Содержание

Вступительное слово 9

Предисловие 10

Глава 1. Теорема Байеса 16

Условная вероятность 16
Совместная вероятность 17
Задача о булочках 17
Теорема Байеса 18
Диахроническая интерпретация 19
Задача M&M 20
Задача Монти Холла 22
Обсуждение 24

Глава 2. Вычислительная статистика 25

Распределения 25
Задача с булочками 26
Байесовская структура 27
Задача Монти Холла 28
Формирование структуры программного пакета 29
Задача M&M 30
Обсуждение 31
Упражнение 32

Глава 3. Оценивание 33

Задача об игральных костях 33
Задача о локомотиве 34
Что насчет этого приора? 36
Альтернативный приор 37
Доверительный интервал 39
Кумулятивные функции распределения 39
Задача о немецком танке 40
Обсуждение 41
Упражнение 41

Глава 4. Больше об оценивании 43

Задача о евро 43
Итоговый постериор 44

Подавление приоров 45
Оптимизация 47
Бета-распределение 48
Обсуждение 50
Упражнения 50

Глава 5. Отношение вероятностей и добавления 52

Отношение вероятностей 52
Теорема Байеса в форме отношения вероятностей 53
Группа крови Оливера 54
Добавления 55
Максимизации 58
Перемешивание 60
Обсуждение 62

Глава 6. Анализ решений 63

Задача «Справедливой цены» 63
Приор 64
Функция плотности вероятности 65
Представление PDF 65
Моделирование участников 67
Правдоподобие 69
Обновление 70
Оптимальное предложение цены 71
Обсуждение 74

Глава 7. Предсказание 75

Задача о Бостон Брюинс 75
Процесс Пуассона 76
Постериоры 77
Распределение голов 78
Вероятность выигрыша 79
Выигрыш в дополнительное время 80
Обсуждение 82
Упражнения 83

Глава 8. Погрешность наблюдения 85

Задача о линии метрополитена 85
Модель 85
Время ожидания 87
Предсказание ожидаемого времени 89
Оценка времени прибытия 92

Включение неопределенности 94
Анализ решений 95
Обсуждение 97
Упражнение 98

Глава 9. Двумерное измерение 99

Пейнтбол 99
Пакет гипотез 99
Тригонометрия 100
Правдоподобие 102
Совместные распределения 102
Условные распределения 104
Доверительные интервалы 105
Обсуждение 107
Упражнение 108

Глава 10. Аппроксимация при байесовских вычислениях 109

Гипотеза изменчивости 109
Среднее и стандартное отклонение 110
Обновление 112
Апостериорное распределение CV 113
Потеря значимости 113
Логарифмическое правдоподобие 115
Небольшая оптимизация 116
Аппроксимация при байесовских вычислениях (ABC) 117
Робастное оценивание 118
Кто более изменчив? 120
Обсуждение 122
Упражнение 122

Глава 11. Проверка гипотез 124

Обратно к задаче о евро 124
Справедливое сравнение 125
Треугольный приор 126
Обсуждение 127
Упражнения 128

Глава 12. Свидетельства 129

Интерпретация оценки SAT 129
Шкала 129
Приор 130
Постериор 132

Улучшенная модель ... 133
Градуировка ... 135
Апостериорное распределение эффективности ... 136
Распределение предсказания ... 138
Обсуждение ... 138

Глава 13. Моделирование ... 140

Проблема опухоли почек ... 140
Простая модель ... 141
Более общая модель ... 143
Реализация ... 144
Кеширование совместного распределения ... 145
Условные распределения ... 146
Последовательная корреляция ... 147
Обсуждение ... 151

Глава 14. Иерархическая модель ... 152

Задача о счетчике Гейгера ... 152
Простое начало ... 153
Создание иерархии ... 154
Небольшая оптимизация ... 155
Извлечение постериоров ... 155
Обсуждение ... 157
Упражнение ... 157

Глава 15. Борьба с размерностью ... 158

Бактерии пупка ... 158
Львы, тигры и медведи ... 158
Иерархическая версия ... 161
Случайная выборка ... 163
Оптимизация ... 164
Сворачивание иерархии ... 165
Еще одна проблема ... 167
Мы сделали еще не все ... 168
Данные пупка ... 170
Прогнозирующее распределение ... 172
Совместный постериор ... 175
Перекрывающая зона ... 176
Обсуждение ... 178

Предметный указатель ... 180

Вступительное слово

Около десяти лет назад, когда изучение байесовских методов впервые заинтересовало меня, я обнаружил острую нехватку книг по данной теме на русском языке. Материала, в котором бы практически, с точки зрения реализации на конкретном языке программирования, описывались как базовые, так и более продвинутые методы анализа данных с помощью байесовских методов. При этом не составляло труда найти огромное количество достойных книг на английском языке, дающих глубокое практическое понимание этого отдельного важного класса методов, которые применимы в самом широком спектре областей: начиная от анализа экспериментальных данных и заканчивая современными системами принятия решений и даже блокчейном. Кстати, если говорить о последнем, то можно привести в пример NeuroChainTech – проект большой международной команды, в котором мне посчастливилось стать научным консультантом. Это реализация умного блокчейна с новым оригинальным алгоритмом консенсуса и элементами машинного обучения, включающими как раз байесовские сети. В этом году проект провел успешное ICO и в настоящее время находится в активной фазе своего развития. Кроме того, в настоящее время на байесовских методах базируются в том числе и современные системы принятия решений и анализа данных, которые активно используются для решения задач цифровизации экономики, выходящих в настоящее время на первый план в государственном и корпоративном развитии.

Несколько лет назад мои статьи на русском языке, опубликованные на популярном ресурсе в сети Интернет, в которых описывались базовые принципы имплементации байесовских методов на Python'е, нашли очень живой отклик читателей. Более того, до сегодняшнего дня, спустя пять лет с момента их публикации, мне по-прежнему поступают вопросы, связанные с практическим воплощением алгоритмов байесовского анализа, что в очередной раз подтверждает неподдельный и неснижающийся интерес широкой аудитории к пониманию и практическим аспектам реализации байесовских методов.

Появление перевода на русский язык отличной книги, подробно описывающей практическое воплощение байесовских методов на Python'е, – это безусловный повод для радости. Настоящая книга включает в себя описание базовых принципов реализации байесовских методов в самом широком спектре их применений, и я очень надеюсь, что она вызовет должный интерес у читателей и придаст новый импульс к изучению, активному применению и дальнейшему развитию байесовских методов.

Желаю читателям успешного овладения инструментарием байесовских методов и интересных его применений в будущих проектах!

Максим Иришкин, PhD,
научный консультант NeuroChainTech,
эксперт по инновационному развитию корпораций

Предисловие

Мой подход

Предпосылкой для этой книги, как и других книг серии *Think X*, является мысль о том, что если вы умеете программировать, вы можете использовать это умение, чтобы овладеть другими знаниями.

Большинство книг по байесовской статистике используют математические формулировки и представляют эти идеи как исчисление в терминах математических концепций. В этой книге вместо математики используются язык программирования Питон (Python, Пайтон) и дискретная аппроксимация вместо непрерывной математики. В результате то, что книгах по математике является интегралом, становится суммированием, а большинство операций с вероятностными распределениями – просто циклами.

Мне кажется, что такое представление более понятно, по крайней мере для людей с навыками программиста. Оно также имеет более общий характер, потому что мы можем выбирать наиболее подходящую модель, не слишком беспокоясь, поддается ли она традиционному анализу реальных проблем. Глава 3 – хороший пример этого. Она начинается с простого примера с игральными костями – одного из основных в базовой теории вероятности. Затем небольшими шагами идет продвижение к задаче о локомотивах, которая позаимствована из книги Фредерика Мостеллера (Frederick Mosteller) «Пятьдесят интересных вероятностных задач с решениями» (Fifty Challenging Problems in Probability with Solutions. Dover, 1987) и затем к задаче о немецком танке, знаменитому успешному применению байесовского метода во время Второй мировой войны.

Моделирование и аппроксимация

Многие задачи в этой книге мотивированы реальными проблемами, что влечет за собой необходимость построения модели. Прежде чем мы применим байесовские методы (как и любой другой анализ), мы должны принять решение о том, какую часть реальной системы мы включим в модель и от каких деталей мы можем абстрагироваться.

Например, в главе 7 мотивирующей проблемой является предсказание победителя в игре в хоккей. Я применил для подсчета голов пуассоновский процесс, который подразумевает, что голы могут быть забиты равновероятно в любой момент игры. Это не совсем так, но эта модель, вероятно, подходит для многих других задач.

В главе 12 мотивацией проблемы является интерпретация экзаменационных оценок SAT (SAT является стандартизированным тестом, используемым при поступлении в колледж в США). Я начинаю с простой модели, в которой

предполагается, что все вопросы на экзамене SAT имеют одинаковую сложность. Однако фактически организаторы SAT сознательно включают часть вопросов, которые являются относительно простыми, а часть вопросов – которые являются более сложными. Я представил и вторую модель, которая учитывает это обстоятельство, и показал, что в конечном счете оно не оказывает существенного влияния на результат.

Я думаю, что важно считать моделирование неотъемлемой частью решения задачи, потому что это заставляет нас думать о модельных ошибках (то есть ошибках, появляющихся вследствие упрощений и предположений в моделях).

Многие методы в этой книге базируются на дискретных распределениях, что заставляет некоторых беспокоиться о численных ошибках. Но для реальных проблем численные ошибки почти всегда меньше модельных ошибок.

С другой стороны, непрерывные методы иногда приводят к преимуществам в производительности – например, путем замены вычислений с линейным или квадратичным временем на решение с постоянной продолжительностью.

Я рекомендую следующую общую процедуру при решении задач:

1) при исследовании проблемы начинайте с простой модели и опишите ее в ясных, читаемых и явно правильных кодах. Сфокусируйте внимание на хороших модельных решениях, не на оптимизации;
2) когда простая модель заработает, определите основные источники ошибок. Возможно, вам потребуется увеличить количество значений при аппроксимации, или увеличить число итераций в модели Монте-Карло, или добавить детали в модель;
3) если это решение приемлемо для вашего приложения, вам, возможно, нет необходимости заниматься его оптимизацией. Но если вы собираетесь это делать, существует два подхода, которые следует рассмотреть. Вы можете провести ревизию вашего кода и поискать возможности его оптимизации. Например, если вы кешировали полученные результаты, возможно, вам удастся избежать излишних вычислений. Или вы можете поискать аналитические методы, которые дают компьютерное ускорение.

Одно из достоинств этой процедуры в том, что шаги 1 и 2 склонны быть быстрыми, что дает возможность исследовать альтернативные модели до того, как вы начнете в них вкладываться.

Другим преимуществом является то, что если вы перейдете к шагу 3, вы уже начнете с некоторой справочной реализации, которая, вероятно, будет правильной, и вы сможете использовать ее для регрессионного тестирования (то есть проверяя, что оптимизированный код дает похожие результаты, по крайней мере приблизительно).

Работа с кодом

Многие примеры в этой книге используют классы и функции, определенные в модуле `thinkbayes.py`. Вы можете загрузить этот модуль из http://thinkbayes.com/thinkbayes.py.

Многие разделы содержат справки о кодах, которые вы можете загрузить из http://thinkbayes.com. Некоторые из этих файлов имеют зависимости, которые вы также должны будете загрузить. Я предлагаю вам держать все эти файлы в одной той же директории, так чтобы они могли импортироваться один в другой без изменения поискового пути Питона.

Вы можете загрузить эти файлы один за другим, когда они потребуются, или вы можете загрузить их все сразу из http://thinkbayes.com/thinkbayes_code.zip. Этот файл также содержит файлы данных, используемые в некоторых программах. Когда вы разархивируете их, это создаст директорию, названную thinkbayes_code, которая содержит все коды, использованные в этой книге.

Или, если вы пользователь программы консольных утилит Git, вы можете сразу получить все файлы посредством ветвления и клонирования репозитория: https://github.com/AllenDowney/ThinkBayes.

Одним из модулей, которые я использовал, является `thinkplot.py`, который обеспечивает упаковку для некоторых функций в `pyplot`. Чтобы использовать, его необходимо загрузить `matplotlib`. Если у вас его еще нет, проверьте ваш менеджер пакетов на его доступность. В противном случае вы можете загрузить инструкции из http://matplotlib.org.

Наконец, некоторые программы в этой книге используют NumPy и SciPy, которые доступны на http://numpy.org и http://scipy.org.

Стиль кодов

Опытные программисты на Питоне могут заметить, что коды в этой книге не соответствуют PEP 8, являющемуся наиболее общим руководством стиля для Питона (http://www.python.org/dev/peps/pep-0008/).

Конкретно PEP 8 требует нижнего регистра названий функций и подчеркивания между словами, `like_this`. В этой книге и сопровождающих кодах названия функции и методов начинаются с заглавной буквы и используют «верблюжий» стиль (camel style) LikeThis.

Я нарушил это правило, потому что я разработал эти коды, когда я был приглашенным ученым в Google, и следовал руководству Google по стилю кодов, который в некоторой части отличается от PEP 8. Используя стиль Google, я нашел его привлекательным, и мне было бы трудно его менять.

Также в духе этого стиля я написал «Bayes's Theorem» с одним «s» после апострофа, что предпочитают в одних руководствах и осуждают в других. У меня нет в этом вопросе каких-либо предпочтений, я должен был выбрать что-либо одно, и здесь то, что я выбрал.

И наконец, одно типографическое замечание: всюду в этой книге я использовал PMF и CDF для математической концепции функции масс вероятности и кумулятивной функции распределения и Pmf и Cdf для ссылки на объекты Питона, которые я использую для их представления.

Предпосылки

Существует несколько замечательных модулей для работы с байесовской статистикой в Питоне, включая pym и OpenBUGS. Я решил не использовать их в этой книге, потому что вам необходимо много предварительной информации, чтобы начать работать с ними, а я хотел свести необходимость предпосылок к минимуму. Если вы знаете Питон и немного о вероятности, вы готовы работать с этой книгой.

Глава 1 знакомит вас с вероятностью и теоремой Байеса. В ней нет кодов. Во второй главе вводится Pmf и Cdf, несколько измененный словарь Питона, который я использовал для представления функции масс вероятности (PMF). Затем в главе 3 вводится Suite, своего рода структура для осуществления байесовских обновлений. И это, пожалуй, все.

Нет, почти все. В некоторых последних главах я использовал распределения, включающие распределение Гаусса (нормальное распределение), экспоненциальное распределение, распределение Пуассона и бета-распределение. В главе 15 я использовал не слишком употребительное распределение Дирихле, но затем я представил его в процессе дальнейшего изложения. Если вы незнакомы с этими распределениями, то можете прочитать о них в Википедии. Вы можете также прочитать о них и в одной из книг этой серии *Think Stats* и книгах по введению в статистику (хотя я боюсь, что большинство из них использует математический подход, что не особенно полезно для практики).

Список участников

Если у вас есть предложения или замечания, пожалуйста, направляйте e-mail по адресу downey@allendow.ney.com. Если я проведу изменения, основанные на ваших откликах, я включу вас в список участников (если вы не попросите этого не делать).

Если вы включите, по крайней мере, часть предложения, в котором обнаружилась ошибка, это облегчит мне поиск ее. Это касается и просто страниц и разделов, но это уже будет для меня более трудным. Спасибо!

- Прежде всего я должен упомянуть замечательную книгу Давида МакКея (David MacKay) «Теория информации, вывод и обучающие алгоритмы» (*Information Theory, Inference, and Learning Algorithms*), благодаря которой я пришел к пониманию байесовского метода. С его разрешения я использовал несколько задач из этой книги в качестве примеров.
- Эта книга улучшилась благодаря консультациям с Sanjoy Mahajan, особенно осенью 2012 года, когда я был аудитором его класса в колледже Олин.
- Часть этой книги я написал во время вечерней работы над проектом с группой Boston Python User Group, и я хотел бы поблагодарить их за сотрудничество и пиццу.
- Jonathan Edwards исправил первые опечатки.

- George Purkins нашел ошибки в разметке текста.
- Olivier Yiptong прислал несколько полезных предложений
- Yuriy Pasichnyk нашел несколько ошибок.
- Kristopher Overholt прислал длинный перечень исправлений и предложений.
- Robert Marcus нашел неправильно проставленное i.
- Max Hailperin предложил более ясное изложение главы 1.
- Markus Dobler указал на то, что вытаскивание булочек из корзины с заменой не вполне реалистичный сценарий.
- Tom Pollard и Paul A. Giannaros обнаружили проблему в версии с некоторыми величинами в задаче о локомотивах.
- Ram Limbu нашел опечатки и предложил исправления.
- Весной 2013 года студенты моего класса Вычислительная байесовская статистика (Computational Bayesian Statistics) сделали много полезных исправлений и предложений: Kai Austin, Claire Barnes, Kari Bender, RachelBoy, Kat Mendoza, Arjun Iyer, Ben Kroop, Nathan Lintz, Kyle McConnaughay, Alec Radford, Brendan Ritter и Evan Simpson.
- Greg Marra и Matt Aasted помогли мне обсуждением задачи о справедливой цене.
- Marcus Ogren указал мне на то, что первоначальное описание задачи о локомотивах было неоднозначным.

Jasmine Kwityn и Dan Fauxsmith из O'Reilly Media сделали корректуру книги и нашли много возможностей для улучшения.

Аллен Б. Дауни

Отзывы и пожелания

Мы всегда рады отзывам наших читателей. Расскажите нам, что вы думаете об этой книге – что понравилось или, может быть, не понравилось. Отзывы важны для нас, чтобы выпускать книги, которые будут для вас максимально полезны.

Вы можете написать отзыв прямо на нашем сайте www.dmkpress.com, зайдя на страницу книги, и оставить комментарий в разделе «Отзывы и рецензии». Также можно послать письмо главному редактору по адресу dmkpress@gmail.com, при этом напишите название книги в теме письма.

Если есть тема, в которой вы квалифицированы, и вы заинтересованы в написании новой книги, заполните форму на нашем сайте по адресу http://dmkpress.com/authors/publish_book/ или напишите в издательство по адресу dmkpress@gmail.com.

Скачивание исходного кода примеров

Скачать файлы с дополнительной информацией для книг издательства «ДМК Пресс» можно на сайте www.dmkpress.com или www.дмк.рф на странице с описанием соответствующей книги.

Список опечаток

Хотя мы приняли все возможные меры для того, чтобы удостовериться в качестве наших текстов, ошибки все равно случаются. Если вы найдете ошибку в одной из наших книг — возможно, ошибку в тексте или в коде, — мы будем очень благодарны, если вы сообщите нам о ней. Сделав это, вы избавите других читателей от расстройств и поможете нам улучшить последующие версии этой книги.

Если вы найдете какие-либо ошибки в коде, пожалуйста, сообщите о них главному редактору по адресу dmkpress@gmail.com, и мы исправим это в следующих тиражах.

Нарушение авторских прав

Пиратство в Интернете по-прежнему остается насущной проблемой. Издательства «ДМК Пресс» и O'Reilly Media очень серьезно относятся к вопросам защиты авторских прав и лицензирования. Если вы столкнетесь в Интернете с незаконно выполненной копией любой нашей книги, пожалуйста, сообщите нам адрес копии или веб-сайта, чтобы мы могли применить санкции.

Пожалуйста, свяжитесь с нами по адресу электронной почты dmkpress@gmail.com со ссылкой на подозрительные материалы.

Мы высоко ценим любую помощь по защите наших авторов, помогающую нам предоставлять вам качественные материалы.

Глава 1

Теорема Байеса

Условная вероятность

Основная идея всей байесовской статистики – это теорема Байеса, которую удивительно легко получить, если понять, что такое условная вероятность. Поэтому мы начнем с вероятности, затем перейдем к условной вероятности, далее к теореме Байеса и, наконец, к байесовской статистике.

Вероятность – это число между 0 и 1 (включая оба), которые представляют собой уровень уверенности в каком-либо факте или предсказании. Числом 1 представляется абсолютная уверенность, что некоторый факт справедлив. Числом 0 – абсолютная уверенность, что этот факт не справедлив. Промежуточные числа в этом интервале определяют степень уверенности. Число 0,5, часто обозначаемое как 50%, означает, что предсказанное событие в одинаковой степени может как осуществиться, так и не осуществиться. Например, при подбрасывании монеты вероятность того, что она упадет лицевой стороной вверх, близка к 50%.

Условная вероятность – это вероятность, основанная на некотором предыдущем знании. Например, я хочу узнать, какова вероятность того, что в следующем году у меня случится сердечный приступ. Данные Центра контроля заболеваний гласят: «Ежегодно сердечный приступ случается у примерно 785 тысяч американцев» (http://www.cdc.gov/heartdisease/facts.htm).

Численность населения США составляет примерно 311 миллионов человек. Значит, вероятность того, что случайным образом выбранный американец получит в следующем году сердечный приступ, составляет примерно 0,3%.

Но я не являюсь случайно выбранным американцем. Эпидемиологи определили много факторов, влияющих на риск получения сердечного приступа, и в зависимости от этих факторов мой риск получить сердечный приступ может сильно отличаться от среднего значения.

Я, 45-летний мужчина, имею пограничо высокий холестерин. Этот факт увеличивает риск, что я получу сердечный приступ. Вместе с тем у меня низкое кровяное давление и я не курю, и это уменьшает мой шанс получения сердечного приступа.

Включая всю такого рода информацию в онлайновый калькулятор, размещенный на интернет-странице http://hp2010.nhlbihin.net/atpiii/calculator.as,

я увижу, что значение риска получения мною сердечного приступа в следующем году равно 0,2%. То есть это значение гораздо меньше, чем в среднем по стране. Вот это значение и есть условная вероятность, поскольку она основана на ряде факторов, которые составляют мои «условия».

Обычное обозначение условной вероятности p(A|B) – вероятность A при условии, что справедливо B. В этом примере A является предсказанием, что я получу сердечный приступ в следующем году, а B – множество условий.

Совместная вероятность

Совместная вероятность – это способ полагать, что оба факта или предсказания окажутся осуществленными. Я напишу p(A и B), если имею в виду, что вероятность выполнения совместно и A, и B существует.

Если вы поняли вероятность в контексте подбрасывания монеты или игральной кости, вы можете понять и содержание следующей формулы:

p(A и B) = p(A) p(B). ПРЕДУПРЕЖДЕНИЕ: это не всегда так.

Например, если я подбрасываю две монеты, значение A означает, что первая монета упала лицевой стороной вверх, а значение B – что вторая тоже упала лицевой стороной вверх, то есть p(A) = p(B) = 0,25.

Но эта формула справедлива, только если A и B независимы, то есть результат первого события не изменяет вероятность второго. Или более формально p(A|B) = p(B).

Приведу другой пример, в котором события не независимы. Предположим, A означает, что сегодня идет дождь, а B – что дождь пойдет завтра. Если я знаю, что сегодня идет дождь, то весьма вероятно, что и завтра будет дождь, поэтому p(A|B) > p(B).

В общем, вероятность при логическом объединении событий

p(A и B) = p(A) p(B|A)

для любых A и B. Таким образом, шанс, что дождь пойдет в любой данный день, равна 0,5, а шанс, что дождь пойдет два дня подряд, равна не 0,25, а несколько выше.

Задача о булочках

Мы вскоре приступим к теореме Байеса. Но прежде я хочу объяснить ее с помощью примера, известного как задача о булочках[1]. Предположим, что имеется две корзины с булочками. В корзине под номером 1 лежит 30 ванильных и 10 шоколадных булочек, а в корзине под номером 2 лежит по 20 булочек обоих сортов.

[1] Основана на примере из http://en.wikipedia.org/wiki/Bayes'__theorem. Эта интернет-страница больше не существует.

Теперь предположим, что вы случайным образом выбираете одну из корзин и, не заглядывая внутрь, берете первую попавшуюся булочку. Ею оказывается ванильная булочка. Какова вероятность, что он выбрана из корзины 1?

Это условная вероятность. Мы хотим определить p(корзина1|ваниль), но не понятно, как это сделать. Проще ответить на другой вопрос – какова вероятность, что ванильная булочка лежит в корзине номер 1:

p(корзина1|ваниль) = 3/4.

Хотя p(*A*|*B*) не то же, что p(*B*|*A*), существует способ, как из одного получить другое. И здесь нам поможет теорема Байеса.

Теорема Байеса

Теперь у нас есть все необходимое, чтобы вывести теорему Байеса. Начнем с коммутативности объединения:

p(*A* и *B*) = p(*B* и *A*).

Это утверждение справедливо для любого из событий *A* и *B*.

Далее напишем соотношение для вероятности объединения:

p(*A* и *B*) = p(*A*) p(*B*|*A*).

Так как нами ничего не было сказано о значении *A* и *B*, можно предположить, что они взаимозаменяемые. Их взаимозаменяемость приводит к утверждению:

p(*A* и *B*) = p(*B*) p(*A*|*B*).

Это все, что нам необходимо. Приравнивая соотношения друг к другу, получаем:

p(*B*) p(*A*|*B*) = p(*A*) p(*B*|*A*).

Здесь можно сказать, что существует два способа объединения. Если у вас есть p(*A*), то вы умножаете его на p(*B*|*A*). Или можете сделать по-другому: если вам известно p(*B*), то вы умножаете его на p(*A*|*B*). Оба способа приводят к одному и тому же результату.

Наконец, мы можем разделить правую часть на p(*B*):

$$p(A \mid B) = \frac{p(A)p(B \mid A)}{p(B)}.$$

Это и есть теорема Байеса! Может быть, она простенькая на вид, но, как окажется, удивительно действенна.

Например, мы можем использовать ее для решения задачи о булочках. Я назову гипотезой B_1 событие, что булочка была вынута из корзины 1, и гипотезой *V*, что она оказалась ванильной. Подставляя в теорему Байеса, мы получаем:

$$p(B_1|V) = \frac{p(B_1)\, p(V|B_1)}{p(V)}.$$

Левая часть формулы – это то, что мы хотим: вероятность корзины 1 при условии, что мы выбрали ванильную булочку. В правой части формулы имеем:

- $p(B_1)$ – вероятность, что мы выбрали корзину 1 независимо от того, какую булочку мы взяли. Поскольку в задаче сказано, что выбор корзины случайный, полагаем, что $p(B_1) = 1/2$;
- $p(V|B_1)$ – это вероятность получения ванильной булочки из корзины 1, которая равна 3/4;
- $p(V)$ – вероятность вынуть ванильную булочку из любой корзины. Поскольку у нас одинаковый шанс выбора из любой корзины и в обеих корзинах находится одинаковое количество булочек, мы имеем одинаковый шанс вытащить любую булочку. В двух корзинах находится 50 ванильных и 30 шоколадных булочек, поэтому $p(V) = 5/8$.

Подставив это значение в формулу, получаем:

$$p(B_1|V) = \frac{(1/2)(3/4)}{5/8},$$

что составляет 3/5. Отсюда следует, что ванильная булочка свидетельствует в пользу гипотезы, что мы вынимали булочку из корзины 1, потому более вероятно, что ванильная булочка вынута из корзины 1.

Этот пример демонстрирует пользу теоремы Байеса: данная теорема обеспечивает стратегию выбора между $p(B|A)$ и $p(A|B)$. Эта стратегия полезна в задачах, подобных задаче о булочках, где проще вычислять значения правой части теоремы Байеса, чем левой.

Диахроническая интерпретация

Существует другой подход к теореме Байеса. Эта теорема дает возможность обновить вероятность гипотезы H при наличии некоторого объема данных D.

Такое представление теоремы Байеса называется **диахронической интерпретацией**. «Диахроническое» означает что-то, происходящее с течением времени. В таком случае при изменении в течение времени старых и появлении новых данных вероятность гипотез меняется.

Переписывая теорему Байеса с H и D, получаем

$$p(H \mid D) = \frac{p(H)\, p(D \mid H)}{p(D)}.$$

В этой интерпретации каждая составляющая формулы имеет свое наименование:

- $p(H)$ – вероятность гипотезы до получения новых данных. Данное значение называется априорной вероятностью, или просто **приор**;

- $p(H|D)$ – это то, что мы хотим определить: вероятность гипотезы после получения новых данных. Это значение называется апостериорной вероятностью, или **постериор**;
- $p(D|H)$ – вероятность данных для этой гипотезы, называемой **правдоподобием**;
- $p(D)$ – вероятность данных для любой из гипотез, носящей название **нормализующей константы**.

Иногда мы можем вычислить приор на основе предварительной информации. Например, в задаче с булочками предопределено, что вероятность случайного выбора корзины одинакова.

В других случаях приор субъективен, то есть разумный человек может с этим не согласиться либо по причине использования другой исходной информации, либо потому, что интерпретирует ту же информацию по-иному.

Правдоподобие вычисляется наиболее просто. В задаче о булочках, если известно, из какой корзины мы достаем булочку, мы находим вероятность ванильной булочки простым подсчетом.

Нормализующая константа может быть ненадежной. Предполагается, что это вероятность полученных данных по любой гипотезе, но в самом общем случае трудно интерпретировать их значение.

Чаще всего мы упрощаем дело, определяя гипотезы как:

- *взаимно исключаемые*: не более чем одна гипотеза из данного множества может быть верной;
- *совместно исчерпывающие*: не существует других возможностей. Хотя бы одна из гипотез должна быть верной.

Я использую термин **suite** (**комплект**) относительно множества гипотез, обладающих этими свойствами.

В задаче с булочками присутствуют только две гипотезы – булочка достается из корзины 1 или из корзины 2. Эти гипотезы взаимно исключающие и совместно исчерпывающие.

В этом случае можно вычислить $p(D)$, используя закон полной вероятности. Данный закон говорит: если имеется два исключающих варианта того, что может случиться, то вы можете добавить такую вероятность:

$$p(D) = p(B_1)\,p(D|B_1) + p(B_2)\,p(D|B_2).$$

Подставляя значения из задачи о булочках, имеем:

$$p(D) = (1/2)\,(3/4) + (1/2)\,(1/2) = 5/8.$$

То есть мы получили то же самое значение, что вычислили ранее, мысленно объединив две корзины.

Задача M&M

M&M – шоколадное драже, поверхность которого окрашена в разные цвета. Компания «Марс» (Mars, Inc.) изготавливает это драже, время от времени меняя цвета.

В 1995 году драже окрашивались в синий цвет. До этого пакетик M&M содержал драже следующих цветов: 30% коричневых, 20% желтых, 20% красных, 10% зеленых, 10% оранжевых и 10% желто-коричневых. В дальнейшем цвета драже были изменены следующим образом: 24% синих, 20% зеленых, 16% оранжевых, 14% желтых, 13% красных, 13% коричневых.

Допустим, у моего приятеля было два пакетика M&M. Один пакетик 1994 года выпуска, а другой – 1996-го. Не сообщив, какой пакетик какого года выпуска, мой приятель дал мне по одному драже из каждого. Одно драже было желтым, другое – зеленым. Какова вероятность, что желтое драже было из пакетика 1994 года выпуска?

Задача аналогична задаче о булочках. С той разницей, что я вынимал один экземпляр булочки из корзины (пакетика).

Задача о драже позволяет мне продемонстрировать табличный метод решения, который полезен для письменного решения задачи. В следующей главе мы будем решать задачу о драже на компьютере.

Сначала перечислим гипотезы. Пакетик, из которого я получил желтое драже, назовем Пакетик 1. Другой назовем Пакетик 2. Итак, мы имеем следующие гипотезы:

- A: Пакетик 1 1994 года выпуска предполагает, что Пакетик 2 выпуска 1996 года;
- B: Пакетик 1 1996 года выпуска, а Пакетик 2, наоборот, выпуска 1994 года.

Теперь составим таблицу со строками для каждой из гипотез и со столбцами для каждой составляющей теоремы Байеса.

	Приор $p(H)$	**Правдоподобие** $p(D\|H)$	$p(H)\ p(D\|H)$	**Постериор** $p(D\|H)$
A	1/2	(20)(20)	200	20/27
B	1/2	(10)(14)	70	7/27

Первый столбец содержит приоры. Исходя из условий задачи, вероятность, что Пакетик 1 1994 года выпуска, такая же, как и то, что он выпущен в 1996 году. То же самое можно сказать и о Пакетике 2. Поэтому разумно выбрать $p(A) = p(B) = 1/2$.

Второй столбец содержит правдоподобия, которые следуют из содержания задачи. Например, если справедливо A, то желтое драже с 20%-ной вероятностью поступило из пакетика 1994 года, а зеленое – с вероятностью 20% из пакетика 1996 года. Поскольку выборки независимы, мы посредством умножения получаем совместную вероятность.

Третий столбец – результат умножения данных из первых двух столбцов. Это нормализующие константы для каждой строки. Сумма нормализующих констант равна 270. Последний столбец содержит постериоры, для получения которых следует разделить нормализующую константу соответствующей строки на сумму констант.

Просто. Не правда ли?

Однако вас может беспокоить одна деталь. Я описал $p(D|H)$ в процентах, а не в значениях вероятности. Это означает увеличение относительно значения

терминов вероятности в 10 000 раз. Но эта деталь нивелируется при делении на нормализующую константу и, следовательно, не влияет на результат.

Если множество гипотез является взаимно исключаемым и совместно исчерпывающим, то, если это удобно, вы можете умножать правдоподобие на любой коэффициент, применив этот коэффициент к значениям во всех колонках.

Задача Монти Холла

Задача Монти Холла, возможно, вызвала наибольшие споры в истории вероятности. Сценарий прост. Но правильный ответ, казалось, настолько противоречит здравому смыслу, что многие никак его не могут воспринять. И даже умные люди ставят себя в неловкое положение, не только отстаивая неверный ответ, но и агрессивно защищая его на публике.

Монти Холл был первым ведущим шоу «*Давай сделаем дело*». Задача Монти Холла основана на одной из игр, регулярно проводимых на этом шоу. Если бы вы участвовали в этом шоу, то с вами происходило бы следующее:

- Монти показывал на три закрытые двери и говорил, что за каждой дверью находится приз: за одной дверью – автомобиль, а за двумя другими дверями – два различных малоценных приза. Например, арахисовое масло и накладные ногти. Призы за дверями размещались случайным образом;
- вам необходимо догадаться, за какой дверью находится автомобиль. Если вы угадывали, то получали автомобиль в качестве приза;
- есть три двери: Дверь A, Дверь B и Дверь C. Вы указывали на выбранную дверь;
- прежде чем открыть выбранную вами дверь, Монти увеличивал неопределенность, открывая либо Дверь B, либо Дверь C. Но обязательно открывал ту дверь, за которой автомобиля не было. Понятно, если автомобиль действительно находился за Дверью A, Монти мог безопасно открыть Дверь B или C, выбирая любую из них случайным образом;
- затем Монти предлагал вам выбрать: или остановиться на выбранной вами двери, или указать на дверь, оставшуюся закрытой.

Вопрос состоит в том, следует ли вам «остаться» на ранее выбранной двери или «переключиться» на оставшуюся закрытой дверь. Или это не важно.

Большинство людей чисто интуитивно полагают, что это не важно. Осталось две двери, рассуждают они. Поэтому шанс, что автомобиль находится за Дверью A, равен 50%.

Но это неверно. Фактически шанс выиграть, если вы останетесь на прежнем выборе Двери A, составит лишь 1/3. А если вы «переключитесь» на другую дверь, то ваш шанс составит 2/3.

Применив теорему Байеса, мы разобьем эту задачу на несколько частей и, возможно, убедимся, что такой ответ действительно корректен.

Для начала нам следует аккуратно выбрать исходные данные. Здесь D состоит из двух частей: Монти выбирает Дверь B, *и* за ней нет автомобиля.

Теперь мы определяем три гипотезы: A, B и C. То есть A, B и C представляют гипотезы, когда автомобиль находится за Дверью A, Дверью B и Дверью C. Вновь применим табличный метод.

	Приор p(*H*)	**Правдоподобие p(*D*|*H*)**	**p(*H*) p(*D*|*H*)**	**Постериор p(*D*|*H*)**
A	1/3	1/2	1/6	1/3
B	1/3	0	0	0
C	1/3	1	1/3	2/3

Заполнение столбца Приор не представляет трудностей, потому что нам сказано, что призы размещены случайным образом. Это предполагает, что автомобиль с одинаковой вероятностью может находиться за любой из трех дверей.

Подсчет правдоподобий потребует некоторых рассуждений. Но, соблюдая известную осторожность, мы сможем определить эти рассуждения правильно:

- если автомобиль действительно находится за дверью A, Монти мог безопасно открыть двери B и C. Поэтому вероятность, что он выберет C, равна 1/2. А поскольку автомобиль в действительности находится за дверью A, вероятность того, что автомобиль не за дверью B, равна 1;
- если автомобиль в действительности находится за дверью B, то Монти должен открыть дверь C. Поэтому вероятность, что он откроет дверь B, равна 0;
- наконец, если автомобиль находится за дверью C, то Монти открывает дверь B с вероятностью 1 и не находит там автомобиля с вероятностью 1.

Самая трудная часть решения позади. Далее простая арифметика. Сумма третьей колонки в таблице равна 1/2. Разделив полученные результаты, имеем p(*A*|*D*) = 1/3 и p(*C*|*D*) = 2/3.

Существует много вариантов задачи Монти Холла. Одна из сильных сторон байесовского подхода – в том, что он обобщает методику решения этих вариантов задач.

Например, предположим, что Монти всегда выбирает B, если есть такая возможность, и только C, если он должен это сделать (поскольку автомобиль за дверью B). В этом случае преобразованная таблица выглядит так:

	Приор p(*H*)	**Правдоподобие p(*D*|*H*)**	**p(*H*) p(*D*|*H*)**	**Постериор p(*D*|*H*)**
A	1/3	1	1/3	1/2
B	1/3	0	0	0
C	1/3	1	1/3	1/2

Единственное, что претерпело изменения, – p(*D*|*A*). Если автомобиль за дверью A, Монти может открыть или дверь B, или дверь C. Но в этом варианте он всегда выбирает B. Поэтому p(*D*|*A*) =1.

В результате правдоподобие всегда одинаково для гипотез A и C, и постериоры тоже одинаковы: p($A|D$) = p($C|D$) = 1/2. В этом случае факт, что Монти открывает B, не дает информации о размещении автомобиля. Поэтому не важно, участник «остается» на ранее выбранной двери или «переключается» на другую дверь.

С другой стороны, если он открыл дверь C, мы знаем, что p($B|D$) = 1.

Я включил задачу Монти Холла в эту главу, так как она мне кажется забавной и потому, что теорема Байеса несколько уменьшает сложность решения задачи. Но эта задача не является типичной для применения теоремы Байеса. Так что если вы нашли ее слегка сбивающей с толку, не унывайте!

Обсуждение

Для многих задач, затрагивающих условные вероятности, теорема Байеса обеспечивает стратегию «разделяй и властвуй». Если вычисление p($A|B$) или ее экспериментальное определение затруднительны, проверьте, не проще ли вычислить другие составляющие теоремы Байеса p($B|A$), p(A), p(B).