ГЛАВА 8

Ищите скрытые группы

«Если вы проанализируете данные достаточно тщательно, то сможете отыскать послания Бога»

_ Дилберт[71]

Представьте, что вам звонит друг и просит помочь категоризовать его музыкальную коллекцию, представляющую собой набор винтажных виниловых пластинок. Вы соглашаетесь.

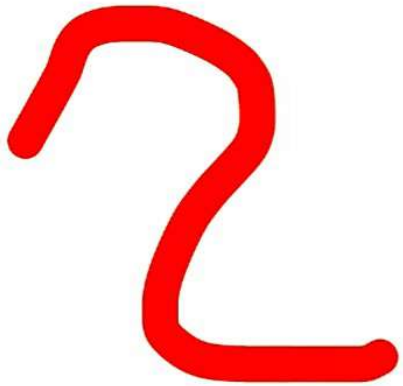


Табл. 8.3. Обучение без учителя. Резюме

Обучение без учителя	Снижение размерности	Кластеризация
Пример	Анализ главных компонент	Метод *k*-средних
Что это?	Группировка и объединение столбцов (признаков)	Группировка строк (наблюдений)
Что делает?	Находит меньший набор новых некоррелированных признаков, содержащий большую часть информации в наборе данных.	Группирует похожие наблюдения, создавая *k*-значимых «кластеров» в данных.
Зачем?	Это позволяет вам визуализировать и исследовать данные или уменьшить размер набора данных для ускорения процесса вычислений. Как правило, АГК является промежуточным этапом анализа.	Данный метод позволяет выявить закономерности и структуру данных и дает возможность по-разному воздействовать на кластеры (например, запускать разные маркетинговые кампании для разных сегментов рынка).
Необходимый контроль	Пользователь должен решить, как масштабировать данные, сколько главных компонент оставить и как их интерпретировать.	Пользователь должен решить, как масштабировать данные, а также выбрать подходящую метрику «расстояния» и необходимое количество кластеров.

Возможность какой-либо группировки данных зависит от выбранного алгоритма, его реализации, качества исходных данных и существующей в них вариации. Это означает, что принятие разных решений может приводить к созданию разных групп. Проще говоря, обучение без учителя требует

ГЛАВА 9

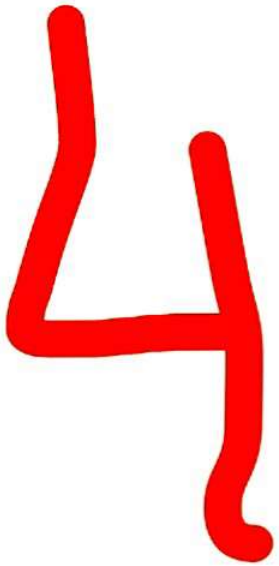
Освойте модели регрессии

«Регрессионный анализ похож на один из тех изощренных мощных инструментов, который относительно легко использовать, но сложно делать это правильно. А его неправильное использование потенциально опасно»

_ Чарльз Уилан, цитата из книги «Голая статистика»[81]

ОБУЧЕНИЕ С УЧИТЕЛЕМ

Предыдущая глава была посвящена обучению без учителя — способу обнаружения закономерностей или кластеров в наборе данных без использования заранее определенных групп. Помните, что к неконтролируемому обучению



ОБУЧЕНИЕ С УЧИТЕЛЕМ

Предыдущая глава была посвящена обучению без учителя — способу обнаружения закономерностей или кластеров в наборе данных без использования заранее определенных групп. Помните, что к неконтролируемому обучению мы подходим без каких-либо предвзятых представлений. Вместо этого мы опираемся на основополагающие аспекты данных, задаем некоторые границы и позволяем данным организоваться самим.

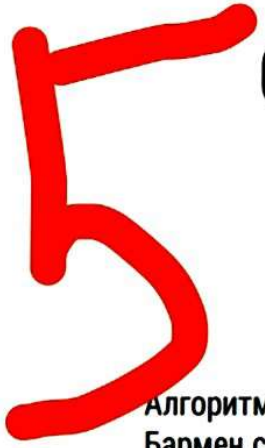
Однако во многих случаях о наборе данных что-то известно. Тогда вы можете использовать обучение с учителем или контролируемое обучение для выявления в нем взаимосвязей с помощью входных и известных выходных данных. В данном случае у вас есть правильные ответы, на которых вы можете «учиться». Затем вы можете оценить надежность модели, сравнив ее результаты с тем, что вам известно о реальном мире. Хорошая модель позволит вам делать точные прогнозы и объяснять некоторые основополагающие взаимосвязи между входными и выходными данными.

---

81 «Голая статистика. Самая интересная книга о самой скучной науке», Чарльз Уилан (Издательство: Манн, Иванов и Фербер, 2022).

ГЛАВА 10

Освойте модели классификации



Алгоритм машинного обучения заходит в бар.
Бармен спрашивает: «Что будете?»
Алгоритм отвечает: «А что заказали остальные?»

_ *Чет Хаасе (@chethaase)*

В предыдущей главе мы говорили о контролируемом обучении с помощью моделей регрессии, которые позволяют предсказывать численные значения (вроде объема продаж) путем подгонки модели к набору признаков. Но что, если вам требуется предсказать конкретный результат? Например, захочет ли человек, обладающий определенным набором

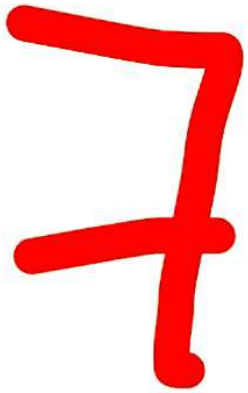
ГЛАВА 11

Освойте текстовую аналитику

«Стремитесь к успеху, но готовьтесь к овощам»

_ *InspireBot™ бот на основе искусственного интеллекта, «предназначенный для создания неограниченного количества уникальных вдохновляющих цитат»*[104]

В нескольких предыдущих главах мы говорили о данных в их традиционном понимании. Для большинства людей наборы данных представляют собой таблицы, состоящие из строк и столбцов. Это структурированные данные. Однако в реальном мире большая часть данных, с которыми вы взаимодействуете каждый день, является неструктурированной. Эти данные содержатся в текстах, которые вы читаете, в словах и предложениях электронных



КАК ТЕКСТ ПРЕВРАЩАЕТСЯ В ЧИСЛА

Читая текст, люди понимают настроение, сарказм, намеки, нюансы и смысл. Иногда это даже невозможно объяснить: стихотворение вызывает в памяти воспоминание, шутка заставляет смеяться.

Так что совсем не удивительно, что компьютер не понимает смысла так же, как это делает человек. Компьютеры могут лишь «видеть» и «считывать» числа. Чтобы проанализировать массу неструктурированных текстовых данных, их необходимо сначала преобразовать в числа и уже знакомые вам структурированные наборы данных. Это преобразование неструктурированного и запутанного текста, содержащего орфографические ошибки, сленг, смайлики или аббревиатуры, в аккуратный структурированный набор данных из строк и столбцов может быть весьма субъективным и трудоемким процессом. Сделать это можно несколькими способами; три из них мы рассмотрим далее.

Большой мешок слов

Самый простой способ преобразования текста в числа предполагает создание модели «мешка слов», которая игнорирует порядок слов и грамматику. В результате фраза «Это предложение является очень большим мешком слов» преобразуется в набор, называемый документом, в котором каждое слово является идентификатором, а количество слов — признаком. Порядок слов не имеет значения, поэтому мы сортируем содержимое мешка по алфавиту: {большим: 1, мешком: 1, очень: 1, предложение: 1, слов: 1, это: 1, является: 1}.

Преимущества технологических гигантов

В отличие от многих других компаний, такие технологические гиганты, как Apple, Amazon, Google и Microsoft, обладают обилием текстовых и голосовых данных (данных, снабженных метками, которые можно использовать для контролируемого обучения моделей), мощными компьютерами, группами преданных делу исследователей мирового уровня и деньгами.

Благодаря таким ресурсам они добились значительного прогресса в области анализа не только текста, но и звука. В последние годы произошли заметные улучшения в следующих сферах:

— Преобразование речи в текст. Голосовые помощники и функции преобразования голоса в текст на смартфонах стали работать более точно.
— Преобразование текста в речь. Голоса в программах для чтения с экрана компьютера теперь больше напоминают человеческие.
— Преобразование текста в текст. Перевод с одного языка на другой выполняется мгновенно и с достаточно высокой точностью.
— Чат-боты. Окна чата, которые теперь автоматически открываются на каждом веб-сайте с вопросом: «Чем я могу вам помочь?», стали (чуть) более полезными.
— Генерация понятного человеку текста. Языковая модель GPT-3[120] от компании OpenAI способна генерировать текст, напоминающий человеческий, отвечать на вопросы, а также генерировать компьютерный код по запросу. На момент написания этой книги данная модель самая продвинутая в своем роде. Согласно оценкам, стоимость ее обучения (здесь имеется в виду только использование компьютеров без

— Генерация понятного человеку текста. Языковая модель GPT-3[120] от компании OpenAI способна генерировать текст, напоминающий человеческий, отвечать на вопросы, а также генерировать компьютерный код по запросу. На момент написания этой книги данная модель самая продвинутая в своем роде. Согласно оценкам, стоимость ее обучения (здесь имеется в виду только использование компьютеров без учета оплаты труда исследователей) составила 4,6 миллиона долларов США[121].

Добавьте к этому наличие доступа к данным и группы экспертов-исследователей, и вы поймете, почему обработка естественного языка (пока) остается недоступной большинству компаний. Хотя алгоритмы имеют открытый исходный код, массовый сбор данных и доступ к суперкомпьютерам остается прерогативой технологических гигантов.

---

[120] Generative Pre-trained Transformer 3

[121] **https://www.forbes.com/sites/bernardmarr/2020/10/05/what-is-gpt-3-and-why-is-it-revolutionizing-artificial-intelligence/?sh=2f45a93b481a**

ГЛАВА 12

10

Концептуализируйте глубокое обучение

«Появление искусственного интеллекта иногда называют новой промышленной революцией. И если глубокое обучение — это паровой двигатель этой революции, то данные — это уголь: топливо, питающее наши интеллектуальные машины, без которого ничего не было бы возможно»

_ *Франсуа Шолле, исследователь ИИ и автор книг*[122]

Поздравляем: вам удалось добраться до главы, которая во многих отношениях является кульминацией вашего пути становления главным по дан-

11

Поздравляем: вам удалось добраться до главы, которая во многих отношениях является кульминацией вашего пути становления главным по данным. В ней мы соберем вместе различные фрагменты мозаики и погрузимся в развивающуюся область машинного обучения, называемую глубоким обучением.

Сегодня использование глубокого обучения стимулирует развитие передовых технологий, а его человекоподобные проявления периодически вызывают восхищение общественности. Сфера глубокого обучения охватывает технологии, лежащие в основе работы систем распознавания лиц, автономного вождения, обнаружения рака и перевода речи. То есть они помогают принимать решения, которые некогда считались прерогативой человека. Однако, как будет показано далее, глубокое обучение не является чем-то новым и не настолько похоже на работу человеческого разума, как может показаться на первый взгляд.

---

[122] Шолле Франсуа, «Глубокое обучение на Python» (Издательство: Питер, 2018).

ний и более широких последствиях применения моделей типа «черный ящик».

НЕЙРОННЫЕ СЕТИ

Прежде чем концептуализировать глубокое обучение, сначала необходимо познакомиться с его строительными блоками — искусственными нейронными сетями.

Чем нейронные сети похожи на мозг?

Человеческий мозг — это сеть, состоящая из биологических нейронов. Считается, что эти нейроны «поглощают информацию» в виде химических сигналов и электрических импульсов. В определенный момент — мы не до конца понимаем, в какой именно — эта информация «активирует» нейрон, то есть заставляет его среагировать. Если вы ведете машину, и на дорогу внезапно выбегает олень, ваш мозг быстро обрабатывает входные данные (вашу скорость, расстояние до оленя, присутствие машин поблизости), активируя миллионы нейронов, которые, в свою очередь, принимают решение (нажать на тормоз или свернуть с дороги)[123].

[123] Разумеется, продемонстрировать резкие и ожидаемые изменения в химии мозга можно не только с помощью такого экстремального примера, как выбегающий на дорогу олень. Дело в том, что ваш мозг обрабатывает входные и выходные данные прямо сейчас. Миллионы нейронов активируются в процессе чтения этих строк.

13

Трансферное обучение, или как работать с небольшими наборами данных

При наличии некоторого количества размеченных данных, например, не тысяч, а сотен изображений, вашу команду может выручить так называемое трансферное обучение.

Идея трансферного обучения заключается в загрузке модели, обученной распознавать такие повседневные объекты, как воздушные шары, кошки, собаки и так далее[137]. Это означает, что тысячи значений параметров в сети были оптимизированы для работы с группой изображений. Как вы помните, первые слои нейронных сетей, обученных на изображениях, изучают такие общие представления, как формы и линии. А последующие, более глубокие слои соединяют эти края и линии, формируя ожидаемое выходное изображение.

Суть трансферного обучения — выделить несколько последних слоев, которые изучают то, как линии и края образуют, например, изображения кошек и собак, и заменить их новыми слоями, которые в результате очередного раунда обучения становятся способны объединять эти формы в очертания опухолей на медицинских изображениях. Имейте в виду то, что трансферное обучение может уменьшить количество размеченных изображений в 10 раз, но оно не позволяет обойтись несколькими десятками.

---

[137] Многие практики используют для трансферного обучения модели, обученные на базе данных ImageNet (**https://ru.wikipedia.org/wiki/ImageNet**).

параметрами и общей архитектурой сети. При построении большой сети старайтесь не допускать ее переобучения; в этом вам помогут уроки из глав 9 и 10.

14

Глубокое обучение для практиков

Если вы хотите научиться самостоятельно создавать модели глубокого обучения, мы настоятельно рекомендуем серию книг Франсуа Шолле, посвященных использованию библиотеки глубокого обучения Keras для языков R и Python.

- Шолле Ф. «Глубокое обучение на Python» (Издательство: Питер, 2018).
- Шолле Ф. и Аллер Дж. Дж. «Глубокое обучение на R». (Издательство: Питер, 2018).

ИСКУССТВЕННЫЙ ИНТЕЛЛЕКТ И ВЫ

В завершение этой главы мы хотим немного поговорить об искусственном интеллекте (ИИ) и его приложениях. Как главный по данным, вы должны знать о существовании двух типов искусственного интеллекта. Первый из них — общий искусственный интеллект (ОИИ), призванный воспроизвести процесс человеческого познания. Здесь вы можете вспомнить свой любимый научно-фантастический фильм. Однако прогресс в области ОИИ столь незначительный, что поводов для беспокойства пока нет.

Тем не менее значительный прогресс был достигнут в области искусственного интеллекта узкого назначения (или слабого ИИ). Она охватывает ком-

15

Некоторые люди используют термин ИИ более свободно, чем другие. Например, в обществе принято называть систему рекомендаций фильмов искусственным интеллектом, тогда как в ее основе лежит скорее машинное или статистическое обучение. Почему это важно? Дело в том, что понимание того, что создание «ИИ», о котором говорится в новостях, требует больших наборов данных, собранных у таких людей, как вы и я, ставит вопрос о качестве этих данных, изменчивости, возможной утечке, переобучении и множестве других практических проблем. ИИ усиливает закономерности, содержащиеся в данных, собранных в прошлом; речь не идет о создании чего-то напоминающего человеческое сознание.

Рис. 12.9. Глубокое обучение — это подраздел машинного обучения, которое является подразделом искусственного интеллекта