ДЖОН КЕЛЛЕХЕР БРЕНДАН ТИРНИ

Наука о данных

альпина ПАБЛИШЕР | бизнес

Джон Келлехер
Брендан Тирни

НАУКА О ДАННЫХ

Базовый курс

Перевод с английского

альпина
паблишер

Москва
2020

Все права защищены. Данная электронная книга предназначена исключительно для частного использования в личных (некоммерческих) целях. Электронная книга, ее части, фрагменты и элементы, включая текст, изображения и иное, не подлежат копированию и любому другому использованию без разрешения правообладателя. В частности, запрещено такое использование, в результате которого электронная книга, ее часть, фрагмент или элемент станут доступными ограниченному или неопределенному кругу лиц, в том числе посредством сети интернет, независимо от того, будет предоставляться доступ за плату или безвозмездно.

Копирование, воспроизведение и иное использование электронной книги, ее частей, фрагментов и элементов, выходящее за пределы частного использования в личных (некоммерческих) целях, без согласия правообладателя является незаконным и влечет уголовную, административную и гражданскую ответственность.

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.

Глава 1

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 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.

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.

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 бóльшая 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 Hadoop. 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 минимальной 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 продолжаются разработки новых подходов 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»[1], в которой показано, как можно использовать нейронные сети для обучения линейных моделей классификации. Через год Хант, Марин и Стоун разработали систему концептуального обучения, породившую целое семейство алгоритмов, которые, в свою очередь, привели к появлению деревьев решений на основе данных нисходящего порядка. Примерно в то же время независимые исследователи разрабатывали и публиковали ранние версии *метода k-средних*, который теперь рутинно используется для сегментации клиентских данных.

Область машинного обучения лежит в основе современной науки о данных, поскольку она предоставляет алгоритмы, способные автоматически анализировать большие наборы данных для выявления потенциально интересных и полезных закономерностей. Машинное обучение и сегодня продолжает развиваться и модернизироваться. В число наиболее важных разработок входят *ансамблевые методы*, прогнозирование в которых осуществляется на основе набора моделей, где каждая модель участвует в каждом из запросов, а также дальнейшее развитие *нейронных сетей глубокого обучения*, имеющих более трех слоев нейронов. Такие глубокие слои в сети способны обнаруживать и анализировать отображения сложных атрибутов (состоящие из нескольких взаимодействующих входных значений, обработанных более ранними слоями), которые позволяют сети изучать закономерности и обобщать их для всех входных данных. Благодаря своей способности исследовать сложные атрибуты сети глубокого обучения лучше других подходят для многомерных данных — именно они произвели переворот в таких областях, как *машинное зрение* и *обработка естественного языка*.

Как уже упоминалось в историческом обзоре баз данных, начало 1970-х гг. ознаменовало приход современной технологии с *реляционной моделью данных* Эдгара Кодда и последующий взрывной рост генерации данных и их хранения, который в 1990-х гг. привел к развитию хранилищ, а позднее — к возникновению феномена больших данных. Однако еще задолго до появления больших данных, фактически к концу 1980-х — началу 1990-х гг., стала очевидной необходимость в исследованиях, направленных на анализ больших наборов данных. Примерно в то же время появился термин «*глубинный анализ данных*». Как мы уже отметили, в ответ на это началась разработка хранилищ данных и технологии OLAP. Кроме того, параллельно велись исследования в других областях. В 1989 г. Григорий Пятецкий-Шапиро провел первый семинар по *обнаружению знаний в базах данных (KDD)*. Следующая цитата из анонса этого семинара дает ясное представление о том, какое внимание на нем уделялось междисциплинарному подходу к проблеме анализа больших баз данных:

> Обнаружение знаний в базах данных ставит много интересных проблем, особенно когда эти базы огромны. Таким базам данных обычно сопутствуют существенные знания предметной области, которые могут значительно облегчить обнаружение данных. Доступ к большим базам данных недешев — отсюда необходимость выборки и других статистических методов. Наконец, для обнаружения знаний в базах данных могут оказаться полезными многие существующие инструменты и методы из различных областей, таких как экспертные системы, машинное обучение, интеллектуальные базы данных, получение знаний и статистика[2].

Фактически термины «KDD» и «глубинный анализ данных» описывают одну и ту же концепцию; различие заключается только в том, что термин «глубинный анализ данных» более распространен в бизнес-сообществах, а «KDD» — в академических кругах. Сегодня эти понятия часто взаимозаменяются[3], и многие ведущие академические центры используют как одно, так и другое. И это закономерно, ведь главная научная конференция в этой сфере так и называется — Международная конференция по обнаружению знаний и глубинному анализу данных.

Возникновение и эволюция науки о данных

Термин «наука о данных» появился в конце 1990-х гг. в дискуссиях, касающихся необходимости объединения статистиков с теоретиками вычислительных систем для обеспечения математической строгости при компьютерном анализе больших данных. В 1997 г. Джефф Ву выступил с публичной лекцией «Статистика = наука о данных?», в которой осветил ряд многообещающих тенденций, в том числе доступность больших и сложных наборов данных в огромных базах и рост использования вычислительных алгоритмов и моделей. В завершение лекции он призвал переименовать статистику в «науку о данных».

В 2001 г. Уильям Кливленд опубликовал план действий по созданию университетского факультета, сфокусированного на науке о данных [1]. В плане подчеркивалось место науки о данных между математикой и информатикой и предлагалось понимать ее как междисциплинарную сферу. Специалистам по данным предписывалось учиться, работать и взаимодействовать с экспертами из этих областей. В том же году Лео Брейман опубликовал статью «Статистическое моделирование: две культуры» [2]. В ней он охарактеризовал традиционный подход к статистике как культуру моделирования данных, которая

предполагает основной целью анализа выявление скрытых стохастических моделей (например, *линейной регрессии*), объясняющих, как были сгенерированы данные. Брейман противопоставляет это культуре алгоритмического моделирования, которая фокусируется на использовании компьютерных алгоритмов для создания более точных моделей прогнозирования, не объясняющих то, как данные были получены. Проведенная Брейманом граница между статистическими моделями, которые объясняют данные, и алгоритмическими, которые могут их точно прогнозировать, подчеркивает коренное различие между статистиками и исследователями машинного обучения. Споры между этими двумя подходами не утихают до сих пор [3]. В целом сегодня большинство проектов, осуществляемых в рамках науки о данных, соответствует подходу машинного обучения к построению точных моделей прогнозирования и все меньше озабочены статистическим объяснением. Таким образом, хотя наука о данных родилась в дискуссиях вокруг статистики и до сих пор заимствует некоторые статистические методы и модели, со временем она разработала свой собственный, особый подход к анализу данных.

С 2001 г. концепция науки о данных значительно расширилась и вышла за пределы модификаций статистики. Например, в последние 10 лет наблюдается колоссальный рост объема данных, генерируемых онлайн-активностью (интернет-магазинами, социальными сетями или развлечениями). Чтобы собрать эту информацию (порой неструктурированную) из внешних веб-источников, подготовить и очистить ее для использования в проектах по анализу данных, специалистам по данным требуются навыки программирования и взлома. Кроме того, появление больших данных означает, что специалист по данным должен уметь работать с такими технологиями, как Hadoop. Фактически сегодня понятие «специалист по данным» стало настолько

широким, что вызвало настоящие дебаты о том, как определить его роль и требуемые опыт и навыки [4]. Тем не менее можно перечислить их, опираясь на мнение большинства людей, как это сделано на рис. 1. Одному человеку трудно овладеть всем перечисленным, и большинство специалистов по данным действительно обладают глубокими знаниями и реальным опытом только в некоторых из этих областей. При этом важно понимать и осознавать вклад каждой из них в проекты по обработке данных.

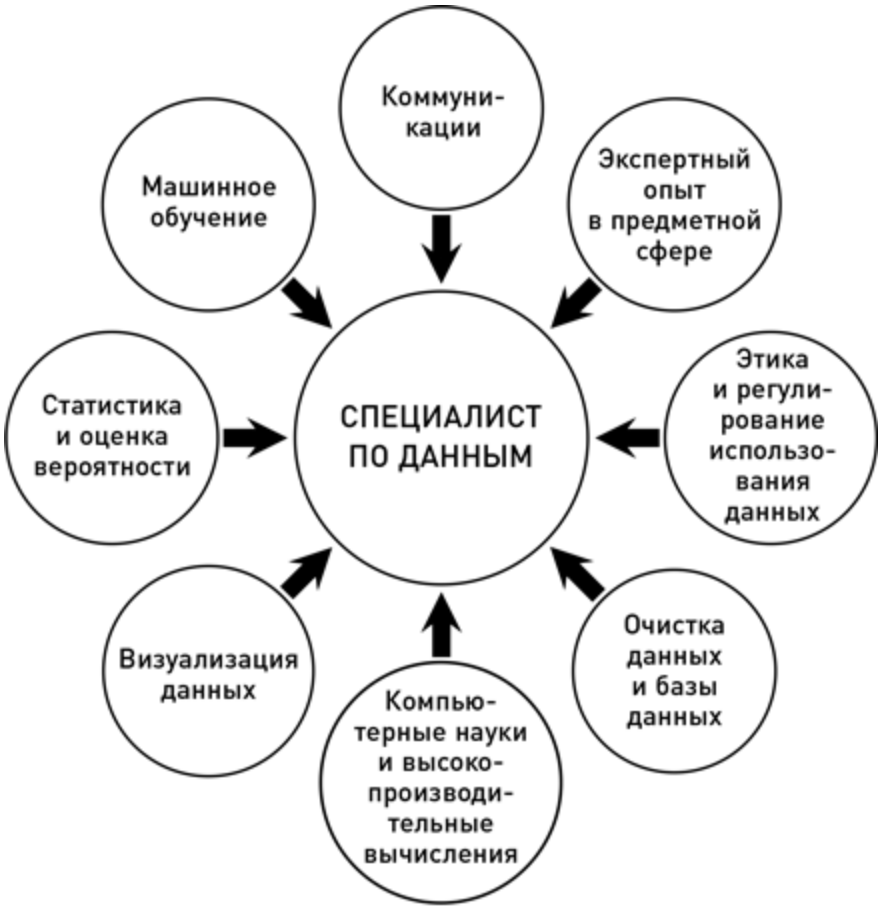

Рис. 1. Навыки специалиста по данным

Специалист по данным должен иметь экспертный опыт в предметной сфере. Большинство проектов начинаются с реальной проблемы и необходимости разработать ее решения. Специалист

по данным должен понимать и проблему, и то, как ее решение могло бы вписаться в организационные процессы. Этот экспертный опыт направляет специалиста при поиске оптимального решения. Он также позволяет конструктивно взаимодействовать с отраслевыми экспертами, чтобы докопаться до самой сути проблемы. Кроме того, специалист по данным может использовать его в работе над аналогичными проектами в той же или смежной областях и быстро определять их фокус и охват.

В центре всех проектов науки о данных находятся сами данные. Однако тот факт, что организация имеет доступ к данным, не означает, что у нее есть формальное или этическое право на их использование. В большинстве юрисдикций существует антидискриминационное законодательство и законы о защите персональных данных. Специалист по данным должен знать и понимать эти правила, а также (в более широком смысле) понимать этические последствия своей работы, если хочет использовать данные на законных основаниях и надлежащим образом. Мы вернемся к этой теме в главе 7, где обсудим правовые нормы и этические вопросы, связанные с наукой о данных.

В большинстве организаций значительная часть данных поступает из баз, размещенных внутри самой организации. Но по мере роста архитектуры данных проекты начнут получать их из множества других источников, в том числе из источников больших данных. Данные в этих источниках могут существовать в различных форматах, но, как правило, представляют собой базы на основе реляционной модели, NoSQL или Hadoop. Эти данные должны быть интегрированы, очищены, преобразованы, нормализованы и т.д. Такие задачи могут называться по-разному, например: ETL (извлечение, преобразование, загрузка), подготовка, слияние, уплотнение данных и др. Результаты обработки должны храниться и управляться, как и исходные

данные. Для этого также используют базы, чтобы результаты можно было легко распределить между частями организации или обеспечить им совместный доступ. Следовательно, специалист по данным должен обладать навыками взаимодействия с базами данных и обработки содержащейся в них информации.

Понятие «компьютерные науки» используется здесь для обозначения целого ряда навыков и инструментов, которые позволяют специалисту работать с большими данными и преобразовывать их в новую значимую информацию. Высокопроизводительные вычисления (HPC) предполагают агрегацию вычислительных мощностей для достижения большей производительности, чем может дать автономный компьютер. Многие проекты имеют дело с очень большими наборами данных и/или алгоритмами машинного обучения, которые требуют дорогостоящих вычислений. В таких ситуациях важно иметь навыки доступа к ресурсам HPC и их использования. Помимо HPC, мы уже упоминали о задачах сбора, очистки и интегрирования веб-данных, стоящих перед специалистом. Сюда же входит умение обрабатывать неструктурированный текст и изображения. Кроме того, неплохо, если специалист по данным способен сам написать приложение для выполнения конкретной задачи или изменить существующее, чтобы настроить его под конкретные данные и сферу деятельности. Наконец, необходима компьютерная грамотность, чтобы понимать и разрабатывать модели машинного обучения и интегрировать их в производственные, аналитические или внутренние приложения организации.

Графическое отображение данных существенно упрощает их просмотр и понимание. Визуализация применяется на всех этапах процесса. Работая с данными в табличной форме, легко пропустить такие вещи, как выбросы, тренды в распределениях или незначительные изменения данных во времени. Правильное графическое отображение выявляет эти и другие аспекты.

Визуализация является важной и растущей областью науки о данных, и мы рекомендуем работы Эдварда Туфта [5] и Стефана Фью [6] как отличное введение в ее принципы и методы.

В процессе обработки данных (от их первоначального сбора и исследования до сравнения результатов различных моделей и типов анализа) используются статистические и вероятностные методы. Машинное обучение применяет их для поиска закономерностей. Специалист по данным не обязан уметь писать алгоритмы машинного обучения, но должен понимать, как и для чего они используются, что означают сгенерированные ими результаты и на каком типе данных могут выполняться конкретные алгоритмы. Иначе говоря, воспринимать их как «серый ящик» — систему с частично известной внутренней структурой. Это позволит сконцентрироваться на прикладных аспектах и провести тестирование различных алгоритмов машинного обучения, чтобы понять, какие из них лучше всего подходят для конкретного сценария.

Наконец, важным аспектом успешности специалиста по данным является умение рассказать с их помощью историю. Это может быть история прозрения, которое дал анализ, или история о моделях, созданных в ходе проекта, которые идеально впишутся в процессы организации и благотворно повлияют на ее функционирование. В потрясающем проекте по обработке данных нет никакого смысла, если его результаты не будут использованы, но для этого надо сообщить о них коллегам, не имеющим технического образования, в такой форме, чтобы они смогли все понять.

Где используется наука о данных?

Наука о данных определяет принятие решений практически во всех сферах современного общества. В этом разделе мы опишем три тематических кейса, которые иллюстрируют ее влияние на потребительские компании, использующие науку о данных в продажах и маркетинге, на правительства, совершенствующие с ее помощью здравоохранение, правосудие и городское планирование, и на профессиональные спортивные клубы, проводящие на ее основе отбор игроков.

Наука о данных в продажах и маркетинге

Компания Walmart (и другие розничные сети) имеет доступ к большим наборам данных о предпочтениях своих покупателей, собирая их через системы торговых точек, отслеживая поведение клиентов в интернет-магазине и анализируя комментарии о компании и ее продуктах в социальных сетях. Уже более 10 лет Walmart использует науку о данных для оптимизации уровня запасов в магазинах. Хорошо известен пример, когда Walmart пополняла ассортимент пирожных с клубникой в магазинах на пути следования урагана «Фрэнсис» в 2004 г. на основе анализа данных о продажах в период прохождения урагана «Чарли» несколькими неделями ранее. Недавно Walmart использовала науку о данных для увеличения розничных доходов, начав внедрять новые продукты на основе анализа тенденций в социальных сетях, анализировать активность по кредитным картам для составления рекомендаций клиентам, а также оптимизировать и персонализировать взаимодействие с клиентами через официальный сайт. Walmart связывает увеличение объема онлайн-продаж на 10–15% именно с использованием науки о данных [7].

В онлайн-мире эквивалентом апселлинга (продажи более дорогих версий товара) и перекрестных продаж являются рекомендательные системы. Если вы смотрели фильмы на Netflix

или покупали что-нибудь на Amazon, то знаете, что эти сайты собирают и используют данные, а затем предлагают вам варианты следующих просмотров или покупок. Одни рекомендательные системы направляют вас к блокбастерам и бестселлерам, а другие — к нишевым продуктам, соответствующим вашим вкусам. В книге Криса Андерсона «Длинный хвост: Эффективная модель бизнеса в интернете» [8] утверждается, что по мере удешевления производства и дистрибуции рынки переходят от продажи большого количества небольшого набора хитов к продажам меньшего количества более разнообразных нишевых продуктов. Этот компромисс между стимулированием продаж популярных и нишевых продуктов лежит в основе разработки рекомендательных систем и влияет на алгоритмы обработки данных, используемые в этих системах.

Использование науки о данных государственными структурами

В последние годы государственные структуры осознали преимущества науки о данных. Например, правительство США в 2015 г. назначило математика Дханурджая Патила первым главным специалистом по данным. Некоторые из крупнейших инициатив в области науки о данных, возглавляемых правительством, были связаны со здоровьем. Наука о данных лежит в основе проектов «Раковый прорыв» (Cancer Moonshot) и «Точная медицина» (Precision Medicine)[4]. «Точная медицина» сочетает секвенирование генома человека и науку о данных при разработке индивидуальных лекарств для отдельных пациентов. Одной из его частей является программа «Все мы» (All of Us)[5], которая занимается сбором информации об окружающей среде, образе жизни и биологических параметрах более миллиона добровольцев для создания крупнейших в мире баз данных точной медицины. Наука о данных радикальным образом меняет

устройство городов, где она применяется для отслеживания, анализа и контроля экологических, энергетических и транспортных систем, а также при долгосрочном городском планировании [9]. Мы вернемся к здоровью и умным городам в главе 9, когда будем обсуждать перспективы науки о данных на ближайшие десятилетия.

Еще одна инициатива правительства США в области данных направлена на то, чтобы департаменты полиции лучше понимали, как они могут помочь местным сообществам[6]. Наука о данных также способствует прогнозированию очагов преступности и рецидивов преступлений, однако правозащитные группы подвергли критике ее использование в уголовном правосудии. В главе 7 мы обсудим вопросы конфиденциальности и этики, поднятые наукой о данных, и одним из факторов в этой дискуссии станет то, что многие люди имеют разное мнение о приватности информации, в зависимости от области, где она применяется. Если ее использование в медицинских исследованиях, финансируемых государством, находит поддержку, то реакция тех же людей меняется на противоположную, когда речь заходит о деятельности полиции и уголовном правосудии. В главе 7 мы также обсудим использование персональных данных для определения размера выплат при страховании жизни, здоровья, автомобиля, дома и путешествий.

Ключом к успеху является получение правильных данных и поиск правильных атрибутов.

Наука о данных в профессиональном спорте

Фильм 2011 г. «Человек, который изменил все» с участием Брэда Питта продемонстрировал растущую роль науки о данных в современном спорте. Фильм основан на книге «Moneyball»[7] 2004 г., в которой рассказана реальная история о том, как бейсбольный клуб «Окленд Атлетикс» использовал науку о данных для улучшения отбора игроков [10]. С ее помощью было выявлено, что процентное соотношение попадания игрока на базу и упущенных возможностей является более информативным показателем его успешности, чем традиционно принятые в бейсболе статистические данные, такие как средний уровень достижений. Это понимание позволило составить список недооцененных игроков и превзойти возможности бюджета. Успех «Окленд Атлетикс» произвел революцию в бейсболе, и сегодня большинство клубов интегрирует аналогичные стратегии, основанные на данных, в процесс найма.

Эта история — яркий пример того, как наука о данных может дать организации преимущество в конкурентном рыночном пространстве. Но с точки зрения самой науки наиболее важным аспектом здесь является то, что иногда на первый план выходит выявление информативных атрибутов. Распространено мнение, что ценность науки о данных заключается в моделях, которые создаются в процессе. Однако, как только мы узнаем важные атрибуты области определения, можно легко создавать модели, управляемые данными. Ключом к успеху является получение правильных данных и поиск правильных атрибутов. В своей книге «Фрикономика»[8] Стивен Левитт и Стивен Дабнер иллюстрируют важность этого на примере широкого круга проблем, поскольку считают, что ключом к пониманию современной жизни является «знание того, что и как измерять» [11]. Используя науку о данных, мы можем выявить важные

закономерности, которые, в свою очередь, помогут идентифицировать нужные атрибуты области определения. Причина, по которой наука о данных используется все шире, заключается в том, что сфера ее приложения не имеет значения: важны только правильные данные и четкая формулировка проблемы.

Почему сейчас?

Есть ряд факторов, способствующих росту науки о данных. Как мы уже говорили, появление больших данных обусловлено относительной легкостью, с которой организации могут собирать информацию. Записи транзакций в точках продаж, клики на онлайн-платформах, публикации в социальных сетях, приложения на смартфонах и прочее — все это каналы, через которые компании теперь могут создавать ценные профили отдельных клиентов. Другим фактором является коммодификация хранилищ данных с экономией на масштабе, что делает хранение информации дешевле, чем когда-либо прежде. На это влияет и колоссальный рост мощности компьютеров. Графические карты и процессоры (GPU) были изначально разработаны для быстрой визуализации графики в компьютерных играх. Отличительная особенность графических процессоров — способность выполнять быстрое умножение матриц, а это полезно не только для рендеринга графики, но и для машинного обучения. В последние годы графические процессоры были адаптированы и оптимизированы для использования в машинном обучении, что способствовало заметному ускорению обработки данных и обучения моделей. Также стали доступны удобные инструменты для обработки данных, которые снизили барьеры для доступа к ним. В совокупности это означает, что

сбор, хранение и обработка данных никогда еще не были такими простыми.

За последние 10 лет появились более мощные модели машинного обучения, известные как глубокое обучение, которые произвели революцию в компьютерной обработке данных языка и изображений. Термин «глубокое обучение» описывает семейство моделей многослойных нейронных сетей. Нейронные сети существуют с 1940-х гг., но лучше всего они проявили себя с большими сложными наборами данных и мощными вычислительными ресурсами для обучения. Таким образом, появление глубокого обучения в последние несколько лет связано с ростом больших данных и вычислительной мощности. Тем не менее не будет преувеличением сказать, что влияние глубокого обучения на целый ряд областей исключительно. История AlphaGo[9] от DeepMind является отличным примером того, как глубокое обучение произвело революцию в области исследований. Го — настольная игра, созданная в Китае 3000 лет назад. Играть в го проще, чем в шахматы: игроки по очереди размещают фигуры на доске с целью захвата фигур противника или окружения пустой территории. Однако простота правил и тот факт, что в го используется доска с бóльшим числом клеточек, означают и большее число возможных конфигураций, нежели в шахматах. Число возможных конфигураций в го больше, чем число атомов во Вселенной, и это делает го гораздо более сложной игрой для компьютера, чем шахматы, в силу огромного пространства для поиска и сложности в оценке всех возможных конфигураций. Команда DeepMind использовала модели глубокого обучения, чтобы AlphaGo смогла оценивать конфигурации на доске и выбирать следующий ход. В результате AlphaGo стала первой компьютерной программой, которая победила профессионального игрока, а в марте 2016 г. она одержала победу над 18-кратным чемпионом мира по го Ли Седолем в матче, который посмотрели более 200 млн человек во

всем мире. Еще совсем недавно, в 2009 г., лучшая компьютерная программа для игры в го оценивалась как соответствующая любительскому уровню, а уже спустя семь лет AlphaGo обыграла чемпиона мира. В 2016 г. в самом престижном академическом журнале *Nature* была опубликована статья, описывающая алгоритмы глубокого обучения, заложенные в AlphaGo [12].

Глубокое обучение также оказало огромное влияние на ряд публичных потребительских технологий. В настоящее время Facebook использует глубокое обучение для распознавания лиц и анализа текста, чтобы подбирать людям рекламу на основе их онлайн-разговоров. Google и Baidu используют глубокое обучение для распознавания изображений, титрования и поиска, а также для машинного перевода. Виртуальные помощники Apple Siri, Amazon Alexa, Microsoft Cortana и Samsung Bixby используют распознавание речи на основе глубокого обучения. Huawei разрабатывает виртуального помощника для китайского рынка, в котором также будет использоваться система распознавания речи с глубоким обучением. В главе 4 мы более подробно расскажем об этом. Хотя глубокое обучение является важной технической разработкой, возможно, с точки зрения роста науки о данных наиболее интересным его аспектом будет демонстрация возможностей и преимуществ самой науки о данных и привлечение внимания организаций к результатам таких успешных историй.

Разоблачение мифов

Наука о данных дает много преимуществ современным организациям, но вокруг нее крутится и масса слухов, поэтому важно понять, каковы реальные ограничения науки о данных. Одним из самых больших мифов является вера в то, что наука о

данных — автономный процесс, который сам найдет решения наших проблем. Но на деле на всех этапах этого процесса требуется квалифицированный человеческий контроль. Люди нужны для того, чтобы сформулировать проблему, спроектировать и подготовить данные, выбрать, какие алгоритмы машинного обучения являются наиболее подходящими, критически интерпретировать результаты анализа и спланировать соответствующие действия, основанные на выявленных закономерностях. Без квалифицированного человеческого надзора проект по обработке данных не сможет достичь своих целей. Лучшие результаты мы видим, когда объединяются человеческий опыт и компьютерная мощь. Как выразились Линофф и Берри: «Глубинный анализ данных позволяет компьютерам делать то, что они умеют лучше всего, — копаться в куче информации. Это, в свою очередь, дает людям делать то, что лучше всего получается у них, — ставить задачу и осмыслять результаты» [13].

Широкое и все возрастающее использование науки о данных означает, что сегодня самая большая проблема для многих организаций заключается в найме аналитиков. Человеческий фактор в науке о данных имеет первостепенное значение, и ограниченный ресурс специалистов является основным узким местом в распространении самой науки. Чтобы лучше представить масштаб нехватки специалистов, заглянем в отчет McKinsey Global Institute (MGI) за 2011 г.: прогноз дефицита сотрудников с навыками обработки данных и аналитики в Соединенных Штатах в ближайшие годы — от 140 000 до 190 000 человек; еще больший дефицит — 1,5 млн человек — менеджеров, способных понимать науку о данных и аналитические процессы на уровне, который позволяет им надлежащим образом запрашивать и интерпретировать результаты [14]. Спустя пять лет в своем отчете за 2016 г. MGI по-прежнему убежден, что наука о данных имеет огромный

неиспользованный потенциал в расширяющемся диапазоне приложений, а дефицит специалистов сохраняется с прогнозируемой нехваткой 250 000 человек в ближайшей перспективе [15].

Второй большой миф заключается в том, что каждый проект непременно нуждается в больших данных и требует глубокого обучения. Как правило, наличие большого объема данных помогает, но гораздо важнее, чтобы данные были правильными. Подобные проекты часто ведутся в организациях, которые располагают значительно меньшими ресурсами с точки зрения данных и вычислительной мощности, чем Google, Baidu или Microsoft. Примеры проектов небольшого масштаба: прогнозирование требований возмещения ущерба в страховой компании, которая обрабатывает около 100 заявок в месяц; прогноз отсева студентов в университете, где обучаются менее 10 000 человек; ожидания ротации членов профсоюза с несколькими тысячами участников. Эти примеры показывают, что организации не нужно обрабатывать терабайты информации или иметь в своем распоряжении огромные вычислительные ресурсы, чтобы извлечь выгоду из науки о данных.

Третий миф заключается в том, что современное программное обеспечение для обработки данных легко в использовании и, следовательно, сама наука о данных тоже не представляет собой ничего сложного. Программное обеспечение для обработки данных действительно стало более удобным для пользователя. Однако такая простота может скрывать тот факт, что для получения правильных результатов требуются как соответствующие знания предметной области, так и знания в области науки о данных, касающиеся свойств данных и допущений, лежащих в основе глубинного анализа и алгоритмов машинного обучения. На самом деле никогда еще не было так легко стать плохим специалистом по данным. Как и в любой сфере жизни, если вы не понимаете, что делаете, то будете

совершать ошибки. Опасность, связанная с наукой о данных, заключается в том, что людей может отпугивать сложность технологии, и тогда они готовы поверить любым результатам, которые выдает им программное обеспечение. Однако всегда высока вероятность неправильной постановки задачи, неверного ввода данных или ненадлежащего использования методов анализа. В этих случаях результаты, представленные программным обеспечением, скорее всего, будут ответом на неправильные вопросы или окажутся основанными на неверных данных или расчетах.

Последний миф, который мы упомянем, — вера в то, что наука о данных быстро окупается. Истинность этого утверждения зависит исключительно от контекста организации. Внедрение науки о данных может потребовать значительных инвестиций с точки зрения инфраструктуры и найма персонала с опытом соответствующей работы. Более того, наука о данных не даст положительных результатов по каждому проекту. Иногда в данных нет искомого бриллианта или организация не в состоянии использовать прозрение, полученное в результате анализа. Однако в тех случаях, когда бизнес-проблема ясна, а соответствующая информация и человеческий опыт доступны, наука о данных, как правило, обеспечивает действенное понимание, которое дает организации конкурентное преимущество.

Источники

1. Shmueli, Galit. 2010. “To Explain or to Predict?” Statistical Science 25 (3): 289–310. doi:10.1214/10-STS330.
2. Breiman, Leo. 2001. “Statistical Modeling: The Two Cultures (with Comments and a Rejoinder by the Author).” *Statistical*

Science 16 (3): 199–231. doi:10.1214/ss/1009213726.

3. Silver, David, Aja Huang, Chris J. Maddison, Arthur Guez, Laurent Sifre, George van den Driessche, Julian Schrittwieser, et al. 2016. “Mastering the Game of Go with Deep Neural Networks and Tree Search.” *Nature* 529 (7587): 484–89. doi:10.1038/nature16961.
4. Manyika, James, Michael Chui, Brad Brown, Jacques Bughin, Richard Dobbs, Charles Roxburgh, and Angela Hung Byers. 2011. “Big Data: The next Frontier for Innovation, Competition, and Productivity.” McKinsey Global Institute. http://www.mckinsey.com/business-functions/digital-mckinsey/our-insights/big-data-thenext-frontier-for-innovation.
5. Henke, Nicolaus, Jacques Bughin, Michael Chui, James Manyika, Tamim Saleh, and Bill Wiseman. 2016. “The Age of Analytics: Competing in a Data-Driven World.” McKinsey Global Institute. http://www.mckinsey.com/business-functions/mckinsey-analytics/ourinsights/the-age-of-analytics-competing-in-a-data-driven-world.
6. Tufte, Edward R. 2001. *The Visual Display of Quantitative Information.* 2nd edition edition. Cheshire, Conn: Graphics Press.
7. Taylor, David. 2016. “Battle of the Data Science Venn Diagrams.” *KDnuggets.* http://www.kdnuggets.com/2016/10/battle-data-science-venn-diagrams.html.
8. Cleveland, William S. 2001. “Data Science: An Action Plan for Expanding the Technical Areas of the Field of Statistics.” *International Statistical Review* 69 (1): 21–26. doi:10.1111/j.1751-5823.2001.tb00477.x.
9. DeZyre. 2015. “How Big Data Analysis Helped Increase Walmart’s Sales Turnover?” *DeZyre.* https://www.dezyre.com/article/how-big-data-analysis-helped-increase-walmarts-salesturnover/109.

10. Kitchin, Rob. 2014. *The Data Revolution: Big Data, Open Data, Data Infrastructures and Their Consequences.* Sage.
11. Anderson, Chris. 2008. *The Long Tail: Why the Future of Business Is Selling Less of More.* Revised edition. New York: Hachette Books.
12. Linoff, Gordon S., and Michael JA Berry. 2011. *Data Mining Techniques: For Marketing, Sales, and Customer Relationship Management.* John Wiley & Sons.
13. Lewis, Michael. 2004. *Moneyball: The Art of Winning an Unfair Game.* 1st edition. New York: W. W. Norton & Company.
14. Дабнер Стивен, Левитт Стивен. Фрикономика. Экономист-хулиган и журналист-сорвиголова исследуют скрытые причины всего. — М.: Альпина Паблишер, 2019.
15. Few, Stephen. 2012. *Show Me the Numbers: Designing Tables and Graphs to Enlighten.* Second edition. Burlingame, CA: Analytics Press.

Глава 2

ЧТО ТАКОЕ ДАННЫЕ И ЧТО ТАКОЕ НАБОР ДАННЫХ?

Как следует из названия, наука о данных фундаментально зависит от самих данных. По существу данные являются абстракцией реальной сущности (человека, объекта или события). Термины «*переменная*», «*признак*» *или* «*атрибут*» часто используются взаимозаменяемо для обозначения отдельно взятой абстракции. Обычно каждый объект описывается рядом атрибутов. Например, книга может иметь следующий набор атрибутов: автор, название, тема, жанр, издатель, цена, дата публикации, количество слов, глав, страниц, издание, ISBN и т.д.

Набор данных состоит из данных, относящихся к совокупности объектов, причем каждый объект описан в терминах набора атрибутов. В своей наиболее простой форме[10] набор данных организован в виде матрицы размером $n \times m$, называемой *аналитической записью*, где n — количество объектов (строк), а m — количество атрибутов (столбцов). В науке о данных термины «набор данных» и «аналитическая запись» часто используются взаимозаменяемо, при этом аналитическая запись является конкретным представлением набора данных. Таблица 1 иллюстрирует аналитическую запись для набора данных нескольких книг. Каждый ряд в таблице описывает одну книгу. Термины «объект», «экземпляр», «пример», «сущность», «кейс» и «запись» используются в науке о данных для обозначения строки.

Таким образом, набор данных содержит набор объектов, и каждый из объектов описывается набором атрибутов.

Таблица 1. Набор данных

ID	Название	Автор	Год	Обложка	Издание	Цена, ₽
1	«Эмма»	Остин	1815	Мягкая обложка	20	373
2	«Дракула»	Стокер	1897	Твердый переплет	15	780
3	«Айвенго»	Скотт	1820	Твердый переплет	8	1625
4	«Похищенный»	Стивенсон	1886	Мягкая обложка	11	325

Построение аналитической записи — необходимое условие работы с данными. Фактически в большинстве проектов по обработке данных бóльшая часть времени и усилий уходит на создание, очистку и обновление аналитической записи. Аналитическая запись часто создается путем объединения информации из множества различных источников: может потребоваться извлечение данных из нескольких баз, хранилищ или компьютерных файлов в разных форматах (например, в виде электронных таблиц и CSV-файлов) или скрапинг[11] в интернете или социальных сетях.

В таблице 1 перечислены четыре книги. Если не считать атрибут ID, который представляет собой простую метку строки и, следовательно, бесполезен для анализа, каждая книга описана с помощью шести атрибутов: название, автор, год, обложка, издание и цена. Мы могли бы включить их намного больше для каждой книги, но, как это обычно и бывает в подобных проектах, нам нужно ограничить набор данных. В нашем случае мы должны просто уместить атрибуты в размер страницы. Однако в

большинстве проектов ограничения касаются того, какие атрибуты доступны, а также какие из них имеют отношение к проблеме, которую мы пытаемся решить в конкретной предметной области. Включение дополнительных атрибутов в набор данных никогда не обходится без затрат. Во-первых, вам потребуются дополнительные время и усилия для сбора и проверки качества данных в атрибутах для каждого объекта и их интеграции в аналитическую запись. Во-вторых, включение нерелевантных или избыточных атрибутов может отрицательно сказаться на производительности многих алгоритмов, используемых для анализа данных. Включение большого количества атрибутов в набор данных увеличивает вероятность того, что алгоритм найдет не относящиеся к делу или ложные закономерности, которые только кажутся статистически значимыми в рамках выборки объектов. С проблемой правильных атрибутов сталкиваются все проекты науки о данных, и иногда ее решение сводится к итеративному процессу проведения экспериментов методом проб и ошибок, где каждая итерация проверяет результаты, полученные с использованием различных подмножеств атрибутов.

Существуют разные типы атрибутов, и для каждого из них подходят разные виды анализа. Их понимание и распознавание является фундаментальным навыком для специалиста по данным. К стандартным типам относятся числовые (включая интервальные и относительные), номинальные и порядковые. Числовые атрибуты описывают измеримые величины, представленные целыми числами или действительными величинами. Числовые атрибуты могут быть измерены как по шкале интервалов, так и по шкале отношений. Интервальные атрибуты измеряются по шкале с фиксированными, но произвольными единицами измерений и произвольным началом отсчета. Примерами интервальных атрибутов могут быть измерения даты и времени. К ним применяют упорядочивание и

вычитание. Умножение, деление и прочие операции в этом случае не подходят. Шкала отношений аналогична шкале интервалов с единственным отличием: ее нулевая точка — истинный нуль. Он указывает на то, что количество, которое могло бы быть измерено, отсутствует. Особенность шкалы отношений состоит в том, что мы можем описать любое значение как кратное другому значению. Температура — прекрасный пример для понимания разницы между шкалой интервалов и шкалой отношений [1]. По шкале Цельсия и по шкале Фаренгейта температура измеряется интервально, поскольку значение 0 на любой из этих шкал не указывает на отсутствие тепла. Таким образом, хотя мы и можем вычислить разницу между температурами на этих шкалах и сравнить различия, мы не можем сказать, что 20 °C — это в два раза теплее, чем 10 °C. В отличие от этого, измерение температуры в кельвинах ведется по шкале отношений, поскольку 0 K (абсолютный нуль) — это температура, при которой прекращается всякое тепловое движение. Другие распространенные примеры измерений по шкале отношений: количество денег, вес, рост и экзаменационные отметки (шкала 0–100). В таблице 1 атрибут года является примером атрибута шкалы интервалов, а атрибут цены — примером атрибута шкалы отношений.

Номинальные (также известные как категориальные) атрибуты принимают значения из ограниченного набора. Эти значения являются именами (поэтому они и называются номинальными) для категорий, классов или обстоятельств. Примеры номинальных атрибутов включают семейное положение (холост, женат, разведен) или тип пива (эль, светлый эль, пильзнер, портер, стаут и т.д.). Бинарный атрибут — это особый случай номинального атрибута, у которого набор возможных значений ограничен только двумя. Примером может служить бинарный атрибут «спам», который описывает, является электронная почта спамом (да) или не является (нет). К

номинальным атрибутам не могут быть применены упорядочивание или арифметические операции. Обратите внимание, что номинальный атрибут может быть отсортирован в алфавитном порядке, но эта операция не тождественна упорядочиванию. В таблице 1 автор и название являются примерами номинальных атрибутов.

Порядковые атрибуты аналогичны номинальным, но с той разницей, что можно ранжировать значения переменных. Например, атрибут, описывающий ответ на вопрос анкетирования, может принимать значения из области определения: «очень не нравится», «не нравится», «нейтрально», «нравится» и «очень нравится». Существует естественное упорядочивание этих значений — от сильной неприязни к сильной симпатии (или, наоборот, в зависимости от условия). Тем не менее важной особенностью порядковых атрибутов является отсутствие понятия равного расстояния между этими значениями. Например, когнитивное расстояние между неприязнью и нейтральным отношением может быть отличным от расстояния между симпатией и сильной симпатией. В результате неуместно применять арифметические операции (такие, как усреднение) к порядковым атрибутам. В таблице 1 атрибут «издание» является примером порядкового атрибута. Граница между номинальными и порядковыми данными не всегда четкая. Для примера возьмем атрибут, который описывает погоду и может принимать значения «солнечно», «дождливо», «пасмурно». Один человек может сказать, что этот атрибут номинальный, значения которого не упорядочены, в то время как другой будет утверждать, что атрибут является порядковым, при этом рассматривая облачность как промежуточное значение между «солнечно» и «дождливо» [2].

Тип атрибута (числовой, порядковый, номинальный) влияет на методы анализа и понимания данных. Эти методы включают в себя как основную статистику, которую мы можем использовать

для описания распределения значений атрибута, так и более сложные алгоритмы, которые мы применяем для выявления закономерностей отношений между атрибутами. На базовом уровне анализа числовые атрибуты допускают арифметические операции, а типичный статистический анализ, применяемый к числовым атрибутам, заключается в измерении центральной тенденции (с использованием среднего значения атрибута) и разброса значений атрибутов (с использованием дисперсии или стандартного отклонения). Однако не имеет смысла применять арифметические операции к номинальным или порядковым атрибутам. Базовый анализ этих типов атрибутов включает в себя подсчет того, сколько раз значение встречается в наборе данных, и/или вычисление процента вхождения этого значения.

Данные генерируются в процессе абстракции, поэтому они всегда являются результатом принятых человеком решений и сделанного им выбора. В основе каждой абстракции конкретный человек или группа людей решают, от чего абстрагироваться и какие категории или измерения использовать в полученном отображении. Поэтому данные никогда не являются объективным описанием мира. Данные всегда частичны и предвзяты. Как заметил Альфред Коржибски: «Карта не является отображаемой ею территорией, но если она верная, то имеет структуру, подобную территории, которая содержит информацию о ее полезности [3]».

Тип атрибута (числовой, порядковый, номинальный) влияет на методы анализа и понимания данных.

Другими словами, данные не являются идеальным отображением сущностей и процессов реального мира, которые мы пытаемся постичь, но если быть аккуратным при моделировании и сборе данных, то результаты анализа могут дать полезную информацию для решения наших реальных проблем. Сюжет фильма «Человек, который изменил все» (Moneyball), о котором упоминалось в главе 1, служит примером того, что определяющим фактором успеха во многих проектах науки о данных являются абстракции (атрибуты), подходящие для использования в данной конкретной области. Напомним, что ключом в этой истории было осознание клубом «Окленд Атлетикс» того, что процентное соотношение попадания игрока на базу и упущенных возможностей является более информативным показателем его успешности, чем традиционно принятые в бейсболе статистические данные, такие как средний уровень достижений. Использование различных атрибутов для описания игроков дало «Окленд Атлетикс» лучшую, нежели у других команд, модель, которая позволила им выявлять недооцененных игроков и конкурировать с крупными клубами при меньшем бюджете.

Эта история иллюстрирует применимость старой поговорки «Что посеешь — то и пожнешь» к науке о данных: если входные данные вычислительного процесса неверны, то выходные данные также будут неправильны. Действительно, наука о данных имеет две особенности, которые всегда необходимо учитывать: а) для успешности проектов необходимо уделять много внимания созданию самих данных (как с точки зрения выбора, который мы делаем при моделировании абстракции, так и с точки зрения качества данных, полученных в процессе) и б) необходимо проверять результаты процесса, хотя бы потому, что выявленная компьютером закономерность может оказаться основанной на

отклонениях модели и увести нас в сторону от реального понимания анализируемых процессов.

Перспективы данных

Помимо типов (числовые, номинальные и порядковые), существуют и другие полезные способы классификации данных. Один из них различает *структурированные* и *неструктурированные* данные. *Структурированными* называются данные, которые могут храниться в таблице, где каждый объект имеет одинаковую структуру (т.е. набор атрибутов). В качестве примера можно привести демографические данные населения, где каждая строка в таблице описывает одного человека и состоит из одного и того же набора атрибутов (имя, возраст, дата рождения, адрес, пол, образование, статус занятости и т.д.). Структурированные данные можно легко хранить, систематизировать, искать, переупорядочивать и объединять с другими структурированными данными. К ним легко применяемы методы науки о данных, поскольку по определению они уже находятся в формате, который подходит для интеграции в аналитическую запись. *Неструктурированные* данные описывают такие данные, где каждый объект в наборе может иметь собственную внутреннюю структуру и эта структура необязательно одинакова для каждого объекта. Представьте себе набор веб-страниц, где у каждой есть структура, но при этом отличная от других. Неструктурированные данные встречаются гораздо чаще, чем структурированные. Например, естественные текстовые массивы (электронные письма, твиты, СМС, посты, романы и т.д.) можно считать неструктурированными данными; то же относится к коллекциям звуковых, графических и видеофайлов. Различия в структуре между отдельными

элементами не позволяют анализировать неструктурированные данные в необработанном виде. Зачастую мы можем извлекать структурированные данные из неструктурированных, используя методы искусственного интеллекта (такие, как обработка естественного языка или машинное обучение), цифровую обработку сигналов или компьютерное зрение. Однако внедрение и тестирование этих процессов преобразования данных является дорогостоящим и трудоемким и может привести к значительным накладным расходам в проекте.

Иногда атрибутами являются *необработанные* абстракции, извлеченные непосредственно из событий или объектов, например рост человека, число слов в электронном письме, температура в комнате, время или место события. Но кроме того данные могут быть *производными*, т.е. полученными из других данных. Например, средняя зарплата в компании или разница температур в комнате за период времени. В обоих случаях результирующие данные являются производными от исходного набора необработанных данных (отдельно взятых зарплат или показаний температуры) путем применения к ним функции. Часто реальная ценность проекта по обработке данных состоит в выявлении одного или нескольких важных производных атрибутов, которые обеспечивают понимание проблемы. В качестве иллюстрации представьте, что мы пытаемся исследовать проблему ожирения и выявить атрибуты, которые идентифицируют потенциально подверженных заболеванию людей. Мы бы начали с необработанных атрибутов отдельных лиц, их роста и веса, но после более подробного исследования вопроса создали бы более информативный производный атрибут, такой как индекс массы тела (ИМТ). ИМТ — это соотношение массы тела и роста человека. Понимание того, что взаимосвязь необработанных атрибутов массы и роста дает больше информации об ожирении, чем любой из этих двух признаков по отдельности, может помочь нам определить людей в группе

населения, которые подвержены риску ожирения. Очевидно, что ИМТ является простейшим примером, который мы используем здесь, чтобы показать важность производных атрибутов. Но давайте рассмотрим ситуации, когда понимание проблемы приходит через несколько производных атрибутов, где каждый, в свою очередь, включает в себя две (или более) характеристики. Именно в таких условиях, когда несколько атрибутов взаимодействуют друг с другом, наука о данных дает нам реальные преимущества, поскольку ее алгоритмы способны извлекать производные атрибуты из необработанных данных.

Часто реальная ценность проекта по обработке данных состоит в выявлении одного или нескольких важных производных атрибутов, которые обеспечивают понимание проблемы.

Существует два основных типа *необработанных* данных по способу их получения: *собранные* и *выхлопные данные* [4]. *Собранные данные* получают посредством прямого измерения или наблюдения, предназначенного для этой цели. Например, основная цель опросов или экспериментов состоит в сборе конкретных данных по конкретной теме. *Выхлопные данные*, напротив, побочный продукт процесса (подобно выхлопным газам), основной целью которого является нечто иное, чем сбор данных. Например, основная цель социальных сетей — дать пользователям возможность общаться друг с другом. Однако для каждого опубликованного изображения, поста, ретвита или лайка создается ряд выхлопных данных: кто поделился, кто просмотрел, какое устройство использовалось, чье устройство использовалось, в какое время суток, сколько людей просматривали / поставили лайк / ретвитнули и т.д. Точно так же основная цель сайта Amazon — дать возможность пользователям совершать покупки. Но это не мешает каждой покупке генерировать выхлопные данные: какие товары пользователь добавил в корзину, сколько времени он провел на сайте, какие другие товары он просматривал и т.д.

Одним из наиболее распространенных типов выхлопных данных являются *метаданные*, т.е. данные, описывающие другие данные. Когда Эдвард Сноуден опубликовал документы АНБ, касающиеся программы тотальной слежки PRISM, он также сообщил, что агентство собирало большое количество метаданных о телефонных звонках людей. Это значит, что АНБ фактически не записывало их содержание (т.е. не вело прослушивания телефонных разговоров), но собирало данные о звонках, например когда был сделан звонок, кому, как долго длился и т.д. [5]. Этот тип сбора данных может показаться не столь зловещим, но исследовательский проект MetaPhone, проведенный в Стэнфорде, обнаружил, что метаданные

телефонного звонка могут раскрыть большой объем личной информации [6]. Тот факт, что многие организации работают в узких сферах, позволяет относительно легко выявлять информацию о человеке на основе его телефонных звонков. Например, некоторые из участников исследования MetaPhone звонили «Анонимным алкоголикам», адвокатам по бракоразводным процессам и в медицинские клиники, специализирующиеся на венерических болезнях. О многом могут говорить и закономерности звонков. Вот два примера закономерностей, выявленных в ходе исследования и раскрывающих очень деликатную информацию:

> «Участник А общался с несколькими местными группами поддержки людей, страдающих неврологическими заболеваниями, специализированной аптекой, службой лечения редких состояний и горячей линией лекарственного средства, применяемого исключительно для лечения рассеянного склероза… В течение трех недель участник В связывался с магазином товаров для ремонта, слесарем, продавцом оборудования для гидропоники и торговцем марихуаной [7]».

Традиционно наука о данных была сосредоточена на получении собранных данных. Однако, как показывает исследование MetaPhone, выхлопные данные также могут быть использованы для выявления скрытого смысла. В последние годы выхлопные данные становятся все более и более полезными, особенно в области взаимодействия с клиентами, где связывание между собой различных наборов выхлопных данных может создать более широкий клиентский профиль, тем самым позволяя бизнесу точнее ориентировать свои услуги и маркетинг. Сегодня одним из факторов, стимулирующих развитие науки о данных,

является признание современным бизнесом ценности выхлопных данных и их потенциала.

Данные накапливаются, мудрость — нет!

Цель науки о данных — использовать их, чтобы получить прозрение и понимание. Библия призывает нас к пониманию через мудрость: «Главное — мудрость: приобретай мудрость, и всем имением твоим приобретай разум» (Притч. 4:7). Этот совет разумен, но он ставит вопрос о том, как именно нужно искать мудрости. Следующие строки из стихотворения Т.С. Элиота «Камень» описывают иерархию мудрости, знаний и информации:

> Где мудрость, которую мы потеряли в знанье?
> Где знанье, которое мы потеряли в сведеньях? [8]

Иерархия Элиота отражает стандартную модель структурных отношений между мудростью, знаниями, информацией и данными, известную как пирамида DIKW (см. рис. 2). В пирамиде DIKW данные предшествуют информации, которая предшествует знаниям, которые, в свою очередь, предшествуют мудрости. Хотя порядок уровней в иерархии, как правило, не вызывает споров, различия между этими уровнями и процессы, необходимые для перехода от одного к другому, часто оспариваются. Но если посмотреть в широком смысле, то можно утверждать следующее:

- данные создаются с помощью абстракции или измерения мира;
- информация — это данные, которые были обработаны, структурированы или встроены в контекст таким образом, что стали значимы для людей;

- знание — это информация, которая была истолкована и понята таким образом, что появилась возможность действовать в соответствии с ней по необходимости;
- мудрость — это умение найти надлежащее применение знанию.

Рис. 2. Пирамида DIKW (источник: Kitchin 2014 [4]).

Последовательные операции в процессе обработки данных могут быть представлены аналогичной пирамидальной иерархией, где ширина пирамиды отображает объем данных, обрабатываемых на каждом уровне, и чем выше уровень, тем результаты действий более информативны для принятия решения. Рис. 3 иллюстрирует иерархию операций науки о данных, начиная с их сбора и генерации посредством предварительной обработки и агрегирования и заканчивая пониманием результатов, обнаружением закономерностей и созданием моделей с использованием машинного обучения для принятия решений в бизнес-контексте.

Рис. 3. Пирамида Data Science (*источник*: Han, Kamber and Pei 2011 [1]).

Процесс CRISP-DM

В научной среде регулярно выдвигаются новые идеи о том, каким способом лучше всего взбираться на вершину пирамиды науки о данных. Наиболее часто используется межотраслевой стандартный процесс исследования данных CRISP-DM. Этот процесс в течение целого ряда лет занимает первые места всевозможных отраслевых опросов. Одно из преимуществ CRISP-DM и причина, по которой он так широко используется, заключается в том, что процесс спроектирован как независимый от программного обеспечения, поставщика или метода анализа данных.

CRISP-DM разрабатывался консорциумом организаций, в который входили ведущие поставщики данных, конечные пользователи, консалтинговые компании и исследователи. Первоначальный проект CRISP-DM был частично спонсирован

Европейской комиссией в рамках программы ESPRIT и представлен на семинаре в 1999 г. С тех пор было предпринято несколько попыток обновить процесс, но оригинальная версия все еще остается наиболее востребованной. В течение многих лет существовал отдельный сайт CRISP-DM, но сейчас он закрыт, и в большинстве случаев вы будете перенаправлены на сайт SPSS компании IBM, которая участвовала в проекте с самого начала. Консорциум участников опубликовал детальную (76 страниц), но вполне понятную пошаговую инструкцию для процесса, которая находится в свободном доступе в интернете [9]. Далее мы кратко изложим основную структуру и задачи процесса.

Жизненный цикл CRISP-DM состоит из шести этапов — *понимание бизнес-целей, начальное изучение данных, подготовка данных, моделирование, оценка* и *внедрение,* — показанных на рис. 4. Данные являются центром всех операций, как это видно из диаграммы CRISP-DM. Стрелки между этапами указывают типичное направление процесса. Сам процесс является частично структурированным, т.е. специалист по данным не всегда проходит все шесть этапов линейно. В зависимости от результата конкретного этапа может потребоваться вернуться к одному из предыдущих, повторить текущий или перейти к следующему.

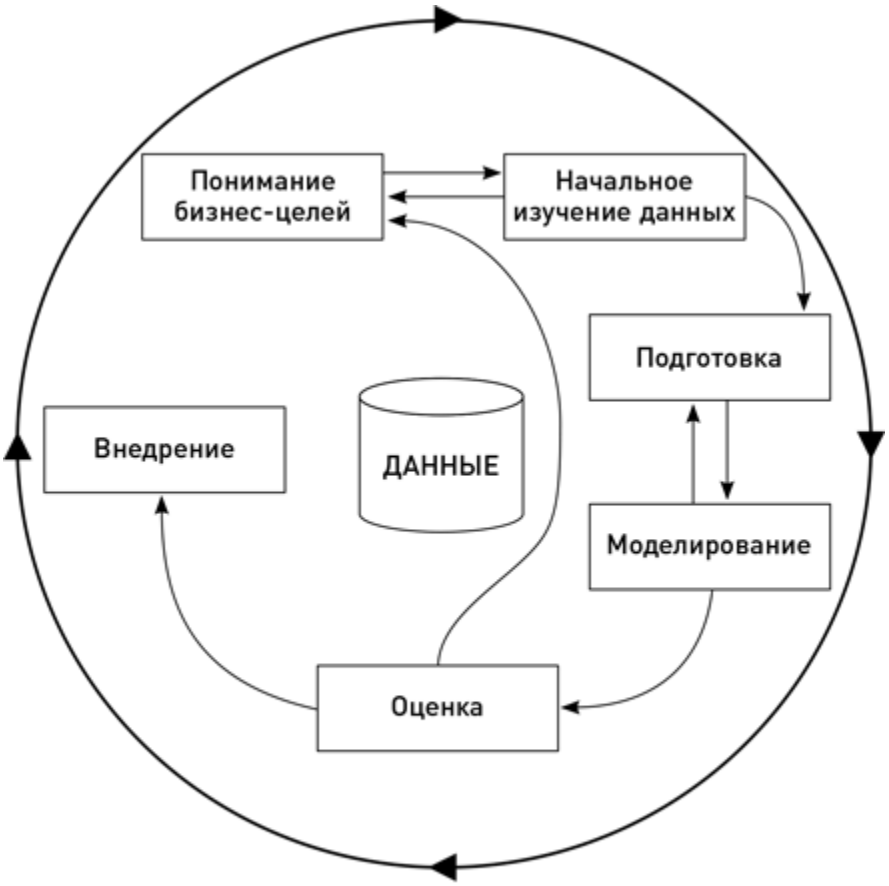

Рис. 4. Жизненный цикл CRISP-DM (основано на рис. 2 из Chapman, Clinton, Kerber 1999 [8]).

На первых двух этапах — понимания бизнес-целей и начального изучения данных — специалист пытается сформулировать цели проекта с точки зрения бизнеса и знакомится с данными, которые тот имеет в своем распоряжении. На ранних стадиях проекта придется часто переключаться между фокусировкой на бизнесе и изучением доступных данных. Это связано с тем, что специалист по данным должен идентифицировать бизнес-проблему, а затем понять, доступны ли соответствующие данные для поиска ее решения. Если они доступны, то проект может продолжаться, в противном случае специалисту придется искать альтернативную проблему. В течение этого периода специалист по данным плотно работает с коллегами из бизнес-отделов организации (продаж, маркетинга,

операций), пытаясь вникнуть в их проблемы, а также с администраторами баз данных, чтобы изучить доступный материал.

Как только бизнес-проблема была четко сформулирована, а специалист убедился в том, что соответствующие данные доступны, происходит переход к очередному этапу CRISP-DM — подготовке данных. Целью этого этапа является создание набора данных, который можно использовать для анализа. Обычно это подразумевает интеграцию источников из нескольких баз данных. Когда в организации существует хранилище данных, эта интеграция значительно упрощается. После создания набора данных необходимо проверить и исправить их качество. Типичные проблемы качества включают выбросы и пропущенные значения. Проверка качества крайне важна, поскольку ошибки в данных могут серьезно повлиять на производительность алгоритмов анализа.

Следующим этапом CRISP-DM является моделирование. На этой стадии используются автоматические алгоритмы для выявления полезных закономерностей в данных и создаются модели, которые кодируют эти закономерности. Алгоритмы для выявления закономерностей также называются алгоритмами машинного обучения. На этапе моделирования специалист по данным обычно использует несколько алгоритмов машинного обучения для подготовки разных моделей в каждом наборе данных. Необходимость в нескольких моделях вызвана тем, что разные типы алгоритмов машинного обучения ищут разные типы закономерностей в данных, и на этапе моделирования специалист, как правило, не знает, какие именно закономерности нужно искать. Таким образом, имеет смысл поэкспериментировать с различными алгоритмами и посмотреть, какой из них работает лучше всего.

В большинстве проектов тестовые результаты испытания моделей позволят выявить проблемы с данными. Иногда эти

ошибки обнаруживаются, когда специалист выясняет, что производительность модели ниже ожидаемой или, наоборот, она подозрительно хороша. Бывает, что, изучая структуру моделей, специалист по данным неожиданно выясняет ее зависимость от каких-либо атрибутов и возвращается к данным, чтобы проверить, правильно ли они кодированы. В результате некоторые этапы в проекте повторяются: за моделированием следует подготовка данных, затем снова моделирование, снова подготовка данных и т.д. Например, Дэн Стейнберг и его команда сообщили, что в ходе одного своего проекта они перестраивали набор данных 10 раз в течение шести недель, причем на пятой неделе этого процесса после ряда итераций по очистке данных и подготовке в них была обнаружена существенная ошибка [10]. Если бы она не была выявлена и исправлена, проект не стал бы успешным.

На двух последних этапах (при оценке и внедрении) вы сосредотачиваетесь на том, каким образом модели будут приспособлены к бизнесу и его процессам. Тесты, выполняемые на этапе моделирования, ориентированы исключительно на точность модели в наборе данных. Этап оценки включает оценку моделей в более широком контексте, определяемом потребностями бизнеса. Соответствует ли модель целям процесса? Адекватна ли она с точки зрения бизнеса? На этом этапе специалист по данным должен провести анализ для обеспечения качества проекта: не было ли что-то упущено, можно ли было сделать лучше и т.д. На основании общей оценки моделей принимается основное решение этого этапа — можно ли внедрять какую-то из них в бизнес или требуется еще одна итерация процесса CRISP-DM для создания моделей более адекватных. Если модели одобрены, проект переходит к финальной стадии процесса — внедрению. На этапе внедрения изучается то, каким образом можно развернуть выбранные модели в бизнес-среде, как интегрировать их в техническую

инфраструктуру и бизнес-процессы организации. Лучшие из моделей — те, которые плавно вписываются в существующую практику. Такие модели ориентированы на конкретных пользователей, столкнувшихся с четко обозначенной проблемой, которую эта модель и призвана решить. Кроме того, на этапе внедрения создается план периодической проверки эффективности модели.

Внешняя окружность диаграммы CRISP-DM подчеркивает тот факт, что весь процесс имеет итеративный характер. При обсуждении проектов науки о данных об их итеративности часто забывают. После разработки и внедрения модель должна регулярно пересматриваться, чтобы удовлетворять задачам бизнеса и оставаться актуальной. Существует масса причин, по которым модель может устареть: изменяются потребности бизнеса, процессы, которые модель имитирует или поясняет (например, поведение клиентов, типы спама и т.д.), или потоки данных, используемые моделью (скажем, новый датчик дает несколько другие показатели, что снижает точность модели). Частота пересмотра зависит от того, как быстро развиваются экосистема бизнеса и данные, используемые моделью. Постоянный мониторинг необходим, чтобы определить наилучшее время для повторного запуска процесса. Это как раз то, что представляет собой внешний круг CRISP-DM. Например, в зависимости от данных, поставленной задачи и сферы деятельности вы можете проходить этот итеративный процесс еженедельно, ежемесячно, ежеквартально, ежегодно или даже ежедневно. На рис. 5 приведена сводная информация об этапах процесса и основных задачах, связанных с ними.

Неопытные специалисты по данным часто допускают ошибку: сосредотачивая усилия на этапе моделирования CRISP-DM, они чересчур поспешно проходят другие этапы. Их логика заключается в том, что наиболее важным результатом проекта должна стать модель, поэтому бóльшую часть своего времени

необходимо посвятить именно ее разработке. Однако маститые специалисты по данным тратят больше времени на то, чтобы задать проекту четкий вектор и обеспечить его правильными данными. Успех в науке о данных достигается ясностью бизнес-задач для специалиста, ведущего проект. Поэтому этап понимания бизнес-целей крайне важен. Что касается получения правильных данных для проекта, то опрос специалистов, проведенный в 2016 г., показал, что 79% своего времени они уделяют именно подготовке данных [11]. Тот же опрос выявил, что распределение времени между основными задачами в проектах выглядит следующим образом:

- сбор данных — 19%;
- очистка и организация данных — 60%;
- построение обучающих моделей — 3%;
- анализ данных для выявления закономерностей — 9%;
- уточнение алгоритмов — 4%;
- другие задачи — 5%.

Показатель 79% для подготовки суммирует время, затраченное на сбор, очистку и организацию данных. Этот показатель — около 80% времени проекта — присутствует в разных отраслевых опросах уже в течение ряда лет. Такой вывод может удивить, поскольку принято считать, что специалист по данным тратит свое время на создание сложных моделей, помогающих получить новые знания. Но простая истина состоит в том, что, как бы ни был хорош ваш анализ, он не найдет полезных закономерностей в неправильных данных.

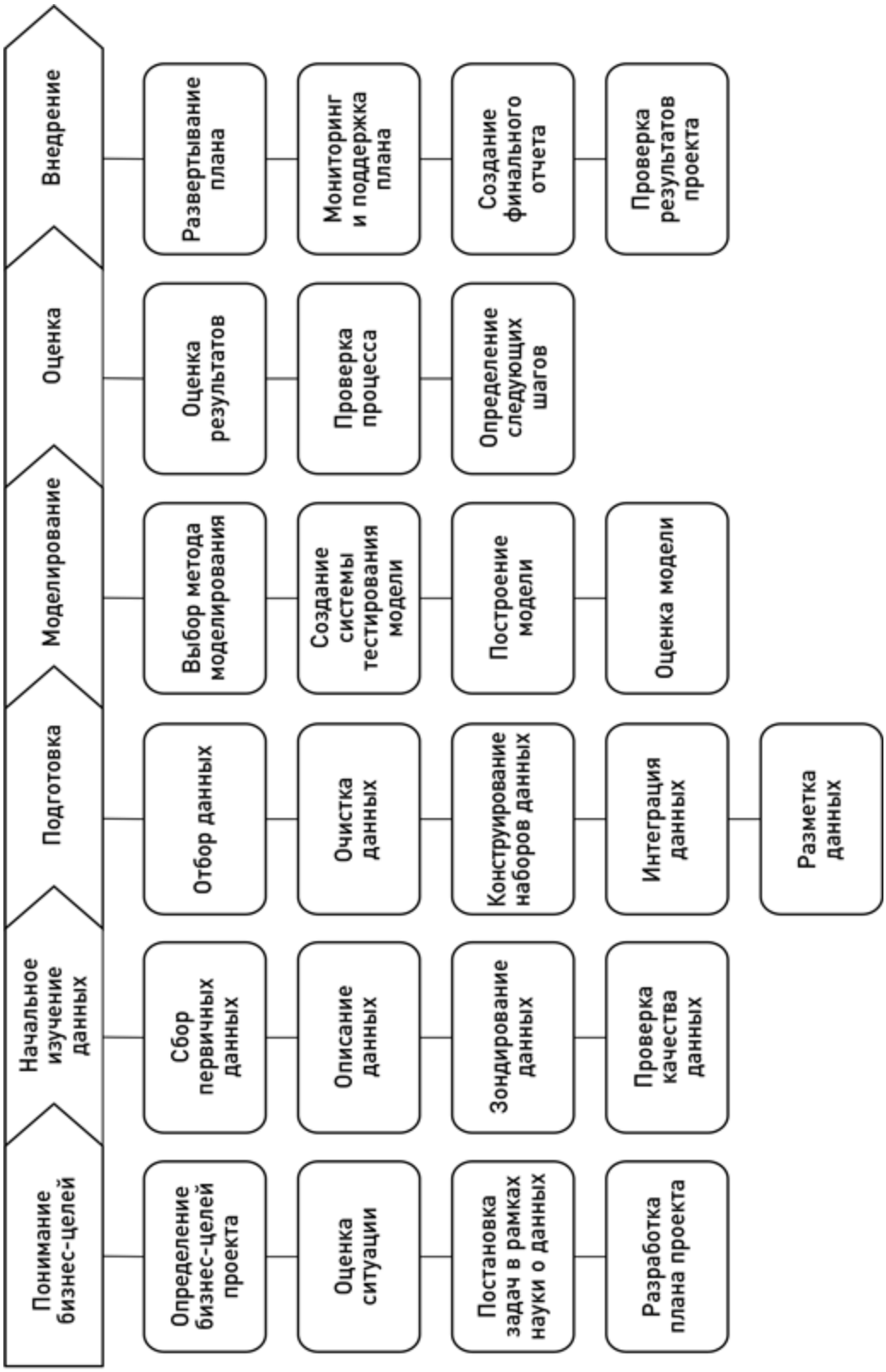

Рис. 5. Этапы и задачи CRISP-DM
(основано на рис. 3 из Chapman, Clinton, Kerber 1999 [8])

Источники

1. Han, Jiawei, Micheline Kamber, and Jian Pei. 2011. Data Mining: Concepts and Techniques, Third Edition. Haryana, India; Burlington, MA: Morgan Kaufmann.
2. Hall, Mark, Ian Witten, and Eibe Frank. 2011. *Data Mining: Practical Machine Learning Tools and Techniques.*
3. Korzybski, Alfred. 1996. “On Structure.” In *Science and Sanity: An Introduction Ot NonAristotelian Systems and General Semantics,* edited by Charlotte Schuchardt-Read, CDROM First Edition. European Society for General Semantics. http://esgs.free.fr/uk/art/sands.htm.
4. Kitchin, Rob. 2014. *The Data Revolution: Big Data, Open Data, Data Infrastructures and Their Consequences.* Sage.
5. Pomerantz, Jeffrey. 2015. *Metadata.* The MIT Press Essential Knowledge Series. https://mitpress.mit.edu/books/metadata-0.
6. Mayer, Jonathan, and Patrick Mutchler. 2014. “MetaPhone: The Sensitivity of Telephone Metadata.” *Web Policy.* http://webpolicy.org/2014/03/12/metaphone-the-sensitivity-oftelephone-metadata/.
7. Mayer, Jonathan, and Patrick Mutchler. 2014. “MetaPhone: The Sensitivity of Telephone Metadata.” *Web Policy.* http://webpolicy.org/2014/03/12/metaphone-the-sensitivity-oftelephone-metadata/.
8. Элиот Т. С. Полые люди. — СПб.: ООО «Издательский Дом «Кристалл»», 2000. (Б-ка мировой лит., Малая серия).
9. Chapman, Pete, Julian Clinton, Randy Kerber, Thomas Khabaza, Thomas Reinartz, Colin Shearer, and Rudiger Wirth. 1999. “CRISP-DM 1.0: Step-by-Step Data Mining Guide.” ftp://ftp.software.ibm.com/software/analytics/spss/support/Modeler/Documentation/14/UserManual/CRISP-DM.pdf.

10. Steinberg, Dan. 2013. “How Much Time Needs to Be Spent Preparing Data for Analysis?” http://info.salford-systems.com/blog/bid/299181/How-Much-Time-Needs-to-be-SpentPreparing-Data-for-Analysis.
11. CrowdFlower. 2016. «Отчет о науке данных за 2016 год». http://visit.crowdflower.com/rs/416-ZBE142/images/CrowdFlower_DataScienceReport_2016.pdf.

Глава 3

ЭКОСИСТЕМА НАУКИ О ДАННЫХ

Набор технологий, используемых для обработки данных, варьируется в зависимости от организации. Чем больше организация и/или объем обрабатываемых данных, тем сложнее технологическая экосистема науки о данных. Обычно эта экосистема содержит инструменты и узлы от нескольких поставщиков программного обеспечения, которые обрабатывают данные в разных форматах. Существует ряд подходов, которые организация может использовать для разработки собственной экосистемы науки о данных. На одном конце этого ряда организация принимает решение инвестировать в готовую систему интегрированных инструментов. На другом — самостоятельно создавать экосистему путем интеграции инструментов и языков с открытым исходным кодом. Между этими двумя крайностями есть несколько поставщиков программного обеспечения, которые предоставляют решения, являющие собой смесь коммерческих продуктов и продуктов с открытым исходным кодом. Однако, хотя конкретный набор инструментов в каждой организации будет свой, наука о данных предусматривает общие компоненты для большинства архитектур.

Рис. 6 дает обзор типичной архитектуры данных. Эта архитектура предназначена не только для больших данных, но и для данных любого размера. Диаграмма состоит из трех основных частей: уровня источников данных, на котором генерируются все

данные в организации; уровня хранения данных, на котором данные хранятся и обрабатываются, и уровня приложений, на котором данные передаются потребителям этих данных и информации, а также различным приложениям.

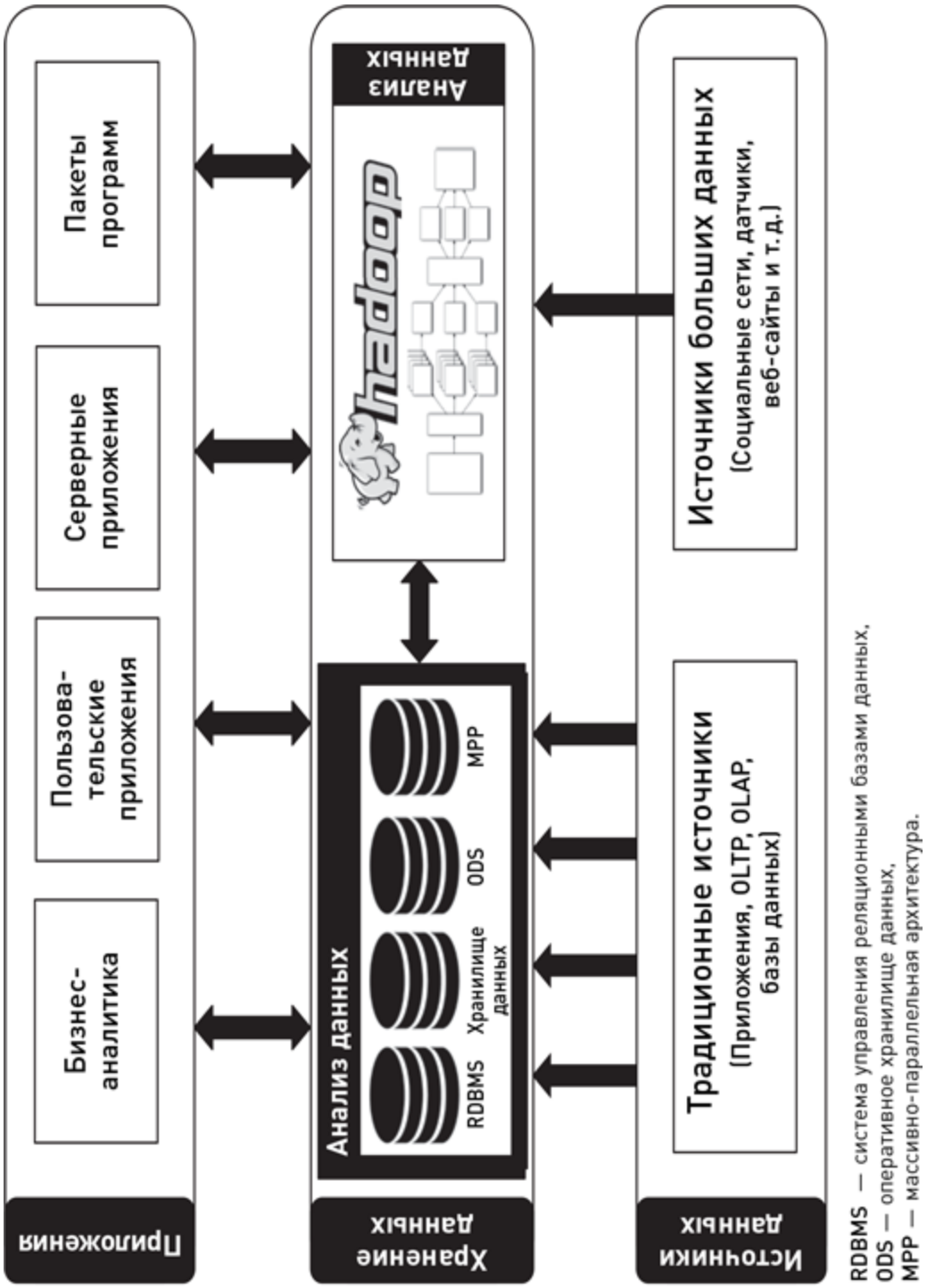

Рис. 6. Типичная архитектура малых и больших данных для проектов науки о данных (на основе рисунка из Hortonworks newsletter, April 23, 2013*)

* https://hortonworks.com/blog/hadoop-and-the-data-warehouse-when-to-use-which.

У каждой организации есть приложения, которые генерируют и собирают данные о клиентах, транзакциях и обо всем, что связано с работой организации (операционные данные). Это источники и приложения следующих типов: управление клиентами, заказы, производство, доставка, выставление счетов, банкинг, финансы, управление взаимодействием с клиентами (CRM), кол-центр, ERP и т.д. Приложения такого типа обычно называют системами обработки транзакций в реальном времени, или OLTP. Во многих проектах науки о данных именно эти приложения используются как источник входного набора данных для алгоритмов машинного обучения. Со временем объем данных, собираемых различными приложениями по всей организации, становится все больше и начинает включать данные, которые ранее были проигнорированы, недоступны или не были собраны по иным причинам. Эти новые источники, как правило, относятся уже к большим данным, поскольку их объем значительно выше, чем объем данных основных операционных приложений организации. Распространенные источники — это сетевой трафик, регистрационные данные различных приложений, данные датчиков, веб-журналов, социальных сетей, веб-сайтов и т.д. В традиционных источниках данные обычно хранятся в базе. Новые источники больших данных часто не предназначены для длительного их хранения, как в случае с потоковыми данными, поэтому форматы и структуры хранения будут варьироваться от приложения к приложению.

По мере увеличения количества источников возрастает проблема использования данных в аналитике и обмена ими между удаленными частями организации. Из истории (см. главу 1) нам известно, что хранение данных стало ключевым компонентом аналитики. В хранилище данные, поступающие со всей организации, интегрируются и становятся доступными для приложений (OLAP) и дальнейшего анализа.

Слой хранения данных, представленный на рис. 6, предназначен для обмена данными и их анализа. При этом он разделен на две части. Первая охватывает программное обеспечение для обмена данными, используемое большинством организаций. Наиболее популярным типом традиционного ПО для интеграции и хранения данных остаются реляционные базы данных (RDBMS). Это ПО часто служит основой для систем бизнес-аналитики (BI) в организациях. BI-системы призваны облегчить процесс принятия решений для бизнеса. Они предоставляют функции агрегирования, интеграции, отчетности и анализа. В основе BI-систем лежат базы данных, которые содержат интегрированные, очищенные, стандартизированные и структурированные данные, поступающие из различных источников. В зависимости от уровня зрелости архитектура BI-систем может состоять из очень разных компонентов — от базовой копии рабочего приложения и оперативного склада данных (ODS) до массивно-параллельных (MPP) решений баз данных BI и хранилищ данных. Аналитику, сгенерированную BI-системой, можно использовать в качестве входных данных для ряда потребителей на уровне приложений (рис. 6).

Вторая часть слоя хранения данных занимается управлением большими данными организации. Архитектура для их хранения и анализа включает платформу с открытым исходным кодом Hadoop, разработанную Apache Software Foundation для обработки больших данных. Эта платформа осуществляет распределенное хранение и обработку данных прямо в кластерах стандартных серверов. Для ускорения обработки запросов в наборах больших данных используется модель программирования MapReduce, которая реализует стратегию *разделения — использования — объединения*: а) большой набор данных разбивается на фрагменты, и каждый блок сохраняется в отдельном узле кластера; б) затем ко всем фрагментам применяется параллельный запрос; в) результат запроса

вычисляется путем объединения результатов, сгенерированных для разных фрагментов. Кроме того, в последние годы платформа Hadoop стала использоваться для расширения корпоративных хранилищ данных. Не так давно хранилища вмещали данные за три года, но теперь это число достигло 10 лет и продолжает расти. Поскольку объемы данных все увеличиваются, требования к хранилищу и обработке баз и сервера также растут. Это может повлечь за собой значительные затраты. В качестве альтернативы некоторые устаревшие данные перемещают из хранилища в кластер Hadoop. В хранилище, таким образом, остаются только последние данные, скажем за три года, которые часто используются и должны быть доступны для быстрого анализа и представления, а старые или редко используемые данные хранятся в Hadoop. Большинство баз данных уровня предприятия имеют соответствующие функции для прямого подключения хранилищ к Hadoop, позволяя специалисту запрашивать на языке SQL любые данные, как если бы они все находились в одной среде. Такой запрос открывает доступ и к хранилищу данных, и к Hadoop. Обработка запроса автоматически разделяет его на две отдельные части, каждая из которых выполняется независимо, а результаты объединяются и интегрируются, прежде чем будут представлены специалисту по данным.

Анализ данных затрагивает и ту и другую части слоя хранения, представленного на рис. 6. Он может выполняться как на основе данных, взятых непосредственно из BI-систем или Hadoop, так и на результатах их анализа, повторенного множество раз. Часто данные из традиционных источников бывают заметно чище и плотнее полученных из источников больших данных. Тем не менее гигантский объем и режим реального времени, свойственные большим данным, означают, что усилия, приложенные для подготовки и анализа их источников, могут окупиться с точки зрения важной информации, недоступной из традиционных источников. Разнообразные методы анализа

данных для тех или иных областей исследования (включая обработку естественного языка, компьютерное зрение и машинное обучение), используются для преобразования неструктурированных больших данных низкой плотности в ценные данные высокой плотности. Такие данные уже могут быть интегрированы с другими ценными данными из традиционных источников для дальнейшего анализа. Описанная структура, проиллюстрированная на рис. 3.1, представляет собой типичную архитектуру экосистемы науки о данных. Она подойдет для большинства организаций независимо от размера, однако по мере масштабирования организации увеличивается и сложность экосистемы науки о данных. Например, для небольших организаций может и не требоваться компонент Hadoop, но для крупных он становится незаменим.

Перемещение алгоритмов в данные

Традиционный подход к анализу данных включает их извлечение из различных баз, интеграцию, очистку, размещение, построение прогнозной модели, а затем загрузку окончательных результатов анализа в базу данных, чтобы их можно было использовать как часть рабочего процесса, отображать в виде отчетности и т.д. На рис. 7 показано, что большая часть процесса обработки данных, включающего их подготовку и анализ, протекает на отдельном сервере, вне баз и хранилища. При этом значительное количество времени может быть затрачено только на перемещение данных из базы и загрузку в нее результатов.

Рис. 7. Традиционный процесс построения моделей прогнозирования и оценки данных

Для примера: в эксперименте по построению модели линейной регрессии около 70–80% времени ушло на извлечение и подготовку данных. На создание моделей было потрачено лишь оставшееся время. В процессе скоринга данных примерно 90% времени было затрачено на их извлечение и сохранение готового набора обратно в базу. Только 10% времени пришлось на сам скоринг. Эти результаты основаны на наборах данных от 50 000 до 1,5 млн записей. Большинство поставщиков корпоративных баз данных уже осознали, сколько времени экономится, если отказаться от их перемещения, и отреагировали на это включением функции анализа и алгоритмов машинного обучения непосредственно в механизмы базы данных. В последующих разделах мы рассмотрим, каким образом алгоритмы машинного обучения были интегрированы в современные базы данных, как хранятся большие данные в Hadoop и как комбинации этих двух подходов позволяют легко работать со всеми данными в

организации, используя SQL как общий язык доступа, анализа, машинного обучения и прогнозной аналитики в режиме реального времени.

Значительное количество времени может быть затрачено только на перемещение данных из базы и загрузку в нее результатов.

Традиционные и современные базы данных

Поставщики баз данных постоянно инвестируют в развитие масштабируемости, производительности, безопасности и функциональности своих продуктов. Современные базы данных намного более продвинуты, чем традиционные реляционные базы данных. Они могут хранить и запрашивать данные в различных форматах. Кроме реляционных форматов, можно определять типы объектов, хранить документы, хранить и запрашивать объекты JSON, геоданные и т.д. Помимо этого, большинство современных баз имеют массу статистических функций, а некоторые поставляются с основными статистическими приложениями. Например, база данных Oracle поставляется с более чем 300 различными встроенными статистическими функциями. Они охватывают бóльшую часть статистического анализа, необходимого для проектов науки о данных, и включают практически все статистические функции, доступные в других инструментах и языках, таких как R. Изучение функционала баз данных организации может позволить аналитику действовать более эффективно и масштабируемо, используя язык SQL. Кроме того, большинство поставщиков баз данных (включая Oracle, Microsoft, IBM и EnterpriseDB) интегрировали в свои базы разнообразные алгоритмы машинного обучения, которые можно запускать на SQL. Использование алгоритмов машинного обучения, встроенных в ядро базы и доступных через SQL, известно как *машинное обучение в базе данных*. Такое машинное обучение способствует более быстрой разработке моделей и скорейшей интеграции результатов анализа с приложениями и панелями мониторинга. Кратко идея размещения алгоритмов машинного обучения непосредственно в базах данных может быть выражена следующим образом: *«Переместить алгоритмы в данные, вместо того чтобы перемещать данные в алгоритмы»*.

Использование алгоритмов машинного обучения в базе данных имеет следующие преимущества:

- **Отсутствие движения данных.** Некоторые продукты для обработки данных требуют их экспорта из базы и конвертации в особый формат, чтобы поместить в алгоритм машинного обучения. Благодаря машинному обучению в базе данных перемещение или преобразование данных не требуется. Это упрощает весь процесс, делает его менее трудоемким и подверженным ошибкам.
- **Скорость.** Для аналитических операций, выполняемых в базе данных без их перемещения, можно использовать вычислительные возможности сервера самой базы, обеспечивая увеличение производительности до 100 раз по сравнению с традиционным подходом. Большинство серверов баз данных имеют высокие спецификации, множество процессоров и эффективное управление памятью для обработки наборов данных, содержащих более миллиарда записей.
- **Высокая безопасность.** База данных обеспечивает контролируемый и поддающийся проверке доступ к данным, увеличивая производительность специалиста и поддерживая нормы безопасности. Кроме того, машинное обучение в базе данных позволяет избежать рисков, присущих процессу извлечения и загрузки данных на альтернативные серверы. К тому же традиционный процесс обработки данных приводит к созданию множества копий (а иногда и разных версий) наборов данных в разных хранилищах организации.
- **Масштабируемость.** База данных может легко масштабировать аналитику по мере увеличения объема данных благодаря алгоритмам машинного обучения.

Программное обеспечение баз предназначено для эффективного управления большими объемами данных с использованием нескольких серверных процессоров и памяти, что позволяет выполнять алгоритмы машинного обучения параллельно другим задачам. Базы данных также очень эффективны при обработке больших наборов данных, которые не помещаются в память. Сорокалетняя история развития баз гарантирует, что наборы данных будут обработаны быстро.

- **Режим реального времени.** Модели, разработанные с использованием алгоритмов машинного обучения в базе данных, могут быть немедленно развернуты и использованы в средах реального времени. Это позволяет интегрировать модели в привычные приложения и предоставлять прогнозы конечным пользователям и клиентам.
- **Развертывание в среде эксплуатации.** SQL — это язык базы данных, который может быть использован для доступа к алгоритмам и моделям машинного обучения в базах. Модели, разработанные с использованием автономного ПО для машинного обучения, возможно, придется перекодировать на другие языки программирования, прежде чем они смогут быть развернуты в корпоративных приложениях. Но это не относится к машинному обучению в базе данных. SQL можно использовать и вызывать любым языком программирования и инструментом науки о данных. Это значительно упрощает задачу включения модели из базы данных в производственные приложения.

Многие организации используют преимущества машинного обучения в базе данных. Среди них встречаются как небольшие

компании, так и крупные. Вот примеры организаций, использующих эту технологию:

- Fiserv — американский поставщик финансовых услуг, который занимается выявлением и анализом мошенничества. Он перешел от работы с несколькими поставщиками технологий хранения данных и машинного обучения к использованию машинного обучения в своей базе данных. В частности, эта технология позволяет сократить время создания/обновления и развертывания модели обнаружения мошенничества с почти недели до нескольких часов.
- Компания 84.51° (формально Dunnhumby, USA) использовала множество различных аналитических решений при создании моделей для своих клиентов. Обычно каждый месяц более 318 часов уходило на перемещение данных из базы на сервера машинного обучения и обратно. При этом на создание моделей тратилось еще как минимум 67 часов. Компания внедрила алгоритмы машинного обучения непосредственно в базу данных. Как только данные перестали покидать базу, экономия времени сразу составила более 318 часов. Поскольку база данных использовалась в качестве вычислительного инструмента, специалисты смогли масштабировать аналитику и время создания или обновления моделей машинного обучения сократилось с 67+ часов до 1 часа. Это дало экономию 16 дней. Теперь они могут получать результаты значительно быстрее и начинать взаимодействие с клиентами намного раньше, вскоре после совершения ими покупки.
- Wargaming — создатели World of Tanks и многих других игр — использует машинное обучение в базе данных, чтобы

моделировать и прогнозировать взаимодействие с более чем 120 млн своих клиентов.

Данные в мире Hadoop

Хотя современная база данных невероятно эффективна для обработки транзакций, в эпоху больших данных для управления разнообразными формами данных и их долгосрочного хранения требуется новая инфраструктура. Современная база данных может справляться с объемами до нескольких петабайт, но при таком масштабе традиционные решения для баз могут стать чрезмерно дорогими. Этот вопрос стоимости обычно упирается в вертикальное масштабирование. В традиционной парадигме чем больше данных должна хранить и обрабатывать организация в течение необходимого срока, тем больший ей требуется сервер, а это увеличивает стоимость его конфигурации и лицензирования баз данных. Традиционная технология позволяет запрашивать и принимать миллиард записей ежедневно, но такой масштаб обработки обойдется в несколько миллионов долларов.

Hadoop — это платформа с открытым исходным кодом, которая была разработана и выпущена Apache Software Foundation. Она хорошо зарекомендовала себя для эффективного приема и хранения больших объемов данных и обходится дешевле, чем традиционный подход. Кроме того, на рынке появился широкий ассортимент продуктов для обработки и анализа данных на платформе Hadoop. Приведенное выше высказывание, касающееся современных баз данных — «переместить алгоритмы в данные, вместо того чтобы перемещать данные в алгоритмы», — также применимо и к Hadoop.

В Hadoop данные делятся на разделы, которые распределяются по узлам кластера. В процессе работы с Hadoop различные аналитические инструменты обрабатывают данные в каждом из

кластеров (часть этих данных может постоянно находиться в оперативной памяти), что обеспечивает быструю обработку данных, поскольку несколько кластеров анализируются одновременно. Ни извлечение данных, ни ETL-процесс не требуются. Данные анализируются там, где они хранятся. Существуют и другие примеры аналогичного подхода, скажем решения от Google и Amazon, где аналитическое программное обеспечение, такое как Spark, разворачивается на распределенных вычислительных архитектурах, позволяя анализировать данные там, где они находятся.

В мире больших данных специалист может запрашивать их массивные наборы с использованием аналитических языков, таких как Spark, Flink, Storm, и широкого спектра инструментов, а также постоянно растущего числа бесплатных и коммерческих продуктов. Эти продукты представляют собой инструменты высокоуровневой аналитики или панели мониторинга, которые упрощают работу специалиста с данными и аналитикой, что позволяет ему сконцентрироваться на анализе данных. Однако современному специалисту по данным приходится анализировать их в двух разных местах: в современных базах данных и в хранилищах больших данных на Hadoop. В следующей части мы рассмотрим, как решается эта проблема.

Мир гибридных баз данных

Если у организации нет данных такого размера и масштаба, которым требуется Hadoop, то для управления данными ей будет достаточно традиционной базы данных. Однако есть мнение, что инструменты хранения и обработки данных, доступные в мире Hadoop, в итоге вытеснят традиционные базы данных. Такое сложно себе представить, и потому в последнее время обсуждается более сбалансированный подход к управлению

данными в так называемом мире гибридных баз, где традиционные базы данных сосуществуют с Hadoop.

Гибридная база данных сама определяет местоположение данных на основе частоты запросов и типа проводимого анализа.

В мире гибридных баз все данные связаны между собой и работают вместе, что позволяет эффективно обмениваться ими, обрабатывать и анализировать их. На рис. 8 показано традиционное хранилище данных, но при этом большая часть данных находится не в базе или хранилище, а перемещена в Hadoop. Между базой данных и Hadoop создается соединение, которое позволяет специалисту запрашивать данные, как если бы они находились в одном месте. Ему не потребуется запрашивать отдельно данные из базы и из Hadoop. Гибридная база автоматически определит, какие части запроса необходимо выполнить в каждом из местоположений, затем объединит результаты и представит их специалисту. Точно так же по мере роста хранилища часть данных устаревает, и гибридное решение автоматически перемещает редко используемые данные в среду Hadoop, а те, что становятся востребованными, наоборот, возвращает обратно. Гибридная база данных сама определяет местоположение данных на основе частоты запросов и типа проводимого анализа.

Одним из преимуществ гибридных решений является то, что специалист по-прежнему запрашивает данные на SQL. Ему не нужно изучать другой язык запросов или применять особые инструменты. Сегодняшние тенденции позволяют предположить, что в ближайшем будущем основные поставщики баз данных, облачных хранилищ и программного обеспечения для интеграции данных будут предлагать именно гибридные решения.

Рис. 8. Совместная работа баз данных, хранилища данных и Hadoop (на основе рисунка из официальных документов к платформе данных Gluent, 2017*).

* https://gluent.com/wp-content/uploads/2017/09/Gluent-Overview.pdf.

Подготовка и интеграция данных

Интеграция данных включает в себя их получение из разных источников и последующее объединение с целью получения единого представления данных по всей организации. Разберем это на примере медицинской карты. В идеале у каждого человека должна быть одна медицинская карта, чтобы каждая больница, поликлиника и врач могли использовать один и тот же идентификатор пациента, единицы измерения, систему оценок и

т.д. К сожалению, почти в каждой больнице имеется собственная независимая система учета пациентов и то же справедливо в отношении внутрибольничных медицинских лабораторий. Представьте себе, как трудно бывает найти историю болезни и назначить правильное лечение пациенту. Такие проблемы возникают в рамках одной больницы. Когда же несколько больниц обмениваются данными пациентов, проблемы их интеграции становятся еще существеннее. Именно поэтому первые три этапа CRISP-DM занимают до 70–80% общего времени проекта, причем бóльшая часть этого времени уходит на интеграцию данных.

Интеграция данных из нескольких источников — непростая задача, даже когда данные структурированы. Если же задействованы современные источники больших данных, в которых частично или вовсе неструктурированные данные являются нормой, то стоимость интеграции и управления архитектурой может значительно увеличиваться. Наглядный пример проблем интеграции — данные клиентов. Они могут находиться в различных приложениях и соответствующих им базах данных. Каждое приложение при этом будет содержать данные о клиентах, немного отличающиеся от тех же данных в других приложениях. Например, внутренние источники данных могут содержать кредитный рейтинг клиента, продажи, платежи, контактную информацию кол-центра и т.д. Внешние источники могут содержать дополнительную информацию о клиентах. В таком контексте создание единого представления клиента требует извлечения и интеграции данных из всех этих источников.

Типичный процесс интеграции данных включает в себя несколько этапов, а именно: извлечение, очистку, стандартизацию, преобразование и, наконец, собственно интеграцию для создания унифицированной версии данных. Извлечение данных из нескольких источников может

осложняться тем, что доступ к ним возможен только через определенный интерфейс или API. Следовательно, специалисту понадобится широкий набор навыков для взаимодействия с каждым из источников данных.

Как только данные извлечены, необходимо проверить их качество. Очистка данных — это процесс, который обнаруживает, очищает или удаляет поврежденные или неточные данные. Например, может потребоваться очистка информации с адресами клиентов, чтобы преобразовать ее в стандартный формат. Кроме того, данные в источниках могут дублироваться. В этом случае необходимо определить запись клиента, подходящую для использования, и удалить все остальные из наборов данных. Важно обеспечить согласованность значений. Например, одно исходное приложение может использовать числовые значения для представления кредитного рейтинга клиента, а другое — иметь комбинацию числовых и символьных значений. В таком сценарии необходимо принять решение о том, какие значения использовать, и привести их к единому стандарту. Представьте, что одним из атрибутов в наборе данных является размер обуви клиента. При этом клиенты покупают обувь из разных регионов мира. Но система нумерации, используемая для описания размеров обуви в Европе, США, Великобритании и других странах, немного различается. Перед этапом анализа данных и моделирования эти значения должны быть стандартизированы.

Преобразование данных включает в себя их изменение или объединение. На этом этапе используются самые разные методы, включая сглаживание данных, объединение, нормализацию и написание пользовательского кода для выполнения конкретного преобразования. Типичным примером преобразования данных является обработка возраста клиента. Во многих задачах науки о данных не требуется знать точный возраст клиентов. Разница между покупателями 42 и 43 лет, как правило, незначительна, в то время как разница в возрасте от 42 до 52 лет уже становится

информативной. Поэтому возраст покупателя часто преобразуется из конкретного значения в диапазон. Процесс преобразования возрастов в диапазоны является примером одного из методов преобразования данных, называемого биннингом. Хотя биннинг относительно прост с технической точки зрения, сложность состоит в том, чтобы определить наиболее подходящие пороговые значения диапазона, которые следует применять.

Интеграция данных из нескольких источников — непростая задача, даже когда данные структурированы.

Последний этап интеграции включает создание выходных данных для алгоритмов анализа, используемых в проекте. Версия данных, которая подается в алгоритм на входе, называется базовой аналитической таблицей.

Создание базовой аналитической таблицы

Первым шагом в создании базовой аналитической таблицы является выбор атрибутов, которые будут включены в анализ. Выбор должен быть основан на знании предметной области и анализе связей между атрибутами. В качестве конкретного примера рассмотрим сценарий анализа, ориентированного на клиентов сервиса. В этом сценарии необходимо создать список часто употребляемых понятий, который будет использован при разработке и выборе атрибутов: детали клиентского контракта, демография, привычки, изменения в привычках, особые привычки, фаза жизненного цикла, сетевые ссылки и т.д. Если будет обнаружена высокая корреляция между двумя атрибутами, вероятнее всего, один из них должен быть исключен. Набор выбранных атрибутов создает так называемую аналитическую запись. Обычно она включает как необработанные, так и производные атрибуты. Каждый объект в базовой аналитической таблице представлен одной записью, поэтому именно набор атрибутов, включенных в нее, определяет отображение анализируемых объектов.

После того как форма аналитической записи разработана, необходимо извлечь и объединить эти записи в набор данных для анализа. Когда записи созданы и сохранены, например, в базе данных, мы получаем то, что и называют базовой аналитической таблицей — набор данных, которые используются в качестве

входных для алгоритмов машинного обучения. Следующая глава познакомит вас с областью машинного обучения и некоторыми из самых распространенных алгоритмов, используемых в науке о данных.

Глава 4

ОСНОВЫ МАШИННОГО ОБУЧЕНИЯ

Наука о данных — это партнерство между специалистом по данным и компьютером. В главе 2 мы описали жизненный цикл процесса CRISP-DM, которому следует специалист по данным. CRISP-DM определяет последовательность принимаемых им решений и действия, которые помогут их воплотить. Основные задачи специалиста по данным в цикле CRISP-DM сводятся к тому, чтобы определить проблему, спроектировать набор данных, подготовить их, принять решение о том, какой тип анализа будет использован, а затем оценить и интерпретировать результаты. Вклад компьютера в этом партнерстве заключается в его способности обрабатывать данные и искать закономерности. Машинное обучение — это область исследований, которая разрабатывает алгоритмы для выявления компьютером закономерностей в данных. Алгоритмы и методы машинного обучения в основном применяются на этапе моделирования в CRISP-DM. Процесс машинного обучения представляет собой два последовательных этапа.

На первом алгоритм машинного обучения применяется к набору данных для выявления в нем закономерностей. Сами закономерности могут быть представлены разными способами. Позже в этой главе мы опишем наиболее популярные из них: деревья решений, регрессионные модели и нейронные сети. Эти представления закономерностей известны как модели, поэтому и сам этап жизненного цикла CRISP-DM называется этапом

моделирования. Проще говоря, все алгоритмы машинного обучения создают модели из данных, но каждый из них разработан для создания моделей, использующих определенный тип представления.

На втором этапе, когда модель создана, она применяется для анализа. В ряде случаев решающее значение имеет структура модели, которая показывает, какие именно атрибуты являются важными для конкретной области определения. Например, мы могли бы применить алгоритм машинного обучения к набору данных пациентов, уже перенесших инсульт, а затем использовать такую структуру модели, которая распознавала бы факторы, тесно связанные с инсультом. Существуют модели для маркировки или классификации новых объектов. К примеру, основная цель модели спам-фильтра состоит в том, чтобы маркировать входящие электронные письма, а не выявлять атрибуты спам-сообщений.

Обучение с учителем и без

Большинство алгоритмов машинного обучения можно отнести либо к обучению с учителем, либо к обучению без учителя. Цель обучения с учителем состоит в том, чтобы научить алгоритм сопоставлять разные значения разных атрибутов объекта со значением заданного атрибута этого же объекта, известного как целевой атрибут. Например, когда обучение с учителем применяется для спам-фильтра, алгоритм пытается изучить функцию, которая сопоставляет атрибуты, описывающие электронную почту, со значением (спам / не спам) целевого атрибута; функция, которую изучает алгоритм, является моделью спам-фильтра. В этом контексте искомая алгоритмом закономерность является функцией, которая сопоставляет

значения входных атрибутов со значением целевого атрибута, а модель, которую возвращает алгоритм, является компьютерной программой, выполняющей эту функцию. По сути, обучение с учителем осуществляется путем поиска одной из множества функций, которая наилучшим образом отображает связь между входными и выходными данными. Однако для любого набора данных разумной сложности существует так много комбинаций входных данных и их возможных сопоставлений с выходными данными, что алгоритм не может испробовать их все. Поэтому каждый алгоритм машинного обучения предпочитает определенные типы функций во время поиска. Эти предпочтения известны как смещение обучения алгоритма. Реальная проблема в использовании машинного обучения состоит в том, чтобы найти алгоритм, смещение обучения которого лучше всего подходит для конкретного набора данных. Как правило, для того, чтобы выяснить, какой из алгоритмов лучше всего работает с конкретным набором данных, требуются эксперименты.

Реальная проблема в использовании машинного обучения состоит в том, чтобы найти алгоритм, смещение обучения которого лучше всего подходит для конкретного набора данных.

Обучение с учителем называется именно так, потому что каждый объект в наборе данных содержит как входные значения, так и выходное (целевое) значение. Таким образом, алгоритм обучения может направлять свой поиск наилучшей функции, проверяя соответствие каждой пробуемой функции набору данных, и в то же время сам набор данных выступает в качестве контролера процесса обучения или учителя, обеспечивая обратную связь. Очевидно, что для обучения с учителем каждый объект в наборе данных должен быть промаркирован значением целевого атрибута. Однако зачастую целевой атрибут бывает сложно измерить в необработанном виде, а значит, и создать набор данных с маркированными объектами. При подобном сценарии много времени и усилий тратится, чтобы создать набор данных с целевыми значениями атрибутов, прежде чем модель можно будет обучать.

При обучении без учителя целевой атрибут отсутствует. Следовательно, алгоритмы обучения без учителя не требуют времени и усилий на маркировку целевым атрибутом объектов в наборе данных. Однако отсутствие целевого атрибута означает и то, что обучение становится более сложным: вместо конкретной задачи поиска соответствующего отображения между входным и выходным значениями, перед алгоритмом ставится более общая задача поиска закономерностей в данных. Самым распространенным типом обучения без учителя является кластерный анализ, когда алгоритм ищет кластеры объектов, схожих друг с другом. Часто эти алгоритмы кластеризации начинают со случайной группы кластеров, а затем итеративно обновляют кластеры (перебрасывая объекты из одного кластера в другой) таким образом, чтобы увеличить подобие внутри каждого кластера и разницу между ними.

Задача кластеризации — выяснить, как измерить подобие. Если все атрибуты в наборе данных являются числовыми и имеют

одинаковые диапазоны, то, вероятно, имеет смысл просто рассчитать евклидово расстояние (или расстояние по прямой) между рядами. Объекты, которые находятся близко друг к другу в евклидовом пространстве, рассматриваются как подобные. Однако существует ряд факторов, которые могут усложнить обнаружение сходства между объектами. В некоторых наборах данных разные числовые атрибуты имеют разные диапазоны, в результате чего разброс значений в одном атрибуте может быть не таким значительным, как в другом. В таких случаях атрибуты должны быть нормализованы путем присвоения им одинакового диапазона. Еще одним усложняющим фактором при расчете сходства является то, что подобие объектов можно определять по-разному. Порой одни атрибуты являются более важными, чем другие, поэтому имеет смысл при расчетах задавать весовой параметр некоторым атрибутам, что бывает необходимо и тогда, когда набор данных содержит нечисловые значения. Эти более сложные сценарии могут потребовать разработки индивидуальных параметров подобия для использования алгоритмом кластеризации.

Чтобы проиллюстрировать обучение без учителя на конкретном примере, представим, что мы проводим анализ причин развития диабета 2-го типа среди взрослых белых американцев мужского пола. Мы начнем с построения набора данных, в котором каждая строка будет представлять одного человека, а столбцы — атрибуты, которые, по нашему мнению, имеют отношение к исследованию. Для этого примера мы возьмем следующие атрибуты: рост человека в метрах, его вес в килограммах, продолжительность тренировок в течение недели в минутах, размер обуви и вероятность развития у него диабета, полученную на основе клинических тестов и изучения образа жизни, выраженную в процентах. Таблица 2 иллюстрирует фрагмент этого набора данных. Очевидно, что есть и другие атрибуты, которые могут быть включены в набор, например

возраст человека, и что среди выбранных атрибутов есть лишние, например размер обуви, который не коррелирует с развитием сахарного диабета. Как мы обсуждали в главе 2, выбор атрибутов для набора данных — ключевая задача науки о данных, но в этом примере мы намеренно будем работать с таким набором данных, какой у нас есть.

Таблица 2. Набор данных для исследования диабета

Объект	Рост (м)	Вес (кг)	Размер обуви	Продолжительность тренировок в неделю (мин.)	Вероятность развития диабета (%)
1	1,70	70	5	130	0,05
2	1,77	88	9	80	0,11
3	1,85	112	11	0	0,18

При обучении без учителя алгоритм кластеризации будет искать группы строк, которые более похожи друг на друга, чем на другие строки. Каждая из этих групп определяет кластер подобных объектов. С точки зрения изучения причин развития диабета выявление кластеров схожих пациентов (объектов) может помочь выявить причины заболевания или сопутствующих диабету заболеваний путем поиска значений атрибутов, которые относительно часто встречаются в кластере. Простая идея поиска кластеров подобных объектов служит мощным инструментом и применима ко многим областям жизни. Другой пример кластеризации строк — предоставление рекомендаций для клиентов. Если клиенту понравилась книга, песня или фильм, он с высокой вероятностью получит удовольствие от другой книги, песни или фильма из того же кластера.

Обучение моделей прогнозирования

Прогнозирование — это задача оценки значения целевого атрибута конкретного объекта на основе значений других его атрибутов. Проблему прогнозирования решают алгоритмы машинного обучения с учителем, которые генерируют модели прогнозирования. Пример спам-фильтра, который мы использовали для иллюстрации обучения с учителем, подойдет и здесь: мы используем обучение с учителем при создании модели спам-фильтра, которая является моделью прогнозирования. Типичным случаем использования модели прогнозирования является оценка целевого атрибута для новых объектов, которых нет в наборе обучающих данных. Продолжая пример со спамом, мы обучаем спам-фильтр (модель прогнозирования) на наборе данных старых писем, а затем используем эту модель, чтобы предсказать, являются ли новые письма спамом или нет. Проблемы прогнозирования, возможно, самый популярный тип проблем, для которых используется машинное обучение, поэтому оставшаяся часть этой главы будет посвящена прогнозированию в качестве примера для введения в машинного обучения. Мы начнем наше знакомство с моделями прогнозирования с фундаментальной прогностической концепции, известной как корреляционный анализ. Затем мы покажем, как алгоритмы машинного обучения с учителем работают над созданием различных типов популярных моделей прогнозирования, в том числе моделей линейной регрессии, моделей нейронных сетей и деревьев решений.

Корреляции — это не причинно-следственные связи, но некоторые из них бывают полезны[1][2]

Корреляция описывает силу взаимосвязи между двумя атрибутами. В общем смысле корреляция может описывать любой тип связи. Термин «корреляция» также имеет конкретное значение в статистике, где он часто используется как

сокращенный вариант «коэффициент корреляции Пирсона». Коэффициент корреляции Пирсона измеряет силу линейных зависимостей между двумя числовыми атрибутами и находится в диапазоне значений от –1 до +1. Для его обозначения используется буква *r*, также называемая коэффициентом корреляции между двумя атрибутами. Коэффициент $r = 0$ указывает, что два атрибута независимы друг от друга. Коэффициент $r = +1$ указывает, что два атрибута имеют идеальную положительную корреляцию, означающую, что любое изменение одного из них сопровождается эквивалентным изменением другого в том же направлении. Коэффициент $r = –1$ указывает, что два атрибута имеют идеальную отрицательную корреляцию, при которой каждое изменение в одном из них сопровождается противоположным изменением в другом. Общие рекомендации по интерпретации коэффициентов корреляции Пирсона состоят в том, что значение $r \approx \pm 0{,}7$ указывает на сильную линейную зависимость между атрибутами, $r \approx \pm 0{,}5$ — на умеренную линейную зависимость, $r \approx \pm 0{,}3$ — на слабую зависимость, а $r \approx 0$ — на отсутствие зависимости между атрибутами.

Но вернемся к исследованию диабета. Исходя из наших знаний о физиологии людей, мы ожидаем, что между некоторыми признаками в табл. 4.1 будут взаимосвязи. Например, обычно чем выше человек, тем больше размер его обуви. Мы можем ожидать, что чем больше кто-то тренируется, тем меньше в нем будет избыточного веса, с учетом того, что более высокий человек, вероятно, будет тяжелее более низкого, который тратит столько же времени на физические упражнения. Мы также ожидаем, что не обнаружим очевидной связи между размером обуви и временем тренировок. На рис. 9 представлены три диаграммы рассеяния, которые иллюстрируют, как эти интуитивные ожидания отражаются в данных. Диаграмма рассеяния вверху показывает, как распределяются данные, если они построены в

зависимости от размера обуви и роста. На этой диаграмме рассеяния наблюдается четкая закономерность, идущая из нижнего левого угла в верхний правый, указывающий на то, что по мере того, как люди становятся выше (движение вправо по оси *y*), размер их обуви тоже увеличивается (движение вверх по оси *x*). Подобная закономерность данных в диаграмме рассеяния указывает на положительную корреляцию между двумя атрибутами. Если мы вычислим коэффициент корреляции Пирсона между размером обуви и ростом, то *r* составит 0,898, т.е. мы имеем сильную положительную корреляцию между этой парой атрибутов. Средняя диаграмма рассеяния показывает, как данные распределяются, когда мы строим график корреляции веса и физических упражнений. Здесь общая схема имеет противоположное направление от левого верхнего угла до нижнего правого, что указывает на отрицательную корреляцию — чем больше люди тренируются, тем меньше их вес. Коэффициент корреляции Пирсона для этой пары признаков равен $r = –0{,}710$, что указывает на сильную отрицательную корреляцию. На последнем графике рассеяния отображается корреляция времени тренировок и размера обуви. Мы видим, что данные распределены на этом графике случайным образом и коэффициент корреляции Пирсона для этой пары атрибутов $r = –0{,}272$, иначе говоря, корреляция отсутствует.

Рис. 9. Диаграммы рассеяния размера обуви и роста, веса и физических упражнений, размера обуви и физических упражнений

Может показаться, что применение статистического коэффициента корреляции Пирсона к анализу данных ограничено только парами атрибутов. К счастью, мы можем обойти эту проблему, применяя функции для групп атрибутов. В главе 2 мы ввели индекс массы тела (ИМТ) — отношение веса человека (в килограммах) к квадрату его роста (в квадратных метрах). ИМТ был изобретен в XIX в. бельгийским математиком Адольфом Кетле для того, чтобы задать значения для каждой из следующих категорий: люди с недостаточным весом, с нормальным, с избыточным или страдающие ожирением. Мы знаем, что вес и рост имеют положительную корреляцию (как правило, кто выше, тот и тяжелее), поэтому, поделив вес на рост, мы можем отслеживать зависимость первого от второго. Есть два аспекта ИМТ, которые представляют интерес для нашего обсуждения корреляции между несколькими атрибутами. Во-первых, ИМТ — это функция, которая принимает ряд атрибутов в качестве входных данных и сопоставляет их с новым значением. По сути, такое отображение создает новый производный атрибут (в отличие от необработанного атрибута) в данных. Во-вторых, поскольку ИМТ человека представляет собой числовое значение, мы можем рассчитать корреляцию между ним и другими атрибутами.

В нашем тематическом исследовании причин развития диабета 2-го типа у белых взрослых американцев мужского пола нам требуется определить, имеет ли какой-нибудь из признаков сильную корреляцию с целевым атрибутом, описывающим вероятность развития диабета у человека. На рис. 10 представлены три диаграммы рассеяния, каждая из которых показывает отношения между целевым атрибутом диабета и одним из следующих признаков (слева направо): ростом, весом и ИМТ. Если посмотреть на диаграмму рассеяния роста и диабета, то в данных не наблюдается какой-либо определенной

закономерности, что свидетельствует об отсутствии реальной корреляции между этими двумя атрибутами ($r = –0{,}277$). Средняя диаграмма рассеяния показывает распределение данных для веса и диабета и указывает на положительную корреляцию между людьми с бóльшей массой тела и вероятностью развития заболевания ($r = 0{,}655$). Нижняя диаграмма рассеяния показывает набор данных, построенный с использованием ИМТ и диабета. Она напоминает среднюю диаграмму, данные так же распределяются снизу слева направо вверх, что указывает на положительную корреляцию. Однако в этой последней диаграмме объекты более тесно связаны, а это означает, что корреляция между ИМТ и диабетом сильнее, чем между диабетом и массой тела. Коэффициент корреляции Пирсона для диабета и ИМТ составляет $r = 0{,}877$.

Пример ИМТ иллюстрирует, что можно создать новый производный атрибут, задав функцию, которая принимает несколько атрибутов в качестве входных данных. Таким же путем можно вычислить корреляцию Пирсона между этим производным атрибутом и другим атрибутом в наборе данных. Производный атрибут может иметь более высокую корреляцию с целевым атрибутом, чем любой из отдельно взятых атрибутов, используемых для его генерации. Для лучшего понимания: ИМТ имеет более высокую корреляцию с признаком диабета, чем рост или вес, потому что вероятность развития диабета зависит от взаимосвязи роста и веса, а атрибут ИМТ моделирует именно эту взаимосвязь. Вот почему врачи интересуются ИМТ людей, это дает им больше информации о вероятности развития диабета 2-го типа, чем рост или вес человека по отдельности.

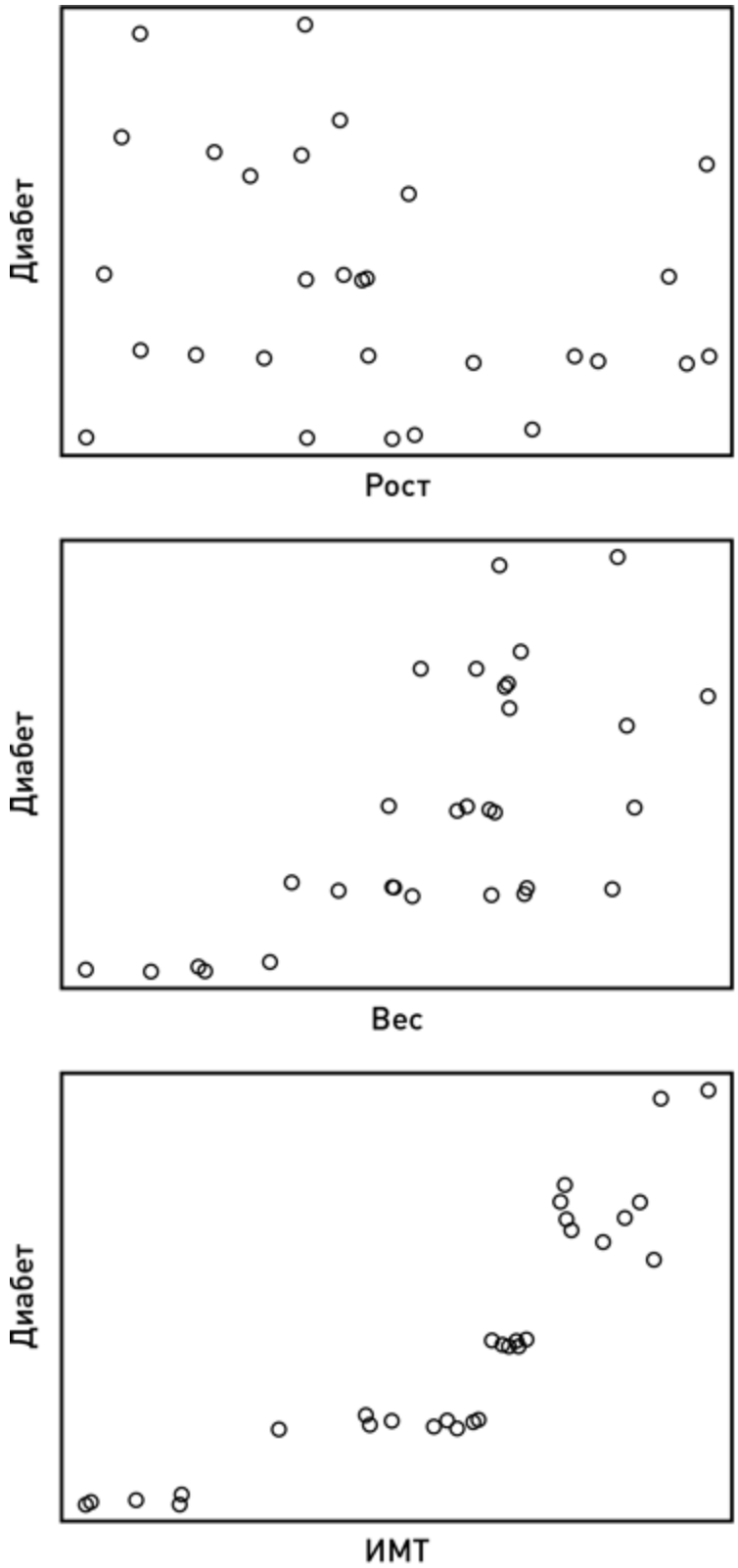

Рис. 10. Диаграммы рассеяния вероятности диабета в зависимости от роста, веса и ИМТ

Мы уже отмечали, что выбор атрибутов — ключевая задача в науке о данных. То же касается и моделирования атрибутов. Часто моделирование производного атрибута, который имеет сильную корреляцию с целевым, — это уже полдела в науке о данных. Когда вы знаете правильные атрибуты для представления данных, вы можете создавать модели точно и быстро. Выбор и моделирование правильных производных атрибутов является непростой задачей. ИМТ был разработан в XIX в., однако сейчас алгоритмы машинного обучения способны изучать взаимодействия между входными атрибутами и создавать полезные производные атрибуты, просматривая различные их комбинации, проверяя корреляцию между ними и целевым атрибутом. Вот почему машинное обучение полезно в тех случаях, когда существует множество атрибутов, имеющих слабо выраженную взаимосвязь с процессом, который мы пытаемся понять.

Выявление атрибута (необработанного или производного), который имеет высокую корреляцию с целевым атрибутом, полезно, поскольку коррелированный атрибут может дать нам понимание процесса, представленного целевым атрибутом. В нашем случае факт сильной корреляции ИМТ с вероятностью развития диабета указывает на то, что не вес сам по себе способствует заболеванию, а его избыточность. Кроме того, если наблюдается сильная корреляция входного атрибута с целевым, скорее всего, будет нелишним ввести его в модель прогнозирования. Подобно корреляционному анализу, прогнозирование включает в себя анализ отношений между атрибутами. Чтобы иметь возможность сопоставлять значения набора с целевым атрибутом, должна существовать корреляция между ним и входными атрибутами (или некоторой производной функцией от них). Если этой корреляции не существует (или она не найдена алгоритмом), то входные атрибуты не имеют

значения при прогнозировании, и лучшее, что может сделать модель, — игнорировать входные данные и всегда прогнозировать центральную тенденцию этой цели[13] в наборе данных. И наоборот, если между входными атрибутами и целью существует сильная корреляция, то весьма вероятно, что алгоритм машинного обучения сможет сгенерировать точную модель прогнозирования.

Линейная регрессия

Когда набор данных состоит из числовых атрибутов, часто используются модели прогнозирования, основанные на регрессии. Регрессионный анализ оценивает ожидаемое (или среднее) значение числового целевого атрибута, когда все входные атрибуты фиксированы. Первый шаг в регрессионном анализе — выдвижение гипотезы о структуре отношений между входными атрибутами и целевым. Затем определяется параметризованная математическая модель предполагаемой взаимосвязи. Эта параметризованная модель называется функцией регрессии. Вы можете представить себе функцию регрессии как машину, которая преобразует входные данные в выходные, а параметры — в виде настроек, управляющих поведением машины. Функция регрессии может иметь несколько параметров, и целью регрессионного анализа является поиск правильных настроек для этих параметров.

С помощью регрессионного анализа можно выдвинуть гипотезу и смоделировать множество различных типов зависимостей между атрибутами. В принципе, единственное ограничение для структуры, которая может быть смоделирована, — это возможность определить соответствующую функцию регрессии. В некоторых областях могут быть веские теоретические причины для использования конкретного типа зависимости, но в иных случаях целесообразно начинать с самого

простого типа, а именно с линейной зависимости, и уже затем, если это требуется, моделировать с более сложными. Одна из причин, по которой следует начинать с линейной зависимости, — простота интерпретации функции линейной регрессии. Другая причина — здравый смысл, который состоит в том, чтобы ничего не усложнять без необходимости.

Регрессионный анализ, использующий линейную зависимость, называется линейной регрессией. Простейшим применением линейной регрессии является моделирование взаимосвязи между двумя атрибутами: входным атрибутом X и целевым атрибутом Y. В этой задаче функция регрессии имеет следующий вид:

$$Y = \omega_0 + \omega_1 X.$$

Это уравнение линейной функции (часто записываемой как $y = mx + c$), которая знакома большинству людей из курса средней школы[14]. Переменные ω_0 и ω_1 и являются параметрами функции регрессии. Изменение этих параметров меняет и то, как функция отображает прямую между входящим X и выходящим Y. Параметр ω_0 (или c из школьной формулы) — это точка пересечения прямой с осью ординат, когда X равен нулю. Параметр ω_1 определяет угол наклона прямой (т.е. он эквивалентен m из школьной формулы).

В регрессионном анализе параметры функции регрессии изначально неизвестны. Установка этих параметров эквивалентна поиску строки, которая наилучшим образом соответствует данным. Стратегия установки этих параметров состоит в том, чтобы начать со случайных значений, а затем итеративно обновлять параметры, уменьшая общее отклонение функции в наборе данных. Общее отклонение рассчитывается в три этапа:

1. Функция применяется к набору данных и для каждого объекта в наборе оценивает значение целевого атрибута.
2. Отклонение функции для каждого объекта вычисляется путем вычитания оценочного значения целевого атрибута из его фактического значения.
3. Отклонение функции для каждого объекта возводится в квадрат, а затем эти возведенные в квадрат значения суммируются.

Отклонение функции для каждого объекта возводится в квадрат на последнем шаге так, чтобы отклонение, когда функция завышает значение, не отменялось отклонением, когда цель недооценена. Возведение в квадрат и в том и в другом случае придает отклонению положительное значение. Этот параметр известен как *сумма квадратов отклонений*, а стратегия подбора линейной функции путем поиска параметров, минимизирующих сумму квадратов отклонений (SSE), называется методом наименьших квадратов. SSE определяется как

$$\text{SSE} = \sum_{i=i}^{n} (target_i - prediction_i)^2,$$

где набор данных содержит n объектов, $target_i$ — это значение целевого атрибута для объекта i в наборе данных, а $prediction_i$ — оценка функцией цели для того же объекта.

Чтобы создать линейную регрессионную модель прогнозирования, которая оценивает вероятность развития диабета у человека с учетом его ИМТ, мы заменяем X на атрибут ИМТ, а Y — на атрибут «Диабет» и применяем алгоритм наименьших квадратов, чтобы найти наиболее подходящую прямую для этого набора данных. Рис. 11 *а* иллюстрирует эту прямую и ее расположение относительно объектов в наборе

данных. На рис. 11 *b* пунктирными линиями показано отклонение (или остаток) для каждого объекта в этой прямой. При использовании метода наименьших квадратов линией наилучшего соответствия будет прямая, которая минимизирует сумму квадратов отклонений. Вот уравнение для этой прямой:

$$Диабет = -7{,}38431 + 0{,}55593 \times \text{ИМТ}.$$

Значение угла наклона прямой = 0,55593 указывает на то, что для каждого увеличения ИМТ на 1 единицу модель увеличивает предполагаемую вероятность развития диабета у человека чуть более чем на 0,5%. Чтобы предсказать вероятность развития диабета у человека, мы просто вводим его значение ИМТ в модель. Например, когда ИМТ = 20, модель возвращает прогноз 3,73% для атрибута «Диабет», а для ИМТ = 21 модель прогнозирует 4,29% вероятности[15].

Линейная регрессия, использующая метод наименьших квадратов, рассчитывает средневзвешенное значение для объектов. Фактически значение сдвига линии по вертикали $\omega_0 = -7{,}38431$ гарантирует, что линия наилучшего соответствия проходит точку, определенную средним значением ИМТ и средним значением диабета для набора данных. Если ввести среднее значение ИМТ в наборе данных (ИМТ = 24,0932), модель оценивает атрибут диабета как 4,29%, что является средним значением для всего набора данных.

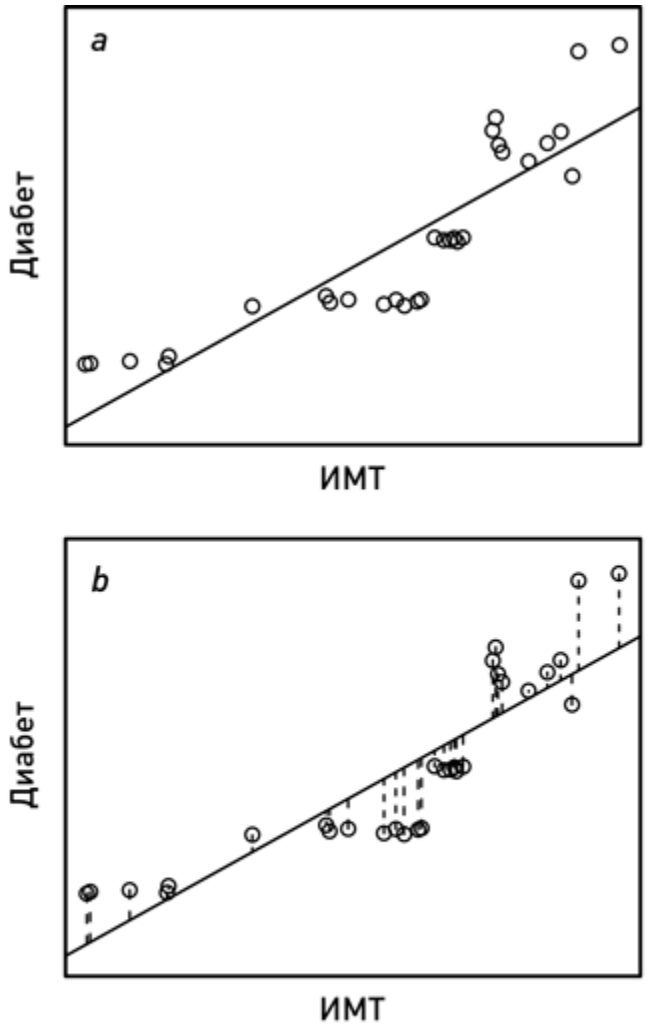

Рис. 11. *a* — линия регрессии, наиболее подходящая для модели Diabetes = –7,38431 + 0,55593 × ИМТ; *b*— пунктирные вертикальные линии иллюстрируют остаточную погрешность для каждого объекта

Взвешивание объектов основано на их расстоянии от линии — чем дальше объект находится от линии, тем его отклонение выше и алгоритм будет взвешивать экземпляр по квадрату этого отклонения. Как следствие, объекты, которые имеют экстремальные значения (выбросы), могут оказать непропорционально большое влияние на процесс вычерчивания линии, в результате чего она будет удалена от других объектов. Поэтому перед использованием метода наименьших квадратов важно проверить наличие выбросов в наборе данных.

Модели линейной регрессии могут быть расширены, чтобы принимать несколько входных значений. Новый параметр добавляется в модель для каждого нового входного атрибута, а уравнение обновляется, чтобы суммировать результат умножения нового атрибута. Например, чтобы расширить модель для включения в нее в качестве входных данных атрибутов веса и времени, затраченного на физические упражнения, структура функции регрессии станет такой:

$$Диабет = \omega_0 + \omega_1 ИМТ + \omega_2 Упражнения + \omega_3 Вес.$$

В статистике функция регрессии, которая прогнозирует переменную на основе нескольких факторов, называется функцией множественной линейной регрессии. Структура функции такой регрессии является основой для ряда алгоритмов машинного обучения, включая и нейронные сети.

Между корреляцией и регрессией наблюдаются сходства, поскольку и та и другая представляют собой техники, сосредоточенные на выявлении зависимостей между столбцами в наборе данных. Корреляция ищет взаимосвязь между двумя атрибутами, а регрессия сосредоточена на прогнозировании значений зависимой переменной при нескольких входных атрибутах. В частных случаях коэффициент корреляции Пирсона измеряет степень линейной зависимости двух атрибутов, а линейная регрессия, обученная по методу наименьших квадратов, представляет собой процесс поиска линии наилучшего соответствия, которая прогнозирует значение одного атрибута при заданном значении другого.

Нейронные сети и глубокое обучение

Нейронная сеть состоит из нейронов, соединенных друг с другом. Нейрон принимает набор числовых значений в качестве входных

данных и сопоставляет их с одним выходным значением. По своей сути нейрон — это функция линейной регрессии с несколькими входами. Единственное существенное различие состоит в том, что в нейроне выходной сигнал определяется другой функцией, которая называется функцией активации.

Функции активации, как правило, отображают выходной сигнал множественной линейной регрессии нелинейно. В качестве функций активации наиболее часто применяются логистическая функция и функция tanh (рис. 12). Обе функции принимают на вход одно значение *x*, являющееся выходным значением функции множественной линейной регрессии, которую нейрон применяет к своим входным данным. Также обе функции используют число Эйлера, приблизительно равное 2,71828182. Эти функции иногда называют функциями сжатия, поскольку они принимают любое значение от «плюс бесконечности» до «минус бесконечности» и отображают его в небольшом заранее определенном диапазоне. Диапазон выходных значений логистической функции составляет от 0 до 1, а функции tanh — от –1 до 1. Следовательно, выходные значения нейрона, который использует логистическую функцию в качестве своей функции активации, всегда находятся в диапазоне от 0 до 1. Тот факт, что обе функция используют нелинейные отображения, ясно по S-образной форме кривых. Причиной введения нелинейного отображения в нейрон является то, что одним из ограничений функции линейной регрессии с несколькими входами является ее линейность по определению, и если все нейроны в сети будут выполнять только линейные отображения, то и сама сеть также будет ограничена изучением линейных функций. Однако нелинейная функция активации в нейронах сети позволяет ей изучать более сложные (нелинейные) функции.

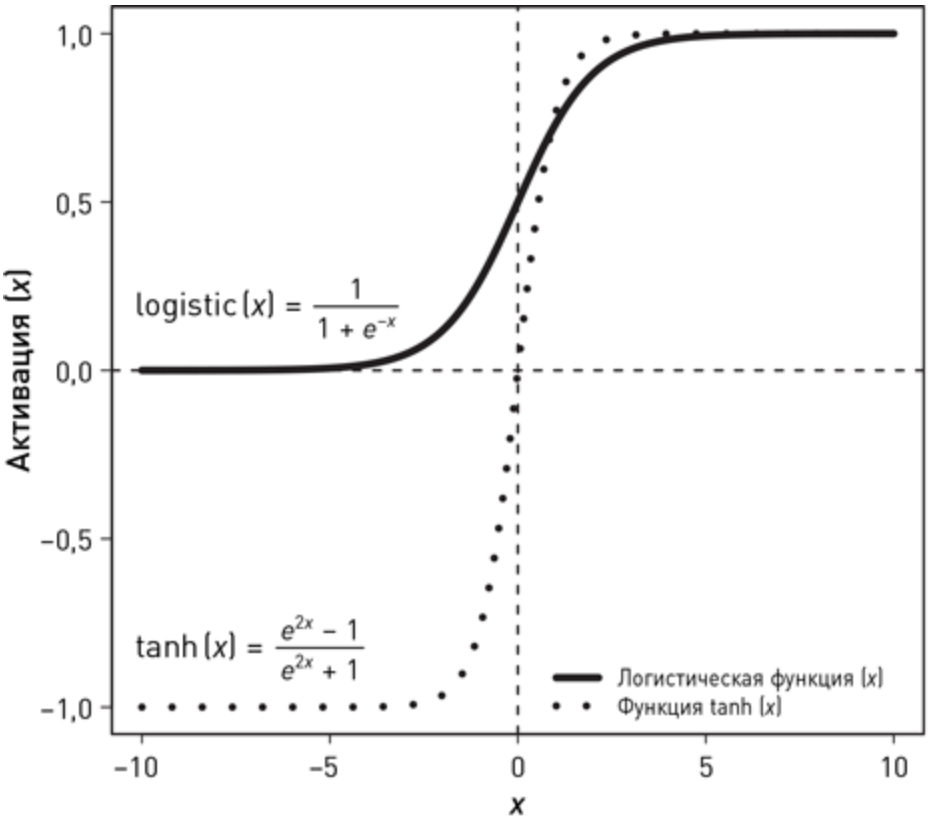

Рис. 12. Отображение логистической функции и функции tanh, применяемое к входящему *x*

Стоит подчеркнуть, что каждый нейрон в нейронной сети выполняет очень простой набор операций:

1. Умножает каждый вход на его вес;
2. Суммирует результаты умножения;
3. Проводит этот результат через функцию активации.

Операции 1 и 2 являются просто вычислением функции регрессии с несколькими входами, а операция 3 использует функцию активации.

Все связи между нейронами в нейронной сети являются направленными, и каждая имеет свой вес. Нейрон применяет вес связи к входящему значению, которое он получает через эту связь, когда вычисляет функцию множественной входной

регрессии. Рис. 13 иллюстрирует топологическую структуру простой нейронной сети. Квадраты *A* и *B* в левой части обозначают зоны памяти, которые мы используем для представления входных данных в сеть. В этих зонах обработка или преобразование данных не выполняются. Эти узлы можно считать входными или сенсорными нейронами, функция активации которых настроена таким образом, чтобы выходное значение равнялось входному[16]. Круги *C*, *D*, *E* и *F* на рисунке обозначают нейроны в сети. Бывает полезно представлять нейроны в сети организованными в слои. Сеть на рисунке имеет три слоя нейронов: входной слой содержит *A* и *B*, скрытый — *C*, *D* и *E*, а выходной слой содержит *F*. Понятие «скрытый слой» указывает на тот факт, что нейроны в этом слое не принадлежат ни входному, ни выходному слоям и в этом смысле недоступны взгляду.

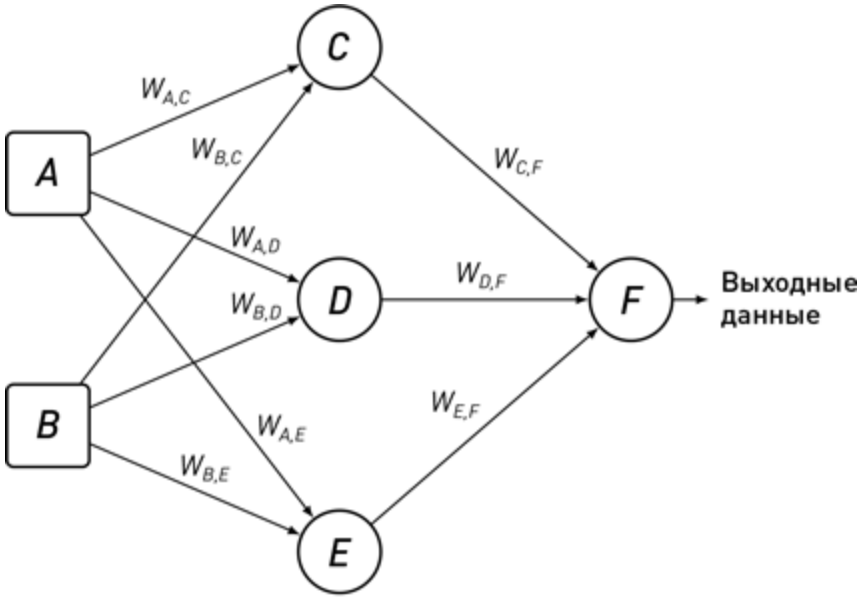

Рис. 13. Простая нейронная сеть

Стрелки, соединяющие нейроны в сети, обозначают поток информации. Технически данная конкретная сеть является нейронной сетью с прямой связью, поскольку в сети нет петель —

все соединения направлены в одну сторону, от входа к выходу. Кроме того, эта сеть считается полностью подключенной, поскольку каждый нейрон связан со всеми нейронами в следующем слое сети. Можно создать множество различных типов нейронных сетей, изменив количество слоев, число нейронов в каждом слое, тип используемых функций активации, направление соединений между слоями и другие параметры. На самом деле разработка нейронной сети для конкретной задачи во многом сводится к экспериментам по поиску наилучшей схемы.

Метки на каждой стрелке показывают вес, который узел применяет к информации, передаваемой по этому соединению. Например, есть стрелка, соединяющая *C* с *F*, которая указывает, что выходные данные из *C* передаются как входные данные для *F* и *F* будет применять к ним вес.

Предположим, что нейроны в сети на рис. 13 используют функцию активации tanh. Тогда вычисление, выполняемое нейроном *F*, может быть представлено как:

$$\text{Выходные данные} = tanh\ (\omega_{C,F}C + \omega_{D,F}D + \omega_{E,F}E).$$

Математическое представление обработки, выполняемой в нейроне F, показывает, что конечное выходное значение сети рассчитывается с использованием набора функций. Компоновка функций означает, что выходные данные одной функции используются в качестве входных данных для другой. В этом случае выходы нейронов *C*, *D* и *E* используются в качестве входов для нейрона *F*, поэтому функция, выполняемая в *F*, скомпонована из функций, выполняемых в *C*, *D* и *E*.

Для наглядности этого описания на рис. 14 показана нейронная сеть, которая принимает значение процентного содержания жира в организме человека и его МПК (максимальное потребление кислорода[17]) в качестве входных данных и вычисляет индивидуальный уровень физической

подготовки[18]. Все нейроны в среднем слое сети вычисляют функцию на основе процентного содержания жира и МПК: $f_1()$, $f_2()$ и $f_3()$. Каждая из этих функций моделирует взаимодействие между входами иначе, чем две другие. Эти функции по существу представляют собой новые атрибуты, которые получены сетью из необработанных входных данных. Они схожи с атрибутом ИМТ, описанным ранее, который был рассчитан как функция веса и роста. Иногда оказывается возможным интерпретировать выходные данные нейрона внутри сети, насколько это позволяет предметная область, и понять, почему этот производный атрибут полезен для сети. Однако чаще производный атрибут, рассчитанный нейроном, не будет нести никакого смысла для человека. Просто эти атрибуты фиксируют взаимодействия между другими атрибутами, которые сеть сочла полезными. Последний узел в сети f_4 вычисляет другую функцию — скомпонованную из $f_1()$, $f_2()$ и $f_3()$, — на выходе которой получается прогноз уровня физической подготовки, возвращаемый сетью. Опять же, эта функция не может быть значимой для человека, кроме того факта, что она определяет взаимодействие, которое, как обнаружила сеть, имеет высокую корреляцию с целевым атрибутом.

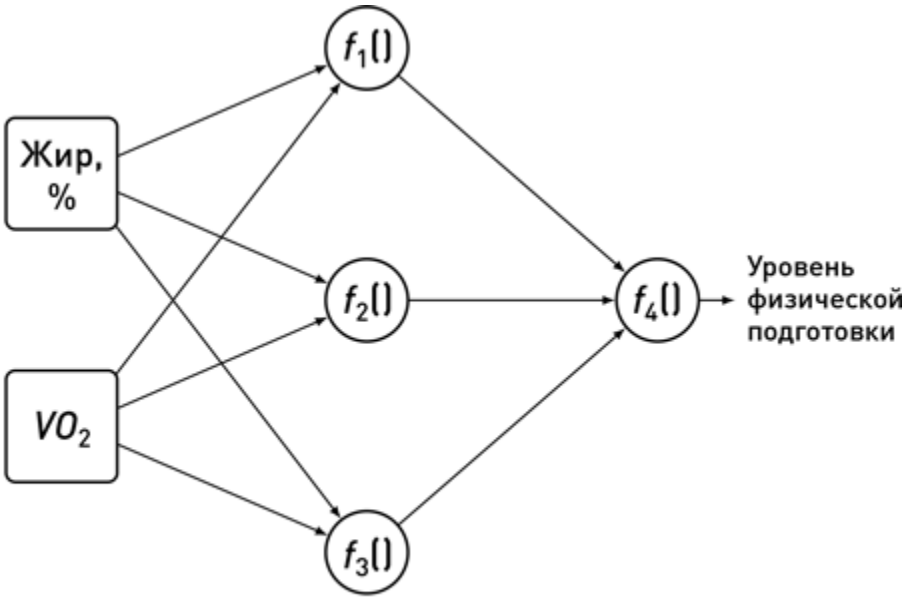

Рис. 14. Нейронная сеть, которая прогнозирует уровень физической подготовки человека

Обучение нейронной сети включает в себя поиск правильных весов для ее связей. Чтобы понять, как обучается сеть, полезно начать с размышлений о том, как отдельный нейрон обучается рассчитывать вес связи. Предположим, что у нас есть обучающий набор данных, который имеет для каждого объекта и входные значения, и целевой атрибут. Также предположим, что входящим связям нейрона уже назначены веса. Если мы возьмем объект из набора данных и представим нейрону значения входных атрибутов этого объекта, он выдаст прогноз для цели. Вычитая этот прогноз из значения целевого атрибута в наборе данных, мы сможем измерить отклонение нейрона для этого объекта. Используя ряд простых вычислений, можно вывести правило для обновления весов входящих связей нейрона с учетом выходного отклонения нейрона, чтобы уменьшить выходное отклонение. Точное определение этого правила будет варьироваться в зависимости от функции активации, используемой нейроном, поскольку она влияет на производную, используемую при выводе

правила. Но можно дать следующее наглядное пояснение того, как работает правило обновления веса:

1. Если отклонение равно 0, не меняйте веса на входах.
2. Если отклонение положительное, требуется увеличить прогнозное значение, поэтому нужно прибавить веса всех связей с положительным входом и понизить веса связей с отрицательным.
3. Если отклонение отрицательное, требуется уменьшить прогнозное значение, поэтому нужно понизить веса всех связей с положительным входом и прибавить веса связей с отрицательным.

Сложность обучения нейронной сети состоит в том, что правило обновления веса требует оценки ошибки в нейроне, и, хотя вычислить ошибку для каждого нейрона в выходном слое сети довольно просто, сделать то же самое для нейронов в более ранних слоях намного сложнее. Стандартный способ обучения нейронной сети заключается в использовании алгоритма, называемого методом обратного распространения ошибки. Алгоритм обратного распространения является алгоритмом машинного обучения с учителем, поэтому он предполагает набор обучающих данных, который бы имел как входные значения, так и целевой атрибут для каждого объекта. Обучение начинается с назначения случайных весов каждой связи в сети. Затем алгоритм итеративно обновляет весовые коэффициенты, показывая сети обучающие объекты из набора данных и обновляя весовые коэффициенты до тех пор, пока сеть не начнет работать как ожидалось. Алгоритму присваивается имя, потому что после того, как каждый обучающий объект представлен сети, ее веса обновляются путем последовательных шагов в направлении назад по сети:

1. Рассчитайте ошибку для каждого из нейронов в выходном слое и обновите согласно правилу веса входящих связей этих нейронов.
2. Поделитесь ошибкой, рассчитанной для нейрона, с каждым из нейронов в предыдущем слое, который связан с ним, пропорционально весу связей между двумя нейронами.
3. Для каждого нейрона на предыдущем уровне вычислите общую ошибку сети, за которую он ответственен, суммируя с теми ошибками, которые были переданы обратно в него, и используйте результат этого суммирования, чтобы обновить веса на связях, входящих в этот нейрон.
4. Пройдите таким же образом остальные слои в сети, повторяя шаги 2 и 3 до тех пор, пока веса между входными нейронами и первым слоем скрытых нейронов не будут обновлены.

При обратном распространении ошибки вес, обновляемый для каждого нейрона, высчитывается так, чтобы уменьшить, но не устранить полностью ошибку нейрона в обучающем экземпляре. Причина этого заключается в том, что цель обучения сети — дать ей возможность сделать выводы, которых нет в данных обучения, а не просто запомнить эти данные. Таким образом, каждое обновление весов продвигает сеть к такому их набору, который лучше всего подходит к набору данных, и на протяжении многих итераций сеть постепенно сужает значения весов в наборе, которые учитывают общее распределение данных больше, чем характеристики обучающих объектов. В некоторых версиях обратного распространения ошибки веса обновляются только после того, как несколько объектов (или пакет объектов) были представлены сети, а не после ввода каждого обучающего объекта. Единственная настройка, необходимая для этого,

заключается в том, чтобы алгоритм использовал среднюю ошибку сети для этих объектов в качестве меры ошибки на выходе для процесса обновления веса.

Одним из наиболее удивительных технических достижений последних 10 лет стало появление глубокого обучения. Сети глубокого обучения — это те же нейронные сети, имеющие несколько[19] слоев скрытых юнитов, — другими словами, они *глубоки* с точки зрения количества скрытых слоев. Нейронная сеть на рис. 15 имеет пять слоев: один входной, три скрытых (черные кружки) и один выходной слой справа, содержащий два нейрона. Эта сеть иллюстрирует то, что в каждом слое может быть разное количество нейронов: входной слой содержит три нейрона, первый скрытый слой — пять, следующие два скрытых слоя — четыре, а выходной слой — два. На примере этой сети видно и то, что выходной слой также может иметь несколько нейронов. Использование нескольких выходных нейронов полезно, если целью является номинальный или порядковый тип данных, имеющий разные уровни. В подобных сценариях сеть настраивают таким образом, чтобы для каждого уровня существовал один выходной нейрон, и обучают ее так, чтобы для каждого входа только один из выходных нейронов выводил высокую активацию (означающую прогнозируемый целевой уровень).

Подобно предыдущим сетям, которые мы рассматривали, это также полностью подключенная сеть с прямой связью. Однако не все сети являются таковыми. Было разработано множество типов сетевых топологий. Например, рекуррентные нейронные сети (РНС) вводят в сетевую топологию петли: выходное значение нейрона возвращается на один из входов в процессе обработки следующего набора входных значений. Этот цикл дает сети память, которая позволяет ей обрабатывать каждый вход в контексте предыдущих, уже обработанных ею раньше. Следовательно, РНС подходят для обработки последовательных

данных, таких как естественный язык[20]. Другой популярной архитектурой глубоких нейронных сетей являются сверхточные нейронные сети (CHC). CHC были первоначально разработаны для использования с данными изображений [1]. Сеть распознавания изображений должна обнаруживать на изображении визуальный признак независимо от того, в какой части изображения он встречается. Например, если сеть выполняет распознавание лиц, она должна уметь распознавать форму глаза, где бы он ни находился — в верхнем правом углу или в центре изображения. CHC достигают этого за счет групп нейронов, которые имеют одинаковый набор весов на своих входах. В этом контексте набор входных весов определяет функцию, которая возвращает истинное значение, если в наборе поступающих в нее пикселей появляется определенный визуальный признак. Это означает, что каждая группа нейронов с одинаковыми весами учится идентифицировать определенный визуальный признак и каждый нейрон в группе действует как детектор этого признака. В CHC нейроны в каждой группе расположены так, чтобы каждый исследовал свой фрагмент изображения, а вместе группа охватывала бы его целиком. Таким образом, если заданный визуальный признак присутствует на изображении, один из нейронов в группе идентифицирует его.

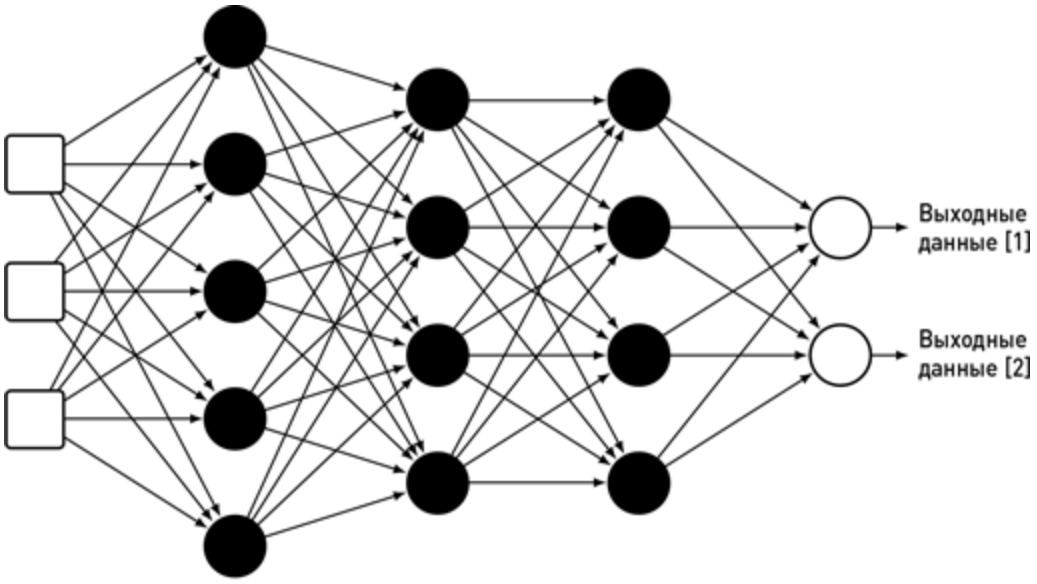

Рис. 15. Глубокая нейронная сеть

Сила глубоких нейронных сетей в том, что они могут автоматически изучать полезные атрибуты, такие как детекторы признаков в СНС. Глубокое обучение иногда так и называют — «обучение признакам», поскольку глубокие сети по сути изучают новое представление входных данных, которое лучше подходит для прогнозирования целевого выходного атрибута, чем исходный необработанный ввод. Каждый нейрон в сети определяет функцию, которая отображает значения в новый входной атрибут. Поэтому нейрон в первом слое сети может изучать функцию, которая преобразует необработанные входные значения (например, вес и рост) в более полезный атрибут (например, ИМТ). Однако выход этого нейрона наравне с его сестринскими нейронами в первом слое подается в нейроны второго слоя, изучающие функции, которые преобразуют выходные данные первого слоя в новые и еще более полезные представления. Этот процесс сопоставления входных данных с новыми атрибутами и передачи этих новых атрибутов в качестве входных данных для следующих функций распространяется по сети, и по мере того, как сеть становится глубже, она может

изучать все более и более сложные сопоставления. Именно способность автоматически изучать сложные сопоставления входных данных с полезными атрибутами делает модели глубокого обучения настолько точными при выполнении задач с многомерным вводом (таких, как обработка изображений и текста).

Давно известно, что чем глубже нейронная сеть, тем более сложные отображения данных она способна изучать. Однако развитие глубокое обучение получило лишь в последние несколько лет, и причина этого заключается в том, что стандартная комбинация случайного веса с последующим алгоритмом обратного распространения ошибки не очень хорошо работала с глубокими сетями. Во-первых, ошибка в этом случае распределяется по мере того, как процесс возвращается со слоя на слой, так что к тому времени, когда алгоритм достигает ранних слоев глубокой сети, оценки ошибок уже не так полезны[21]. В результате слои в ранних частях сети не учатся полезным преобразованиям данных. Однако в последние годы были разработаны новые типы нейронов и адаптации к алгоритму обратного распространения, которые помогают решить эту проблему. Также было обнаружено, что требуется осторожная инициализация весов сети. Два других фактора, которые усложняли обучение глубоких сетей, заключались в том, что для обучения нейронной сети требуется большая вычислительная мощность и к тому же нейронные сети показывают максимальную эффективность на большом количестве обучающих данных. В последние годы большие вычислительные мощности стали доступнее, и это сделало обучение глубоких сетей осуществимым.

Деревья решений

Линейная регрессия и нейронные сети лучше всего работают с числовыми входными данными. Если входные атрибуты в наборе данных в основном номинальные или порядковые, лучше использовать другие алгоритмы и модели машинного обучения, такие как деревья решений.

Дерево решений кодирует условный оператор *если-то-иначе* в древовидной структуре. Рис. 16 иллюстрирует дерево решений для проблемы, стоит ли смотреть фильм. Прямоугольники с закругленными углами представляют собой тесты атрибутов, а квадраты обозначают узлы решения, или классификации. Это дерево кодирует следующие правила: *если фильм — комедия, то смотреть; если фильм не комедия, а триллер, то тоже смотреть; если он не комедия и не триллер, то не смотреть*. Процесс решения для объекта в структуре дерева решений начинается с его вершины и спускается вниз, последовательно тестируя атрибуты объекта. Каждый узел дерева устанавливает один атрибут для тестирования, и процесс спускается вниз узел за узлом, выбирая следующую ветвь по метке, соответствующей значению теста атрибута. Финальное решение — это метка конечного (или листового) узла, к которому спускается объект.

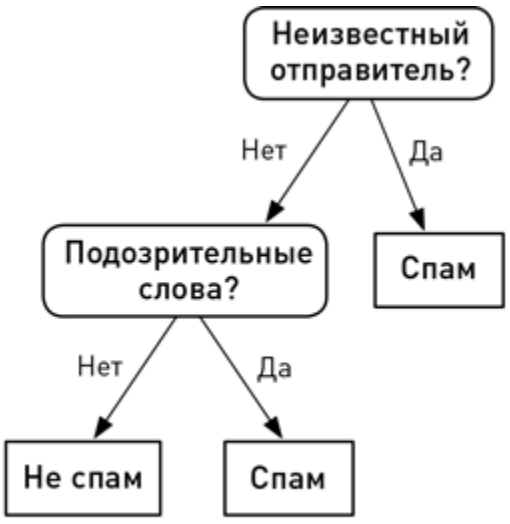

Рис. 16. Дерево решений для фильтрации спама

Все пути в структуре дерева решений от корня до листа определяются правилом классификации, состоящим из последовательных тестов. Цель обучения дерева решений состоит в том, чтобы найти такие правила классификации, которые делят обучающий набор данных на группы объектов, имеющих одинаковое значение целевого атрибута. Идея состоит в том, что если правило классификации может отделить от набора данных подмножество объектов с одинаковым целевым значением и если оно истинно для нового объекта (т.е. такого, который идет по этому пути в дереве), то вероятно, что правильный прогноз для этого нового объекта — целевое значение, общее для всех обучающих объектов, соответствующих этому правилу.

Прародителем большинства современных алгоритмов машинного обучения деревьев решений является алгоритм ID3 [3]. Он строит деревья решений рекурсивным способом в глубину, добавляя один узел зараз, начиная с корневого узла. ID3 начинается с выбора атрибута для проверки в корневом узле. Ветвь вырастает для каждого значения из области определения этого тестового атрибута и помечается этим значением. Узел с бинарным тестовым атрибутом будет иметь две ветви, исходящие от него. Затем набор данных разделяется: каждый объект из этого набора перемещается вниз по той ветви, метка которой соответствует значению тестового атрибута для объекта. Затем ID3 наращивает каждую ветвь, используя тот же процесс, что и для создания корневого узла: выбрать тестовый атрибут — добавить узел с ветвями — разделить данные, направив объекты по соответствующим ветвям. Этот процесс повторяется до тех пор, пока все объекты каждой ветви не будут иметь одинаковое значение целевого атрибута, тогда в дерево добавляется конечный узел и помечается значением целевого атрибута, общего для всех объектов ветви[22].

ID3 выбирает атрибут для тестирования в каждом узле дерева, чтобы минимизировать количество тестов, необходимых для

создания очищенных наборов (т.е. таких групп объектов, которые имеют одинаковое значение целевого атрибута). Одним из способов измерения чистоты набора является использование информационной энтропии — меры неопределенности информации Клода Шеннона. Минимально возможная энтропия для множества равна нулю, поэтому очищенное множество имеет энтропию, равную нулю. Максимальное значение возможной энтропии для набора зависит от его размера и разнообразия представленных типов элементов. Набор будет иметь максимальную энтропию, когда все элементы в нем разного типа[23]. ID3 выбирает для тестирования в узле атрибут, который приводит к наименьшему значению взвешенной энтропии после разделения набора данных с использованием этого атрибута. Взвешенная энтропия для атрибута рассчитывается следующим путем: 1) разделение набора данных по атрибуту; 2) вычисление энтропии результирующих множеств; 3) взвешивание каждой энтропии по ее доле в наборе данных; 4) суммирование результатов.

Таблица 3. Набор данных электронных писем

Вложение	Подозрительные слова	Неизвестный отправитель	Спам
Нет	Нет	Да	Нет
Нет	Нет	Да	Нет
Нет	Да	Нет	Нет
Нет	Нет	Нет	Нет
Нет	Нет	Нет	Нет

В табл. 3 приведен список электронных писем, в котором каждое описывается рядом атрибутов и тем, является оно спамом или нет. Атрибут «Вложение» имеет значение «Истина» для

электронных писем, содержащих вложение, и значение «Ложь» в ином случае (в этой выборке ни одно из электронных писем не имело вложений). Атрибут «Подозрительные слова» имеет значение «Истина», если электронное письмо содержит одно или несколько слов в предварительно определенном списке подозрительных слов. Атрибут «Неизвестный отправитель» истинен, если отправитель электронного письма отсутствует в адресной книге получателя. Этот набор данных использовался для обучения дерева решений на рис. 16. В нем атрибуты «Вложение», «Подозрительные слова» и «Неизвестный отправитель» были входными атрибутами, а атрибут «Спам» — целью. Атрибут «Неизвестный отправитель» разбивает набор данных на более чистые группы, чем другие атрибуты (одна из них содержит большинство объектов, где Спам = Истина, а другая, где Спам = Ложь). Затем «Неизвестный отправитель» помещается в корневой узел, как на рис. 17. После этого начального разделения все объекты в правой ветви имеют одинаковое целевое значение, а объекты в левой — разное. Разделение объектов в левой ветви с использованием атрибута «Подозрительные слова» приводит к образованию двух чистых наборов, где также для одного Спам = Ложь, для другого Спам = Истина. Таким образом, «Подозрительные слова» выбраны в качестве тестового атрибута для нового узла в левой ветви, как на рис. 18. На этом этапе подмножества данных в конце каждой ветви являются чистыми, поэтому алгоритм завершает работу и возвращает дерево решений, показанное на рис. 16.

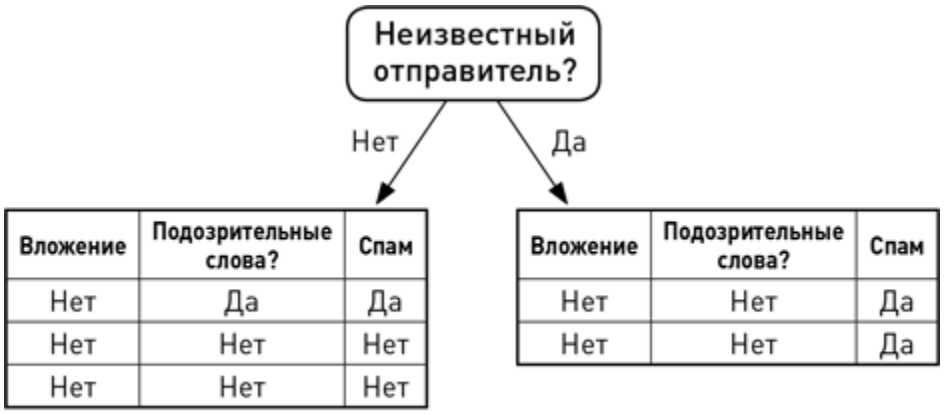

Рис. 17. Создание корневого узла в дереве

Одно из преимуществ деревьев решений заключается в том, что они понятны, притом с их помощью можно создавать очень точные модели. Например, модель «случайного леса» состоит из набора деревьев решений, где каждое дерево обучается случайной подвыборке обучающих данных, а прогноз, возвращаемый моделью для отдельного запроса, является прогнозом большинства деревьев в лесу. Хотя деревья решений хорошо работают с номинальными и порядковыми данными, они испытывают трудности с числовыми данными. В дереве решений существуют отдельные ветви, исходящие из каждого узла к каждому значению в области определения атрибута, проверяемого на узле. Числовые атрибуты, однако, могут иметь неограниченное число значений в своих областях определений, а это означает, что дереву потребуется бесконечное число ветвей. Одним из решений этой проблемы является преобразование числовых атрибутов в порядковые атрибуты, хотя для этого нужно определить соответствующие пороговые значения, что также может быть непросто.

Рис. 18. Добавление второго узла в дерево

Наконец, поскольку алгоритмы обучения дерева решений многократно разветвляют набор данных, то, когда дерево становится большим, повышается его чувствительность к шуму (например, к ошибочно помеченным объектам). Это происходит потому, что подмножества примеров в каждой ветви становятся все меньше и, следовательно, выборка данных, на которой основано правило классификации, тоже уменьшается. Чем меньше выборка данных, тем более чувствительным к шуму становится правило. Поэтому неглубокие деревья решений — это хорошая идея. Один метод заключается в том, чтобы остановить рост ветви, как только число объектов в ней станет меньше предварительно определенного порога (например, 20 объектов). Другие методы позволяют дереву расти, но потом обрезают его. В таких методах для определения ветвей в нижней части дерева, которые следует удалить, обычно используются статистические тесты или производительность модели на наборе данных, специально предназначенных для этой задачи.

Смещения в науке о данных

Цель машинного обучения — создание моделей, которые кодируют соответствующие обобщения из наборов данных. Есть два основных фактора, которые делают генерируемое алгоритмом обобщение (или модель) более ценным. Первый — это набор данных, на котором работает алгоритм. Если он не является репрезентативным для совокупности, то модель, которую генерирует алгоритм, не будет точной. Например, ранее мы разработали регрессионную модель, которая предсказывала вероятность того, что у человека разовьется диабет 2-го типа, на основе его ИМТ. Эта модель была сгенерирована из набора данных белых американских мужчин. Следовательно, эта модель вряд ли будет точной, если она используется для прогнозирования вероятности развития диабета для женщин или мужчин, принадлежащих другим расам или этническим группам. Термин «*смещение выборки*» описывает то, как процесс, используемый для формирования набора данных, может внести искажения в последующую аналитику, будь то статистический анализ или создание прогностических моделей с использованием машинного обучения.

Вторым фактором, который влияет на модель, генерируемую из набора данных, является выбор алгоритма машинного обучения. Их существует множество, и каждый кодирует свой способ обобщения набора данных. Тип обобщения, который кодирует алгоритм, известен как *смещение обучения* (или смещение выбора) алгоритма. Например, алгоритм линейной регрессии кодирует линейное обобщение и в результате игнорирует другие нелинейные отношения, которые могут более точно соответствовать данным. Смещение обычно понимается как нечто нежелательное. Например, рассмотренное выше смещение выборки является примером смещения, которого

специалист по данным постарается избежать. Тем не менее без смещения обучение невозможно, поскольку алгоритм сможет только запоминать данные.

Однако, поскольку алгоритмы машинного обучения смещены для поиска различных типов закономерностей, нет одного наилучшего алгоритма, потому что не может быть наилучшего смещения для всех ситуаций. Так называемая *теорема об отсутствии бесплатных завтраков* [5] утверждает, что не существует ни одного алгоритма МО, в среднем превосходящего все другие алгоритмы по всем возможным наборам данных. Поэтому этап моделирования процесса CRISP-DM обычно включает в себя построение нескольких моделей с использованием разных алгоритмов и их последующее сравнение, чтобы определить, какой из алгоритмов генерирует наилучшую модель. В сущности, эти эксперименты тестируют, какое смещение обучения в среднем дает лучшие модели для конкретных задач и набора данных.

Оценка моделей: обобщение, а не запоминание

После того как набор алгоритмов выбран, следующая основная задача — создать план тестирования для оценки моделей, сгенерированных этими алгоритмами. Цель плана тестирования — обеспечить, чтобы оценка производительности модели на новых данных была реалистичной. Модель прогнозирования, которая просто запоминает набор данных, вряд ли справится с оценкой значений для новых примеров. Во-первых, при простом запоминании данных большинство наборов будут содержать шум и модель прогнозирования также будет запоминать шум в данных. Во-вторых, простое запоминание сводит процесс

прогнозирования к поиску в таблице и оставляет нерешенной проблему, как обобщить обучающие данные для работы с новыми примерами, которых нет в таблице.

Одна часть плана связана с тем, как набор данных используется для обучения и тестирования моделей. По сути, набор данных предназначен для двух разных целей. Первая состоит в том, чтобы выявить алгоритм, который генерирует лучшие модели. Вторая — оценить эффективность обобщения наилучшей модели, т.е. насколько хорошо она может справиться с новыми данными. Золотое правило оценки моделей заключается в том, что их никогда не следует тестировать на тех же данных, на которых они были обучены. Использование одних и тех же данных для моделей обучения и тестирования равносильно тому, чтобы показать ученикам экзаменационные вопросы за ночь до экзамена. Естественно, студенты сдадут его на «отлично», но их результаты не будут отражать реальное знание материала курса. То же самое относится и к моделям машинного обучения: если модель оценивается по тем же данным, на которых она обучалась, то оценка будет более оптимистичной по сравнению с реальной эффективностью модели. Стандартный процесс для обеспечения этого правила таков: данные разбиваются на три части — обучающий набор, оценочный набор и тестовый набор. Пропорции, используемые для этого разбиения, будут различаться в зависимости от проекта, но обычно они составляют 50:20:30 или 40:20:40. Размер набора данных является ключевым фактором для определения пропорций разбиения: как правило, чем больше весь набор данных, тем больше и тестовый набор. Учебный набор используется для обучения начальной группы моделей. Оценочный набор — для сравнения эффективности этих моделей на новых данных. Сравнение эффективности начальных моделей на оценочном наборе позволяет нам определить, какой алгоритм сгенерировал лучшую модель. После его выявления обучающий и оценочный наборы могут быть объединены в

больший обучающий набор, и этот набор данных подается в лучший алгоритм для создания окончательной модели. Важно отметить, что тестовый набор не использовался ни во время процесса выбора наилучшего алгоритма, ни для обучения окончательной модели. По этой причине он может быть использован для оценки ее эффективности на новых данных.

Золотое правило оценки моделей заключается в том, что их никогда не следует тестировать на тех же данных, на которых они были обучены.

Другим ключевым компонентом плана тестирования является выбор подходящих показателей для оценки. Обычно модели оцениваются на основе того, насколько часто их выходные данные соответствуют выходным данным из тестового набора. Если целевой атрибут является числовым значением, то одним из способов измерения точности модели на тестовом наборе будет сумма квадратов ошибок. Если целевой атрибут является номинальным или порядковым, то самый простой способ оценить точность модели — вычислить долю в тестовом наборе правильно полученных ею примеров. Однако в некоторых случаях важно включить анализ ошибок в процесс оценки. Например, если модель используется для медицинской диагностики, при диагностировании больного пациента как здорового последствия могут быть гораздо серьезнее, чем при обратной ошибке. Диагностика больного пациента как здорового может привести к тому, что пациента отправят домой без соответствующей медицинской помощи. Если же модель диагностирует здорового пациента как больного, то с большой вероятностью эта ошибка будет обнаружена в ходе последующего обследования, назначенного пациенту. Таким образом, при оценке производительности модели требуется придать одному типу ошибки больший вес, чем другому. После создания плана тестирования специалист по данным может начать обучение и оценку моделей.

Выводы

Эта глава началась с того, что наука о данных — партнерство между специалистом по данным и компьютером. Машинное обучение представляет собой набор алгоритмов, которые

генерируют модели из большого набора данных. Однако пригодность этих моделей зависит от опыта специалиста по данным. Для успешного выполнения проекта набор данных должен быть репрезентативным для исследуемой области и включать в себя соответствующие атрибуты. Специалист по данным оценивает ряд алгоритмов машинного обучения, чтобы найти, какие из них генерируют лучшие модели. Процесс оценки модели должен следовать золотому правилу, согласно которому модель нельзя тестировать на тех же данных, на которых она была обучена.

В большинстве проектов науки о данных основным критерием выбора модели является ее точность. Однако в ближайшем будущем на выбор алгоритмов машинного обучения могут повлиять правила использования данных и конфиденциальности. Так, например, 25 мая 2018 г. в Евросоюзе вступил в силу Общий регламент по защите данных (General Data Protection Regulation, GDPR). Подробнее мы обсудим GDPR в главе 7, а сейчас просто отметим, что в нем есть отдельные статьи, которые наделяют человека «правом на получение разъяснений» в отношении автоматизированных процессов принятия решений [6]. Потенциальное значение такого права состоит в том, что использование труднообъяснимых моделей, таких как нейронные сети, для принятия решений, касающихся отдельных лиц, может стать проблематичным. При таких условиях прозрачность и простота объяснения других моделей, например деревьев решений, могут сделать их использование более подходящим.

И последнее: мир меняется, а модели — нет. В процессе построения набора данных, обучения модели и ее оценки предполагается, что будущее будет таким же, как и прошлое. Это так называемое предположение о стационарности, которое, по сути, означает, что моделируемые процессы или модели поведения являются постоянными во времени (т.е. не меняются). Наборы данных изначально имеют исторический характер в том

смысле, что представляют собой наблюдения, сделанные в прошлом. Поэтому алгоритмы машинного обучения ищут в прошлом закономерности, которые можно обобщить и интерполировать в будущее. Очевидно, что это предположение не всегда работает. Для описания того, как процесс или поведение могут со временем изменяться, специалисты по данным используют понятие дрейфа. По причине дрейфа модели перестают работать и нуждаются в переподготовке, именно поэтому процесс CRISP-DM включает в себя внешний круг, подчеркивающий итеративность науки о данных. Эти процессы должны внедряться после развертывания модели, чтобы проверить ее на устаревание, и, если устаревание имеет место, модель должна пройти переподготовку. Большинство подобных решений не могут быть автоматизированы и требуют человеческой проницательности и знаний. Компьютер ответит на поставленный вопрос, но сам вопрос может оказаться неверным.

Источники

1. Le Cun, Yann. 1989. “Generalization and Network Design Strategies.” Technical Report CRGTR-89-4. Univeristy of Toronto Connectionist Research Group.
2. Kelleher, John D. 2016. “Fundamentals of Machine Learning for Neural Machine Translation.” In Proceedings of the European Translation Forum.
3. Quinlan, J. R. 1986. “Induction of Deciston Trees.” Machine Learning 1 (1): 81–106. doi:10.1023/A:1022643204877.
4. Kelleher, John D., Brian Mac Namee, and Aoife D’Arcy. 2015. Fundamentals of Machine Learning for Predictive Data Analytics. MIT Press.

5. Wolpert, D. H., and W. G. Macready. 1997. “No Free Lunch Theorems for Optimization.” IEEE Transactions on Evolutionary Computation 1 (1): 67–82. doi:10.1109/4235.585893.
6. Burt, Andrew. 2017. “Is There a ‘Right to Explanation’ for Machine Learning in the GDPR?” https://iapp.org/news/a/is-there-a-right-to-explanation-for-machine-learning-in-the-gdpr/.

Глава 5

СТАНДАРТНЫЕ ЗАДАЧИ НАУКИ О ДАННЫХ

Одним из важнейших навыков специалиста по данным является способность сформулировать насущную проблему как стандартную задачу науки о данных. Большинство проектов в этой области можно отнести к одному из четырех основных классов задач:

- кластеризация (или сегментация);
- обнаружение аномалий (или выбросов);
- поиск ассоциативных правил;
- прогнозирование (включая подзадачи классификации и регрессии).

Понимание того, на какую задачу нацелен проект, облегчит принятие многих проектных решений. Например, для обучения модели прогнозирования требуется, чтобы каждый из объектов в наборе данных содержал значение целевого атрибута, а это, в свою очередь, дает ориентиры (через запросы) с точки зрения проектирования набора данных. Понимание задачи также определяет, какие алгоритмы машинного обучения использовать. Существует множество алгоритмов машинного обучения, и каждый предназначен для конкретной задачи глубинного анализа данных. Например, алгоритмы, генерирующие модели дерева решений, в первую очередь предназначены для решения задач прогнозирования. Каждой задаче соответствует множество

алгоритмов машинного обучения, поэтому знание задачи определяет не конкретный алгоритм, а их набор. Поскольку задача влияет как на структуру набора данных, так и на выбор алгоритмов машинного обучения, определиться с ее типом необходимо на раннем этапе жизненного цикла проекта, в идеале — на этапе понимания бизнес-целей CRISP-DM. Чтобы лучше понять типы задач, ниже мы покажем, как некоторые стандартные проблемы бизнеса соотносятся с ними.

Кто наши клиенты? (Кластеризация)

Распространенная область применения науки о данных в бизнесе — поддержка маркетинговых кампаний и продаж. Разработка целевой маркетинговой кампании требует понимания целевого клиента. Круг клиентов большинства предприятий довольно широк, в нем присутствуют потребители с разными предпочтениями и запросами, поэтому подход «один размер для всех», скорее всего, окажется провальным. Правильнее будет определить количество клиентских профилей, каждый из которых описывает значительный сегмент клиентской базы, а затем разработать целевые маркетинговые кампании для каждого. Эти профили могут быть созданы вручную с привлечением опыта в предметной сфере, но, как правило, наилучшим решением будет создать их на основе данных, которые бизнес накопил о своих клиентах. Интуиция зачастую может пропустить важные неочевидные моменты или не обеспечить требуемый для тонкой маркетинговой настройки уровень детализации. Например, Браун сообщает, что в одном проекте науки о данных известный стереотип о *маме футболиста* (домохозяйке из пригорода, которая много времени посвящает тому, чтобы возить детей на футбольные тренировки) не нашел соответствия в клиентской

базе. Однако при использовании процесса кластеризации, основанного на данных, были выявлены более сфокусированные профили, такие как *матери, работающие полный рабочий день вне дома с маленькими детьми в детском саду,* или *матери старшеклассников, которые работают неполный рабочий день,* или *женщины без детей, которые заинтересованы в здоровом питании*. Эти клиентские профили определяют более четкие цели для маркетинговых кампаний и могут выявить неизвестные сегменты в базе клиентов.

Интуиция зачастую может пропустить важные неочевидные моменты или не обеспечить требуемый для тонкой маркетинговой настройки уровень детализации.

Стандартный наукоемкий подход к этому типу анализа состоит в том, чтобы сформулировать проблему как задачу *кластеризации*. Кластеризация включает в себя сортировку объектов в наборе данных на подгруппы по принципу схожести. Кластеризация обычно проводится аналитиком, который вводит произвольное значение количества подгрупп, после чего алгоритм создает их путем объединения объектов на основе сходства значений их атрибутов. Затем эксперт в данной области определения просматривает полученные кластеры, чтобы понять, являются ли они значимыми. В контексте разработки маркетинговой кампании такой обзор нужен, чтобы проверить, насколько адекватно клиентские профили отражают действительность, или выявить новые профили, которые ранее не рассматривались.

Диапазон атрибутов, которые можно использовать для описания клиентов в процессе кластеризации, огромен, но есть наиболее типичные: демографическая информация (возраст, пол и т.д.), место жительства (почтовый индекс, адрес и т.д.), транзакционная информация, например какие продукты или услуги приобретал клиент, доход, который компания получает от него, как долго он является клиентом, участвует ли в программах лояльности, возвращал ли когда-нибудь продукт или жаловался на услугу и проч. Как и во всех проектах науки о данных, в кластеризации одна из самых больших проблем — определить, какие атрибуты должны быть включены, а какие исключены, чтобы добиться наилучших результатов. Принятие решения о выборе атрибутов основано на итерациях экспериментов, их анализе специалистом и пересмотре результатов каждого проекта.

Наиболее известным алгоритмом машинного обучения для кластеризации является метод *k-средних*. Буква *k* в названии указывает количество кластеров, которые алгоритм ищет в

данных. Значение *k* задается заранее и часто устанавливается экспериментальным путем, методом проб и ошибок. Алгоритм *k-средних* предполагает, что все атрибуты, описывающие клиентов в наборе данных, являются числовыми. Если набор данных содержит нечисловые атрибуты, то они должны быть соотнесены с числовыми значениями для использования метода *k-средних*, иначе потребуется другой алгоритм. Данный алгоритм рассматривает каждого клиента как точку в облаке точек (или в диаграмме рассеяния), где позиция клиента определяется значениями атрибутов в его профиле. Цель алгоритма — найти положение центра каждого кластера в облаке точек. Задавая количество *k* кластеров, мы задаем и количество кластерных центров (или средних), отсюда и название алгоритма.

Алгоритм работает, выполняя двухэтапный процесс: сначала каждый объект назначают ближайшему к нему кластерному центру, а затем обновляют этот центр таким образом, чтобы он оказался в середине назначенных ему объектов. Процесс начинается с выбора *k* объектов, которые будут действовать в качестве начальных кластерных центров. В настоящее время для выбора начальных кластерных центров оптимальным является так называемый алгоритм *k-средних++*. Логическое обоснование его использования состоит в максимально возможном распределении исходных кластерных центров. Первый центр устанавливается путем выбора случайного объекта в наборе данных. Второй, третий (и последующие) центры кластеров — путем выбора объектов с вероятностью, пропорциональной квадрату расстояния от ближайшего существующего кластерного центра. Как только все *k* кластерных центров инициализированы, происходит первая итерация назначения объектов ближайшему центру. После этого центры перемещаются так, чтобы совпасть с центром назначенных им объектов. Перемещение кластерных центров сместит их ближе к одним объектам и отодвинет от других, в том числе и от объектов, им назначенных. Затем

объекты переназначаются снова ближайшему обновленному кластерному центру. Некоторые объекты останутся назначенными одному и тому же центру, другие могут быть переназначены новому. Этот процесс назначения объектов и обновления центра продолжается до тех пор, пока при очередной итерации никакие объекты не будут переназначены новому кластерному центру. Алгоритм *k-средних* недетерминирован, т.е. разные начальные позиции кластерных центров, вероятно, будут давать и разные кластеры. В результате алгоритм обычно запускается несколько раз, а затем результаты этих прогонов сравниваются, чтобы увидеть, какие кластеры выглядят наиболее адекватными с учетом предметной области и ее понимания специалистом по данным.

Как и во всех
проектах науки
о данных,
в кластеризации
одна из самых
больших проблем —
определить, какие
атрибуты должны
быть включены,
а какие исключены,
чтобы добиться
наилучших
результатов.

Часто, когда кластеры в наборе находят полезными, им присваивают имена, отражающие основные характеристики профилей. Каждый кластерный центр определяет отдельный профиль клиента с описанием, сгенерированным из значений атрибутов назначенных ему объектов. В алгоритме *k-средних* нет обязательного условия, что все кластеры должны быть одного размера. Размеры кластеров могут дать полезную информацию для управления маркетингом. Например, процесс кластеризации может выявить небольшие целевые кластеры клиентов, которые отсутствуют в текущих маркетинговых кампаниях. Другая стратегия может заключаться в том, чтобы сосредоточиться на кластерах с клиентами, приносящими наибольший доход. Стратегии могут быть разными, но при любой из них понимание сегментов клиентской базы является предпосылкой успеха маркетинга.

Одним из преимуществ кластеризации как аналитического подхода является то, что она может применяться к большинству типов данных. Благодаря своей универсальности кластеризация часто используется как инструмент исследования данных на этапе их понимания во многих проектах науки о данных. Кроме того, хотя в нашем примере кластеризация применяется для разбиения клиентов на группы, она также бывает полезна и для других задач. Например, для анализа учебных курсов с целью выявления групп студентов, которые нуждаются в дополнительной поддержке или предпочитают разные методы обучения; для идентификации групп похожих документов в корпусе текстов; в биоинформатике для анализа последовательностей генов в процессе, называемом микрочиповым анализом.

Мошенничество ли это? (Обнаружение аномалий)

Обнаружение аномалий (или анализ выбросов) включает в себя поиск и выявление объектов, которые не соответствуют типичным данным в наборе. Эти несоответствующие объекты часто называют аномалиями или выбросами. Обнаружение аномалий используется в том числе при анализе финансовых транзакций с целью выявления потенциальных мошеннических действий и запуска расследований. Например, оно позволяет определить мошеннические действия по кредитным картам путем выявления транзакций, происходящих в необычном месте или на необычно большую сумму по сравнению с другими транзакциями по этой кредитной карте.

Первый подход, который большинство компаний использует для обнаружения аномалий, состоит в том, чтобы вручную определить ряд правил, основанных на экспертных знаниях в конкретной области, которые помогают идентифицировать аномальные события. Часто набор этих правил описывают на SQL или на других языках и запускают в базах или хранилищах данных. Некоторые языки программирования уже включают специальные команды для облегчения кодирования этих типов правил. Например, версии SQL для базы данных теперь включают функцию MATCH_RECOGNIZE, упрощающую обнаружение закономерности в данных. Распространенная схема мошенничества с кредитными картами заключается в том, что вор проверяет, работает ли украденная карта, совершая по ней небольшую покупку, а затем, если транзакция проходит, как можно быстрее покупает что-нибудь дорогое, прежде чем карта будет аннулирована. Функция MATCH_RECOGNIZE в SQL позволяет программистам баз данных писать сценарии, которые выявляют последовательности транзакций по кредитной карте,

соответствующие этой закономерности, и либо автоматически блокируют карту, либо предупреждают компанию-эмитента. Со временем, когда накапливается опыт выявления более сложных аномалий (например, благодаря клиентам, которые сообщают о мошенничестве), набор идентифицирующих правил расширяется, чтобы включить обработку этих новых объектов.

Основным недостатком подхода, основанного на правилах, является то, что он может идентифицировать аномальные события только после того, как они произошли и попали в поле внимания организации. В идеале большинство организаций хотели бы иметь возможность выявлять аномалии, когда они происходят впервые или если они произошли, но остались незафиксированными в отчетах. В некотором смысле обнаружение аномалий является противоположностью кластеризации: цель кластеризации состоит в том, чтобы найти группы схожих элементов, тогда как цель обнаружения аномалий — поиск элементов, непохожих на остальную часть набора данных. Такая интуитивная кластеризация может быть использована для автоматической идентификации аномалий, при этом существует два метода. Первый группирует нормальные данные вместе, а аномальные помещает в отдельные кластеры. Эти кластеры содержат небольшое число объектов по сравнению с основной частью записей. Второй метод заключается в измерении расстояния между объектом и центром кластера. Чем дальше объект находится от центра кластера, тем выше вероятность того, что он окажется аномальным и требует расследования.

Другой подход к обнаружению аномалий состоит в обучении модели прогнозирования, такой как дерево решений, для классификации объектов на нормальные и аномальные. Однако для создания такой модели обычно требуется набор обучающих данных, который содержит как аномальные, так и нормальные записи. Кроме того, нескольких экземпляров аномальных записей

недостаточно, чтобы обучить модель прогнозирования — набор данных должен содержать определенное количество объектов каждого класса. В идеале он должен быть сбалансирован на выдачу бинарного результата, что подразумевает разделение данных 50:50. Как правило, получение таких обучающих данных для обнаружения аномалий не представляется возможным: по определению аномалии являются редкими событиями, составляющими 1–2% всех данных или менее. Это ограничение препятствует нормальному использованию моделей прогнозирования. Однако существуют алгоритмы машинного обучения, известные как одноклассные классификаторы, которые предназначены для работы с несбалансированными данными при обнаружении аномалий.

Метод опорных векторов (SVM) является хорошо известным одноклассным классификатором. В общих чертах алгоритм SVM анализирует данные как одну единицу (т.е. один класс) и выявляет основные характеристики и ожидаемое поведение объектов. Затем алгоритм маркирует каждый объект, чтобы указать, насколько он похож или отличен от основных характеристик и ожидаемого поведения. С помощью этой информации выявляют аномалии, требующие дальнейшего расследования. Чем больше объект не похож на остальные, тем выше необходимость его исследования.

Тот факт, что аномалии редки, означает, что их легко можно упустить и трудно идентифицировать. По этой причине специалисты по данным часто комбинируют друг с другом модели для обнаружения аномалий. Идея состоит в том, что разные модели улавливают разные типы аномалий. Как правило, новые модели используют в дополнение к уже известным, выявляющим аномальную активность. Модели интегрируют вместе в единое решение. Это решение позволяет использовать прогнозы каждой модели при формировании окончательного результата прогноза. Например, если транзакция

идентифицирована как мошенническая только одной из четырех моделей, то система принятия решений не будет определять ее как случай мошенничества и игнорирует. И наоборот, если три или четыре модели из четырех идентифицируют транзакцию как возможное мошенничество, она будет помечена для обработки аналитиком данных.

Обнаружение аномалий может применяться во многих проблемных областях помимо мошенничества с кредитными картами. Оно используется клиринговыми центрами при мониторинге финансовых транзакций для выявления любых действий, которые требуют дальнейшего расследования, — от потенциально мошеннических до отмывания денег. Обнаружение аномалий применяется при анализе страховых претензий для выявления нетипичных. В кибербезопасности оно используется для обнаружения возможных взломов или нетипичного поведения сотрудников в сети. В области медицины выявление аномалий в историях болезней пациентов может быть полезно для диагностики заболеваний и для изучения методов лечения и их воздействия на организм. Наконец, с распространением датчиков и технологии интернета вещей обнаружение аномалий будет играть важную роль при мониторинге данных и формировании предупреждений, когда происходят нештатные ситуации и требуется вмешательство.

Добавить картофель фри? (Поиск ассоциативных правил)

Одна из стандартных стратегий продаж — перекрестные продажи, т.е. предложение клиентам дополнительных продуктов, которые они могут захотеть приобрести. Идея состоит в том, чтобы увеличить общий чек клиента, заставляя его покупать

больше и в то же время улучшая обслуживание за счет напоминания о продуктах, которые тот, возможно, хотел купить, но забыл. Классический пример перекрестных продаж — когда сотрудник ресторана быстрого питания спрашивает клиента, который только что заказал гамбургер: «Добавить картофель фри?» Супермаркетам и предприятиям розничной торговли хорошо известно, что покупатели приобретают товары группами, и они используют эту информацию для настройки перекрестных продаж. Например, клиенты супермаркетов, покупающие хот-доги, часто берут с ними кетчуп и пиво. Используя эту информацию, магазин может планировать расположение продуктов в торговом зале. Разместив хот-доги, кетчупы и пиво рядом друг с другом, магазин помогает клиентам быстрее собрать эту группу товаров, а также увеличивает свои продажи, поскольку клиенты могли забыть о кетчупе и пиве. Понимание этих связей между продуктами является основой перекрестных продаж.

Поиск ассоциативных правил — это метод анализа данных при обучении без учителя. Его суть состоит в поиске групп элементов, часто встречающихся вместе. Ассоциативные правила применяются при *анализе покупательской корзины*, когда розничные компании пытаются выявить наборы товаров, приобретаемых вместе, например хот-дог, кетчуп и пиво. Для такого анализа данных бизнес отслеживает корзину товаров каждого покупателя при каждом посещении магазина. При поиске ассоциативных правил каждая строка в наборе данных описывает содержимое корзины, оплаченной конкретным покупателем в конкретное время. Атрибуты в этом наборе данных — приобретенные товары. На основе данных алгоритм поиска ассоциативных правил ищет товары, которые встречаются в каждой корзине. В отличие от кластеризации и обнаружения аномалий, которые фокусируются на выявлении сходств или различий между объектами (или строками) в наборе данных, поиск ассоциативных правил фокусируется на рассмотрении

связей между атрибутами (или столбцами) в наборе данных. В общем смысле этот тип анализа ищет корреляции — т.е. совместные вхождения — между продуктами. Используя поиск ассоциативных правил, компания может изучить поведение своих клиентов, выявляя закономерности в данных. Вот некоторые из вопросов, на которые анализ корзины может дать ответы: «Работает ли маркетинговая кампания?», «Меняются ли закономерности покупок конкретного клиента?», «Когда клиент отмечает главные для себя праздники?», «Влияет ли местоположение конкретного магазина на покупательское поведение?», «На кого мы должны ориентировать наш новый продукт?».

Основным алгоритмом создания ассоциативных правил является алгоритм Apriori, состоящий из двух этапов:

1. Найти все комбинации товаров в наборе транзакций, которые случаются с заданной минимальной частотой. Эти комбинации называются *частыми предметными наборами*.
2. Рассчитать правила, которые отражают совместное вхождение товаров в частые предметные наборы. Алгоритм Apriori вычисляет вероятность появления элемента в частом предметном наборе с учетом присутствия в нем других предметов.

Алгоритм Apriori генерирует ассоциативные правила, которые выражают вероятностные отношения между элементами в часто встречающихся наборах элементов. Ассоциативное правило имеет форму: ЕСЛИ *{предпосылка}* — ТО *{следствие}*. Оно гласит, что предмет или группа предметов (предпосылка) подразумевает наличие с некоторой вероятностью другого предмета в той же корзине (следствие). Например, правило, выведенное из частых предметных наборов, содержащих предметы *A*, *B* и *C*, может

утверждать, что если предметы *A* и *B* включены в транзакцию, то, вероятно, в нее будет включен и предмет *C*:

ЕСЛИ *{хот-доги, кетчуп}* — ТО *{пиво}*.

Это указывает на то, что клиенты, покупающие *хот-доги* и *кетчуп*, также могут купить и *пиво*. Часто в качестве примера поиска ассоциативных правил приводят историю о том, как неизвестный американский супермаркет в 1980-х гг. одним из первых использовал компьютерную систему для анализа своих данных и выявил неожиданную ассоциацию клиентов, покупающих вместе подгузники и пиво. Теоретическое обоснование этого правила заключалось в том, что семьи с маленькими детьми готовились к уик-энду и знали, что им нужно запастись подгузниками и купить пиво, чтобы дома было что выпить. Магазин разместил эти два товара рядом, и продажи выросли. И хотя история о пиве и подгузниках теперь считается мифом, она остается ярким примером преимуществ ассоциативных правил для предприятий розничной торговли.

Ассоциативные правила имеют два основных статистических показателя: *поддержка* и *достоверность*. Процент *поддержки* ассоциативного правила указывает, как часто элементы встречаются вместе. Поддержка — это отношение транзакций, которые включают в себя элементы (и предпосылки, и следствия) к общему числу транзакций. Процент *достоверности* ассоциативного правила указывает на вероятность появления предпосылки и следствия в одной и той же транзакции. Достоверность — это условная вероятность, с какой следствие наступает в случае предпосылки. Достоверность рассчитывается как отношение поддержки к количеству транзакций, в которые входит предпосылка. Так, например, показатель достоверности 75% для ассоциативного правила, касающегося хот-догов, кетчупа и пива, указывает на то, что в 75% случаев, когда

покупатель покупал хот-доги и кетчуп, он также покупал и пиво. Значение поддержки указывает процент корзин в наборе данных, в которых выполняется правило. Например, поддержка 5% для того же примера будет показывать, что 5% всех корзин в наборе данных содержали все три элемента правила.

Даже небольшой набор данных может содержать большое количество ассоциативных правил. Чтобы упростить их анализ, набор обычно ограничивают только теми правилами, которые имеют высокие значения поддержки и достоверности. Правила, не отвечающие этим требованиям, не интересны либо потому, что охватывают очень небольшой процент корзин (низкая поддержка), либо потому, что взаимосвязь между предпосылкой и следствием низкая (низкая достоверность). Правила, которые являются тривиальными или их невозможно объяснить, также не принимаются во внимание. Тривиальные правила представляют собой ассоциации, которые очевидны и известны каждому, кто разбирается в данной сфере. Необъяснимые правила представляют собой ассоциации настолько странные, что трудно понять, как из этого правила вывести полезное действие. Вполне вероятно, что необъяснимое правило является результатом выброса данных (представляет собой ложную корреляцию). После сокращения набора правил специалист по данным может проанализировать оставшиеся и понять, какие продукты связаны друг с другом. Обычно организации используют эту информацию для составления планограмм торговых точек или проведения целевых маркетинговых кампаний. Последние могут включать в себя рекомендации продуктов на сайтах, рекламу в магазинах, прямые рассылки, перекрестные продажи на выезде и т.д.

Поиск ассоциативных правил становится более эффективным, если корзины товаров связаны с демографическими данными клиента. По этой причине многие ритейлеры используют программы лояльности, которые позволяют связывать разные корзины не только с одним клиентом, но и с его

демографическими данными. Например, ассоциативные правила, основанные на демографии, могут применяться к новым клиентам, о привычках и предпочтениях которых у компании нет информации. Вот пример ассоциативного правила, учитывающего демографические данные:

ЕСЛИ *пол (мужской), возраст (<35) и {хот-доги, кетчуп}* — ТО *{пиво} [поддержка = 2%, доверие = 90%].*

Привычная область поиска ассоциативных правил — содержимое покупательских корзин. Она охватывает товары, приобретенные за одно посещение магазина или сайта. Этот сценарий работает для большинства ритейлеров и аналогичных бизнесов, однако поиск ассоциативных правил полезен и в ряде других областей за пределами розничной торговли. К примеру, в индустрии телекоммуникаций применение ассоциативных правил в отношении клиента помогает компаниям проектировать различные сервисы и объединять их в пакеты. В страховании ассоциативные правила используются, чтобы обнаруживать связи между страховыми продуктами и требованиями клиентов. В области медицины с их помощью проверяют взаимосвязь между существующими и новыми методами лечения и лекарственными средствами. А в банковских и финансовых услугах используют ассоциативные правила для определения соответствия продуктов и конкретных клиентов, чтобы применить их к новым клиентам. Анализ ассоциативных правил также может быть использован для исследования поведения покупателей в течение определенного периода времени. Например, клиенты, как правило, одновременно покупают продукты *X* и *Y*, а через три месяца — продукт *Z*. Этот период времени можно рассматривать как одну корзину покупок, хотя он охватывает три месяца. Поиск ассоциативных правил в такой расширенной во времени корзине расширяет и области применения найденных правил, включая

графики обслуживания и замены деталей, сервисные вызовы, предложение финансовых продуктов и прочее.

Уйдет иль не уйдет, вот в чем вопрос (Классификация)

Стандартной бизнес-задачей в сфере управления взаимоотношениями с клиентами является оценка вероятности того, что отдельный клиент предпримет какое-либо действие. Для описания этой задачи используют термин «*моделирование склонности*», поскольку цель состоит в том, чтобы смоделировать склонности человека. Это могут быть реакция на маркетинг, дефолт по кредиту или отказ от услуг. Возможность идентифицировать тех, кто может покинуть сервис, особенно важна для операторов мобильной связи, которым требуются значительные инвестиции для привлечения новых клиентов. Фактически привлечение нового клиента в этой области обычно обходится в пять-шесть раз дороже, чем удержание постоянного. В результате многие операторы готовы биться за сохранение своих нынешних клиентов, стараясь при этом минимизировать затраты. Поэтому удержание клиентов за счет снижения тарифов для всех и замены старых телефонов на новые — неподходящий вариант. Вместо этого компании нацеливают свои предложения на тех клиентов, которые могут уйти в ближайшем будущем. Если идентифицировать клиента, который собирается сменить оператора, и попытаться убедить его остаться, предлагая новый телефон взамен старого или выгодный тарифный план, то компания может даже сэкономить на разнице между щедрым предложением и стоимостью привлечения нового клиента.

Термин «отток клиентов» применяется для описания группы потребителей, которые покидают один сервис и присоединяются к другому. Соответственно, проблема выявления клиентов, которые могут уйти в ближайшем будущем, называется *прогнозированием оттока*. Как следует из названия, эта задача

прогнозирования и состоит в том, чтобы классифицировать клиента, подпадает он под риск оттока или нет. Многие компании в телекоммуникационной, коммунальной, банковской, страховой и других отраслях используют этот вид анализа для прогнозирования оттока клиентов. Еще одна растущая сфера применения — прогнозирование текучести кадров или оттока персонала, т.е. того, какие сотрудники, скорее всего, покинут компанию в течение определенного периода времени.

Когда модель прогнозирования возвращает метку или категорию для входных данных, она называется моделью классификации. Обучение модели классификации требует исторических данных, где для каждого объекта указано, произошло целевое событие в его случае или нет. Процесс обучения модели классификации обычно описывают таким высказыванием:

«Мы учимся на прошлом, чтобы предсказывать будущее».

Классификация — это метод машинного обучения с учителем, в ходе которого берется набор данных с помеченными экземплярами и строится модель классификации с использованием одного или нескольких алгоритмов. Помеченный набор данных называется обучающим. Он состоит из объектов, целевой результат которых уже известен. Например, для анализа оттока клиентов требуется набор данных (по одной строке на каждого), в котором клиентам будут присвоены метки, указывающие на возможность смены ими поставщика услуг. Такой набор данных будет включать в себя целевой атрибут, который перечисляет эту метку для каждого клиента. В одних случаях назначить метку оттока для записи клиента несложно. Например, клиент сам связался с компанией и недвусмысленно отменил свою подписку или контракт. В других случаях вероятность оттока может быть неявной. К примеру, не все

абоненты имеют ежемесячный контракт с оператором мобильной связи. Некоторые предпочитают договор предоплаты, который позволяет пополнять счет не регулярно, а только по необходимости. Определить, собирается ли клиент с таким типом контракта прекратить пользование услугами, бывает непросто, поскольку неясно, что считать признаком: отсутствие звонков в течение двух недель, нулевой баланс, прекращение активности на три недели или что-то еще. После того как факт оттока был установлен с точки зрения бизнеса, необходимо реализовать это определение в коде, чтобы назначить целевую метку клиенту в наборе данных.

Другим фактором, усложняющим прогнозирование оттока, является необходимость учета временных задержек. Цель прогнозирования оттока состоит в том, чтобы смоделировать склонность (или вероятность) клиента к уходу в определенный момент в будущем. Следовательно, этот тип модели имеет временное измерение, которое необходимо учитывать при создании набора данных. Атрибуты в наборе данных для модели склонности взяты из двух разных периодов времени — периода наблюдения и итогового периода. Период наблюдения — это период времени, на основе которого рассчитываются значения входных атрибутов. Итоговый период — период, на основе которого рассчитывается целевой атрибут. Цель создания модели оттока клиентов состоит в том, чтобы дать возможность бизнесу провести вмешательство до события оттока, чтобы побудить клиента остаться. Это означает, что прогноз относительно оттока клиентов должен быть сделан до того, как клиент фактически покинет сервис. Продолжительность периода, необходимого для попытки удержания клиента, равна продолжительности итогового периода, и прогноз, который возвращает модель оттока, по факту состоит в том, расстанется ли клиент с компанией в течение этого итогового периода. Например, модель может быть обучена предсказывать, что клиент уйдет в течение

одного или двух месяцев, в зависимости от скорости предпринятых бизнесом мер по его удержанию.

Определение итогового периода влияет на то, какие данные следует использовать в качестве входных для модели. Если модель предназначена для прогнозирования оттока клиентов в ближайшие два месяца начиная с сегодняшнего дня, то при ее обучении нельзя использовать данные клиентов, описывающие их активность за последние два месяца. Таким образом, при построении набора обучающих данных входные атрибуты для каждого потерянного клиента должны рассчитываться только с использованием данных, полученных не позднее, чем за два месяца до того, как он отказался от услуг. Точно так же входные атрибуты, описывающие активных в настоящий момент клиентов, должны рассчитываться на основе данных, полученных не ранее двух месяцев назад. Это гарантирует, что все объекты набора данных, включая как ушедших, так и активных клиентов, позволяют сделать прогноз на ближайшие два месяца.

Почти во всех моделях склонности клиентов в качестве атрибутов используется демографическая информация: *возраст, пол, род занятий* и т.д. Сценарии продолжительного обслуживания могут также включать в себя атрибуты, описывающие фазы жизненного цикла клиента, например *адаптацию, середину цикла, приближение к концу контракта*. В телекоммуникационных моделях оттока клиентов также могут присутствовать атрибуты, характерные для этой отрасли. Например, средний счет клиента, изменения сумм счетов, привычки, превышение количества минут тарифного плана, соотношение вызовов внутри сети и за ее пределами, подробности, касающиеся телефонного аппарата и проч.[24] Тем не менее конкретные атрибуты, используемые в каждой модели, будут варьироваться в зависимости от проекта. Линофф и Берри рассказывают об одном проекте прогнозирования, реализованном в Южной Корее, где полезным оказался атрибут,

описывающий зависимость оттока клиентов от возраста телефонного аппарата (т.е. какой процент клиентов с телефоном определенного возраста отказались от услуг компании). Однако, когда они создавали аналогичную модель оттока клиентов в Канаде, этот атрибут стал бесполезным. Причина такой разницы заключалась в том, что в Южной Корее оператор мобильной связи предлагал большие скидки на мобильные телефоны только новым клиентам, тогда как в Канаде такие же скидки предлагались как новым, так и действующим клиентам. В результате в Южной Корее устаревание телефона приводило к оттоку клиентов, которые были заинтересованы в том, чтобы перейти к другому оператору за новыми скидками, а в Канаде такого стимула для оттока не было [1].

После создания маркированного набора данных начинается построение модели классификации с использованием алгоритма машинного обучения. В процессе моделирования рекомендуется экспериментировать с различными алгоритмами машинного обучения, чтобы выяснить, какой из них лучше работает с конкретным набором данных. После выбора окончательной модели вероятная точность ее прогнозов для новых объектов оценивается путем тестирования на подмножестве набора данных, не использованном ранее на этапе обучения модели. Если модель оценивается как достаточно точная и удовлетворяющая бизнес-потребности, она развертывается и применяется к новым данным. Этот процесс может происходить как периодически, так и в режиме реального времени. Важной частью развертывания модели является внедрение соответствующих бизнес-процессов и ресурсов для ее эффективного использования. Нет смысла создавать модель оттока клиентов, если не существует процесса, позволяющего бизнесу вмешаться для их удержания.

Кроме вышеперечисленного, модели прогнозирования могут также определять степень достоверности прогноза. Этот

показатель называется *вероятностью прогноза* и принимает значение от нуля до единицы. Чем оно выше, тем выше вероятность того, что прогноз верен. Значение вероятности прогноза можно использовать для определения приоритетов клиентов. Например, при прогнозировании оттока клиентов организация хочет сфокусироваться на тех из них, кто, скорее всего, уйдет. Используя вероятность прогноза, сортируя потоки данных на основе этого значения, компания может приоритетно сосредотачивать свои усилия на ключевых клиентах, прежде чем переходить к клиентам с более низким показателем вероятности прогноза.

Сколько это будет стоить? (Регрессия)

Ценовое прогнозирование — это задача оценки стоимости товара в определенный момент времени. Товаром может быть автомобиль, дом, баррель нефти, акции или медицинская процедура. Очевидно, что качественное ценовое прогнозирование будет востребовано любым, кто рассматривает возможность покупки товара. Точность модели напрямую зависит от предметной области. Например, из-за нестабильности фондовых рынков прогнозировать цену акций на завтра очень сложно. Для сравнения: предсказать цену дома на аукционе проще, поскольку цены на жилье колеблются меньше, чем цены акций.

Тот факт, что ценовое прогнозирование включает в себя оценку значения непрерывного атрибута, означает, что оно решается как проблема регрессии. Структурно проблема регрессии похожа на проблему классификации — в обоих случаях наука о данных предполагает построение модели, которая может предсказать недостающее значение на основании набора входных

атрибутов. Единственное отличие состоит в том, что классификация оценивает значения категориального атрибута, а регрессия — значения непрерывного. Регрессионный анализ требует набора данных, в котором указано значение целевого атрибута для каждого из объектов. Модель линейной регрессии с несколькими входами из предыдущей главы является базовой — большинство других представляют собой варианты этого подхода. Базовая структура регрессионных моделей прогнозирования цены одинакова независимо от товара — меняется только имя и количество атрибутов. Например, для прогнозирования цены на дом входные данные должны включать в себя такие атрибуты, как размер дома, количество комнат, этажность, средняя цена квадратного метра в этом районе, средний размер дома в этом районе и т.д. Для сравнения: чтобы предсказать цену автомобиля, атрибуты должны включать марку, возраст автомобиля, пробег, объем двигателя, количество дверей и т.д. В любом случае при наличии соответствующих данных алгоритм регрессии определяет, какое влияние каждый из атрибутов оказывает на окончательную цену.

Как и все примеры, приведенные в этой главе, пример применения регрессионной модели для прогнозирования цен иллюстрирует лишь тип проблемы, которую целесообразно решать с помощью регрессионной модели. Регрессионный анализ может быть использован в самых разных областях, в том числе для решения таких задач, как расчет прибыли, стоимости, объема продаж, спроса, размеров, расстояний, дозировок и объемов.

Источники

1. Linoff, Gordon S., and Michael JA Berry. 2011. *Data Mining Techniques: For Marketing, Sales, and Customer Relationship*

Management. John Wiley & Sons.

Глава 6

КОНФИДЕНЦИАЛЬНОСТЬ И ЭТИКА

Самый большой вопрос, стоящий сегодня перед наукой о данных, — как найти баланс между свободой частной жизни отдельных лиц и меньшинств и безопасностью и интересами всего общества. В контексте науки о данных этот старый вопрос формулируется с точки зрения того, что считать разумными способами сбора и использования персональных данных в таких разнообразных контекстах, как борьба с терроризмом, улучшение медицины, исследования государственной политики, борьба с преступностью, выявление мошенничества, оценка кредитного риска, страхование и таргетированная реклама.

Наука о данных предлагает свой способ для того, чтобы понять мир. В нынешнюю эпоху больших данных это предложение очень заманчиво, и действительно существует целый ряд аргументов в поддержку разработки и внедрения инфраструктуры и технологий, основанных на данных. Первый аргумент связан с повышением эффективности, экономичности и конкурентоспособности — аргумент, который в контексте бизнеса подтверждается научными исследованиями. Например, исследование, проведенное в 2011 г. с участием 179 крупных публичных компаний, показало, что чем больше решений принимается на основе данных, тем выше производительность: «Мы видим, что фирмы, которые принимают решения на основе данных, имеют производительность на 5–6% выше, чем можно

было бы ожидать, учитывая другие их инвестиции и использование информационных технологий» [1].

Еще один аргумент в пользу широкого внедрения методов науки о данных связан с безопасностью. Правительства часто его используют, оправдывая наблюдение и слежку долгосрочным повышением уровня безопасности. Как бы то ни было, начиная с 11 сентября 2001 г. и с каждым новым терактом этот аргумент набирал силу. Его использовали в публичных дебатах, которые начались после того, как Эдвард Сноуден раскрыл информацию о программе наблюдения PRISM Агентства национальной безопасности (АНБ) США, регулярно собирающей данные об американских гражданах. Красноречивым примером силы этого аргумента являются $1,7 млрд, инвестированные АНБ в центр обработки данных в Блаффдейле, штат Юта, который способен хранить огромное количество перехваченных сообщений [2].

В то же время общество, государственные структуры и бизнес пытаются понять долгосрочные последствия применения науки о данных в мире больших данных. Учитывая быстрое развитие технологий сбора, хранения и анализа данных, неудивительно, что действующая правовая база и более широкие дискуссии вокруг этой темы, в том числе о неприкосновенности частной жизни, пытаются идти в ногу с достижениями прогресса. Несмотря на это, существуют основные правовые принципы сбора и использования данных, которые применимы почти всегда и которые важно понимать. Кроме того, дискуссии об использовании данных и конфиденциальности выявили ряд тревожных тенденций, о которых мы должны знать.

Коммерческие интересы против частной жизни

Науку о данных можно представить как процесс создания процветающего и безопасного мира. Но одни и те же аргументы могут использовать организации, имеющие очень разные повестки. Мы видим противоположные призывы: с одной стороны, от групп защиты гражданских свобод к большей открытости правительств в отношении данных, чтобы граждане могли привлекать правительства к ответственности, и с другой — призывы деловых кругов использовать эти же данные для увеличения прибыли [3]. Поэтому наука о данных — это палка о двух концах. Она может быть использована для улучшения жизни за счет повышения эффективности государственного управления, развития медицины и здравоохранения, удешевления страховки, создания умных городов, снижения уровня преступности и прочего. Но в то же время ее можно использовать для слежки, таргетирования нежелательной рекламы и поведенческого контроля — как в открытую, так и тайно (страх слежки может влиять на нас не менее негативно, чем сама слежка).

Часто противоречивость аспектов науки о данных можно увидеть в одном и том же приложении. Например, в андеррайтинге в сфере медицинского страхования используются сторонние маркетинговые наборы данных, которые содержат такую информацию, как покупательские привычки, история веб-поиска, а также сотни других атрибутов, касающихся частной жизни людей [4]. Использование таких данных от третьих сторон вызывает беспокойство, поскольку может привести к тому, что люди начнут избегать определенных видов активности, скажем посещения сайтов экстремальных видов спорта из-за боязни повышения страховых взносов [5]. В оправдание использования этих данных приводится тот факт, что они выступают в роли аналога более агрессивных и дорогих источников информации, таких как анализы крови, и в долгосрочной перспективе сокращают расходы и страховые премии, таким образом увеличивая количество застрахованных людей [6].

Линия раскола между сторонниками коммерческих преимуществ и сторонниками этических соображений становится особенно очевидной в дискуссиях об использовании персональных данных для целевого маркетинга. С точки зрения рекламного бизнеса стимулом к такому использованию является наличие связи между персонализацией услуг и продуктов и эффективностью маркетинга. Было показано, что использование персональных данных из социальных сетей, например, для идентификации потребителей, связанных с действующими клиентами, повышает эффективность прямой почтовой рассылки от телекоммуникационных компаний в 3–5 раз по сравнению с традиционным подходом [7]. Аналогичные заявления были сделаны о персонализации интернет-маркетинга на основе данных. Например, проведенное в 2010 г. исследование стоимости и эффективности таргетированной онлайн-рекламы в США сравнило *сетевой маркетинг*[25] с *поведенческим таргетингом*[26] [8]. Исследование показало, что поведенческий таргетинг в среднем обходится в 2,68 раза дороже, но и коэффициент конверсии в этом случае превышает аналогичный показатель сетевого маркетинга более чем в два раза. Другое совместное исследование эффективности интернет-рекламы, основанной на данных, было проведено учеными из Университета Торонто и Массачусетского технологического института [9]. В этом исследовании эффективность онлайн-рекламы в пределах Европейского союза, где был введен новый закон о защите конфиденциальности[27], ограничивающий возможность рекламных агентств отслеживать действия пользователей в интернете, сравнивалась с эффективностью онлайн-рекламы в США и других странах, где не действовали новые ограничения. Исследование показало, что из-за новых ограничений эффективность интернет-рекламы значительно снизилась: падение покупательской активности участников исследования составило 65%. Результаты этого исследования были оспорены

(см., например, [10]), но они продолжают использоваться в поддержку аргумента, что чем больше доступно информации о человеке, тем более эффективна направленная на него реклама. Зачастую сторонники целевого маркетинга подают этот аргумент как беспроигрышный и для рекламодателя, и для потребителя, утверждая, что рекламодатели снижают маркетинговые затраты за счет сокращения расходов на рекламу и достижения лучших показателей конверсии, а потребители получают более релевантную рекламу.

Этот утопический взгляд на использование персональных данных для целевого маркетинга в лучшем случае основан на избирательном понимании проблемы. Вероятно, одна из самых тревожных историй, связанных с целевой рекламой, была опубликована в *The New York Times* в 2012 г. и касалась американского сетевого ритейлера — компании Target [11]. Маркетологи знают, что одна из причин, радикально меняющих покупательские привычки человека, — рождение ребенка. Из-за этого беременность рассматривается маркетологами как потенциальная смена привычек покупателя и приверженности брендам. Это хорошо известное явление, поэтому многие ритейлеры используют общедоступные сведения о рождениях, чтобы инициировать персонализированный маркетинг для молодых родителей, отправляя им предложения, касающиеся детских товаров. Чтобы получить конкурентное преимущество, Target решила выявлять беременность клиентов на ранней стадии (в идеале во втором триместре), но без ведома будущих матерей[28]. Это понимание должно было позволить Target начать персональный маркетинг прежде, чем другие ритейлеры узнают, что ребенок уже на подходе. Для достижения этой цели Target инициировала проект науки о данных с целью прогнозирования беременности на основе анализа покупательских привычек. Отправной точкой проекта стал анализ покупательских привычек женщин, скачавших составленный Target список покупок для

будущего ребенка. Анализ показал, что в начале второго триместра беременные женщины, как правило, покупали большое количество лосьона без запаха, а в течение первых 20 недель беременности часто приобретали определенные пищевые добавки. На основе результатов анализа Target создала модель, использующую около 25 товаров и показателей, и присвоила каждому клиенту оценку «прогноз беременности». *Успех* этой модели, если можно так выразиться, стал очевидным, когда в магазин Target пришел мужчина, который пожаловался, что его дочь-старшеклассница получила по почте именные купоны на детскую одежду и кроватки. Он обвинил Target в том, что компания пыталась убедить его дочь забеременеть. Однако через несколько дней выяснилось, что его дочь на самом деле была беременна, просто держала это в секрете. Модель прогнозирования Target смогла распознать беременную старшеклассницу и использовать эту информацию еще до того, как та решилась открыться своей семье.

Этические последствия науки о данных: профилирование и дискриминация

История о том, как Target выявила беременность старшеклассницы без ее согласия и ведома, показывает, каким образом наука о данных может использоваться для социального профилирования не только отдельных лиц, но и меньшинств. Изучая конкретные кейсы целевой рекламы, Джозеф Туроу в своей книге «Ежедневный ты» (The Daily You) рассказывает, как маркетологи используют цифровое профилирование для классификации потребителей на *целевых* и *нецелевых*, после чего персонализируют предложения и рекламные акции, адресованные конкретным лицам: «Нецелевые потребители

игнорируются или перемещаются на другие продукты, которые маркетологи сочтут более подходящими их вкусам или доходам» [12]. Такая персонализация может привести к привилегиям для одних и понижению социального статуса других. Ярким примером этого является дифференцированное ценообразование на сайтах, где с одних клиентов взимают больше, чем с других, за один и тот же продукт, основываясь на их профилях [13].

Часто эти профили создаются путем получения данных из нескольких отрывочных источников с высоким содержанием шума. Поэтому профиль может вводить в заблуждение относительно личности человека. Хуже всего то, что такие маркетинговые профили рассматриваются как продукты и продаются другим компаниям, в результате чего негативная маркетинговая оценка может преследовать человека в разных областях. Мы уже обсуждали использование маркетинговых наборов данных в качестве основы для страхового андеррайтинга [14], но эти же профили могут влиять и на решения, касающиеся оценки кредитного риска, и на многие другие процессы, влияющие на человеческую жизнь. Два аспекта маркетинговых профилей делают их особенно проблематичными: их природа «черного ящика» и устойчивость. Природа «черного ящика» не позволяет человеку узнать, что о нем записано в профиле, где и когда это было записано и как работают процессы принятия решений, использующие эти данные. В результате если человек попадает в черный список заемщиков или пассажиров авиакомпаний, то «весьма затруднительно докопаться до причин такой дискриминации и оспорить их» [15]. Более того, в современном мире, где компьютерная память обходится дешево, данные часто хранятся в течение длительного срока. Поэтому записи о событиях в жизни человека продолжают существовать еще долго после самого события. Туроу предупреждает: «Превращение персональных профилей в персональные оценки

— это результат того, что профиль начинают воспринимать как репутацию» [16].

Такая персонализация может привести к привилегиям для одних и понижению социального статуса других.

Кроме того, если использовать науку о данных неосторожно, она может увековечить и усилить подобное предубеждение. Часто утверждается, что наука о данных объективна: она основана на числах, поэтому предвзятости, влияющие на человеческие решения, в ней не используются и не кодируются. Однако правда в том, что алгоритмы науки о данных скорее аморальны, чем объективны. Наука о данных выявляет закономерности в данных, однако если данные кодируют предвзятые отношения в обществе, то алгоритм, скорее всего, идентифицирует эту закономерность и будет основывать свои выводы на ней. В самом деле, чем последовательнее предубеждение в обществе, тем сильнее оно будет отражено в данных и тем вероятнее алгоритм извлечет и воспроизведет эту модель предубеждения. Например, проведенное академическое исследование в системе онлайн-рекламы Google, показало, что система чаще предлагала рекламу высокооплачиваемой работы участникам исследования, чей профиль идентифицировала как мужской, по сравнению с участниками, идентифицируемыми как женщины [17].

Тот факт, что алгоритмы науки о данных могут усилить предубежденность, особенно заметен при их применении полицией. PredPol[29] (сокращенно от Predictive Policing) — это инструмент, предназначенный для прогнозирования места и времени вероятного преступления. При развертывании в городе PredPol генерирует ежедневный отчет с указанием на карте горячих точек (небольших участков размером 150 на 150 метров), где, по мнению системы, могут быть совершены преступления, а также помечает каждую горячую точку временным отрезком, в который это преступление вероятно произойдет. Многие полицейские управления в Соединенных Штатах и Великобритании уже используют PredPol. Идея этого типа интеллектуальной системы контроля заключается в более эффективном управлении ресурсами. На первый взгляд такое

применение науки о данных кажется разумным, ведь оно может привести к предупреждению преступлений и снижению затрат на работу полиции. Однако встают вопросы о точности предсказаний PredPol и эффективности аналогичных инициатив прогнозирования в полицейской деятельности [18], [19], [20]. Также отмечается потенциал этих типов систем для кодирования профилей по расовым или классовым признакам [21]. Развертывание отрядов полиции на основе исторических данных может привести к увеличению полицейского присутствия в определенных районах — как правило, экономически неблагополучных, — что, в свою очередь, приведет к росту ответной криминальной активности в этих районах. Другими словами, прогнозирование преступности становится самоисполняющимся пророчеством. Результатом этого цикла является то, что отдельные районы будут подвергаться избыточному контролю со стороны полиции, что повлечет снижение к ней доверия у жителей этих районов [22].

Если использовать науку о данных неосторожно, она может увековечить и усилить предубеждения.

Другим примером полицейского контроля на основе данных является Стратегический список подозреваемых (SSL), который используется отделом полиции Чикаго для снижения уровня преступности, связанной с применением огнестрельного оружия. Этот список был создан в 2013 г. и на тот момент включал 426 человек, которые были признаны вероятными участниками преступлений с применением огнестрельного оружия. В попытке предотвратить эти преступления полицейское управление Чикаго связалось со всеми людьми из списка, чтобы предупредить их, что они находятся под наблюдением. При этом как минимум несколько человек были крайне удивлены, что попали в эту категорию: у них были судимости, но за мелкие, ненасильственные правонарушения [23]. Отсюда вытекает первый вопрос: насколько точна эта технология? Недавнее исследование показало, что люди, попавшие в SSL в 2013 г., «с той же степенью вероятности могут стать жертвами убийства или стрельбы, что и случайные участники контрольной группы» [24]. В то же время в докладе указывалось, что лица, включенные в список, с большей вероятностью будут арестованы за стрельбу, причем это может быть вызвано самим фактом их наличия в списке, что подразумевает повышенную осведомленность полиции об этих людях [25]. Отвечая на это исследование, полицейское управление Чикаго заявило, что оно обновляет алгоритм, используемый для регулярной компиляции SSL, а его эффективность заметно улучшилась с 2013 г. [26]. Второй вопрос, который следует задать: как человек попадает в этот список? В версии SSL 2013 г., по-видимому, кроме прочих атрибутов личности, использовался анализ социальных сетей, включая истории арестов за стрельбу среди друзей [27], [28]. С одной стороны, идея анализа социальных сетей имеет смысл, но при этом она вскрывает серьезную проблему виновности и связей. Один из аспектов этой проблемы заключается в том, что бывает

очень сложно определить наличие прочной связи между людьми. Достаточно ли для этого жить на одной улице? Кроме того, в Америке, где подавляющее большинство заключенных — мужчины африканского и латиноамериканского происхождения, очевидно, что алгоритмы полицейского прогнозирования будут ориентироваться на цвет кожи [29].

Предупреждающий характер полицейского прогнозирования означает, что отношение к человеку определяется не тем, что он сделал, а выводами, основанными на данных, о том, что он может сделать. В результате такие типы систем способны усиливать дискриминацию, копируя закономерности из исторических данных, и создавать самосбывающиеся пророчества.

Этические последствия науки о данных: создание паноптикума

Если вы потратите некоторое время на изучение коммерческой пропаганды, которая окружает науку о данных, у вас появится ощущение, что любая проблема может быть решена с использованием ее технологий при наличии достаточного объема корректных данных. Этот маркетинг возможностей порождает иллюзию того, что подход к управлению, основанный на данных, является наилучшим решением сложных социальных проблем, таких как преступность, бедность, образование и здравоохранение: все, что от нас требуется, — это внедрить повсеместно датчики, затем объединить данные и запустить алгоритмы, чтобы сгенерировать ключевые идеи, которые обеспечат решение.

Но после того, как эти аргументы приняты, на первый план выходят два процесса. Во-первых, общество становится более технократическим по своей природе, и многие аспекты жизни начинают регулироваться системами на основе данных. Такое технологическое регулирование уже существует: так, в некоторых

юрисдикциях наука о данных используется на слушаниях об условно-досрочном освобождении [30] и при вынесении приговоров [31]. Из других примеров, за пределами судебной системы, можно привести технологии умного города, которые регулируют потоки городского трафика с помощью алгоритмов, динамически определяющих, какой из потоков получит приоритет на перекрестке в зависимости от часа дня [32]. Побочным продуктом этого технократического всплеска является засилье датчиков, поддерживающих автоматизированные системы регулирования. Второй процесс, получающий развитие, — «расползание контроля», когда данные, собранные для одной цели, перенаправляются и используются для другой [33]. Например, дорожные камеры, установленные в Лондоне для регулирования заторов, были в дальнейшем использованы для задач безопасности [34]. Или пример технологии под названием ShotSpotter, состоящей из общегородской сети микрофонов, предназначенных для распознавания выстрелов из огнестрельного оружия и их локализации. Она также используется для записи разговоров, часть из которых уже стали основанием для вынесения приговоров по уголовным делам [35]. Или использование автомобильных навигационных систем для отслеживания водителей, которые выезжают за пределы штата, и их штрафования [36], [37].

Одним из аспектов расползания контроля является стремление объединить данные из разных источников, чтобы обеспечить более полную социальную картину и таким образом попытаться достичь более глубокого понимания системных проблем. Часто приводятся веские причины для перепрофилирования данных, звучат призывы объединить данные из разных ветвей власти с целью, например, поддержки исследований в области здравоохранения или удобства для государства и граждан. Однако с точки зрения гражданских свобод эти тенденции вызывают беспокойство. Усиленный надзор, интеграция данных из

нескольких источников, расползание контроля и упреждающее управление (например, использование программ полицейского прогнозирования) могут привести к тому, что человек станет вызывать подозрение только потому, что последовательность не связанных между собой невинных действий и/или встреч совпадет с моделью, которую система управления данными считает подозрительной. Жизнь в таком обществе превратит нас из свободных граждан в заключенных паноптикума Бентама[30], заставляя самодисциплинироваться из опасения этих неверных выводов. Различие между людьми, которые верят в то, что они свободны от надзора и ведут себя соответственно, и людьми, которые самодисциплинируются внутри паноптикума, является главным различием между свободным обществом и тоталитарным государством.

В поисках утраченной приватности

Поскольку современные люди взаимодействуют и живут в технически развитом обществе, они неизбежно оставляют цифровые следы. Повсеместное внедрение видеонаблюдения означает, что данные о человеке могут собираться в любое время и где бы он ни был — на улице, в магазине, на парковке, не говоря уже о возможности отслеживания мобильных телефонов. Реальные примеры сбора данных включают учет покупок по кредитным картам, использование схем лояльности в супермаркетах, снятие наличных в банкоматах, звонки по мобильному телефону и проч. В интернете данные о людях собираются, когда они посещают сайты или входят в систему, отправляют электронную почту, совершают онлайн-покупки, назначают даты, посещают ресторан или магазин, пользуются устройством для чтения электронных книг, смотрят лекцию на

открытых онлайн-курсах или публикуют что-то в социальной сети. Чтобы можно было составить представление об объеме данных, собираемых в среднем на одного человека, отметим, что, согласно отчету голландского управления по защите данных за 2009 г., среднестатистический гражданин Нидерландов был включен в 250–500 различных баз, а для более социально активных людей этот показатель достигает 1000 [38]. Собранные вместе точечные данные и определяют цифровой след человека.

Персональные данные собираются двумя способами, и оба являются проблематичными с точки зрения конфиденциальности. Во-первых, данные могут собираться без ведома человека. Во-вторых, даже если человек решил оставить свои данные или мнение, он не может знать, как они будут использованы, будут ли переданы третьим сторонам и каким образом. Термины «цифровая тень» и «цифровой след» используются для различения этих двух типов сбора данных: цифровая тень — это данные, собранные о человеке без его ведома, согласия или осведомленности, а цифровой след состоит из данных, сознательно опубликованных человеком [39].

Сбор персональных данных без ведома или согласия, конечно, вызывает беспокойство. Но не стоит забывать и того, что мощь современных методов выявления закономерностей в сочетании с интеграцией и многократным использованием данных из нескольких источников означает, что даже собранные с ведома и согласия человека они могут иметь для него негативные и непредсказуемые последствия. Эти методы науки о данных способны на основе открытой информации, которую мы охотно публикуем, например, в социальных сетях, вывести другую информацию, сугубо личную, которой мы делиться не планировали. Например, многие пользователи Facebook ставят лайки просто потому, что хотят продемонстрировать поддержку своим друзьям. Тем не менее, используя эти данные, модели, разработанные для Facebook, могут довольно точно

предсказывать сексуальную ориентацию, политические и религиозные взгляды человека, умственные способности, особенности личности, склонность к употреблению вызывающих привыкание веществ, таких как алкоголь, наркотики и сигареты, и даже то, например, оставались ли вместе его родители до его совершеннолетия [40]. Среди неконтекстных связей, установленных этими моделями, встречаются, например, поддержка кампании за права человека как предиктор гомосексуальности (и для мужчин, и для женщин) или симпатия к бренду Honda как вероятный признак того, что человек не курит [41].

Вычислительные методы сохранения конфиденциальности

В последние годы растет интерес к вычислительным методам сохранения конфиденциальности на протяжении всего процесса анализа данных. Наиболее известны два из этих методов: *дифференциальная приватность* и *федеративное машинное обучение*.

Дифференциальная приватность — это математический метод получения полезной информации о населении, без изучения отдельных людей. Дифференциальная приватность использует узкое определение конфиденциальности, когда та не считается нарушенной включением персональных данных в процесс анализа, если выводы, сделанные в результате этого анализа, совпадают с выводами, сделанными без включения данных отдельных лиц. Существует ряд процессов, реализующих дифференциальную приватность. В их основе лежит идея добавления шума либо на этапе сбора данных, либо в ответы на запросы к базе данных. Шум защищает конфиденциальность

отдельных лиц, но может быть удален из данных на агрегированном уровне так, чтобы можно было рассчитать полезную статистику по населению в целом. Хорошим примером введения шума в данные является метод случайного ответа. Например, при анкетировании респондентам предлагается ответить «Да» или «Нет» на деликатный вопрос (касающийся нарушения закона, состояния здоровья и т.д.), используя следующую процедуру:

1. Подбросьте монету и держите результат в секрете.
2. Если выпал орел, отвечайте «Да».
3. Если выпала решка, отвечайте правдиво.

Половине респондентов выпадет орел, и она ответит «Да», другая половина ответит правдиво. Таким образом, истинное число респондентов, ответивших «Нет» в общей численности населения приблизительно вдвое превысит количество данных ответов «Нет» (монета выпадает случайным образом, поэтому соотношение ответов «Да» и «Нет» среди респондентов, которым выпал орел, должно быть таким же, как и среди ответивших правдиво). Зная истинное число ответов «Нет», мы можем вычислить истинное число ответов «Да». Однако, несмотря на то что теперь мы относительно точно знаем долю ответивших «Да», невозможно определить, для кого конкретно из респондентов это условие выполняется. Существует компромисс между количеством шума, вводимого в данные, и полезностью данных для анализа. Дифференциальная приватность устраняет этот компромисс, оценивая необходимый уровень шума с учетом таких факторов, как распределение данных в базе, типы обрабатываемых запросов и их количество. Хорошим введением в дифференциальную приватность и знакомством с методами ее реализации может стать книга Синтии Дворк и Аарона Рота «Алгоритмические основы дифференциальной приватности» [42].

В настоящее время техники дифференциальной приватности уже используются при создании потребительских продуктов. Например, Apple внедрила дифференциальную приватность в iOS 10, чтобы защитить конфиденциальность отдельных пользователей, но в то же время сохранить возможность выявлять закономерности в данных для совершенствования функции поиска и интеллектуального набора текста в мессенджерах.

Правда в том, что алгоритмы науки о данных скорее аморальны, чем объективны.

В некоторых сценариях данные поступают в проект из нескольких разнородных источников. Например, несколько больниц участвуют в общем исследовательском проекте, или компания собирает данные от большого числа пользователей приложения для мобильного телефона. Вместо того чтобы централизовать данные в одном хранилище и проводить анализ в единой базе, альтернативный метод предлагает обучать различные модели подмножеств непосредственно в источниках данных (т.е. в отдельно взятых больницах или в телефонах пользователей), а затем объединить уже обученные модели. Google использует этот федеративный метод машинного обучения, чтобы улучшить советника запросов, сделанных с помощью клавиатуры Google на Android [43]. Сперва мобильное устройство загружает в матрицу федеративного машинного обучения Google копию текущего приложения. Данные по его использованию собираются непосредственно на устройстве, и к ним применяется алгоритм обучения, который действует локально до обновления. В процессе обновления полученные модели загружаются в облако, где они усредняются с такими же моделями, загруженными с других телефонов пользователей. Затем базовая модель обновляется с использованием полученной усредненной модели. Используя этот процесс, компания улучшает базовую модель и в то же время сохраняет конфиденциальность пользователей.

Правовые рамки регулирования использования данных и защиты конфиденциальности

В разных юрисдикциях существуют разные правовые методы защиты конфиденциальности и допустимого использования данных. Тем не менее в большинстве демократических юрисдикций присутствуют два основных законодательства: антидискриминационное и о защите личных данных.

Антидискриминационное законодательство, как правило, запрещает дискриминацию на основании некоторого подмножества следующих признаков: инвалидность, возраст, пол, раса, этническая принадлежность, национальность, сексуальная ориентация и религиозные или политические убеждения. В США Закон о гражданских правах 1964 г. [44] запрещает дискриминацию по расовым, половым, религиозным или национальным признакам. Позднее этот список был расширен: Закон об американцах-инвалидах 1990 г. защищает людей от дискриминации по признаку инвалидности [45]. Подобная законодательная база существует и во многих других юрисдикциях. Например, Хартия основных прав Европейского союза запрещает дискриминацию по любым признакам, включая расу, цвет кожи, этническое или социальное происхождение, генетические особенности, пол, возраст, место рождения, инвалидность, сексуальную ориентацию, религию или убеждения, имущество, принадлежность к национальным меньшинствам, а также политическое или любое другое мнение [46].

Ситуация схожести и частичного совпадения наблюдается и в отношении законодательств о конфиденциальности. В США Принципы честной работы с информацией [47] послужили основой для большей части государственного Закона о конфиденциальности. Аналогично в Евросоюзе Директива о защите данных [48] стала основой европейского законодательства о конфиденциальности, последним воплощением которого является Общий регламент по защите данных [49]. Однако наиболее широко принятыми являются

Руководящие принципы по защите частной жизни и трансграничных потоков персональных данных, опубликованные Организацией экономического сотрудничества и развития [50]. В рамках этих руководящих принципов персональные данные определяются как данные, относящиеся к идентифицируемому лицу или субъекту данных. Этот документ устанавливает восемь частично перекрывающихся принципов, которые предназначены для защиты конфиденциальности субъекта данных:

1. **Принцип ограничения сбора данных:** персональные данные должны быть получены только законным образом, с ведома и согласия субъекта данных.
2. **Принцип качества данных:** любые собираемые персональные данные должны соответствовать цели использования и быть точными, полными и актуальными.
3. **Принцип детализации цели:** во время или до момента сбора личных данных субъект данных должен быть проинформирован о цели их использования. Кроме того, изменения цели допустимы, но они не должны быть произвольными (новая цель должна быть совместима с первоначальной) и требуют согласия субъекта данных.
4. **Принцип ограничения использования:** использование персональных данных ограничивается целью, о которой субъект данных был проинформирован, и они не должны раскрываться третьим лицам без его согласия.
5. **Принципы обеспечения безопасности:** персональные данные должны быть защищены мерами безопасности от удаления, кражи, разглашения, изменения или несанкционированного использования.
6. **Принцип открытости:** субъект данных должен иметь возможность легко получать информацию, касающуюся

сбора, хранения и использования его персональных данных.

7. **Принцип индивидуального участия:** субъект данных имеет право на доступ к своим персональным данным и их оспаривание.
8. **Принцип подотчетности:** ответственность за соблюдение принципов несет оператор данных.

На пути к этической науке о данных

Хорошо известно, что, несмотря на существующие правовые рамки, государства часто собирают персональные данные своих и иностранных граждан без их ведома. Часто это делается под предлогом безопасности и в целях разведки: программа PRISM АНБ США, программа Tempora Центра правительственной связи Великобритании [51] и программа СОРМ правительства России [52]. Эти программы влияют на общественное мнение о правительствах и на использование телекоммуникационных технологий. Результаты опроса Pew Research Centre 2015 г. на тему стратегий приватности американцев после заявлений Сноудена показали, что 87% респондентов были осведомлены о государственном надзоре за телефонной связью и интернетом; среди тех, кто знал об этом, 61% заявили о своей неуверенности в том, что эти программы служат общественным интересам; а 25% сообщили, что стали иначе использовать технологии в ответ на эту информацию [53]. Аналогичные результаты были получены и в европейских опросах: более половины европейцев знают о крупномасштабном сборе данных государственными учреждениями, и большинство респондентов утверждают, что такой надзор негативно влияет на уровень их доверия

правительству в отношении использования персональных данных [54].

В то же время многие частные компании избегают соблюдения правил в вопросах персональных данных и приватности, утверждая, что используют производные, агрегированные или анонимные данные. Переупаковывая данные таким образом, компании утверждают, что данные больше не являются персональными и их можно собирать без ведома или согласия людей, не имея четкой непосредственной цели, чтобы хранить в течение длительных периодов времени, перепрофилировать эти данные или продавать их с выгодой. Мотивация такой позиции состоит в том, что многие сторонники науки о данных с точки зрения коммерческих возможностей утверждают, что реальная ценность данных заключается в их повторном использовании или «необязательном значении» [55]. Сторонники повторного использования данных любят говорить о двух технических инновациях, которые делают сбор и хранение данных разумной бизнес-стратегией: во-первых, сегодня данные можно собирать пассивно без особых усилий или осведомленности со стороны отслеживаемых лиц, и, во-вторых, хранение данных стало относительно дешевым. В этом контексте имеет коммерческий смысл записывать и хранить данные на случай, если будущие (непредсказуемые сегодня) коммерческие возможности сделают их ценными.

Современная практика коммерческого накопления, перепрофилирования и продажи данных полностью противоречит принципу детализации цели и Руководящим принципам защиты частной жизни и трансграничных потоков персональных данных ОЭСР. Кроме того, принцип ограничения сбора данных нарушается всякий раз, когда компания представляет потребителю соглашение о конфиденциальности, которое разработано так, что его невозможно понять и/или оно оставляет за компанией право вносить изменения без

дальнейших консультаций или уведомлений. Поэтому любое уведомление превращается в бессмысленное действие «для галочки». Как и в случае с государственным надзором ради безопасности, общественное мнение весьма негативно настроено по отношению к коммерческим сайтам, собирающим и повторно использующим личные данные. В качестве лакмусовой бумажки опять возьмем американские и европейские опросы. К примеру, анкетирование американских интернет-пользователей, проведенное в 2012 г., показало, что 62% опрошенных взрослых заявили, что не знают, как ограничить информацию, собираемую о них сайтами, а 68% — что им не нужна целевая реклама, потому что она отслеживает и анализирует их поведение в интернете [56]. Недавний опрос европейских граждан обнаружил схожие результаты: 69% респондентов считают, что для сбора персональных данных необходимо получить их явное одобрение, но только 18% респондентов полностью читают заявления о конфиденциальности [57]. Кроме того, 67% респондентов заявили, что они не читают заявления о конфиденциальности, потому что находят их слишком длинными, а 38% считают их запутанными или слишком сложными для понимания [58]. Опрос также показал, что 69% респондентов были обеспокоены тем, что их информация используется для целей, отличных от той, для которой она была собрана, а 53% не устраивали интернет-компании, использующие их персональную информацию для адаптации рекламы.

Таким образом, на данный момент общественное мнение в целом негативно относится как к государственному надзору, так и к интернет-компаниям, собирающим, хранящим и анализирующим персональные данные. Сегодня большинство комментаторов согласны с тем, что законодательство о конфиденциальности данных необходимо обновить и внести изменения. В 2012 г. Европейский союз и Федеральная торговая комиссия в США опубликовали обзоры и обновления,

касающиеся защиты данных и политики конфиденциальности [59], [60], [61]. В 2013 г. Руководящие принципы ОЭСР были расширены, чтобы включить, помимо прочего, более подробную информацию о принципе подотчетности. В частности, новые Руководящие принципы определяют обязанности оператора данных с точки зрения наличия программы управления конфиденциальностью и четко определяют, что влечет за собой такая программа и как ее следует рассматривать с точки зрения управления рисками в отношении персональных данных [62].

В 2014 г. гражданин Испании Марио Костеха Гонсалес выиграл дело в Европейском суде против Google, отстаивая свое право на забвение [63]. Суд постановил, что физическое лицо может при определенных условиях запросить поисковую интернет-систему, чтобы та удалила ссылки на веб-страницы, полученные в результате поиска от имени физического лица. Основанием для такого иска было то, что данные могут быть неточными, устаревшими или храниться дольше, чем это необходимо для исторических, статистических или научных целей. Это решение имеет серьезные последствия для всех поисковых систем в интернете, но может оказать влияние и на другие накопители больших данных. Например, пока неясно, как это повлияет на сайты социальных сетей, таких как Facebook или Twitter [64].

Концепция права на забвение была утверждена и в других юрисдикциях. Например, в калифорнийском законе закреплено право несовершеннолетнего на удаление по запросу материалов, которые он или она разместили в интернете или на мобильном сервисе; также закон запрещает интернет-операторам и операторам мобильной связи собирать персональные данные несовершеннолетних в целях адресной рекламы или для передачи третьей стороне [65].

В качестве заключительного примера происходящих изменений стоит упомянуть, что в 2016 г. был подписан и принят

Щит конфиденциальности ЕС–США [66]. Щит конфиденциальности — это соглашение об обязательствах в отношении персональных данных двух юрисдикций. Подход заключается в том, чтобы усилить защиту данных граждан ЕС, когда эти данные оказываются перемещены за пределы ЕС. Этим соглашением введены более строгие обязательства для компаний в отношении прозрачности использования данных наряду с жесткими механизмами надзора и возможными санкциями, а также соглашения между правительством США и Европейским союзом об ограничениях и механизмах надзора для государственных органов, осуществляющих запись или доступ к персональным данным.

Тем не менее на момент написания этой книги сила и эффективность Щита конфиденциальности ЕС–США проверяется в ходе судебного разбирательства в ирландском суде. Причина, по которой ирландская правовая система находится в центре этих дебатов, заключается в том, что многие крупные американские транснациональные интернет-компании (Google, Facebook, Twitter и другие) расположили свои EMEA[31] штаб-квартиры в Ирландии. В результате уполномоченный по защите данных в Ирландии несет ответственность за соблюдение правил ЕС о трансграничной передаче данных этими компаниями. Недавняя история показывает, что судебные дела могут привести к значительным и быстрым изменениям в регулировании работы с персональными данными. Фактически Щит конфиденциальности ЕС–США является прямым следствием дела, возбужденного Максом Шремсом, австрийским юристом и активным защитником приватности, против Facebook. Результатом дела Шремса в 2015 г. стало немедленное вступление в силу действующего соглашения о «Безопасной гавани» между ЕС и США, и в качестве экстренного реагирования был разработан Щит конфиденциальности. По сравнению с первоначальным соглашением о «Безопасной гавани» Щит конфиденциальности

укрепил права граждан ЕС в отношении их данных [67], и вполне может быть, что в скором времени укрепит еще больше.

С точки зрения науки о данных важно то, что эти примеры демонстрируют постоянные изменения правил, касающихся конфиденциальности и защиты данных. Примеры, приведенные здесь, взяты из контекстов США и Европейского союза, но они свидетельствуют о более широких тенденциях. Сложно предсказать, как эти изменения будут развиваться в долгосрочной перспективе. В этой области много заинтересованных сторон, включая крупные рекламные, страховые и интернет-компании, спецслужбы, органы полиции, правительства, медицинские и социальные организации, а также правозащитные группы. Каждый из этих секторов общества использует данные в своих целях, и, следовательно, мы имеем разные взгляды на проблему конфиденциальности данных. Кроме того, мы сами, как частные лица, вероятно, будем менять свои взгляды в зависимости от того, чью точку зрения разделяем. Например, вы одобряете использование и повторное использование ваших персональных данных для медицинских исследований. Однако, как показывают опросы общественного мнения, о которых говорилось выше, многие не поддерживают сбор, повторное использование и обмен данными для целевой рекламы. Вообще говоря, в дискуссии о будущем конфиденциальности данных прослеживаются две темы. Согласно одной точке зрения, необходимо ужесточить правила сбора персональных данных и в некоторых случаях дать людям возможность контролировать их сбор, обмен и использование. Другая точка зрения требует ослабить регулирование, но законодательно ужесточить требования компенсации за ненадлежащее использование персональных данных. С таким количеством различных заинтересованных сторон и точек зрения не может быть простых и очевидных ответов. Вполне вероятно, что возможные решения, которые только предстоит разработать,

будут учитывать интересы секторов и состоять из компромиссов, согласованных между соответствующими сторонами.

В таком изменчивом контексте лучше всего действовать консервативно и этично. Если мы разрабатываем новые решения бизнес-задач в области данных, то должны учитывать и этические вопросы. Для этого есть веские деловые причины. Во-первых, этичные и прозрачные действия с персональными данными обеспечат хорошие отношения с клиентами, а вот ненадлежащая практика их использования способна нанести серьезный ущерб репутации бизнеса и привести к оттоку клиентов [68]. Во-вторых, существует риск того, что по мере усиления интеграции, повторного использования, профилирования и нацеливания данных общественное мнение в ближайшие годы будет консолидироваться вокруг идеи защиты конфиденциальной информации, что приведет к ужесточению правил. Прозрачные и этичные действия — лучший способ гарантировать, что разрабатываемые решения в области науки о данных не будут противоречить действующим нормам или тем, которые могут появиться в ближайшие годы.

В качестве примера того, что несоблюдение этических норм может иметь серьезные последствия для разработчиков и поставщиков технологий, можно привести случай 2015 г. [69]. Дело закончилось тем, что Федеральная торговая комиссия США (FTC) оштрафовала разработчиков и издателей игр в соответствии с Законом о защите конфиденциальности детей в интернете (COPPA). Разработчики интегрировали стороннюю рекламу в свои бесплатные игры. Интеграция сторонней рекламы является стандартом бесплатных бизнес-моделей. Однако возникла проблема, поскольку игры были предназначены для детей в возрасте до 13 лет. В результате, предоставляя данные своих пользователей рекламным сетям, разработчики нарушили COPPA. Кроме того, в одном эпизоде разработчики даже не сообщили рекламным сетям, что приложения предназначены для

детей. Детям могла быть показана неуместная реклама, поэтому FTC постановила, что издатели игр несут ответственность за обеспечение соответствия контента и рекламы возрасту пользователей. В последние годы число подобных случаев увеличилось, и ряд организаций, в том числе FTC [70], призвали компании принять *принципы «проектируемой конфиденциальности»* [71]. Эти принципы были разработаны в 1990-х гг. и признаются в качестве основы для защиты частной жизни во всем мире. Принципы утверждают, что защита конфиденциальности должна быть режимом по умолчанию для проектирования технологий и информационных систем. Чтобы следовать этим принципам, разработчик должен сознательно и активно стремиться воплотить идеи конфиденциальности в технологиях, организационных методах и архитектурах сетевых систем.

Хотя доводы в пользу этической науки о данных ясны, действовать этично не всегда так просто. Чтобы конкретизировать этические вопросы науки о данных, представьте себе, что вы работаете в критически важном для бизнеса проекте специалистом по данным. Анализируя данные, вы обнаружили ряд зависимых атрибутов, которые в совокупности оказались замещающей переменной расы, религии, сексуальной ориентации или иного конфиденциального признака. Вы знаете, что по закону не можете использовать расовый атрибут в вашей модели, но вы уверены, что эти замещающие переменные позволят вам обойти антидискриминационное законодательство. Вы также полагаете, что включение этих атрибутов в модель заставит ее работать, хотя, конечно, обеспокоены тем, что это может усилить дискриминацию, которая уже присутствует в системе. Спросите себя: «Что я буду делать?»

Источники

1. Brynjolfsson, Erik, Lorin M. Hitt, and Heekyung Hellen Kim. 2011. “Strength in Numbers: How Does Data-Driven Decisionmaking Affect Firm Performance?” SSRN Scholarly Paper ID 1819486. Rochester, NY: Social Science Research Network. https://papers.ssrn.com/abstract=1819486.
2. Carroll, Rory. 2013. “Welcome to Utah, the NSA’s Desert Home for Eavesdropping on America.” The Guardian, June 14, sec. US news. https://www.theguardian.com/world/2013/jun/14/nsa-utah-data-facility.
3. Kitchin, Rob. 2014a. The Data Revolution: Big Data, Open Data, Data Infrastructures and Their Consequences. Sage.
4. Batty, Mike, Arun Tripathi, Alice Kroll, Cheng-sheng Peter Wu, David Moore, Chris Stehno, Lucas Lau, Jim Guszcza, and Mitch Katcher. 2010. “Predictive Modeling for Life Insurance: Ways Life Insurers Can Participate in the Business Analytics Revolution.” Society of Actuaries. https://www.soa.org/files/pdf/research-pred-mod-life-batty.pdf.
5. Mayer-Schönberger, Viktor, and Kenneth Cukier. 2014. Big Data: A Revolution That Will Transform How We Live, Work, and Think. Reprint edition. Boston: Eamon Dolan/Mariner Books.
6. Batty, Mike, Arun Tripathi, Alice Kroll, Cheng-sheng Peter Wu, David Moore, Chris Stehno, Lucas Lau, Jim Guszcza, and Mitch Katcher. 2010. “Predictive Modeling for Life Insurance: Ways Life Insurers Can Participate in the Business Analytics Revolution.” Society of Actuaries. https://www.soa.org/files/pdf/research-pred-mod-life-batty.pdf.
7. Hill, Shawndra, Foster Provost, and Chris Volinsky. 2006. “Network-Based Marketing: Identifying Likely Adopters via

Consumer Networks." Statistical Science 21 (2): 256– 276. doi:10.1214/088342306000000222.
8. Beales, Howard. 2010. "The Value of Behavioral Targeting." Network Advertising Initiative. http://www.networkadvertising.org/pdfs/Beales_NAI_Study.pdf.
9. Goldfarb, Avi, and Catherine E. Tucker. 2011. "Online Advertising, Behavioral Targeting, and Privacy." Communications of the ACM 54 (5): 25–27.
10. Mayer, J. R., and J. C. Mitchell. 2012. "Third-Party Web Tracking: Policy and Technology." In 2012 IEEE Symposium on Security and Privacy, 413–27. doi:10.1109/SP.2012.47.
11. Duhigg, Charles. 2012. "How Companies Learn Your Secrets." The New York Times, February 16. http://www.nytimes.com/2012/02/19/magazine/shopping-habits.htm.
12. Turow, Joseph. 2013. The Daily You: How the New Advertising Industry Is Defining Your Identity and Your Worth. Yale University Press.
13. Clifford, Stephanie. 2012. "Supermarkets Try Customizing Prices for Shoppers." The New York Times, August 9. http://www.nytimes.com/2012/08/10/business/supermarkets-trycustomizing-prices-for-shoppers.html.
14. Batty, Mike, Arun Tripathi, Alice Kroll, Cheng-sheng Peter Wu, David Moore, Chris Stehno, Lucas Lau, Jim Guszcza, and Mitch Katcher. 2010. "Predictive Modeling for Life Insurance: Ways Life Insurers Can Participate in the Business Analytics Revolution." Society of Actuaries. https://www.soa.org/files/pdf/research-pred-mod-life-batty.pdf.
15. Kitchin, Rob. 2014a. The Data Revolution: Big Data, Open Data, Data Infrastructures and Their Consequences. Sage.
16. Clifford, Stephanie. 2012. "Supermarkets Try Customizing Prices for Shoppers." The New York Times, August 9. http://www.nytimes.com/2012/08/10/business/supermarkets-trycustomizing-prices-for-shoppers.html.

17. Datta, Amit, Michael Carl Tschantz, and Anupam Datta. 2015. “Automated Experiments on Ad Privacy Settings.” Proceedings on Privacy Enhancing Technologies 2015 (1): 92–112.
18. Hunt, Priscillia, Jessica Saunders, and John S. Hollywood. 2014. Evaluation of the Shreveport Predictive Policing Experiment. Rand Corporation. http://www.rand.org/pubs/research_reports/RR531.
19. The Oakland Privacy Working Group. 2015. “PredPol: An Open Letter to the Oakland City Council,” June 25. https://www.indybay.org/newsitems/2015/06/25/18773987.php.
20. Harkness, Timandra. 2016. Big Data: Does Size Matter? Bloomsbury Sigma.
21. Baldridge, Jason. 2015. “Machine Learning And Human Bias: An Uneasy Pair.” TechCrunch. http://social.techcrunch.com/2015/08/02/machine-learning-and-human-bias-an-uneasypair/.
22. Harkness, Timandra. 2016. Big Data: Does Size Matter? Bloomsbury Sigma.
23. Gorner, Jeremy. 2013. “Chicago Police Use Heat List as Strategy to Prevent Violence.” Tribunedigital-Chicagotribune. August 21. http://articles.chicagotribune.com/2013-0821/news/ct-met-heat-list-20130821_1_chicago-police-commander-andrew-papachristosheat-list.
24. Saunders, Jessica, Priscillia Hunt, and John S. Hollywood. 2016. “Predictions Put into Practice: A Quasi-Experimental Evaluation of Chicago’s Predictive Policing Pilot.” Journal of Experimental Criminology 12 (3): 347–71. doi:10.1007/s11292-016-9272-0.
25. Saunders, Jessica, Priscillia Hunt, and John S. Hollywood. 2016. “Predictions Put into Practice: A Quasi-Experimental Evaluation of Chicago’s Predictive Policing Pilot.” Journal of Experimental Criminology 12 (3): 347–71. doi:10.1007/s11292-016-9272-0.

26. Rhee, Nissa. 2016. “Study Casts Doubt on Chicago Police’s Secretive ‘Heat List.’” Chicago Magazine. August 17. http://www.chicagomag.com/city-life/August-2016/ChicagoPolice-Data.
27. Gorner, Jeremy. 2013. “Chicago Police Use Heat List as Strategy to Prevent Violence.” Tribunedigital-Chicagotribune. August 21. http://articles.chicagotribune.com/2013-0821/news/ct-met-heat-list-20130821_1_chicago-police-commander-andrew-papachristosheat-list.
28. Dokoupil, Tony. 2013. “‘Small World of Murder’: As Homicides Drop, Chicago Police Focus on Social Networks of Gangs.” NBC News. December 17. http://www.nbcnews.com/news/other/small-world-murder-homicides-drop-chicagopolice-focus-social-networks-f2D11758025.
29. Baldridge, Jason. 2015. “Machine Learning And Human Bias: An Uneasy Pair.” TechCrunch. http://social.techcrunch.com/2015/08/02/machine-learning-and-human-bias-an-uneasypair.
30. Berk, Richard A., and Justin Bleich. 2013. “Statistical Procedures for Forecasting Criminal Behavior.” Criminology & Public Policy 12 (3): 513–544.
31. Barry-Jester, Anna Maria, Ben Casselman, and Dana Goldstein. 2015. “Should Prison Sentences Be Based On Crimes That Haven’t Been Committed Yet?” FiveThirtyEight. August 4. https://fivethirtyeight.com/features/prison-reform-risk-assessment/.
32. Kitchin, Rob. 2014b. “The Real-Time City? Big Data and Smart Urbanism.” GeoJournal 79 (1): 1–14. doi:10.1007/s10708-013-9516-8.
33. Innes, Martin. 2001. “Control Creep.” Sociological Research Online 6 (3). https://ideas.repec.org/a/sro/srosro/2001-45-2.html.
34. Dodge, Martin, and Rob Kitchin. 2007. “The Automatic Management of Drivers and Driving Spaces.” Geoforum 38 (2): 264–275.

35. Weissman, Cale Gutherie. 2015. “The NYPD’s Newest Technology May Be Recording Conversations.” Business Insider. http://uk.businessinsider.com/the-nypds-newesttechnology-may-be-recording-conversations-2015-3.
36. Elliott, Christopher. 2004. “BUSINESS TRAVEL; Some Rental Cars Are Keeping Tabs on the Drivers.” The New York Times, January 13. http://www.nytimes.com/2004/01/13/business/business-travel-some-rental-cars-arekeeping-tabs-on-the-drivers.html.
37. Kitchin, Rob. 2014a. The Data Revolution: Big Data, Open Data, Data Infrastructures and Their Consequences. Sage.
38. Koops, Bert-Jaap. 2011. “Forgetting Footprints, Shunning Shadows: A Critical Analysis of the ‘Right to Be Forgotten’ in Big Data Practice.” SCRIPTed, Tilburg Law School Legal Studies Research Paper No. 08/2012, 8 (3): 229–56. doi:10.2139/ssrn.1986719.
39. Koops, Bert-Jaap. 2011. “Forgetting Footprints, Shunning Shadows: A Critical Analysis of the ‘Right to Be Forgotten’ in Big Data Practice.” SCRIPTed, Tilburg Law School Legal Studies Research Paper No. 08/2012, 8 (3): 229–56. doi:10.2139/ssrn.1986719.
40. Kosinski, Michal, David Stillwell, and Thore Graepel. 2013. “Private Traits and Attributes Are Predictable from Digital Records of Human Behavior.” Proceedings of the National Academy of Sciences 110 (15): 5802–5. doi:10.1073/pnas.1218772110.
41. Kosinski, Michal, David Stillwell, and Thore Graepel. 2013. “Private Traits and Attributes Are Predictable from Digital Records of Human Behavior.” Proceedings of the National Academy of Sciences 110 (15): 5802–5. doi:10.1073/pnas.1218772110.
42. Dwork, Cynthia, and Aaron Roth. 2014. “The Algorithmic Foundations of Differential Privacy.” Foundations and Trends®

in Theoretical Computer Science 9 (3–4): 211–407.
43. McMahan, Brendan, and Daniel Ramage. 2017. “Federated Learning: Collaborative Machine Learning without Centralized Training Data.” Google Research Blog. Accessed July 30. https://research.googleblog.com/2017/04/federated-learning-collaborative.html.
44. Celler, Emanuel. 1964. Civil Rights Act of 1964. 78 United States Statutes at Large. Vol. 241. https://www.gpo.gov/fdsys/pkg/STATUTE-78/pdf/STATUTE-78-Pg241.pdf.
45. Harkin, Tom. 2017. Americans with Disabilities Act of 1990. 104 Statutes at Large. Vol. 327. Accessed September 1. https://www.gpo.gov/fdsys/pkg/STATUTE-104/pdf/STATUTE104-Pg327.pdf.
46. Convention, European. 2000. “Charter of Fundamental Rights of the European Union.” Official Journal of the European Communities C (364): 1–22.
47. US Dept of Health, Education, and Welfare. 1973. “Records, Computers and the Rights of Citizens.” https://www.ncjrs.gov/App/Publications/abstract.aspx?ID=9994.
48. EU, Council and Parliment. 1995. “95/46/EC of the European Parliament and of the Council of 24 October 1995 on the Protection of Individuals with Regard to the Processing of Personal Data and on the Free Movement of Such Data.” Official Journal of the EC 23 (6).
49. EU. 2016. General Data Protection Regulation Of the European Council and Parliment. Vol. L 119. http://ec.europa.eu/justice/data-protection/reform/files/regulation_oj_en.pdf.
50. OECD. 1980. “Guidelines on the Protection of Privacy and Transborder Flows of Personal Data.” Organisation for Economic Co-Operation and Development. https://www.oecd.org/sti/ieconomy/oecdguidelinesontheprotectionofprivacyandtransborderflowsofpersonaldata.htm.

51. Shubber, Kadhim. 2013. “A Simple Guide to GCHQ’s Internet Surveillance Programme Tempora.” WIRED UK. July 24. http://www.wired.co.uk/article/gchq-tempora-101.
52. Soldatov, Rei, and Irina Borogan. 2012. “In Ex-Soviet States, Russian Spy Tech Still Watches You.” WIRED. https://www.wired.com/2012/12/russias-hand/all/.
53. Rainie, Lee, and Mary Madden. 2015. “Americans’ Privacy Strategies Post-Snowden.” Pew Research Center. http://www.pewinternet.org/files/2015/03/PI_AmericansPrivacyStrategies_0316151.pdf.
54. Eurobarometer. 2015. “Data Protection.” Special Eurobarometer 431. http://ec.europa.eu/COMMFrontOffice/publicopinion/index.cfm/Survey/index#p=1&instruments=SPECIAL.
55. Mayer-Schönberger, Viktor, and Kenneth Cukier. 2014. Big Data: A Revolution That Will Transform How We Live, Work, and Think. Reprint edition. Boston: Eamon Dolan/Mariner Books.
56. Purcell, Kristen, Joanna Brenner, and Lee Rainie. 2012. “Search Enging Use 2012.” Pew Research Center. http://www.pewinternet.org/2012/03/09/main-findings-11/.
57. Eurobarometer. 2015. “Data Protection.” Special Eurobarometer 431. http://ec.europa.eu/COMMFrontOffice/publicopinion/index.cfm/Survey/index#p=1&instruments=SPECIAL.
58. Eurobarometer. 2015. “Data Protection.” Special Eurobarometer 431. http://ec.europa.eu/COMMFrontOffice/publicopinion/index.cfm/Survey/index#p=1&instruments=SPECIAL.
59. European Commission. 2012. “Commission Proposes a Comprehensive Reform of the Data Protection Rules — European Commission.” January 25. http://ec.europa.eu/justice/newsroom/data-protection/news/120125_en.htm.

60. Federal Trade Commission. 2012. “Protecting Consumer Privacy in an Era of Rapid Change.” https://www.ftc.gov/sites/default/files/documents/reports/federal-trade-commissionreport-protecting-consumer-privacy-era-rapid-changerecommendations/120326privacyreport.pdf.
61. Kitchin, Rob. 2014a. The Data Revolution: Big Data, Open Data, Data Infrastructures and Their Consequences. Sage.
62. OECD. 2013. “2013 OECD Privacy Guidelines — OECD.” https://www.oecd.org/internet/ieconomy/privacy-guidelines.htm.
63. CJEU. 2014. C-131/12. Court of Justice of the European Union.
64. Marr, Bernard. 2015. Big Data: Using SMART Big Data, Analytics and Metrics To Make Better Decisions and Improve Performance. 1 edition. Chichester, West Sussex, United Kingdom ; Hoboken, New Jersey: Wiley.
65. Senate of California. 2013. SB-568 Privacy: Internet: Minors. Business and Professions Code, Relating to the Internet. Vol. Division 8, Chapter 22.1 (commencing with Section 22580). https://leginfo.legislature.ca.gov/faces/billNavClient.xhtml?bill_id=201320140SB568.
66. European Commission. 2016. “The EU-U.S. Privacy Shield.” December 7. http://ec.europa.eu/justice/dataprotection/international-transfers/eu-us-privacy-shield/index_en.htm.
67. O’Rourke, Carol, and Aphra Kerr. 2017. “Privacy Schield for Whom? Key Actors and Privacy Discourse on Twitter and in Newspapers.” Westminster Papers in Communication and Culture, Special Issue : Redesigning or Redefining Privacy? 12 (2).
68. Buytendijk, Frank, and Jay Heiser. 2013. “Confronting the Privacy and Ethical Risks of Big Data.” Financial Times. https://www.ft.com/content/105e30a4-2549-11e3-b34900144feab7de.

69. Kerr, Aphra. 2017. Global Games: Production in the Digital Game Industry. New York: Routledge.
70. Eurobarometer. 2015. “Data Protection.” Special Eurobarometer 431. http://ec.europa.eu/COMMFrontOffice/publicopinion/index.cfm/Survey/index#p=1&instruments=SPECIAL.
71. Cavoukian, Ann. 2013. “Privacy by Design: The 7 Foundation Principles (Primer).” Information and Privacy Commissioner Ontario, Canda. https://www.ipc.on.ca/wpcontent/uploads/2013/09/pbd-primer.pdf.

Глава 7

БУДУЩИЕ ТЕНДЕНЦИИ И ПРИНЦИПЫ УСПЕШНОСТИ

Очевидная тенденция современных обществ — распространение систем, способных воспринимать мир и реагировать на него: смартфоны, умные дома и города, беспилотные автомобили и прочее. Такое засилье интеллектуальных устройств и датчиков ставит под угрозу нашу конфиденциальность. Но также оно способствует росту объема данных и разработке новых технологических концепций, таких как интернет вещей. В этом контексте наука о данных будет оказывать все большее влияние на разные сферы нашей жизни. Однако есть две области, в которых она приведет к значительным переменам уже в ближайшее десятилетие, — это персонализированная медицина и развитие умных городов.

Наука о данных и медицина

В последние годы медицина изучает и внедряет науку о данных и прогнозную аналитику в целый ряд областей. Традиционно, диагностируя состояние пациента или принимая решение о последующем лечении, врачи полагались на свой опыт и интуицию. Такие направления, как доказательная медицина и точная медицина, утверждают, что врачебные решения должны основываться на данных, в идеале связывая наилучшие из

имеющихся данных с ситуацией и предпочтениями отдельного пациента. Например, в случае точной медицины технология быстрого секвенирования генома позволяет анализировать геномы пациентов с редкими заболеваниями, чтобы выявить мутации, которые их вызывают, и выбрать подходящие именно для этого пациента методы лечения. Еще одним фактором, стимулирующим науку о данных в медицине, является стоимость медицинского обслуживания. Наука о данных и, в частности, прогнозная аналитика могут быть использованы для автоматизации некоторых процессов здравоохранения. Существует множество примеров использования прогностической аналитики для определения момента, когда следует начинать вводить антибиотики и другие лекарства детям и взрослым, и уже имеется много сообщений о спасенных жизнях.

Разрабатываются медицинские датчики, которые пациенты могут носить на себе или которые можно проглатывать либо имплантировать. Эти датчики непрерывно контролируют жизненные показатели и реакции пациентов, а также работу их органов в течение дня. Непрерывно собираемые данные передаются обратно на централизованный сервер мониторинга. Здесь, на сервере мониторинга, медицинские работники получают доступ к данным пациентов, оценивают их состояние, определяют, какое влияние оказывает лечение, и сравнивают результаты с результатами других аналогичных пациентов, чтобы наметить следующие шаги лечения. Данные с датчиков интегрируются с другими данными из разных областей медицины и фармацевтической промышленности для изучения действия существующих и новых лекарств. Наука о данных используется при разработке персонализированных программ лечения с учетом типа пациента, его состояния и того, как его организм реагирует на различные лекарства. Кроме того, специалисты по данным в области медицины проводят исследования лекарственных средств и их взаимодействий, разрабатывают

более эффективные и подробные системы мониторинга и помогают глубже понять результаты клинических испытаний.

Умные города

Города по всему миру внедряют новые технологии, чтобы иметь возможность собирать и использовать данные, сгенерированные их жителями, и улучшить управление городскими организациями, коммунальными службами и сервисами. Существует три основных фактора этой тенденции: наука о данных, сами большие данные и интернет вещей. Интернет вещей основан на взаимодействии физических устройств и датчиков таким образом, чтобы эти устройства могли обмениваться информацией. Это может показаться тривиальным, если бы не одно преимущество — теперь мы можем удаленно управлять интеллектуальными устройствами (например, нашим домом, если он оснащен должным образом). Наряду с этим интернет вещей на основе сетевого взаимодействия между компьютерами открывает возможность интеллектуальным средам автономно прогнозировать наши потребности и реагировать на них. Вы уже можете купить интеллектуальный холодильник, который предупредит вас об истечении срока годности продуктов или закажет свежее молоко через смартфон.

Проекты умных городов объединяют данные в режиме реального времени из множества различных источников в единый хаб, где их анализируют и используют для планирования и принятия управленческих решений. Некоторые проекты предполагают создание абсолютно новых умных городов с нуля. Масдар в Объединенных Арабских Эмиратах и Сонгдо в Южной Корее — это совершенно новые города, построенные с использованием интеллектуальных технологий и

ориентированные на экологичность и энергоэффективность. Однако наиболее умные проекты предусматривают модернизацию существующих городов с использованием новых сенсорных сетей и центров обработки данных. Например, в рамках проекта SmartSantander[32] в Испании было установлено более 12 000 сетевых датчиков по всему городу для измерения температуры, уровней шума, освещения, угарного газа и наличия парковочных мест. Часто проекты умного города затрагивают энергоэффективность, дорожное движение и маршрутизацию, а также планирование коммунальных услуг в соответствии с потребностями растущего населения.

Япония приняла концепцию умного города, сделав особый акцент на сокращении потребления энергии. Токийская электроэнергетическая компания (TEPC) установила более 10 млн интеллектуальных счетчиков в домах, находящихся в зоне обслуживания[33]. В то же время она разрабатывает и внедряет приложения для смартфонов, которые позволяют клиентам отслеживать расход электроэнергии в их домах в режиме реального времени и вносить изменения в контракт на поставку электроэнергии. Эти приложения для смартфонов также позволяют TEPC отправлять клиентам индивидуальные рекомендации по энергосбережению. Вне дома технологии умного города могут быть использованы для снижения потребления электроэнергии интеллектуальным уличным освещением. Модель, демонстрирующая возможности городов будущего, установленная в Глазго, управляет уличным освещением, включая и выключая его в зависимости от присутствия людей. Энергоэффективность также является главным приоритетом для всех новых зданий, особенно административных и офисных. Она может быть оптимизирована за счет автоматического управления климат-контролем зданий с помощью комбинации сенсорной технологии, больших данных и науки о данных. Дополнительным преимуществом этих

интеллектуальных систем мониторинга является то, что они могут отслеживать уровни загрязнения и качество воздуха и при необходимости активировать необходимые средства контроля и предупреждения в режиме реального времени.

Транспорт — еще одна область, где используется наука о данных. Во многих городах внедрены системы мониторинга и управления движением. Эти системы используют данные в реальном времени для управления потоком городского трафика. К примеру, они могут управлять переключением светофоров, отдавая приоритет общественному транспорту. Данные об использовании городских транспортных сетей полезны для их дальнейшего планирования. Изучая маршруты, расписания и движение транспортных средств, администрация добивается того, чтобы обслуживать максимальное количество людей, одновременно снижая затраты на предоставление транспортных услуг. Помимо моделирования сети общего пользования, наука о данных также применяется для мониторинга принадлежащих городу транспортных средств и обеспечения их оптимального использования. Датчики, установленные вдоль дорог, на светофорах и в других местах, собирают данные об условиях дорожного движения для оптимизации планирования и динамических корректировок маршрута, которые поступают на транспортные средства в режиме реального времени.

Помимо энергетики и транспорта, наука о данных используется в коммунальном хозяйстве и для долгосрочного планирования инфраструктурных проектов. Эффективность предоставляемых коммунальных услуг контролируется путем мониторинга их текущего, прогнозирования ожидаемого и изучения предыдущего потребления при аналогичных условиях. Коммунальные службы используют науку о данных по-разному. Например, для управления сетью коммунального снабжения, включая контроль поставок для коммунальных предприятий, их качества, оценку возникающих проблем, выявление областей,

требующих более интенсивного снабжения, автоматическое изменение маршрутов доставки и мониторинг любых аномалий в сети. Другой пример использования науки о данных коммунальными службами — мониторинг клиентов. Отклонения параметров потребления могут указывать на криминальную активность (например, наличие домашней плантации конопли), на нелегальную перенастройку измерительного оборудования, а также на клиентов, которые с большой вероятностью не будут платить по счетам. Науку о данных применяют и в городском планировании для поиска оптимального варианта застройки и сопутствующих ей услуг. Симуляции, основанные на моделях прогнозирования прироста населения, позволяют планировщикам оценить, когда и где понадобятся те или иные услуги, например общеобразовательные школы.

Проектные принципы науки о данных: почему одни проекты успешны, а другие нет

Порой проекты науки о данных терпят неудачу, поскольку не оправдывают ожиданий, увязают в технических или политических вопросах, не приносят полезных результатов и, как правило, после этого больше не запускаются. Подобно утверждению о счастливых семьях Льва Толстого[34], успех проекта науки о данных зависит от ряда факторов. Успешные проекты требуют целенаправленности, хорошего качества данных, нужных людей, готовности экспериментировать с несколькими моделями, интеграции в архитектуру и процессы ИТ-бизнеса, поддержки со стороны высшего руководства и признания организацией необходимости регулярного пересмотра моделей в силу меняющегося мира. Сбой в любом из этих аспектов может

привести к провалу всего проекта. Далее мы подробно опишем общие факторы, влияющие на успешность проектов науки о данных, а также типичные причины, которые приводят к их провалу.

Фокусировка. Каждый успешный проект науки о данных начинается с четкого определения проблемы, которую он должен помочь решить. Этот шаг подсказывает обычный здравый смысл — проекту сложно достичь успеха, если у него нет четкой цели. Наличие четкой цели определяет решения относительно того, какие данные и алгоритмы машинного обучения использовать, как оценивать результаты, как будут применяться анализ и развертываться модели и когда может потребоваться повторный процесс для обновления моделей.

Данные. Точно сформулированная задача позволяет определить, какие данные необходимы для проекта. Ясность в этом вопросе помогает направить проект туда, где эти данные находятся. Если какие-то данные в настоящее время недоступны, следует запустить вспомогательные проекты, которые изучат возможность сбора и доступность этих данных. При этом крайне важно обеспечить их высокое качество. Потеря качества данных может произойти в силу плохо спроектированных приложений или плохих моделей, имеющихся у организации, персонала, не обученного правильно вводить данные, или по иным причинам. На самом деле существует масса факторов, которые снижают качество данных в системах, а потребность в данных хорошего качества настолько важна, что некоторые организации нанимают специалистов, которые постоянно проверяют данные, оценивая их качество и внося предложения о его улучшении. Без качественных данных добиться успеха трудно.

Каждый успешный проект науки о данных начинается с четкого определения проблемы, которую он должен помочь решить.

Прежде чем привлекать сторонние источники данных, стоит проверить, какие данные уже собраны и используются в организации. К сожалению, подход некоторых наукоемких проектов заключается в том, чтобы сразу взять доступные данные из транзакционных баз или других источников, очистить и интегрировать их, а затем приступить к исследованию и анализу. Такой подход полностью игнорирует группу бизнес-аналитики и возможное наличие хранилища данных. Во многих организациях бизнес-аналитики и специалисты по организации хранилища данных уже собирают, очищают, трансформируют и интегрируют данные организации в один центральный репозиторий. Если хранилище уже существует, то, вероятно, оно содержит все или бóльшую часть данных, необходимых для проекта, что может сэкономить значительное время на их интеграцию и очистку. Кроме того, в хранилище будет гораздо больше данных, чем в текущих транзакционных базах. Используя хранилище данных, можно вернуться на несколько лет назад и построить прогнозные модели, а затем прокрутить их на разных временных периодах и измерить уровень точности прогнозов для каждой из моделей. Это позволяет отслеживать изменения в данных и их влияние на модели. Кроме того, можно отслеживать, как эти изменения происходят и развиваются с течением времени. Использование такого подхода облегчает демонстрацию поведения моделей в долгосрочном периоде, что помогает укрепить доверие клиентов. Например, в одном проекте на основе пятилетних исторических данных из хранилища было продемонстрировано, как именно компания могла сэкономить более $40 млн за этот период.

Люди. В успешных проектах науки о данных часто принимают участие люди, обладающие сочетанием разных компетенций и навыков. В большинстве организаций такие люди уже есть, они играют различные роли, но могут и должны внести свой вклад в

проекты науки о данных. К ним относятся специалисты по базам данных, по процессу ETL, по интеграции данных, менеджеры проектов, бизнес-аналитики, отраслевые эксперты и т.д. Иногда организациям необходимо нанимать специалистов, обладающих навыками работы с большими данными, применения машинного обучения и точной постановки задач. Успешные специалисты по данным должны быть готовы и способны сотрудничать и общаться с командой менеджеров, конечными пользователями и всеми заинтересованными сторонами, чтобы показать и объяснить, как именно наука о данных может помочь в их работе. Трудно найти людей, которые обладают одновременно необходимыми техническими навыками и способностью общаться и работать с людьми на разных уровнях организации. Тем не менее эта комбинация компетенций имеет решающее значение для успеха проектов науки о данных.

Модели. Экспериментируйте с различными алгоритмами машинного обучения, чтобы выяснить, какой из них лучше работает с набором данных. В литературе часто встречаются примеры использования одного-единственного алгоритма. Возможно, авторы заранее выяснили, какой алгоритм предпочтительнее в их случае, или же просто применили особо любимый алгоритм. Наибольший интерес вызывают нейронные сети и глубокое обучение, однако, как правило, существуют и другие алгоритмы, которые следует рассмотреть и протестировать. Кроме того, на выбор алгоритмов и моделей в проектах науки о данных, ведущихся на территории ЕС, может повлиять Общий регламент по защите данных (GDPR), который вступил в силу с апреля 2018 г. Статьи, затрагивающие «право человека на объяснение» автоматизированных процессов принятия решений, могут ограничить использование в ряде областей сложных моделей, трудно поддающихся интерпретации и объяснению.

В успешных проектах науки о данных часто принимают участие люди, обладающие сочетанием разных компетенций и навыков.

Интеграция с бизнесом. При определении цели проекта науки о данных важно определить, каким образом выходные данные и результаты проекта будут развернуты в ИТ-архитектуре и бизнес-процессах организации. Для этого нужно определить, где и как модель должна быть интегрирована в существующие приложения и как полученные результаты будут использоваться его конечными пользователями или передаваться в другой процесс. Чем более автоматизирован процесс, тем быстрее организация сможет реагировать на изменения профилей своих клиентов, снижая затраты и увеличивая потенциальную прибыль. К примеру, модель определения уровня риска клиента для выдачи банковских кредитов должна быть встроена во внешнее приложение, которое обрабатывает кредитные заявки от клиентов. Таким образом, когда сотрудник банка вводит заявку на кредит, он сразу получает обратную связь от модели. В дальнейшем эта обратная связь может использоваться в реальном времени для решения любых возникших вопросов с клиентом. Другой пример — обнаружение мошенничества. Может потребоваться от четырех до шести недель, чтобы выявить случай потенциального мошенничества, требующий расследования. Используя науку о данных и встраивая ее в системы мониторинга транзакций, компании могут выявлять потенциальные случаи мошенничества в реальном времени. Благодаря автоматизации и интегрированию моделей на основе данных уменьшается время отклика, и действия могут быть предприняты своевременно. Если выходные данные и модели, созданные проектом, не интегрированы в бизнес-процессы, то они просто не будут использоваться, и в итоге проект потерпит неудачу.

Поддержка проекта. Поддержка со стороны высшего руководства имеет решающее значение для успеха большинства проектов науки о данных. Однако старшие ИТ-менеджеры

бывают слишком сосредоточены на происходящем здесь и сейчас, следя за работой повседневных приложений, наличием резервных копий, проверяя процессы восстановления и корректируя приложения на будущее. В успешных проектах науки о данных часто спонсорами выступают старшие бизнес-руководители, а не ИТ-менеджеры. Преимущество этого состоит в том, что бизнес-руководители сосредоточены не на технологии, а на процессах, происходящих вокруг проекта, и на том, как можно использовать его результаты. Чем более сфокусирован на этом спонсор проекта, тем успешнее будет проект. По его завершении такой спонсор станет ключом к информированию остальной части организации об успехе проекта. Но даже когда в проекте в качестве лидера задействован старший руководитель, общая стратегия науки о данных в компании в долгосрочной перспективе может потерпеть неудачу, если начальные проекты будут восприняты как нечто «для галочки». Организация не должна рассматривать науку о данных как разовые проекты. Чтобы получить долгосрочные выгоды, необходимо создать потенциал для науки о данных на постоянной основе, а также использовать результаты ее проектов. Это требует долгосрочных обязательств со стороны высшего руководства и принятия науки о данных как части стратегии.

Итерация. Большинство проектов науки о данных требуют более или менее регулярных обновлений и актуализации. При каждом обновлении или итерации процесса можно добавлять новые данные, корректировки, а возможно, и новые алгоритмы. Модели оттока необходимо обновлять на регулярной основе. Частота этих итераций будет варьироваться от проекта к проекту, от ежедневных до одного раза каждые 3, 4, 6 или 12 месяцев. Для определения необходимости обновления моделей может быть встроен контроль генерируемых выходных данных.

Мысли напоследок

Люди всегда абстрагировались от мира и пытались понять его, выявляя закономерности в собственном опыте. Наука о данных — последнее воплощение этого поиска, этой модели поведения. И хотя она имеет такую долгую предысторию, сила ее влияния на современную жизнь беспрецедентна. Слова «точный», «умный», «целевой» и «персонализированный» являются частью отраслевых названий науки о данных: *точная медицина, точный полицейский контроль, точное сельское хозяйство, умные города, умный транспорт, целевая реклама, персонализированные развлечения*. Все эти сферы человеческой жизни объединяет необходимость принятия решений. Какое лечение использовать для этого пациента? Как распределить полицейские ресурсы? Сколько удобрений нужно внести? Сколько школ необходимо построить в ближайшие четыре года? Кому мы должны отправить это дополнение? Какой фильм или книгу порекомендовать этому человеку? Именно наука о данных помогает принимать такие решения. Успешный проект науки о данных обеспечивает актуальное понимание вопроса, которое помогает принять наилучшее решение и достигнуть наилучших результатов.

Наука о данных в ее современном виде представляет собой смесь больших данных, компьютерных мощностей и человеческой изобретательности в целом ряде технологических областей (от глубинного анализа данных и исследования баз до машинного обучения). Эта книга призвана дать обзор основных идей и концепций, которые необходимы для понимания науки о данных. Жизненный цикл проекта CRISP-DM делает процесс обработки данных открытым и обеспечивает структуру для перехода от данных к мудрости: формулируйте проблему, подготавливайте данные, используйте машинное обучение для выявления закономерностей и создания моделей, применяйте

модели для проникновения в суть. В книге также затрагиваются этические проблемы, связанные с конфиденциальностью. У нас есть искренние и обоснованные опасения, что наука о данных может быть использована правительствами и/или заинтересованными лицами для манипулирования нашим поведением и контроля над нашими действиями. Нам необходимо выработать обоснованное мнение о том, в каком мире мы хотим жить, и подумать о законах, которые бы направили науку о данных в соответствующих направлениях. Говоря о будущем, при всех возможных этических проблемах джинн уже выпущен из бутылки: наука о данных оказывает и будет оказывать существенное влияние на нашу повседневную жизнь. При правильном использовании она сможет улучшить ее. Но для того чтобы организации, в которых мы работаем, сообщества и семьи, в которых мы живем, получали выгоду от науки о данных, нам нужно понять и изучить, что она собой представляет, как работает, что умеет и чего не умеет. Мы надеемся, что эта книга поможет вам в этом.

ГЛОССАРИЙ

CRISP-DM
Межотраслевой стандартный процесс, определяющий жизненный цикл проекта исследования данных. Часто используется в науке о данных.

Пирамида DIKW (DIKW Pyramid)
Модель структурных отношений между данными, информацией, знаниями и мудростью. В пирамиде DIKW данные предшествуют информации, которая предшествует знаниям, которые предшествуют мудрости.

Hadoop
Платформа с открытым исходным кодом, разработанная Apache Software Foundation, предназначенная для обработки больших данных. Использует распределенное хранение и обработку по кластерам аппаратного обеспечения.

OLAP — интерактивная аналитическая обработка
Операции OLAP генерируют сводки исторических данных и включают агрегирование данных из нескольких источников. Они предназначены для генерации сводок по типам отчетов и позволяют пользователям распределять, фрагментировать и переворачивать данные в хранилище, используя предопределенный набор атрибутов, например продажи по магазинам, продажи по кварталам.

SQL — язык структурированных запросов
Международный стандарт для определения запросов к базе данных.

Анализ данных (Data Analysis)
Общий термин, используемый для описания любого процесса извлечения полезной информации из данных. Типы анализа данных включают визуализацию, сводную статистику, корреляционный анализ и моделирование с использованием машинного обучения.
Аналитическая базовая таблица (Analytics Base Table, ABT)
Таблица, в которой каждая строка содержит данные, относящиеся к конкретному объекту, а каждый столбец — параметры определенного атрибута объектов в таблице. Это основной способ ввода информации для глубинного анализа данных и алгоритмов машинного обучения.

Атрибут (Attribute)
Каждый объект набора данных описывается рядом атрибутов (также называемых признаками или переменными). Атрибут фиксирует один фрагмент данных, относящихся к объекту. Атрибут может быть базовым или производным.

База данных (Database)
Центральное хранилище данных. Наиболее распространена реляционная структура базы данных, которая хранит данные в таблицах, где каждая строка отведена одному

объекту, а каждый столбец — одному атрибуту. Это представление идеально подходит для хранения данных с четкой структурой, которые могут быть разложены на базовые атрибуты.

Большие данные (Big Data)
Большие данные часто определяют как «3V»: экстремальный объем (**V**olume), разнообразие типов (**V**ariety) и скорость обработки данных (**V**elocity).

Высокопроизводительные вычисления (High Performance Computing, или HPC)
Нацелены на разработку и реализацию моделей для объединения большого количества компьютеров в кластер, способный эффективно хранить и обрабатывать большие объемы данных.

Выхлопные данные (Exhaust Data)
Данные, являющиеся побочным продуктом процесса, основной целью которого является нечто иное, чем сбор данных. Например, для каждого перепоста, ретвита или лайка в соцсетях создается ряд «выхлопных данных»: кто поделился, кто просмотрел, какое устройство использовалось, какое время суток и т.д. (В отличие от намеренно собранных данных.)

Выявление аномалий (Anomaly Detection)
Включает поиск и идентификацию экземпляров данных, которые являются нетипичными в наборе. Эти отклонения часто называют аномалиями или выбросами. Часто применяется при анализе финансовых транзакций для обнаружения потенциальных мошеннических действий и запуска расследований.

Глубинный анализ данных (Data Mining)
Процесс выявления в наборах данных полезных закономерностей для решения конкретной проблемы. CRISP-DM определяет стандартный жизненный цикл проекта глубинного анализа данных. Тесно связан с наукой о данных, но охватывает меньший круг задач.

Глубокое обучение (Deep Learning)
Модель глубокого обучения — это нейронная сеть, которая имеет несколько (больше двух) слоев скрытых элементов (или нейронов). Глубокие сети являются глубокими именно в смысле количества слоев нейронов. Сегодня большинство глубоких сетей имеют от 10 до 100 слоев. Сила глубокого обучения состоит в том, что на более поздних уровнях нейроны способны изучать производные атрибуты, составляя их из атрибутов, изученных нейронами на более ранних уровнях.

Данные (Data)
В самом общем смысле данные — это набор характеристик (или измерение) некоей реальной сущности (человека, объекта или события).

Дерево решений (Decision Tree)
Тип модели прогнозирования, которая кодирует правила условного оператора (если — тогда — иначе) в древовидной структуре. Каждый узел дерева определяет один атрибут для тестирования, и объект должен пройти путь от корневого узла до конечного, чтобы метка конечного узла в дальнейшем могла быть предсказана для этого объекта.

Интернет вещей (Internet of Things, IoT)
Межсетевой обмен информацией между физическими устройствами и датчиками. Включает в себя область разработки «машина — машина» (м2м) по созданию систем, которые не только позволяют машинам обмениваться информацией, но и реагировать на нее, инициируя действия без участия человека.

Классификация (Classification)
Задача прогнозирования значения целевого атрибута объекта на основе набора значений входных атрибутов, где целевой атрибут отражает номинальный или порядковый тип данных.

Кластеризация (Clustering)
Выявление групп схожих объектов в наборе данных.

Обучение с учителем (Supervised Learning)
Форма машинного обучения, целью которой является изучение функции, отображаемой набором значений входных атрибутов объекта для вычисления отсутствующего значения целевого атрибута того же объекта.

Корреляция (Correlation)
Описывает силу, связывающую атрибуты.

Линейная регрессия (Linear Regression)
Когда в регрессионном анализе предполагается линейная зависимость, анализ называется линейной регрессией. Этот термин часто используется для описания моделей прогнозирования машинного обучения, которые применяют этот вид анализа для вычисления значения числового целевого атрибута.

Машинное обучение (Machine Learning)
Область компьютерных исследований, которая фокусируется на разработке и оценке алгоритмов, способных выявлять полезные закономерности в наборах данных. Алгоритм машинного обучения принимает на вход набор данных и возвращает модель, которая кодирует закономерности, выявленные алгоритмом.

Машинное обучение в базе данных (In-Database Machine Learning)
Использование алгоритмов машинного обучения, встроенных в решение для базы данных. Преимущество машинного обучения в базе данных состоит в том, что оно сокращает время, затрачиваемое на перемещение данных для анализа.

Метаданные (Metadata)
Данные, описывающие структуры и свойства других данных, например, временна́я метка, которая содержит информацию о том, когда фрагмент данных был собран. Метаданные являются одним из наиболее распространенных типов данных о выбросах.

Набор данных (Dataset)
Совокупность данных, относящихся к набору объектов, каждый из которых описан в терминах набора атрибутов. В своей основной форме набор данных организован в виде матрицы $n \times m$, где n — количество объектов (строк), а m — количество атрибутов (столбцов).

Наука о данных (Data Science)
Развивающаяся область знаний, которая использует набор алгоритмов, процессов и методов постановки проблемы для анализа больших данных с целью извлечь из них полезную информацию. Тесно связана с глубинным анализом данных, но имеет более широкую сферу применения и круг проблем. Занимается анализом как структурированных, так и неструктурированных больших данных и базируется на принципах целого ряда научных отраслей, включая машинное обучение, статистику, высокопроизводительные вычисления, а также этические вопросы использования данных и их регулирование.

Нейрон (Neuron)
Нейрон принимает на вход несколько значений (или активаций) и отображает их в качестве выходного сигнала. Это отображение обычно обеспечивается функцией линейной регрессии, примененной к входным данным, и последующим выводом результата этой функции через нелинейные функции активации, такие как логистическая функция или функция TANH.

Нейронная сеть (Neural Network)
Тип модели машинного обучения, которая реализована в виде сети процессорных блоков, называемых нейронами. Можно создавать различные типы нейронных сетей, изменяя в них топологию нейронов. Наиболее часто встречаются полностью подключенные нейронные сети с прямой связью, которые обучают методом обратного распространения ошибки.

Обучение без учителя (Unsupervised Learning)
Форма машинного обучения, целью которой является выявление закономерностей в базе данных, которые включают кластеры похожих объектов или регулярность атрибутов. В отличие от контролируемого обучения в наборе данных не определен целевой атрибут.

Необработанный атрибут (Raw Attribute)
Абстракция сущности, которая является ее прямым измерением, например рост человека (в отличие от производного атрибута).

Неструктурированные данные (Unstructured Data)
Данные, где каждый объект в наборе может иметь собственную внутреннюю структуру, отличающуюся от внутренних структур других объектов. Например, текстовые данные часто не структурированы и требуют, чтобы к ним применялась последовательность операций для извлечения структурированного представления каждого объекта.

Объект (Instance)
Каждая строка в наборе данных содержит информацию, относящуюся к одному объекту (также называемому экземпляром, сущностью, случаем или записью).

Поиск ассоциативных правил (Association Rule Mining)
Техника анализа данных при неконтролируемом обучении, которая ищет группы элементов, часто встречающихся вместе. Классическим примером использования является анализ рыночной корзины, когда розничные компании пытаются

идентифицировать наборы товаров, которые часто покупают вместе, к примеру хот-дог, кетчуп и пиво.

Прогнозирование (Prediction)
В контексте науки о данных и машинного обучения прогнозирование — это задача вычисления значения целевого атрибута для данного объекта на основе значений других атрибутов (или входных атрибутов) этого же объекта.

Производный атрибут (Derived Attribute)
Атрибут, значение которого генерируется путем применения функции к данным, а не путем прямого измерения объекта (в отличие от базового атрибута). Примером производного атрибута является атрибут, который описывает среднее значение выборки.

Регрессионный анализ (Regression Analysis)
Вычисляет ожидаемое (или среднее) значение числового целевого атрибута при всех заданных значениях входного атрибута. Регрессионный анализ предполагает параметризованную математическую модель гипотетической взаимосвязи между входами и выходами, известную как функция регрессии. Функция регрессии может иметь множество параметров, и целью регрессионного анализа является поиск правильных настроек для них.

Собранные данные (Captured Data)
Данные, которые зафиксированы непосредственно в процессе сбора данных (в отличие от аномалий).

Структурированные данные (Structured Data)
Данные, которые могут храниться в таблице, каждый объект которой имеет одинаковый набор атрибутов (в отличие от неструктурированных данных).

Транзакционные данные (Transactional Data)
Включают информацию о событиях, таких как продажа товара, выставление счета, доставка груза, оплата кредитной картой и т.д.

Умный город (Smart City)
Проекты умных городов, как правило, пытаются интегрировать данные в режиме реального времени из множества различных источников в единый центр данных, где они анализируются и используются для принятия управленческих решений и планирования.

Хранилище данных (Data Warehouse)
Централизованный репозиторий, который содержит данные из разных источников со всех уровней организации. Данные структурированы так, чтобы поддерживать генерацию сводных отчетов. Интерактивная аналитическая обработка (OLAP) — термин, используемый для описания типичных операций в хранилище данных.

Целевой атрибут (Target Attribute)
В задаче прогнозирования целевой атрибут — это атрибут, которому модель прогнозирования обучается для вычисления значений.

[1] Нильсон, Н. Дж. Обучающиеся машины. — М.: Мир, 1967.

[2] Цитата взята из приглашения на семинар «KDD — 1989». — *Здесь и далее прим. авт.*

[3] Некоторые специалисты все же проводят границу между глубинным анализом данных и KDD, рассматривая первый как подраздел второго и определяя его как один из методов обнаружения знаний в базах данных.

[4] https://www.cancer.gov/research/key-initiatives.

[5] https://allofus.nih.gov/.

[6] https://www.policedatainitiative.org/.

[7] Льюис М. MoneyBall. — М.: Манн, Иванов и Фербер, 2013.

[8] Дабнер С., Левитт С. Фрикономика. — М.: Альпина Паблишер, 2018.

[9] https://deepmind.com/research/alphago/.

[10] Хотя многие наборы данных можно описать как плоскую матрицу $n \times m$, существуют сценарии, в которых набор данных представлен в более сложной форме: например, если набор

данных описывает эволюцию нескольких атрибутов во времени, то каждый момент времени в наборе данных будет представлен двухмерной плоской матрицей $n \times m$, перечисляющей состояние атрибутов в данный момент времени, но общий набор данных будет трехмерным, где время используется для связывания двумерных срезов момента. В таком контексте термин «тензор» иногда используется для придания идее матрицы дополнительного измерения.

[11] Скрапинг (англ. scraping) — в широком смысле сбор данных с интернет-ресурсов. — *Прим. пер*.

[12] Интерпретация высказывания Джорджа Бокса: «По сути, все модели ошибочны, но некоторые бывают полезны».

[13] Для числового целевого атрибута наиболее распространенным показателем центральной тенденции является среднее значение, а для номинальных или порядковых данных — диапазон (или наиболее часто встречающееся значение).

[14] Здесь мы используем более сложную запись, включающую и , поскольку далее мы будем расширять эту функцию и включать в нее более одного входного атрибута, а для этого понадобятся индексированные переменные.

[15] Предостережение: приведенные здесь числовые значения следует воспринимать только как иллюстрацию, а не как окончательные оценки взаимосвязи между ИМТ и вероятностью развития диабета.

[16] Обычно нейронные сети работают лучше, когда все входные данные имеют небольшие значения. Если заданы широкие диапазоны входных атрибутов, то атрибуты с большими значениями имеют тенденцию доминировать при обработке сетью. Чтобы этого не происходило, лучше всего нормализовать входные атрибуты под одинаковые диапазоны.

[17] МПК — наибольшее количество кислорода, выраженное в миллилитрах, которое человек способен потреблять в течение одной минуты.

[18] Для простоты мы не стали обозначать вес связей на рис. 14.

[19] Не существует единого мнения относительно минимального количества скрытых слоев, необходимых для того, чтобы сеть считалась глубокой. Некоторые полагают, что для этого достаточно даже двух слоев. Однако большинство глубоких сетей имеют десятки слоев, а некоторые — сотни и даже тысячи.

[20] Доступное введение в РНС, а также об их использовании при обработке естественного языка см. [2] по адресу: https://tinyurl.com/RecurrentNeuralNetworks.

[21] Технически это известно как проблема исчезающего градиента, поскольку градиент стремится к нулю при реализации алгоритма обратного распространения.

[22] Существует два особых случая, которые также завершают алгоритм: ветвь сворачивается в отсутствие объектов после

разделения набора данных или все входные атрибуты уже были использованы в узлах между корнем и ветвью. В обоих случаях добавляется завершающий узел, который помечается доминирующим значением целевого атрибута в родительском узле ветви.

[23] Для ознакомления с энтропией и ее использованием в алгоритмах дерева решений см. [4] по адресу: http://www.machinelearningbook.com.

[24] Подробное тематическое исследование на тему оттока клиентов (Kelleher, Mac Namee, D'Arcy 2015) можно найти по адресу: http://www.machinelearningbook.com.

[25] При проведении сетевого маркетинга рекламная кампания распространяется на широкий спектр веб-сайтов без узкого таргетинга на пользователей.

[26] В поведенческом таргетинге используются данные об онлайн-активности пользователей — посещении страниц, кликах, времени, проведенном на сайте, и т.д. — и прогнозное моделирование для выбора рекламных объявлений, показываемых пользователю.

[27] Директива ЕС о конфиденциальности и электронных коммуникациях [2002/58/EC].

[28] Некоторые женщины, впрочем, открыто сообщают ритейлерам, что они беременны, регистрируясь в программах лояльности для будущих мам.

[29] http://www.predpol.com/.

[30] Паноптикум — проект, разработанный в XVIII в. юристом Джереми Бентамом для тюрем и психиатрических больниц. Отличительная особенность паноптикума состоит в том, что персонал может постоянно вести наблюдение без ведома заключенных. Основная идея этого проекта в том, чтобы заставить заключенных вести себя так, будто они находятся под постоянным наблюдением.

[31] EMEA (European, the Middle East and Africa) — Европа, Средний Восток и Африка.

[32] http://smartsantander.eu/.

[33] http://www.tepco.co.jp/en/press/corp-com/release/2015/1254972_6844.html.

[34] Роман Льва Толстого «Анна Каренина» начинается фразой: «Все счастливые семьи похожи друг на друга, каждая несчастливая семья несчастлива по-своему». Идея Толстого заключается в том, что для достижения счастья семья должна быть успешной по ряду критериев (любовь, финансы, здоровье, родственники), но неудача в любом из этих аспектов ведет к несчастью семьи. Таким образом, все счастливые семьи одинаковы, поскольку успешны по всем критериям, а несчастливые могут стать таковыми по разным причинам.

Переводчик *Михаил Белоголовский*
Научный редактор *Заур Мамедьяров*
Главный редактор *С. Турко*
Руководитель проекта *А. Василенко*
Корректоры *Е. Аксенова, Т. Редькина*
Компьютерная верстка *А. Абрамов*
Художественное оформление и макет *Ю. Буга*
Иллюстрация на обложке *shutterstock.com*

Права на публикацию на русском языке получены при содействии Агентства Александра Корженевского (Москва).
© 2018 Massachusetts Institute of Technology
© Издание на русском языке, перевод, оформление. ООО «Альпина Паблишер», 2020
© Электронное издание. ООО «Альпина Диджитал», 2020

Келлехер Дж.
Наука о данных: Базовый курс / Джон Келлехер, Брендан Тирни; Пер. с англ. — М.: Альпина Паблишер, 2020.

ISBN 978-5-9614-3378-4