложности закономерностей, мы тоже можем определить ее относительно человеческих возможностей. Люди неплохо справляются с распознаванием правил, которые связывают один, два или даже три атрибута, но, когда их становится больше трех, мы начинаем перегорать. Наука о данных, напротив, применяется как раз тогда, когда мы хотим найти закономерности среди 10, 100, 1000 или даже миллиона атрибутов.

Закономерности, которые мы выявляем с помощью науки о данных, полезны только в том случае, если они ведут к прозрению, позволяющему что-то сделать для решения проблемы. То, ради чего мы выявляем закономерность, иногда называют «действенные прозрения». Слово «прозрение» подчеркивает, что закономерность должна дать нам важную информацию о проблеме, которая до этого была скрыта. Слово «действенный» говорит о том, что это прозрение должно быть применимо. Например, мы работаем в компании мобильной связи, которая пытается решить проблему *оттока* клиентов (когда слишком много клиентов переключаются на другие компании). Один из способов, каким наука о данных может помочь в решении этой проблемы, — использование данных

бывших клиентов для выявления закономерностей, которые позволят нам выявить среди текущих клиентов группу, наиболее подверженную риску оттока, после чего с этими клиентами можно связаться и постараться заинтересовать их. Закономерности, которые позволят нам идентифицировать вероятную группу оттока, будут полезны только в том случае, если: а) они выявляют клиентов достаточно рано для того, чтобы можно было связаться с ними и предотвратить потенциальное действие с их стороны, и б) компания способна выделить команду для работы с этой группой клиентов. Соблюдение этих параметров необходимо для того, чтобы компания могла действовать в соответствии с полученным прозрением.

Краткая история науки о данных

История термина «наука о данных» начинается в 1990-е гг. Однако области, которые он охватывает, имеют более долгую историю. Одна из них — сбор данных, другая — их анализ. Далее мы рассмотрим, как развивались эти отрасли знаний, а затем опишем, как и почему они сплелись воедино в науке о данных. В этом обзоре будет введено много новых понятий, поскольку он описывает и называет важные технические новшества по мере их возникновения. Для каждого нового термина мы дадим краткое объяснение его значения, однако позже мы еще вернемся ко многим из них и приведем более подробные объяснения. Мы начнем с истории сбора данных, продолжим историей анализа данных и закончим эволюцией науки о данных.

История сбора данных

Первыми из известных нам методов записи данных были зарубки на столбах, вкопанных в землю, чтобы отмечать

восходы солнца и узнавать количество дней до солнцестояния. Однако с развитием письменности наша способность фиксировать опыт и события окружающего мира значительно увеличила объем собираемых нами данных. Самая ранняя форма письма была разработана в Месопотамии около 3200 г. до н.э. и использовалась для коммерческого учета. Этот тип учета фиксирует так называемые *транзакционные данные*. Транзакционные данные включают в себя информацию о событиях, таких как продажа товара, выставление счета, доставка, оплата кредитной картой, страховые требования и т. д. *Нетранзакционные данные*, например демографические, также имеют долгую историю. Первые известные переписи населения прошли в Древнем Египте около 3000 г. до н.э. Причина, по которой древние правители вкладывали так много усилий и ресурсов в масштабные проекты по сбору данных, заключалась в том, что им нужно было повышать налоги и увеличивать армии. Это согласуется с утверждением Бенджамина Франклина о том, что в жизни есть только две несомненные вещи: смерть и налоги.

В последние 150 лет изобретение компьютера, появление электронных датчиков и оцифровка данных способствовали стремительному росту объемов сбора и хранения данных. Ключевое событие в этой сфере произошло в 1970 г., когда Эдгар Кодд опубликовал статью с описанием *реляционной модели данных*, которая совершила переворот в том, как именно данные хранятся, индексируются и извлекаются из баз. Реляционная модель позволила извлекать данные из базы путем простых запросов, которые определяли, что нужно пользователю, не требуя от него знания о внутренней структуре данных или о том, где они физически хранятся. Документ Кодда послужил основой для современных баз данных и разработки SQL (языка структурированных запросов), международного стандарта формулировки запросов

к базам данных. Реляционные базы хранят данные в таблицах со структурой из одной строки на объект и одного столбца на атрибут. Такое отображение идеально подходит для хранения данных с четкой структурой, которую можно разложить на базовые атрибуты.

Базы данных — это простая технология, используемая для хранения и извлечения структурированных транзакционных или *операционных* данных (т. е. генерируемых текущими операциями компании). Но по мере того, как компании росли и автоматизировались, объем и разнообразие данных тоже резко возрастали. В 1990-х гг. стало ясно, что, хотя компании накопили огромные объемы данных, они испытывают трудности с их анализом. Частично проблема была в том, что данные обычно хранились в многочисленных разрозненных базах в рамках одной организации. Другая трудность заключалась в том, что базы были оптимизированы для хранения и извлечения данных — действий, которые характеризуются большими объемами простых операций, таких как SELECT, INSERT, UPDATE и DELETE. Для анализа данных компаниям требовалась технология, которая могла бы объединять и согласовывать данные из разнородных баз и облегчать проведение более сложных аналитических операций. Решение этой бизнес-задачи привело к появлению *хранилищ данных*. Организация хранилищ данных — это процесс агрегирования и анализа данных для поддержки принятия решений. Основная задача этого процесса — создание хорошо спроектированного централизованного банка данных, который тоже иногда называется хранилищем. В этом смысле хранилище данных является мощным ресурсом науки о данных, с точки зрения которой основное преимущество хранилища данных — это сокращение времени выполнения проекта. Ключевым компонентом любого процесса обработки данных являются сами данные, поэтому неудивительно, что во многих проектах

большая часть времени и усилий направляется на поиск, сбор и очистку данных перед анализом. Если в компании есть хранилище данных, то усилия и время, затрачиваемые на подготовку данных, значительно сокращаются. Тем не менее наука о данных может существовать и без централизованного банка данных. Создание такого банка не ограничивается выгрузкой данных из нескольких операционных баз в одну. Объединение данных из нескольких баз часто требует сложной ручной работы для устранения несоответствий между исходными базами данных. *Извлечение, преобразование и загрузка (ETL)* — это термин, используемый для описания стандартных процессов и инструментов для сопоставления, объединения и перемещения данных между базами. Типичные операции, выполняемые в хранилище данных, отличаются от операций в стандартной реляционной базе данных. Для их описания используется термин *интерактивная аналитическая обработка (OLAP)*. Операции OLAP, как правило, направлены на создание сводок исторических данных и включают сбор данных из нескольких источников. Например, запрос OLAP, выраженный для удобства на естественном языке, может выглядеть так: «*Отчет о продажах всех магазинов по регионам и кварталам и разница показателей по сравнению с отчетом за прошлый год*». Этот пример показывает, что результат запроса OLAP часто напоминает стандартный бизнес-отчет. По сути, операции OLAP позволяют пользователям распределять, фрагментировать и переворачивать данные в хранилище, а также получать их различные отображения. Операции OLAP работают с отображением данных, называемым *кубом данных*, который построен поверх хранилища. Куб данных имеет фиксированный, заранее определенный набор измерений, где каждое измерение отображает одну характеристику данных. Для приведенного выше примера запроса OLAP необходимы следующие измерения куба данных: *продажи*

по магазинам, продажи по регионам и продажи по кварталам. Основное преимущество использования куба данных с фиксированным набором измерений состоит в том, что он ускоряет время отклика операций OLAP. Кроме того, поскольку набор измерений куба данных предварительно запрограммирован в систему OLAP, эти системы могут быть отображены дружественным пользовательским интерфейсом (GUI) для формулирования запросов OLAP. Однако отображение куба данных ограничивает типы анализа набором запросов, которые могут быть сгенерированы только с использованием определенных заранее измерений. Интерфейс запросов SQL сравнительно более гибок. Кроме того, хотя системы OLAP полезны для исследования данных и составления отчетов, они не позволяют моделировать данные или автоматически выявлять в них закономерности.

За последние пару десятилетий наши устройства стали мобильными и подключенными к сети. Многие из нас ежедневно часами сидят в интернете, используя социальные технологии, компьютерные игры, медиаплатформы и поисковые системы. Эти технологические изменения в нашем образе жизни оказали существенное влияние на количество собираемых данных. Подсчитано, что объем данных, собранных за пять тысячелетий с момента изобретения письма до 2003 г., составляет около пяти эксабайт. С 2013 г. люди генерируют и хранят такое же количество данных ежедневно. Однако резко вырос не только объем данных, но и их разнообразие. Достаточно взглянуть на список сегодняшних онлайн-источников данных: электронные письма, блоги, фотографии, твиты, лайки, публикации, веб-поиск, загрузка видео, онлайн-покупки, подкасты и т. д. Также не забудьте о метаданных этих событий, описывающих структуру и свойства необработанных данных, и вы начнете понимать, что называется *большими данными*. Большие данные часто описываются по

схеме «3V»: экстремальный объем (**V**olume), разнообразие типов (**V**ariety) и скорость обработки данных (**V**elocity).

Появление больших данных привело к разработке новых технологий создания баз данных. Базы данных нового поколения часто называют базами *NoSQL*. Они имеют более простую модель, чем привычные реляционные базы данных, и хранят данные в виде объектов с атрибутами, используя язык представления объектов, такой как *JavaScript Object Notation (JSON)*. Преимущество использования объектного представления данных (по сравнению с моделью на основе реляционной таблицы) состоит в том, что набор атрибутов для каждого объекта заключен в самом объекте, а это открывает дорогу к гибкому отображению данных. Например, один из объектов в базе данных может иметь сокращенный набор атрибутов по сравнению с другими объектами. В структуре реляционной базы данных, напротив, все значения в таблице должны иметь одинаковый набор атрибутов (столбцов). Эта гибкость важна в тех случаях, когда данные (из-за их разнообразия или типа) не раскладываются естественным образом в набор структурированных атрибутов. К примеру, сложно определить набор атрибутов для отображения неформального текста (скажем, твитов) или изображений. Однако, хотя эта гибкость представления позволяет нам собирать и хранить данные в различных форматах, для последующего анализа их все равно приходится структурировать.

Большие данные также привели к появлению новых платформ для их обработки. При работе с большими объемами информации на высоких скоростях может быть полезным с точки зрения вычислений и поддержания скорости распределять данные по нескольким серверам, затем обрабатывать запросы, вычисляя их результаты по частям на каждом из серверов, а затем объединять их в сгенерированный ответ. Такой подход использован в модели *MapReduce* на платформе

Надоор. В этой модели данные и запросы отображаются на нескольких серверах (распределяются между ними), а затем рассчитанные на них частичные результаты объединяются.

История анализа данных

Статистика — это научная отрасль, которая занимается сбором и анализом данных. Первоначально статистика собирала и анализировала информацию о государстве, такую как демографические данные и экономические показатели. Со временем количество типов данных, к которым применялся статистический анализ, увеличивалось, и сегодня статистика используется для анализа любых типов данных. Простейшая форма статистического анализа — обобщение набора данных в терминах *сводной (описательной) статистики* (включая средние значения, такие как *среднее арифметическое*, или показатели колебаний, такие как *диапазон*). Однако в XVII–XVIII вв. работы Джероламо Кардано, Блеза Паскаля, Якоба Бернулли, Абрахама де Муавра, Томаса Байеса и Ричарда Прайса заложили основы теории вероятностей, и в течение XIX в. многие статистики начали использовать распределение вероятностей как часть аналитического инструментария. Эти новые достижения в математике позволили выйти за рамки описательной статистики и перейти к *статистическому обучению*. Пьер-Симон де Лаплас и Карл Фридрих Гаусс — два наиболее видных математика XIX в. Оба они внесли заметный вклад в статистическое обучение и современную науку о данных. Лаплас использовал интуитивные прозрения Томаса Байеса и Ричарда Прайса и превратил их в первую версию того, что мы сейчас называем *теоремой Байеса*. Гаусс в процессе поиска пропавшей карликовой планеты Цереры разработал *метод наименьших квадратов*. Этот метод позволяет нам найти наилучшую модель, которая соответствует набору данных, так что ошибка в ее подборе сводится к минималь-

ной сумме квадратов разностей между опорными точками в наборе данных и в модели. Метод наименьших квадратов послужил основой для статистических методов обучения, таких как *линейная регрессия* и *логистическая регрессия*, а также для разработки моделей *нейронных сетей* искусственного интеллекта.

Между 1780 и 1820 гг., примерно в то же время, когда Лаплас и Гаусс вносили свой вклад в статистическое обучение, шотландский инженер Уильям Плейфер изобрел статистические графики и заложил основы современной *визуализации данных* и *поискового анализа данных (EDA)*. Плейфер изобрел *линейный график* и *комбинированную диаграмму* для временных рядов данных, *гистограмму*, чтобы проиллюстрировать сравнение значений, принадлежащих разным категориям, и *круговую диаграмму* для наглядного изображения долей. Преимущество визуализации числовых данных заключается в том, что она позволяет использовать наши мощные зрительные возможности для обобщения, сравнения и интерпретации данных. Следует признать, что визуализировать большие (с множеством опорных точек) или сложные (с множеством атрибутов) наборы данных довольно трудно, но визуализация по-прежнему остается важной составляющей науки о данных. В частности, она помогает ученым рассматривать и понимать данные, с которыми они работают. Визуализация также может быть полезна для презентации результатов проекта. Со времен Плейфера разнообразие видов графического отображения данных неуклонно росло, и сегодня продолжают развиваться новые подходы в области визуализации больших многомерных наборов данных. В частности, не так давно был разработан *алгоритм стохастического вложения соседей с t -распределением (t -SNE)*, который применяется при сокращении многомерных данных до двух или трех измерений, тем самым облегчая их визуализацию.

Развитие теории вероятностей и статистики продолжилось в XX в. Карл Пирсон разработал современные методы проверки гипотез, а Рональд Фишер — статистические методы для *многомерного анализа* и предложил идею *оценки максимального правдоподобия* статистических заключений как метод, позволяющий делать выводы на основе относительной вероятности событий. Работа Алана Тьюринга во время Второй мировой войны привела к изобретению компьютера, который оказал исключительно сильное влияние на статистику, позволив совершать существенно более сложные вычисления. В течение 1940-х гг. и в последующие десятилетия были разработаны важные вычислительные модели, которые до сих пор широко применяются в науке о данных. В 1943 г. Уоррен Мак-Каллок и Уолтер Питтс предложили первую математическую модель *нейронной сети*. В 1948-м Клод Шеннон опубликовал статью под названием «Математическая теория связи» и тем самым основал *теорию информации*. В 1951 г. Эвелин Фикс и Джозеф Ходжес предложили модель *дискриминантного анализа* (который сейчас более известен как *теория распознавания образов*), ставшую основой современных алгоритмов *ближайших соседей*. Послевоенное развитие сферы достигло кульминации в 1956 г. с появлением отрасли *искусственного интеллекта* на семинаре в Дартмутском колледже. Даже на этой ранней стадии ее развития термин «*машинное обучение*» уже начал использоваться для описания программ, которые давали компьютеру возможность учиться на основе данных. В середине 1960-х гг. были сделаны три важных вклада в машинное обучение. В 1965 г. Нильс Нильсон опубликовал книгу «Обучающиеся машины»*, в которой показано, как можно использовать нейронные сети для обучения линейных моделей классификации. Через год Хант, Марин и Стоун

* Нильсон, Н. Дж. Обучающиеся машины. — М.: Мир, 1967.

Конец ознакомительного фрагмента.

Приобрести книгу можно

в интернет-магазине

«Электронный универс»

e-Univers.ru