Глава 9

Подготовка данных

В ЭТОЙ ГЛАВЕ...

- Документирование бизнес-целей
- Обработка данных
- Выборка данных
- Преобразование данных
- Извлечение признаков
- Выбор признаков

План построения успешной прогностической модели включает в себя определение бизнес-целей, подготовку данных, а также построение и развертывание модели. В этой главе рассматривается подготовка данных, в частности следующие процедуры.

- Получение данных
- Изучение данных
- Очистка данных
- Выбор переменных, представляющих интерес
- Генерация производных переменных
- Извлечение, загрузка и преобразование данных
- Выборка данных в обучающие и тестовые наборы данных

“Данные” — это слово всего из шести букв. Удивительно, что такое маленькое слово может описать триллионы гигабайтов информации: имена клиентов, адреса, товары, цены со скидкой по сравнению с исходными ценами, коды магазинов, время покупки, расположение поставщиков, цены на печатную рекламу, цвет ваших фургонов для доставки… И это только небольшая часть данных. Данные — это буквально все.

Не каждый источник или тип данных имеет отношение к бизнес-вопросу, на который вы пытаетесь ответить. Модели аналитического прогнозирования строятся на основе нескольких источников данных, и один из первых важных шагов — определить, какие источники включить в модель. Если вы пытаетесь определить, например, будут ли клиенты, подписывающиеся на электронные журналы весной, с большей вероятностью приобретать книги в твердом переплете осенью, можете отказаться от январских отчетов о продажах книг в мягкой обложке.

Затем вы должны проверить конкретные записи и атрибуты каждого возможного источника по формату, количеству и качеству. Возможно, *данные* — это маленькое слово, но оно влечет за собой много больших задач.

Перечисление бизнес-целей

На этом этапе, по-видимому, вы уже провели совещание с руководителями компании и выяснили цели, к которым они стремятся. Теперь необходимо углубиться в детали, оценить, какие источники информации помогут достичь поставленных целей, и выбрать переменные, которые вы будете анализировать для оперативного использования. Понимание того, чего на самом деле хотят заинтересованные стороны от проекта, является сложной задачей. Вы можете столкнуться с множеством конкурирующих потребностей, а также с ограничениями на то, что можно реально сделать.

На этом этапе вы и предполагаемый получатель результатов должны засучить рукава и провести мозговой штурм потенциальных источников информации. Цель состоит в том, чтобы определить, какая информация, из каких источников, будет иметь отношение к достижению типа конкретного результата, который обеспечит истинную ценность как для бизнеса, так и для клиента. Без этого ваши результаты могут быть не более чем академическими — они не будут иметь практической ценности для вашего клиента. Вы можете сделать захватывающие выводы из, например, записей о продажах подразделения аксессуаров во втором квартале и выяснить, насколько вероятно, что люди, которые носят туфли на плоской подошве, приобретут сумки из искусственной кожи.

Но это невозможно, если подразделение аксессуаров планирует прекратить выпуск продукции из искусственной кожи в следующем квартале.

Бизнес-цель может быть количественной и объективной, например “определить две неизвестные в настоящее время основные группы клиентов с вероятностью оттока более 50% в течение следующих шести месяцев” или “определить три группы поставщиков из Азии с сокращающимися сроками поставки в течение следующих пяти лет”. Вы также можете указать более субъективные цели, например “предоставить ценную информацию об эффективности программ поощрения клиентов”.

СОВЕТ

В субъективных случаях обязательно определите, что вы подразумеваете под словом “ценный”.

Определение связанных целей

Как правило, возникает много частных бизнес-вопросов, которые клиент хотел бы решить, — любой из них может дать представление о главном вопросе. Например, основной целью бизнеса может быть выявление недовольных клиентов еще до их *оттока* (перехода на конкурирующий товар). Связанные бизнес-вопросы могут быть такими: “Сколько раз покупатель аннулирует покупательскую корзину онлайн, прежде чем перейти в другой интернет-магазин?” “Предотвращает ли отток снижение суммы товаров с бесплатной доставкой от 100 до 75 долл.?” Несколько типичных примеров первичных и вторичных бизнес-вопросов перечислены в табл. 9.1.

Таблица 9.1. Первичные и вторичные бизнес-вопросы

Первичный бизнес-вопрос	Вторичный бизнес-вопрос
Как увеличить продажи печатных книг?	Какой процент людей купили или скачали художественную электронную книгу в 2012 финансовом году, затем приобрели печатную художественную литературу в 2013 финансовом году?
Как точнее предсказать влияние решений, связанных со здоровым образом жизни, на товары, связанные с фитнесом?	Если в этом году клиенты будут покупать меньше картофеля фри, будут ли они покупать больше или меньше ковриков для йоги в следующем году?
Как новый планшет повлияет на продажи существующих цифровых товаров?	Правда ли, что пользователи iPad реже покупают ноутбуки?

Сбор требований пользователей

Хорошо, предположим, что цели высокого уровня документированы и теперь вы переходите к деталям. Какие требования и сроки проекта вам необходимо выполнить и соблюдать? Каковы требования к вашему бизнесу, проекту, системе, моделям и данным?

Чтобы избежать несогласованных ожиданий, руководители проектов должны встретиться со всеми соответствующими группами в отделе обслуживания клиентов. При планировании маркетинговой кампании это могут быть менеджеры по маркетингу в социальных сетях, специалисты по анализу данных или менеджеры по маркетингу баз данных. Источниками информации могут быть, например, списки клиентов, сметы, графики и другая логистическая документация.

Тщательное документирование и одобрение руководства имеют решающее значение для того, чтобы все сотрудники прилагали максимум усилий, одинаково понимали цели и стремились их достичь.

Обработка данных

Не удивляйтесь, если подготовка данных для прогностической модели окажется такой же утомительной задачей, как и главная. Понимание качества данных, их доступности, знание их источников и любых существующих ограничений будет иметь прямое влияние на успешную реализацию вашего проекта по аналитическому прогнозированию.

Необработанные данные обычно необходимо очистить и, возможно, интегрировать, т.е. объединить с другими наборами данных и использовать для получения новых данных. Следовательно, качество и количество данных должны быть тщательно и внимательно изучены во всех источниках, использованных для построения модели.

На этом этапе исследования вы получите глубокие знания своих данных, что, в свою очередь, поможет вам выбрать соответствующие переменные для анализа. Это понимание также поможет вам оценить результаты вашей модели.

Идентификация данных

Для успеха вашего аналитического проекта необходимо определить соответствующие источники данных, объединить данные из этих источников и представить их в структурированном, хорошо организованном формате. Эти задачи могут быть очень сложными и, вероятно, потребуют тщательной координации между различными управляющими данными в вашей организации.

Вам также необходимо выбрать переменные, которые вы собираетесь анализировать. В этом процессе следует принимать во внимание ограничения данных, проектные ограничения и бизнес-цели.

ЗАПОМНИ!

Переменные, которые вы выбираете, должны иметь предсказательную силу. Кроме того, вы должны учитывать переменные, которые являются ценными и доступными для вашего проекта в рамках бюджета и сроков. Например, если вы анализируете банковские операции в рамках уголовного расследования, телефонные записи для всех вовлеченных сторон могут иметь отношение к анализу, но не будут доступны для аналитиков.

Не удивляйтесь, если вам придется потратить значительное время на этот этап проекта. Сбор данных, их анализ и выяснение их содержания, качества и структуры представляют собой трудоемкие задачи.

Идентификация данных помогает понять их смысл и свойства. Эти знания помогут вам выбрать алгоритм для построения вашей модели. Например, данные в виде временны*х* рядов могут быть проанализированы с помощью алгоритмов регрессии, а для анализа дискретных данных можно использовать алгоритмы классификации.

Выбор переменных зависит от того, насколько хорошо вы понимаете данные. Не удивляйтесь, если вам придется посмотреть и оценить сотни переменных, по крайней мере поначалу. К счастью, поработав с этими переменными и начав получать ключевые сведения, вы сможете сузить их до нескольких десятков. Кроме того, выбор переменных изменится, когда вы поймете, как изменяются данные в течение всего проекта.

СОВЕТ

Может оказаться полезным провести инвентаризацию данных, чтобы отслеживать то, что вы знаете, чего не знаете, а что может отсутствовать. Инвентаризация данных должна включать в себя список различных элементов данных и любых атрибутов, которые имеют отношение к последующим этапам процесса. Например, вы можете документировать, отсутствуют ли в каких-либо сегментах почтовые индексы или записи за определенный период времени.

ЗАПОМНИ!

Специалисты по бизнесу, которых часто называют *экспертами в предметной области*, помогут вам выбрать ключевые переменные, способные положительно повлиять на результаты вашего проекта. Они могут помочь вам объяснить важность этих переменных, а также где и как их получить.

Очистка данных

Необходимо убедиться, что данные очищены от помех, прежде чем использовать их в своей модели. Это включает в себя поиск и исправление любых записей, которые содержат ошибочные значения, и попытку заполнить все пропущенные значения. Вам также необходимо решить, включать ли дубликаты записей (например, две учетные записи клиентов). Общая цель — обеспечить целостность информации, которую вы используете для построения своей прогностической модели. Обратите особое внимание на полноту, правильность и своевременность данных.

СОВЕТ

Полезно вычислить *описательные статистики* (количественные характеристики) для различных областей, например вычислить минимальное и максимальное значения, проверить *распределение частот* (как часто что-то происходит) и ожидаемые диапазоны. Простая проверка может помочь вам пометить любые данные, которые находятся за пределами ожидаемого диапазона для дальнейшего изучения. Этим способом можно пометить любые записи о пенсионерах с датами рождения в 1990-х годах. Кроме того, перекрестная проверка информации важна для обеспечения точности данных. Для более глубокого анализа характеристик данных и определения взаимосвязи между записями данных можно использовать профилирование данных (анализ доступности данных и сбор статистики об их качестве), а также инструменты визуализации.

Отсутствие данных может быть связано с тем, что конкретная информация не была записана. В таком случае вы можете попытаться заполнить столько, сколько сможете. Можно легко добавить подходящие значения по умолчанию, чтобы заполнить пробелы в определенных полях. Например, для пациентов в родильном отделении больницы, где в поле “пол” отсутствует значение, заявка может быть заполнена как женская. В то же время для любого мужчины, который был помещен в больницу с отсутствующей записью о статусе беременности, такое заполнение записи должно считаться неприемлемым. Отсутствующий почтовый индекс для адреса может быть выведен из названия улицы и города, указанного в этом адресе.

В случаях, когда информация неизвестна или не может быть выведена, необходимо использовать значения, *отличные* от пробела, чтобы указать, что данные отсутствуют, но это не влияет на правильность анализа. Пробел в данных может означать несколько вещей, большинство из которых не являются хорошими или полезными. Всякий раз, когда это возможно, вы должны указать характер этого пробела с помощью значимого наполнителя. Для числовых данных, состоящих исключительно из маленьких и положительных

чисел (значения от 0 до 100), пользователь, например, может определить число –999,99 как заполнитель для пропущенных данных.

Точно так же, как розу на кукурузном поле можно считать сорняком, для разных анализов выбросы могут означать разные вещи. Обычно для отслеживания этих выбросов и их разметки создаются специальные модели. Модели выявления мошенничества и мониторинга преступной деятельности интересуют те выбросы, которые в таких случаях указывают на нечто нежелательное. В подобных случаях рекомендуется сохранять выбросы в наборе данных. Однако если выбросы считаются аномалиями и лишь искажают результаты анализа и приводят к ошибочным результатам, то удаляйте их из ваших данных. Совершенно нежелательно, чтобы модель предсказывала выбросы, но не могла предсказать что-либо еще.

Дублирование в данных также может быть как полезным, так и вредным; некоторые из дубликатов могут быть необходимыми, указывать значение и отражать точное состояние данных. Например, клиент может быть представлен несколькими записями, которые (во всяком случае, формально) дублируют и повторяют одна другую. Примером также может служить клиент, у которого есть рабочий и персональный телефон одной и той же компании с одним и тем же адресом — в общем-то, такая информация является полезной. В то же время, если дубликаты записей не вносят ценности в анализ и не являются необходимыми, их удаление может иметь огромное значение. Это особенно верно для больших наборов данных, в которых удаление дублирующих записей может уменьшить сложность данных и сократить время, необходимое для анализа.

СОВЕТ

Вы можете предотвратить попадание неверных данных в ваши системы, выполнив некоторые конкретные процедуры.

- Организовать проверку качества всех собираемых данных.
- Позволить вашим клиентам проверять и самостоятельно исправлять свои личные данные.
- Предоставить своим клиентам возможные и ожидаемые значения на выбор.
- Регулярно запускать проверки целостности, согласованности и точности данных.

Генерация любых производных данных

Производные атрибуты — это совершенно новые записи, построенные из одного или нескольких существующих атрибутов. Примером может служить создание записей, идентифицирующих книги, которые являются бестселлерами на книжных ярмарках. Необработанные данные могут не содержать такие записи, но для целей моделирования эти производные записи могут быть

важны. Отношение цены к прибыли и скользящее среднее за 200 дней являются двумя примерами производных данных, которые интенсивно используются в финансовых приложениях.

Производные атрибуты могут быть получены путем простых вычислений, например выведения возраста из даты рождения. Производные атрибуты также могут быть вычислены путем суммирования информации из нескольких записей. Например, преобразование таблицы клиентов и купленных ими книг в отдельную таблицу может позволить вам отслеживать количество книг, проданных с помощью системы рекомендаций, целевого маркетинга и на книжной ярмарке, и определить демографический состав клиентов, купивших эти книги.

Генерация таких дополнительных атрибутов дает дополнительную предсказательную силу для анализа. На самом деле многие такие атрибуты создаются для того, чтобы исследовать их потенциальную предсказательную силу. Некоторые прогностические модели могут использовать больше производных атрибутов, чем в их исходном состоянии. Если некоторые производные атрибуты оказываются особенно сильными и их сила доказывает свою актуальность, то имеет смысл автоматизировать процесс, который их генерирует.

ЗАПОМНИ!

Производные записи — это новые записи, которые вводят новую информацию и предоставляют новые способы представления необработанных данных; они могут иметь огромное значение для прогностического моделирования. Их часто считают самым важным вкладом, который разработчик может внести в процесс.

Уменьшение размерности данных

Данные, используемые в прогностических моделях, обычно объединяются из нескольких источников. Ваш анализ может основываться на данных, разбросанных по нескольким форматам данных, файлам и базам данных или по нескольким таблицам в одной базе данных. Объединение данных воедино и преобразование их в единый формат, который могут использовать разработчики моделей данных, — очень важный процесс.

Если ваши данные содержат какой-либо иерархический контент, то его, возможно, необходимо *сгладить*. Некоторые данные имеют иерархические характеристики, такие как отношения “родитель–ребенок”, или записи, которые состоят из других записей. Например, такой товар, как автомобиль, может иметь несколько производителей; в данном случае сглаживание данных означает включение каждого производителя в качестве дополнительного признака анализируемой записи. Другим примером является то, что один клиент может выполнять несколько транзакций.

Сглаживание данных важно, когда они объединяются из нескольких связанных записей, чтобы сформировать лучшее представление. Например, анализ

побочных эффектов от нескольких препаратов, производимых несколькими компаниями, может потребовать сглаживания данных на уровне действующего вещества. Тем самым вы в конечном итоге удаляете *отношения "один ко многим"* (в данном случае много производителей и много веществ для одного лекарства), которые могут вызвать слишком большое дублирование данных, повторяя несколько записей с одинаковой информацией о товаре и производителе.

Сглаживание заставляет задуматься об уменьшении *размерности* данных, т.е. количества признаков. Например, клиент может иметь следующие признаки: имя, возраст, адрес, приобретенные товары. Начиная анализ, вы можете использовать много признаков, хотя лишь некоторые из них могут быть важными. Таким образом, вы должны исключить все признаки, кроме очень немногих, которые имеют наибольшую предсказательную силу для вашего конкретного проекта.

Уменьшение размерности данных может быть достигнуто путем помещения всех данных в одну таблицу, которая содержит несколько столбцов, представляющих интересующие нас атрибуты. В начале анализа, конечно, необходимо оценить большое количество столбцов, но по мере анализа это число может быть уменьшено. Этому процессу можно помочь, перестроив наборы данных, например сгруппировав данные в категории, которые имеют сходные характеристики.

Результирующий, очищенный набор данных обычно помещается в отдельную базу данных для использования аналитиками. В процессе моделирования эти данные должны легко извлекаться, обрабатываться и обновляться.

Применение анализа главных компонентов

Анализ основных компонентов (PCA — principal component analysis) является ценным методом, который широко используется в интеллектуальном анализе данных. Он изучает набор данных, чтобы узнать наиболее значимые переменные, ответственные за наибольшую изменчивость в этом наборе данных. PCA в основном используется как метод сокращения объема данных.

При построении прогностических моделей часто желательно уменьшить количество признаков, описывающих ваш набор данных. Очень полезно уменьшить эту высокую размерность данных с помощью методов аппроксимации, к которым относится PCA. Аппроксимированные данные отображают все важные вариации исходных данных.

Например, набор признаков акций может включать цены на акции, дневные максимумы и минимумы, объемы торгов, 200-дневные скользящие средние, отношения цены к доходу, относительную силу по отношению к другим рынкам, процентные ставки и валютный курс.

В основе построения прогностической модели лежит поиск наиболее важных прогностических переменных. Многие делают это, используя метод

полного перебора. Идея состоит в том, чтобы начать с как можно большего количества релевантных переменных, а затем использовать воронкообразный подход для устранения признаков, которые не оказывают влияния или не имеют прогностического значения. Интеллект и знания привносятся в этот метод путем привлечения заинтересованных сторон бизнеса, потому что у этих людей есть определенные догадки о том, какие переменные окажут наибольшее влияние на анализ. Опыт специалистов по анализу данных, участвующих в проекте, также важен для понимания того, с какими переменными следует работать и какие алгоритмы использовать для конкретного типа данных или конкретной предметной задачи.

Чтобы помочь в этом процессе, специалисты по анализу данных используют множество инструментов аналитического прогнозирования, которые облегчают и ускоряют запуск нескольких перестановок и анализов в наборе данных, чтобы измерить влияние каждой переменной на этот набор данных.

Если вам предстоит обработать большой объем данных, попробуйте использовать PCA.

СОВЕТ

Уменьшение количества переменных является достаточной причиной для использования PCA. Кроме того, используя PCA, вы автоматически защищаете себя от переобучения модели (см. главу 15).

Конечно, вы можете найти корреляцию между погодными данными в конкретной стране и показателями ее фондового рынка. Или между цветом обуви человека, маршрутом, по которому он идет в офис, и доходностью его портфеля инвестиций за этот день. Однако включение этих переменных в прогностическую модель — это не просто переобучение, но и заблуждение, которое приводит к ложным прогнозам.

Метод PCA использует математически обоснованный подход для определения подмножества исходного набора данных, который включает в себя наиболее важные признаки. Основываясь на этом меньшем наборе данных, вы получите модель, которая будет иметь прогностическое значение для всего более крупного набора данных, с которым вы работаете. Короче говоря, метод PCA должен помочь вам определить подмножество переменных, ответственных за наибольшую вариацию в исходном наборе данных. Это поможет вам определить избыточность и узнать, что две (или более переменных) содержат одну и ту же информацию.

Кроме того, на основе исходного многомерного набора данных анализ главных компонентов создает новый набор данных, состоящий их переменных, представляющих собой линейные комбинации переменных из исходного набора данных. Кроме того, переменные в результирующем наборе данных являются попарно некоррелированными, а их дисперсия упорядочена по главным

компонентам, и первым является компонент, соответствующий самой большой дисперсии, и т.д. В этом отношении PCA также можно рассматривать как метод конструирования признаков.

ЗАПОМНИ!

Используя PCA или другие подобные методы, помогающие уменьшить размерность исходного набора данных, необходимо всегда проявлять осторожность, чтобы не снизить точность модели. Уменьшение размера данных не должно происходить за счет снижения точности прогностической модели. Поступайте аккуратно и управляйте своим набором данных с осторожностью.

ЗАПОМНИ!

Повышение сложности модели не приводит к повышению качества результата.

Чтобы сохранить точность модели, вам может потребоваться тщательно оценить значимость каждой переменной, измерить ее полезность при формировании окончательной модели.

Поскольку метод PCA может быть особенно полезен, когда переменные сильно коррелированы в пределах исходного набора данных, наличие набора данных с некоррелированными переменными может только усложнить задачу уменьшения размерности многомерных данных. Помимо PCA, в этой ситуации можно использовать другие методы, такие как прямой и обратный выбор признаков (они рассматриваются в этой главе).

ВНИМАНИЕ!

Метод PCA — это не волшебная палочка, которая решает все проблемы с многомерными данными. Его успех во многом зависит от данных, с которыми вы работаете. Статистическая значимость может не совпадать с прогностической силой переменных, даже если с такими приближениями можно безопасно работать.

Использование сингулярного разложения

Сингулярное разложение (SVD — singular value decomposition) представляет собой метод исключения менее важных фрагментов данных и генерирования более точного приближения исходного набора данных. В связи с этим SVD и PCA являются методами сокращения данных.

Входными данными для метода SVD является матрица, которую он раскладывает в произведение трех более простых матриц.

Матрица M размером *m* на *n* может быть представлена как произведение трех других матриц следующим образом.

$$M = U*S*V^T$$

Здесь U — матрица *m* на *r*, V — матрица *n* на *r*, а *S* — матрица *r* на *r*, где *r* — ранг матрицы M, символ * обозначает умножение матрицы, а буква *T* указывает на транспонирование матрицы.

В матрице, в которой данные можно описать с помощью меньшего количества понятий или столбцы матрицы связаны с ее строками, метод SVD является очень полезным инструментом для извлечения этих понятий. Например, набор данных может содержать рейтинги книг, в которых обзоры — это строки, а книги — столбцы. Книги могут быть сгруппированы по жанрам или темам, например литература и художественная литература, история, биографии, книги для детей или подростков. Это будут концепции, которые метод SVD может помочь извлечь.

Эти понятия должны быть осмысленными и убедительными. Если мы используем только несколько понятий или измерений для описания большего набора данных, наше приближение не будет таким точным. Именно поэтому важно исключать только те понятия, которые менее важны и не имеют отношения к общему набору данных.

Латентно-семантическое индексирование — это метод анализа данных и обработки естественного языка, который используется при поиске документов и оценке подобия слов. Латентно-семантическое индексирование использует метод SVD для группировки документов по понятиям, которые могут состоять из разных слов, найденных в этих документах. Множество слов может быть очень большим, и в понятии могут быть сгруппированы различные слова. Метод SVD помогает уменьшить зашумленную корреляцию между этими словами и их документами и дает представление об этой области, используя гораздо меньше измерений, чем исходный набор данных.

СОВЕТ

Легко видеть, что в документах, в которых обсуждаются похожие темы, могут использоваться разные слова для описания этих же тем. Документ, описывающий львов в Зимбабве, и документ, описывающий слонов в Кении, должны быть сгруппированы вместе. Поэтому при группировке этих документов мы полагаемся на концепции (в данном случае на дикую природу в Африке), а не на слова. Связь между документами и их словами устанавливается на основе понятий или тем.

Методы SVD и PCA используются для классификации и кластеризации (см. главы 6 и 7). Генерирование понятий — это просто форма классификации и группировки данных. Оба метода также используются для совместной фильтрации (см. главу 2).

Работа с признаками

Выбор наиболее подходящих признаков на основе набора исходных данных может как создать, так и разрушить модель. Чем больше предсказательная сила ваших признаков, тем успешнее будет ваша модель.

ЗАПОМНИ!

В своих попытках построить прогностические модели специалисты по интеллектуальному анализу данных проводят большую часть своего времени, подготавливая данные и выбирая соответствующие признаки.

Существуют алгоритмы и инструменты, которые помогут вам с выбором и извлечением признаков, и вам, возможно, даже придется подумать о ранжировании признаков в зависимости от их важности. Конечно, это всегда можно сделать, полагаясь исключительно на полный перебор. Некоторые ученые применяют воронкообразный подход, один за другим перебирают признаки, выбирая наиболее подходящие. Однако это отнимет много времени, может привести к отсутствию сходимости итераций и создает сложности, если признаки сильно зависят друг от друга.

Очень часто вы не знаете, какой признак включить, а какой игнорировать. Если вы используете метод проб и ошибок, добавляя или удаляя признак (по одному за раз), то можете увидеть, что это оказывает серьезное влияние на модель, которую вы строите. Результат будет существенно различаться в зависимости от того, включаете ли вы тот или иной признак, и этот подход становится еще более сложным, если один признак имеет смысл только в присутствии другого. Подобный подход является сложной задачей, если характеристики или переменные сильно коррелированы. Признак может оказывать большое влияние на анализ при группировании с другим признаком, в то же время сам по себе не оказывая никакого влияния. Влияние признака может проявляться только в сочетании с другими признаками и отсутствовать, если их нет.

Допустим, вы строите дерево решений. Это дерево может увеличиваться или уменьшаться в зависимости от того, какие признаки вы включаете или исключаете. Кроме того, часто вы не знаете, какая модель лучше, особенно если ваш набор данных небольшой и у вас недостаточно данных для тестирования или принятия обоснованного решения о результате. Кроме достаточного объема времени, необходимого для правильного выполнения этой части процесса, именно на этом этапе необходимо иметь опыт и нужные инструменты. Именно на этом этапе мы называем аналитическое прогнозирование бизнесом и дисциплиной, которая отчасти является искусством, а отчасти наукой.

СОВЕТ

Ниже приведены рекомендации, которые следует учитывать при подготовке данных и построении модели.

- Будьте готовы к тому, что ваши данные будут иметь много признаков.
- Выделите достаточно времени для подготовки и анализа ваших данных.
- Изучите предметную область, представленную в ваших данных.
- Выделите время, чтобы выбрать соответствующие признаки.
- Используйте инструменты и алгоритмы для выбора и извлечения признаков.
- Избегайте переобучения.
- Избегайте упрощения.
- Будьте готовы к большому количеству итераций при выборе признаков и сосредоточьтесь на построении модели.
- Пусть анализ данных и понимание модели помогут вам принять правильное решение.

Приготовьтесь к тому, что объем данных будет огромным. Редко бывает, что данных недостаточно для построения точных моделей. Большинство проектов страдают от избытка данных. В настоящее время мы наблюдаем экспоненциальный рост данных. Это изобилие относится как к размеру выборки, так и к ее размерности. Таким образом, данные могут содержать много шума. Дифференциация сигнала от шума лежит в основе того, что делают ученые.

В некоторых приложениях, таких как биоинформатика или классификация документов, набор данных обычно имеет тысячи признаков. Не все признаки важны для всех задач. Выбор и извлечение признаков — это два метода, которые могут помочь уменьшить размерность набора данных и определить соответствующие признаки для обработки.

Как извлечение признаков, так и их выбор улучшат предсказательную силу вашей модели и увеличат ее точность.

Выбор признаков

Выбор признаков — это процесс выбора подмножества признаков из множества исходных признаков. Подмножество выбирается без каких-либо преобразований и при сохранении свойств исходных признаков. Например, ученый, изучающий несколько белков и их влияние на заболевание, пытается определить, какие белки наиболее важны для анализа. Для заявки на ссуду ваш кредитный рейтинг, вероятно, является наиболее важным решающим фактором.

В задаче классификации (см. главу 7), когда данные обучения уже помечены и классы известны (например, электронные письма со спамом и без спама), выбор наиболее важных признаков, определяющих, является ли электронная почта спамом, может быть итеративным. Пока признаки, которые вы выбираете, все еще приводят к правильной классификации, вы движетесь в правильном направлении.


ЗАПОМНИ!

Цель состоит в том, чтобы определить, какие признаки необходимо сохранить, и при этом правильно распознать помеченные данные на этапе обучения.

Выбор признаков для классификации направлен на выбор подмножества исходных признаков без ущерба для точности классификатора. Подмножество признаков должно все еще быть хорошим предиктором классификации, если включены все доступные признаки.

Выбор признаков очень сложен, и степень сложности зависит от размерности данных, уровня корреляции между признаками, независимо от того, сильно ли они зависят друг от друга, а также от структуры данных.


ЗАПОМНИ!

Определение правильных признаков поможет повысить точность модели с точки зрения скорости и точности ее прогнозирования.

Существует два широко используемых метода выбора признаков.

- Прямой выбор начинается с одного признака и добавляет по одному признаку на каждой итерации. Он продолжает добавлять одну переменную за раз, что помогает уменьшать ошибку до тех пор, пока дальнейшие добавления не перестанут улучшать модель или не будут иметь никакого значения для уменьшения ошибки.
- Обратный выбор начинается со всего множества признаков в наборе данных и удаляет по одному признаку на каждом шаге. При этом необходимо убеждаться, что удаление либо уменьшает ошибку, либо лишь ненамного увеличивает ее. Признак удаляется, если его удаление приводит к наименьшему увеличению частоты ошибок. Если дальнейшее улучшение модели больше не достигается или если ошибка начинает значительно увеличиваться, процесс останавливается.

Существует три подхода к выбору признаков.

- **Фильтры**: методы предварительной обработки, которые вычисляют оценку для каждого признака, а затем выбирают признаки с высокими показателями.

- **Упаковщики**: методы, которые используют алгоритм обучения для оценки подмножеств признаков в соответствии с их полезностью для данного предиктора. При этом создаются и оцениваются несколько подмножеств признаков.
- **Встроенные**: методы, которые выполняют выбор как часть алгоритма или процедуры обучения. В этом случае поиск признаков встроен в сам классификатор.

Извлечение признаков

Извлечение признаков трансформирует ваши исходные признаки и создает небольшое подмножество новых, что приводит к гораздо меньшей размерности. Как показано в предыдущих разделах, уменьшение размерности может помочь вам избавиться от лишних признаков и шума в ваших данных. Извлечение признаков преобразует исходные признаки в новый набор, который намного меньше, чем оригинал.

Мы обсудили идею создания концепций при проведении анализа книг и создании значимых групп, таких как научная фантастика, художественная литература, история и биографии.

Затем мы можем использовать эти новые концепции для анализа наших данных. Эта трансформация из отдельных книг в концепции, или *логическая группировка*, является своего рода трансформацией, которая приводит к уменьшению размерности. Однако эти новые сгенерированные признаки, которые создаются путем извлечения признаков, все еще нуждаются в дополнительном анализе, чтобы мы могли полностью понять данные и в конечном итоге построить прогностическую модель.

Другим примером извлечения признаков, часто используемым в текстовом анализе, является возможность преобразовывать текст в числовое представление. Для этого используются следующие показатели.

- Частота термина (TF — term frequency)
- Частота термина — обратная частота документа (TFI–DF — term frequency –– inverse document frequency)

  Показатель TF–IDF часто используется для корректировки того факта, что одни слова используются чаще, чем другие; частота термина, или количество слов, смещается в зависимости от частоты этого слова в документах.

ЗАПОМНИ!

Основное различие между выбором признаков и извлечением признаков заключается в том, что извлечение уменьшает размерность без необходимости сохранения фактических атрибутов, таких как единицы исходных признаков. Кроме того, извлечение признаков

может быть методом преобразования данных, если взять набор исходных признаков и преобразовать их или извлечь новые.

Выбор признаков сохраняет природу исходных признаков; он только снижает их количество. Выбор признаков направлен на устранение избыточности и максимизацию релевантности.

Ранжирование признаков

Признаки необходимо ранжировать. Разве вы не хотите знать, какой признак является наиболее важным в вашем наборе данных? Какая особенность или набор признаков является абсолютным показателем данного класса или ярлыка? Рассматривая биологические приложения, целесообразно сосредоточиться на одном гене или подгруппе генов, ответственных за биологическое состояние. Затем модель может легко отследить существование этого гена или его экспрессию, чтобы предсказать ожидаемое поведение.

Методы ранжирования помогают выбирать признаки и уменьшать размерность набора данных. Для ранжирования признаков можно выбрать один из следующих методов ранжирования.

- Коэффициент усиления
- Информационный выигрыш
- Хи-квадрат
- SVM Ranker

Эти алгоритмы можно разделить на две большие категории.

- **Статистические методы**, такие как хи-квадрат, основаны на вычислении значения статистики "хи-квадрат" для атрибутов с целью их ранжирования.
- **Методы, основанные на энтропии**, измеряют объем информации в признаке. Релевантность признака, о которой свидетельствует его рейтинг, оценивается путем вычисления ожидаемого значения информации, содержащейся в сообщении, относительно вывода этой переменной.
  - Высокое значение энтропии указывает, что переменная принадлежит к равномерному распределению.
  - Низкое значение энтропии указывает на то, что переменная принадлежит к вариабельному распределению.

В модели дерева решений важность атрибута измеряется с помощью энтропийного подхода, а для выбора признаков модели используется информационный

выигрыш. В этом случае дерево решений фокусируется на признаках, которые приводят к данному решению.

ЗАПОМНИ!

Два признака с высоким рейтингом не обязательно являются двумя лучшими признаками для всей модели. Иначе говоря, полагаясь только на алгоритмы ранжирования при сортировке признаков по объему информации, которой они обладают (т.е. энтропии), вы не всегда будете выбирать лучшие признаки. В сочетании с другими признаками признак с большой энтропией может не обеспечивать высокую точность.

Алгоритм прямого выбора использует более эффективный подход к выбору признаков. На каждой итерации алгоритм ищет лучшие признаки, которые обеспечивают высокую точность при объединении. Алгоритмы выбора признаков, такие как прямой или обратный выбор, получили широкое распространение, несмотря на то что для их выполнения требуется относительно большое время.

СОВЕТ

Сначала можно использовать алгоритмы ранжирования, например на основе информационного выигрыша, чтобы исключить неинформативные признаки (с очень низкими значениями информационного выигрыша), а затем к оставшемуся подмножеству применить алгоритмы выбора признаков.

Структурирование данных

Необработанные данные являются потенциальным ресурсом, но их бесполезно анализировать, пока им не будет придана согласованная структура. В ходе подготовки к анализу данные, находящиеся в нескольких системах, должны быть собраны и преобразованы. Собранные данные должны храниться в отдельной системе, чтобы не мешать работе реального производства. При построении модели следует разделить ваше множество данных на обучающий и тестовый наборы.

Извлечение, преобразование и загрузка данных

После первоначального сбора данные обычно находятся в рассредоточенном состоянии — в нескольких системах или базах данных. Прежде чем использовать их для аналитического прогнозирования, их необходимо объединить в одном месте. Кроме того, не следует работать с данными, находящимися в эксплуатационной среде, — это создает проблемы. Вместо этого поместите часть данных где-нибудь, где вы можете работать над ними свободно, не влияя

на операции. *ETL* (извлечение, преобразование и загрузка — *extract*, *transform and load*) — это процесс, который приводит данные в желаемое состояние.

Многие организации имеют несколько баз данных. Ваша прогностическая модель, вероятно, будет использовать данные из всех этих баз. ETL — это процесс, собирающий всю необходимую информацию и помещающий ее в отдельную среду, в которой вы можете выполнить свой анализ. Однако ETL — это непрерывный процесс, который постоянно обновляет данные. Обязательно запускайте процессы ETL ночью или в другое время, когда нагрузка на эксплуатационную среду низкая.

- **На этапе извлечения** необходимые данные в необработанном виде собираются из эксплуатационных сред.
- **На этапе преобразования** собранные данные практически готовы для использования в вашей прогностической модели: объедините их, сгенерируйте требуемые производные атрибуты и поместите преобразованные данные в соответствующий формат, соответствующий бизнес-требованиям.
- **На этапе загрузки** данные помещаются в назначенное место, где можно выполнить их анализ, например в киоск данных, хранилище данных или другую базу данных.

Необходимо следовать системному подходу для построения ETL-процессов, удовлетворяющих бизнес-требованиям. Рекомендуется хранить копию исходных данных отдельно, чтобы всегда можно было вернуться к ней в случае, если ошибка нарушит этапы преобразования или загрузки процессов. Копия исходных данных служит резервной копией, которую можно использовать для перестройки всего набора данных, используемого вашим анализом, если это необходимо. Цель состоит в том, чтобы отойти от закона Мерфи и быстро встать на ноги, если вам придется перезапустить весь процесс ETL “с нуля”.

Ваш процесс ETL должен использовать *модульность* — разделять задачи и выполнять работу поэтапно. Этот подход имеет преимущества в том случае, если необходимо повторно обрабатывать или перезагружать данные либо использовать некоторые из этих данных для другого анализа или для построения различных прогностических моделей. Дизайн вашего ETL-процесса должен быть в состоянии учесть даже существенные изменения требований бизнеса — с минимальными изменениями в вашем процессе ETL.

Поддержание данных в актуальном состоянии

После шага загрузки в процессе ETL, на котором вы помещаете свои данные в отдельную базу данных, в малое или крупное хранилище, необходимо поддерживать актуальность данных, чтобы разработчики моделей могли повторно запускать ранее построенные модели на новых данных.

Создайте малое хранилище данных, которые необходимо проанализировать, и поддерживайте их в актуальном состоянии, чтобы уточнять модель. В связи с этим вам следует регулярно обновлять рабочие модели после их развертывания. Новые данные могут увеличить предсказательную силу ваших моделей. Новые данные могут позволить модели отображать новые идеи, тенденции и взаимосвязи.

Наличие отдельной среды для данных также позволяет повысить производительность систем, используемых для запуска моделей. Это объясняется тем, что вы не перегружаете эксплуатационные среды интенсивными запросами или анализом, необходимым для запуска моделей.

Данные продолжают поступать — больше, быстрее и с большим разнообразием. Внедрение автоматизации и разделение задач и сред может помочь вам управлять этим потоком данных и поддерживать отклик ваших прогностических моделей в режиме реального времени.

ЗАПОМНИ!

Чтобы обеспечить захват потоков данных и обновление моделей при поддержке автоматизированных процессов ETL, аналитическая архитектура должна быть высокомодульной и адаптивной. Если вы будете помнить об этой цели проектирования для каждой части общего проекта по аналитическому прогнозированию, постоянное улучшение и настройка, которые сопровождают аналитическое прогнозирование, будут более удобными в обслуживании и обеспечат лучший успех.

Планирование тестирования и организация тестовых данных

Когда ваши данные приготовлены и вы собираетесь приступить к построению своей прогностической модели, полезно наметить методологию тестирования и составить его план. Тестирование должно основываться на бизнес-целях, которые вы сформулировали, зафиксировали в документе и для которых собраны все необходимые данные.

Сразу же следует разработать метод проверки успешности достижения бизнес-цели. Поскольку аналитическое прогнозирование измеряет вероятность будущих результатов, а единственный способ быть готовым к выполнению такого теста — это обучение модели на прошлых данных, вам все еще нужно увидеть, как она будет работать на новых данных. Конечно, вы не можете проверить модель на несуществующих пока новых данных, поэтому для реалистичного моделирования будущих данных необходимо использовать существующие данные. Для этого необходимо разделить данные, над которыми вы работаете, на обучающие и тестовые наборы.

СОВЕТ

Убедитесь, что вы выбрали эти два набора данных случайным образом и что оба набора данных содержат и охватывают все параметры данных, которые вы измеряете.

Разделяя данные на обучающие и тестовые, вы избегаете любых проблем с переобучением, которые могут возникнуть из-за настройки модели на основе полного набора данных и выбора всех шаблонов шума или специфических признаков, которые относятся только к обучающим данным и не применимы к другим наборам. (Читайте главу 15 для получения более подробной информации о ловушках переобучения.)

ЗАПОМНИ!

Разделение данных на обучающие и тестовые наборы данных (около 70% и 30% соответственно) обеспечивает точное измерение производительности модели аналитического прогнозирования, которую вы строите. Необходимо сравнить вашу модель с тестовыми данными, потому что это простой способ измерить точность ее прогнозов. Успех является показателем того, что после развертывания модель будет работоспособной. Тестовый набор данных будет служить независимым набором данных, которых модель еще не видела. Запуск вашей модели с набором тестовых данных обеспечивает предварительный просмотр того, как будет работать модель после запуска.

Глава 10

Создание прогностической модели

В ЭТОЙ ГЛАВЕ...

- Определение бизнес-цели
- Подготовка данных
- Разработка, тестирование и оценка модели
- Развертывание и поддержка модели

Некоторые претензии являются мошенническими. Некоторые клиенты отказываются от услуг компании. Некоторые транзакции оказываются мошенническими. Некоторые инвестиции бывают убыточными. Некоторые сотрудники увольняются. Возникает один важный вопрос — какие именно?

Модель аналитического прогнозирования может помочь вашему бизнесу ответить на него. Например, модель будет анализировать имеющиеся у вас данные о клиентах и вычислять вероятность их оттока. Но такие вопросы — лишь небольшая часть того, что может сделать аналитическое прогнозирование. Потенциальные возможности применения этой увлекательной дисциплины безграничны.