Томас Нилд. Математика для data science

Растущая доступность данных привела к тому, что data science и машинное обучение стали востребованными профессиональными областями. Если вы стремитесь сделать карьеру в области data science, искусственного интеллекта или инженерии данных, вам просто необходимо разбираться в основах теории вероятностей, линейной алгебры, математической статистики и машинного обучения. В книге ровно столько высшей математики, математического анализа и статистики, сколько нужно, чтобы лучше понимать, как работают библиотеки, с которыми вы встретитесь.

Томас Нилд. Математика для data science. – Астана: Спринт Бук, 2025. – 352 с.

Купить в Ozon

Чтобы запускать примеры из этой книги, вам нужна среда разработки Python 3. Основные библиотеки Python, которые нам понадобятся, — это numpy, scipy, sympy и sklearn. Если вы не знакомы с Python, рекомендую книгу Джоэла Граса Data Science. Наука о данных с нуля. Вторая глава этой книги — лучший экспресс-курс по Python, который мне встречался. Даже если вы никогда раньше не писали код, Джоэл проделал фантастическую работу, чтобы вы смогли эффективно освоить Python в кратчайшие сроки. Вспомогательные материалы (примеры кода, упражнения и т. д.) можно скачать здесь.

> Книга содержит массу примеров кода на Python, но, как обычно, я предпочитаю примеры в Excel. Они набраны с отступом.

SymPy — это мощная и фантастическая система компьютерной алгебры для Python, которая использует точные символьные вычисления, а не приближенные расчеты в десятичных дробях. Она полезна в тех ситуациях, когда вы бы решали задачи по математике и математическому анализу на бумаге, но отличается тем, что предлагает знакомый синтаксис языка Python. Вместо того чтобы представлять √2 в виде приближенного значения 1,4142135623730951, SymPy сохранит его в формате sqrt(2).

Глава 2. Теория вероятностей

Вероятность — это степень уверенности в том, что событие произойдет, часто выражаемая в процентах. Вот некоторые примеры вопросов, ответ на которые можно оценить как вероятность.

- Какова вероятность того, что, если подбросить монету 10 раз, 7 раз выпадет орел?
- Каковы мои шансы победить на выборах?
- Опоздает или нет мой рейс?
- Насколько я уверен в том, что товар бракованный?

Мы будем обозначать эту вероятность P(X) , где X — интересующее нас событие. Вероятность может быть выражена в виде отношения шансов O(X), например 7:3. Чтобы преобразовать отношение шансов O(X) в вероятность P(X) , воспользуйтесь формулой:

$$(1)\ P(X) = \frac{O(X)}{1 + O(X)} = \frac{7/3}{1 + 7/3} = 0{,}7$$

Шансы полезны, чтобы количественно оценить субъективную уверенность, особенно в контексте азартных игр или ставок. Отношение шансов играет важную роль в байесовской статистике (в том числе в вычислении коэффициента Байеса), а также в логистической регрессии.

Теория вероятностей чисто теоретически оценивает, насколько вероятно наступление того или иного события, и не требует привлекать дополнительные данные. *Статистика* же, напротив, не может существовать без данных и использует их, чтобы выявить вероятности, а также предоставляет инструменты для того, чтобы описывать данные.

Биномиальное распределение

В оставшейся части главы мы изучим два распределения вероятностей: биномиальное и бета-распределение. Они служат полезными инструментами и принципиально важны для того, чтобы выяснить, как наступают события при определенном количестве испытаний.

Допустим, вы разрабатываете новый турбореактивный двигатель и провели 10 испытаний, в результате получив восемь успешных исходов и два неудачных. Вы надеялись получить 90% успешных испытаний, но на основании полученных данных пришли к выводу, что испытания провалились, ведь только 80% из них оказались успешными. Каждое испытание отнимает много времени и средств, поэтому вы решили, что пора вернуться к чертежной доске и перепроектировать конструкцию.

Однако одна из ваших инженеров настаивает, что необходимо провести дополнительные испытания.

— Единственный способ узнать наверняка — это провести больше испытаний, — утверждает она. — А что, если при большем количестве испытаний окажется, что не менее 90% из них будут успешными? В конце концов, если подбросить монету 10 раз и получить 8 орлов, это не значит, что монета «настроена» на 80% орлов.

Недолго думая, вы соглашаетесь с доводами инженера. Даже если честно подбрасывать монету, не всегда будет выпадать одинаковое количество орлов и решек, особенно когда ее подбрасывают всего 10 раз. Скорее всего, выпадет пять орлов, но может выпасть также три, четыре, шесть или семь. Может выпасть даже 10 орлов, хотя это крайне маловероятно. Как же оценить правдоподобность того, что 80% испытаний завершились успешно, притом что настоящая вероятность успеха равна 90%?

Одним из инструментов, который может здесь пригодиться, является биномиальное распределение. Оно позволяет оценить правдоподобность того, что в серии из n испытаний, вероятность успеха в каждом из которых равна p, может произойти всего k успехов:

	A	B	E	F
1	k	Вероятность k успехов из 10 попыток	p	0,9
2	0	0,0000%	n	10
3	1	0,0000%		
4	2	0,0000%		
5	3	0,0009%		
6	4	0,0138%		
7	5	0,1488%		
8	6	1,1160%		
9	7	5,7396%		
10	8	19,3710%		
11	9	38,7420%		
12	10	34,8678%		
13				
14	**Ячейка**	**Формула**		
15	A2	=ПОСЛЕД(11;;0)		
16	B2	=БИНОМ.РАСП(A2#;F2;F1;ЛОЖЬ)		
17				

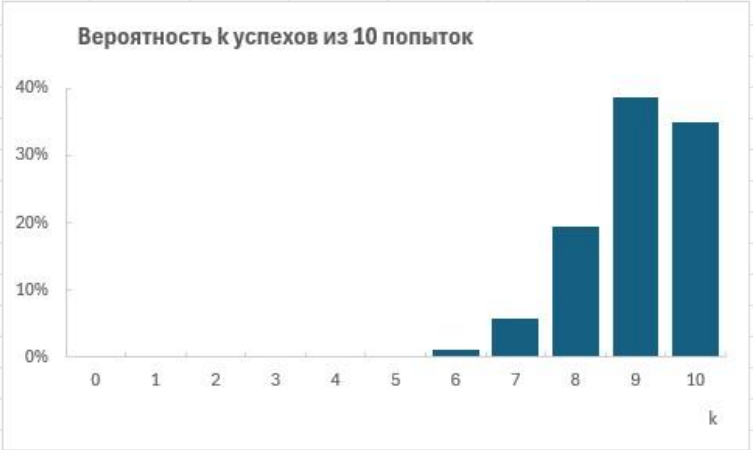


Рис. 2.1. Биномиальное распределение

Здесь для каждого значения k от 0 до 10 показана вероятность того, что k из 10 испытаний будут успешными. Это биномиальное распределение предполагает, что вероятность успеха каждого отдельного испытания p = 90%. Если это так, то вероятность получить 8 успехов из 10 испытаний равна 0,1937.

Чтобы вычислить вероятность восьми или менее успехов, можно сложить значения всех столбиков до восьмого включительно. В результате получится 0,2639. Как реализовать биномиальное распределение?

В Excel можно воспользоваться функцией =БИНОМ.РАСП(k;n;p;ЛОЖЬ)

Как видите, мы задаем n — количество испытаний, p — вероятность успеха для каждого испытания и k — количество успехов, вероятность которого мы хотим найти. Мы перебираем все значения k и для каждого из них вычисляем вероятность того, что мы получим k успехов. Наиболее вероятное количество успехов равно девяти.

Но если сложить вероятности того, что произойдет от 0 до 8 успехов, то получится 0,2639. То есть существует вероятность 26,39% того, что произойдет восемь или меньше успехов, даже если вероятность успеха каждого испытания составляет 90%. Так что не исключено, что инженер была права: вероятность 26,39% — не мелочь и вполне возможна.

Однако в этой модели мы сделали одно допущение, которое рассмотрим далее на примере бета-распределения.

Бета-распределение

Слабое место нашего биномиального распределения в том, что мы предположили, будто вероятность успеха каждого отдельного испытания составляет 90%. Если это так, то вероятность получить 8 или меньше успехов из 10 испытаний составляет 26,39%.

Но давайте перевернем вопрос на 180°: а что, если существуют другие значения базовой вероятности, помимо 90%, которые дают 8 успехов из 10 испытаний? Можно ли получить такой результат, если вероятность успеха каждого испытания равна 80%? 70%? 30%? Зафиксировав условие «8 успехов из 10 попыток», можно ли исследовать вероятности вероятностей?

Вместо того чтобы в поисках ответа на этот вопрос плодить несметное количество биномиальных распределений, воспользуемся одним инструментом. Бета-распределение позволяет оценить правдоподобность того, что при a успехов и b неудач базовая вероятность успеха равна тому или иному значению.

	A	B	C	D	E	F	G	H	I	J	K
1	p	Правдоподобность		число успехов	a или α	8					
2	0,000	0,0000		число неудач	b или β	2					
3	0,010	0,0000									
4	0,020	0,0000		**Ячейка**	**Формула**						
5	0,030	0,0000		A2	=ПОСЛЕД(100;;0;1/100)						
6	0,040	0,0000		B2	=БЕТА.РАСП(A2#;F1;F2;ЛОЖЬ)						
7	0,050	0,0000									
8	0,060	0,0000									
9	0,070	0,0000									
10	0,080	0,0000									
11	0,090	0,0000									
12	0,100	0,0000									
13	0,110	0,0000									
14	0,120	0,0000									
15	0,130	0,0000									
16	0,140	0,0001									
17	0,150	0,0001									
18	0,160	0,0002									
19	0,170	0,0002									
20	0,180	0,0004									
21	0,190	0,0005									
22	0,200	0,0007									

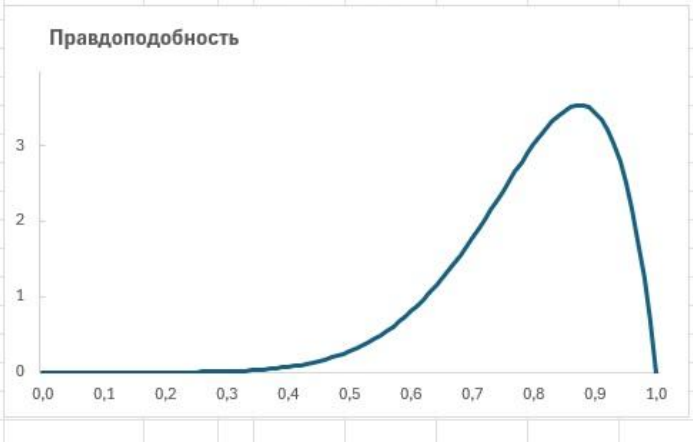


Рис. 2.2. Плотность вероятности бета-распределения

Если вам интересно поэкспериментировать с бета-распределением, можно воспользоваться графическим калькулятором Desmos.

Обратите внимание, что ось x отражает все базовые вероятности успеха от 0,0 до 1,0, а ось y — правдоподобность этой вероятности при восьми успехах и двух неудачах. Другими словами, бета-распределение позволяет увидеть вероятность вероятностей при 8 успехах из 10 попыток. Эту величину можно считать метавероятностью.

Вы наверняка заметили, что бета-распределение — непрерывная функция, то есть ее график — непрерывная кривая из дробных значений (в отличие от четких дискретных целых чисел в биномиальном распределении). Это несколько усложняет математические вычисления, потому что значения по оси y — не вероятность, а плотность вероятности. Чтобы узнать ту или иную вероятность, нужно вычислить соответствующую площадь под графиком.

Бета-распределение — это один из видов распределения вероятностей. Это значит, что площадь под всем графиком равна 1, или 100%. Чтобы найти вероятность, которая нас интересует, нужно вычислить площадь под кривой в определенном интервале. Например, если мы ищем вероятность того, что 8 успехов из 10 попыток произойдут при базовой вероятности успеха 90% или более, нам нужно вычислить площадь области между 0,9 и 1, которая равна 0,225, как показано на рис. 2.3.

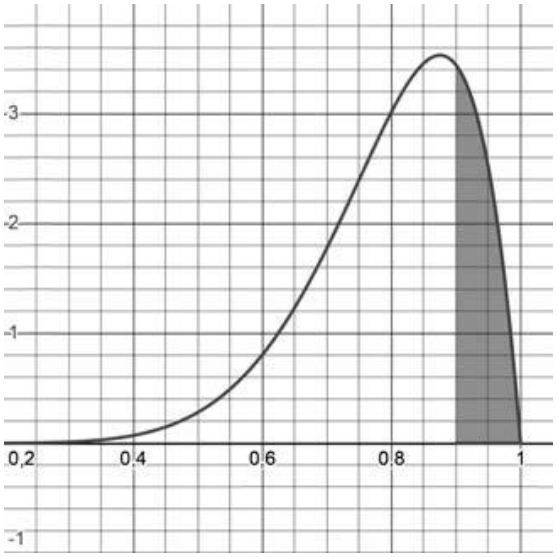


Рис. 2.3. Площадь под графиком в интервале от 90 до 100% соответствует базовой вероятности успеха 22,5%

В Excel площадь под кривой для любого x можно найти с помощью интегральной функции распределения, заменив последний аргумент функции =БЕТА.РАСП(A2#;F1;F2;**ИСТИНА**)

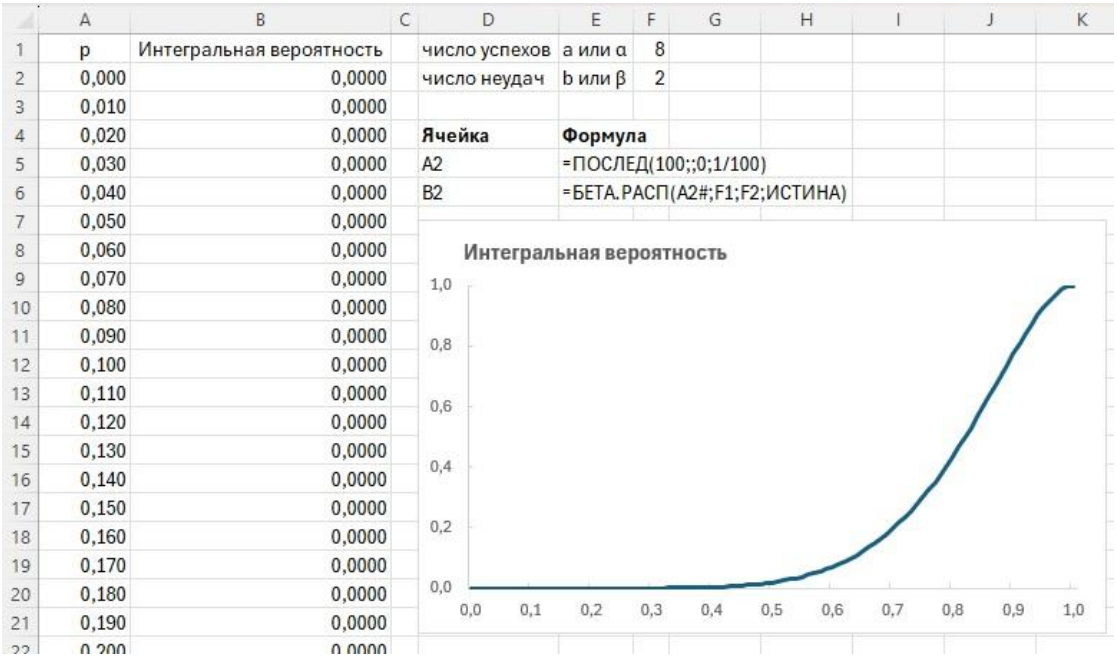
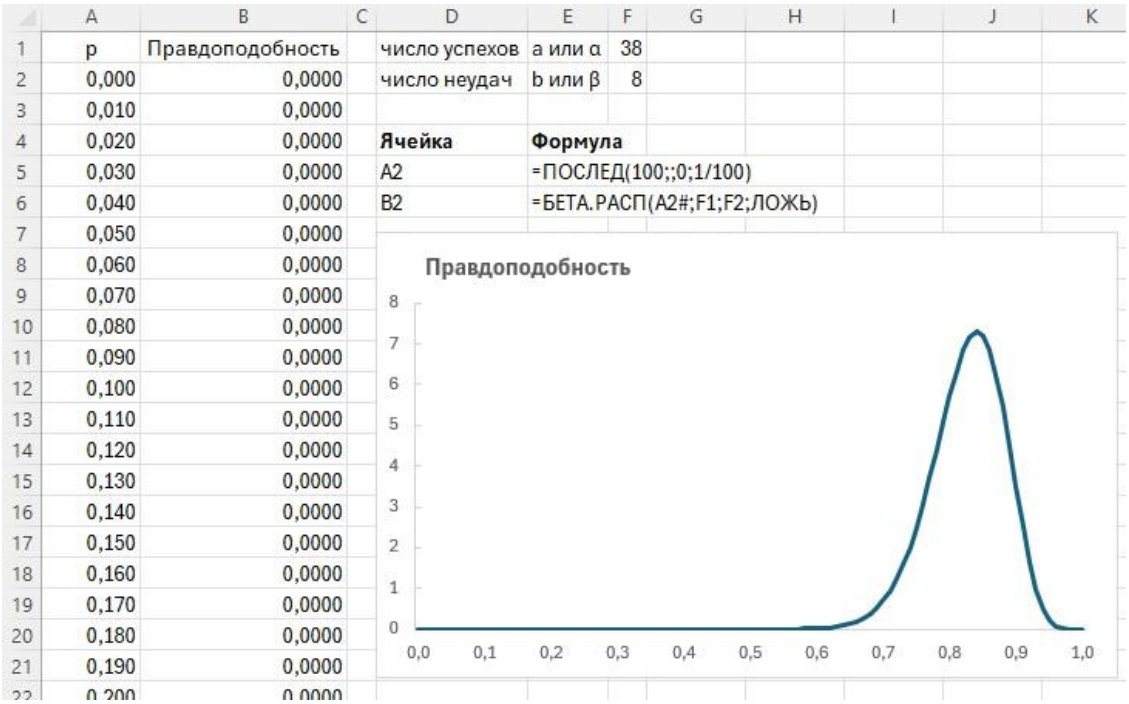


Рис. 2.4 Функция бета-распределения [интегральная]

А как вычислить вероятность того, что базовая вероятность успеха равна 90% или более? Функция распределения (CDF, рис. 2.4) определяет площадь слева от граничного значения.

Поэтому воспользуемся формулой: =БЕТА.РАСП(1;8;2;ИСТИНА)-БЕТА.РАСП(0,9;8;2;ИСТИНА)

Это значит, что при 8 успехах из 10 испытаний двигателя вероятность того, что базовая вероятность успеха составляет 90% или более, равна 22,5%. Шансы на то, что испытания следует признать успешными, не в нашу пользу, но можно попытаться использовать этот 22,5%-ный шанс как повод для того, чтобы провести дополнительные испытания, надеясь на то, что нам повезет. Если финансовый директор выделит средства еще на 36 испытаний, из которых 30 окажутся успешными, а 6 — неудачными, то бета-распределение изменится:



	A	B	C	D	E	F	G	H	I	J	K
1	p	Правдоподобность		число успехов	a или α	38					
2	0,000	0,0000		число неудач	b или β	8					
3	0,010	0,0000									
4	0,020	0,0000		**Ячейка**	**Формула**						
5	0,030	0,0000		A2	=ПОСЛЕД(100;;0;1/100)						
6	0,040	0,0000		B2	=БЕТА.РАСП(A2#;F1;F2;ЛОЖЬ)						
7	0,050	0,0000									
8	0,060	0,0000									
9	0,070	0,0000									
10	0,080	0,0000									
11	0,090	0,0000									
12	0,100	0,0000									
13	0,110	0,0000									
14	0,120	0,0000									
15	0,130	0,0000									
16	0,140	0,0000									
17	0,150	0,0000									
18	0,160	0,0000									
19	0,170	0,0000									
20	0,180	0,0000									
21	0,190	0,0000									
22	0,200	0,0000									

Рис. 2.7. Бета-распределение после дополнительных 36 испытаний

Обратите внимание, что распределение стало уже, а значит, мы больше уверены в том, что базовая вероятность успеха находится в меньшем диапазоне. К сожалению, вероятность того, что мы укладываемся в заданный минимальный показатель 90%, уменьшилась с 22,5 до 7,57%:

=БЕТА.РАСП(1;38;8;ИСТИНА)-БЕТА.РАСП(0,9;38;8;ИСТИНА)

Бета-распределение — замечательный инструмент, с помощью которого можно на основе ограниченного множества наблюдений оценивать вероятность того, что событие произойдет или не произойдет. Оно позволяет рассуждать о вероятностях вероятностей, и его можно корректировать по мере того, как поступают новые данные. Бета-распределение также помогает проверять гипотезы, но в главе 3 мы уделим больше внимания тому, как использовать для этого нормальное распределение и распределение Стьюдента.

Глава 3. Описательная статистика и статистический вывод

Данные — это моментальные снимки, на которых отражено то, что «попало в объектив» в определенный момент времени. Нужно четко обозначать цели проекта, потому что это помогает следить за тем, чтобы собирать актуальные и полные данные. А если ставить перед собой чересчур широкие и неопределенные задачи, то можно столкнуться с проблемами из-за ложных выводов и неполных наборов данных. Дата-майнинг (добыча данных) имеет смысл в определенных ситуациях, но этим инструментом не стоит злоупотреблять.

Данные не отражают контекст и не предлагают объяснений. Это очень важный момент, потому что данные дают подсказки, а не истину. Эти подсказки могут привести к истине, но могут и ввести в заблуждение, если сделать из них ошибочные выводы. Вот почему так важно уметь интересоваться тем, откуда берутся данные.

Существует широко распространенное мнение, будто можно загрузить данные в алгоритм машинного обучения и ждать, что компьютер сам все сделает. Но, как гласит пословица, «мусор на входе — мусор на выходе». Неудивительно, что, по данным сайта VentureBeat только 13% проектов в области машинного обучения оказываются успешными. В успешных проектах осмысляются и анализируются как сами данные, так и их источники.

Типы смещения

Как ни странно, мы, люди, склонны к смещениям. Мы ищем закономерности даже там, где их нет. Возможно, это было эволюционно необходимо, чтобы выживать на заре человечества, потому что поиск закономерностей помогал эффективнее заниматься охотой, собирательством и земледелием.

Существует множество типов смещения, но все они приводят к одному и тому же — к искажению результатов. *Склонность к подтверждению* проявляется, когда вы собираете только те данные, которые подтверждают вашу точку зрения. *Смещение из-за самоотбора* определенные категории респондентов с большей вероятностью включат себя в исследование. *Ошибка выжившего* происходит, когда исследование затрагивает только тех, кто остался в живых, и не учитывает умерших.

Многие консалтинговые компании и книжные издательства любят выявлять качества успешных компаний и руководителей и представлять эти качества как залог будущих успехов. Эти работы — чистой воды ошибка выжившего. В них не учитываются компании и деятели, которые обладали теми же «качествами успеха», но потерпели неудачу и остались неизвестными. Просто мы не слышали о них, потому что они никогда не бывали в центре внимания (см. также По следам великих компаний Джима Коллинза. Каковы их результаты сегодня, и что из этого следует…).

Математика и компьютеры не способны распознать смещение в ваших данных. Если вы хороший специалист по data science, то сами должны его обнаружить! Всегда уточняйте, как именно были получены данные, а затем тщательно анализируйте, как конкретно метод сбора данных мог их исказить.

Статистический вывод

Люди устроены так, что склонны к смещениям и стремятся быстро делать выводы. Чтобы быть хорошим специалистом в data science, необходимо подавлять это природное желание и задумываться о том, что то или иное явление может объясняться по-разному. Иногда вполне допустимо (возможно, даже лучше всего) предположить, что объяснения вообще нет, и вы наблюдаете лишь случайное совпадение.

Обсудим проверку гипотез на примере. Проведенные ранее исследования показали, что среднее время восстановления после простуды составляет 18 дней со стандартным отклонением в 1,5 дня и соответствует нормальному распределению. Это значит, что с вероятностью около 95% восстановление займет от 15 до 21 дня:

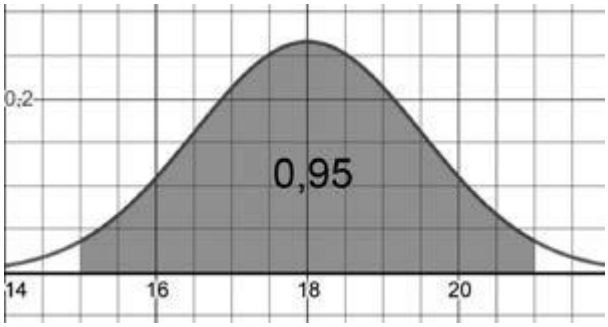


Рис. 3.16. Вероятность восстановления в течение 15–21 дня составляет 95%

Поскольку оставшаяся вероятность составляет 5%, можно сделать вывод, что существует вероятность 2,5% того, что восстановление займет больше 21 дня, и 2,5% того, что оно займет меньше 15 дней.

Теперь предположим, что группа из 40 пациентов получила новое экспериментальное лекарство, и им потребовалось в среднем 16 дней, чтобы восстановиться после простуды. Показало ли лекарство статистически значимый результат? Или лекарство не подействовало, а выздоровление за 16 дней оказалось случайным совпадением в испытуемой группе? Первый из этих вопросов определяет альтернативную гипотезу, а второй — нулевую.

Существует два способа узнать значимость: односторонний и двусторонний тест . Начнем с одностороннего. Когда используется односторонний тест, нулевую и альтернативную гипотезы обычно формулируют в виде неравенств. Мы выдвигаем гипотезы относительно генерального среднего и рассматриваем варианты, что оно либо больше или равно 18 (нулевая гипотеза H_0), либо строго меньше 18 (альтернативная гипотеза H_1):

H_0: генеральное среднее ≥ 18

H_1: генеральное среднее < 18

Чтобы отвергнуть нулевую гипотезу, нужно показать, что выборочное среднее пациентов, которые принимали лекарство, не было результатом случайного стечения обстоятельств. Поскольку p-значение 0,05 или меньше традиционно считается статистически значимым, мы возьмем его в качестве порогового. Вычислим функцию обратную функции распределения (Inverse Cumulative Distribution Function, Inverse CDF или Percent-Point Function, PPF)…

	A	B	C	D
1	p-значение	α	0,05	
2	среднее	μ	18	
3	станд. откл.	σ	1,5	
4	односторонний тест			
5	граница	z критическое	15,53	=НОРМ.ОБР(C1;C2;C3)

Рис. 3.17а. Вычисления в Excel

Любопытно замечание научного редактора русского перевода.

В книгу вкралась досадная ошибка: в примерах 3.19–3.22 и в сопутствующем тексте ни расчеты, ни конечный результат не зависят от размера выборки (n = 40 пациентов), в то время как здравый смысл подсказывает, что доверительный интервал и p-значение должны уменьшаться с ростом n.

На самом деле в этих расчетах вместо стандартного отклонения должна фигурировать стандартная ошибка среднего

$$(2)\ SE_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

Поскольку стандартное отклонение генеральной совокупности σ неизвестно, его можно оценить выборочным стандартным отклонением s. Впрочем, в этом случае результаты тестов будут другими: и односторонний, и двусторонний тест покажут статистическую значимость на уровне 95%, и это позволит отвергнуть нулевую гипотезу и сделать вывод, что лекарство действительно помогает восстанавливаться после простуды.

Я сталкивался с аналогичной дилеммой – что использовать: стандартное отклонение или стандартную ошибку среднего – на занятиях с магистрами по курсу статистики в МФТИ, и могу сказать, что… не согласен с замечанием научного редактора. На мой взгляд у автора всё верно.

Таким образом, если бы среднее выборочное время восстановления в нашей группе пациентов составило 15,53 дня или меньше, то можно было бы считать, что лекарство показало статистически значимый эффект. Однако в нашей выборке среднее составляет 16 дней и не попадает в область отклонения нулевой гипотезы. Таким образом, тест на статистическую значимость не прошел.

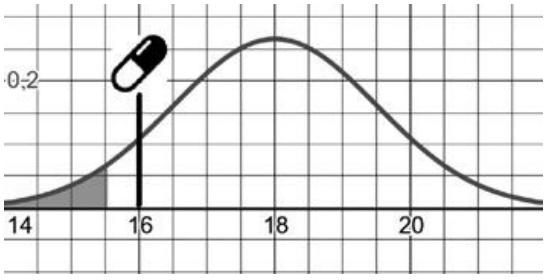


Рис. 3.19. Мы не смогли доказать, что результат испытания лекарства является статистически значимым

Кое-что о больших данных и ошибке меткого стрелка

Когда мы обосновываем свои выводы, всегда следует учитывать роль случайных совпадений. К сожалению, с тех пор, как появились большие данные, машинное обучение и другие инструменты обработки данных, научный метод внезапно превратился в практику, которая работает задом наперед. Это может быть опасно. Почему это так — позвольте продемонстрировать на примере из книги Gary Smith Standard Deviations.

Представим, что я вытягиваю четыре игральные карты из обычной колоды. Это не какая-то игра, и нет никакой другой цели, кроме как вытянуть четыре карты и рассмотреть их. Мне достаются две десятки, тройка и двойка. «Интересно, — говорю я. — Я получил две десятки, тройку и двойку. Нет ли в этом закономерности? Если я вытяну еще четыре карты, будут ли это тоже две последовательные очковые карты и пара? Какая модель лежит в основе этого эксперимента?»

Видите, что я сделал? Я взял совершенно случайное явление и не только предположил закономерности, но и попытался построить на их основе прогностическую модель. Между тем я изначально не ставил своей задачей вытянуть именно такую структуру из четырех карт. Я пронаблюдал ее после того, как она образовалась.

Это именно то, чем постоянно грешит дата-майнинг: он выявляет случайно образовавшиеся структуры в случайных событиях. Когда вам доступны огромные объемы данных и быстрые алгоритмы, которые ищут закономерности, нетрудно отыскать результаты, которые выглядят как закономерности, но на самом деле являются случайными совпадениями.

Это все равно, как если бы я стрелял из пистолета по стене, а потом рисовал мишень вокруг отверстия и приглашал друзей, чтобы продемонстрировать свою потрясающую меткость. Глупо, правда? Однако многие деятели в области data science фактически занимаются этим день ото дня, и эта практика известна как *ошибка меткого стрелка*. Они начинают перемалывать данные без определенной цели, натыкаются на что-то редкое, а затем провозглашают, будто найденное ими каким-то образом создает предсказательную ценность.

Глава 4. Линейная алгебра

В области работы с данными вектор — это массив чисел, в котором хранятся данные.

Если вы хотите больше узнать о линейной алгебре, то вам не найти лучшего ресурса, чем серия видеороликов на YouTube-канале Essence of Linear Algebra (Главное о линейной алгебре) от 3Blue1Brown. Также полезны видео по линейной алгебре на канале PatrickJMT. А если вы хотите освоить NumPy, рекомендую прочитать книгу Уэса Маккинни. Python и анализ данных: Первичная обработка данных с применением pandas, NumPy и Jupiter. В ней не так много внимания уделяется линейной алгебре, зато это прекрасное практическое руководство о том, как работать с наборами данных с помощью NumPy, pandas и Python.

Глава 5. Линейная регрессия

Регрессия пытается подогнать функцию к наблюдаемым данным, чтобы спрогнозировать новые данные. *Линейная* регрессия подгоняет к данным прямую линию, пытаясь установить линейную связь между переменными и предсказать новые данные, которые еще предстоит наблюдать.

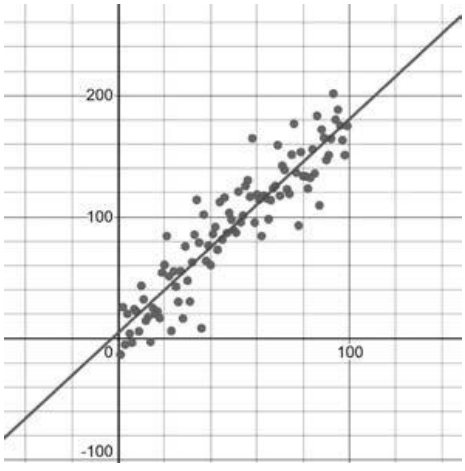


Рис. 5.1. Пример линейной регрессии, которая подгоняет прямую к наблюдаемым данным

Переобучение и дисперсия

Подумайте вот о чем: если бы мы хотели полностью минимизировать потери, то есть уменьшить сумму квадратов до 0, как бы мы поступили? Есть ли другие варианты, кроме линейной регрессии? Одно из решений, которое приходит в голову, — это построить кривую, которая соединяет все точки. Действительно, почему бы просто не соединить точки отрезками, и не использовать эту модель для прогнозов? Это даст нам нулевые потери!

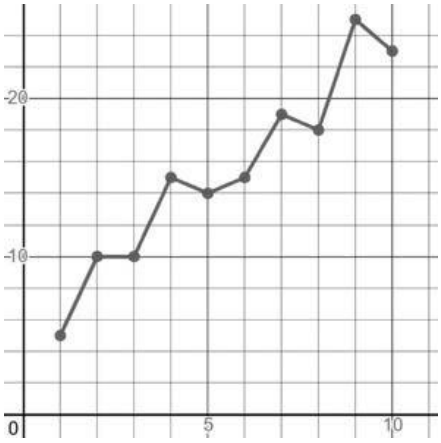


Рис. 5.10. Регрессия путем простого соединения точек приводит к нулевым потерям

И правда, зачем мы так долго возились с линейной регрессией, а не поступили по-простому? Помните, что наша стратегическая цель — не минимизировать сумму квадратов, а делать точные предсказания на новых входных данных. Эта модель «рисования по точкам» сильно переобучена, то есть она слишком точно подстраивает регрессию под обучающие данные и в результате будет плохо работать с новыми данными. Переобученная модель чувствительна к выбросам, которые находятся далеко от остальных точек, а значит, в прогнозах будет высокая дисперсия. Хотя в этом примере точки расположены относительно близко к прямой линии, проблема будет намного хуже, если работать с другими наборами данных, где наблюдается больший разброс и выбросы. Поскольку из-за переобучения увеличивается дисперсия, прогнозируемые значения могут оказаться где ни попадя!

Когда вы слышите, будто регрессия «запомнила» (или «вызубрила») данные, а не обобщила их, — речь идет о *переобучении*. Как вы уже догадались, модель строится затем, чтобы найти эффективные обобщения, а не зубрить данные. Иначе регрессия превратится просто в базу данных, которая годится только для того, чтобы искать в ней уже имеющиеся значения.

Именно поэтому в машинном обучении к модели добавляется смещение, а линейная регрессия считается сильно смещенной моделью. Это не то же самое, что смещение в данных, о котором мы говорили в главе 3. Смещение модели означает, что мы отдаем предпочтение определенной схеме (например, тому, чтобы поддерживать прямую линию) в противовес тому, чтобы изгибать график и точно подгонять модель под данные. Смещенная модель оставляет некоторое пространство для маневра, благодаря чему можно минимизировать потери на новых данных и получить более точные прогнозы вместо того, чтобы минимизировать потери на данных, на которых модель была обучена. Можно сказать, что, если добавить смещение в модель, мы предотвращаем переобучение, ради чего допускаем недообучение, то есть меньшую подгонку к обучающим данным.

Статистическая значимость

В линейной регрессии нужно учитывать еще один аспект: не является ли корреляция данных случайной? В главе 3 мы познакомились с проверкой гипотез и p -значениями, а теперь рассмотрим эти понятия на примере линейной регрессии.

Начнем с основополагающего вопроса: может ли быть так, чтобы линейная зависимость в данных наблюдалась в результате случайного совпадения? Как добиться 95%-ной уверенности в том, что корреляция между двумя переменными значима, а не случайна? Если это напомнило вам проверку гипотез из главы 3, то знайте, что это она и есть! Нам нужно не просто вычислить коэффициент корреляции, но и количественно оценить, насколько мы уверены в том, что он возник не случайно.

Здесь мы оцениваем не среднее арифметическое, как в главе 3 на примере с испытанием лекарства, а коэффициент корреляции генеральной совокупности на основе выборки. Коэффициент корреляции совокупности мы обозначим греческой буквой ρ, а выборки — r. Как и в главе 3, у нас будет нулевая и альтернативная гипотезы:

H_0: $\rho = 0$ (нет взаимосвязи);

H_0: $\rho \neq 0$ (есть взаимосвязь).

Вернемся к набору данных из 10 точек, который представлен на рис. 5.10. Насколько правдоподобно, что эти точки образовались случайно и при этом их конфигурация напоминает линейную зависимость?

Коэффициент корреляции составил 0,957586. Это сильная и убедительная положительная корреляция. Но по-прежнему нужно оценить, не объясняется ли она случайным везением. Давайте проверим нашу гипотезу с помощью двустороннего теста на доверительном уровне 95% и выясним, есть ли связь между этими двумя переменными.

Для проверки гипотез в контексте линейной регрессии мы будем использовать распределение Стьюдента. Построим график распределения Стьюдента и выделим на нем 95%-ный критический интервал. В нашей выборке 10 записей, поэтому у распределения будет 9 степеней свободы.

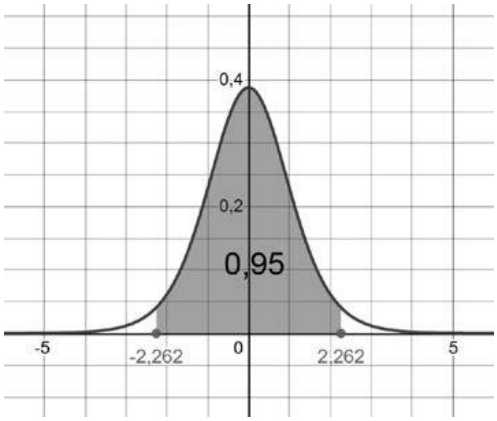


Рис. 5.14. Распределение Стьюдента с 9 степенями свободы

Критическое значение составляет примерно ±2,262.

В Excel можно использовать формулу =СТЬЮДЕНТ.ОБР.2Х(0,05;9)

Если тестовое значение (так называемая t-статистика) окажется вне интервала (-2,262, 2,262), можно будет отвергнуть нулевую гипотезу. Чтобы рассчитать t-статистику, воспользуемся следующей формулой. Здесь r — коэффициент корреляции, а n — размер выборки:

$$(3)\ t = \frac{r}{\sqrt{\frac{1-r^2}{n-2}}} = \frac{0{,}957586}{\sqrt{\frac{1-0{,}957586^2}{10-2}}} = 9{,}339958$$

Тестовое значение составляет 9,39958, что определенно находится за пределами диапазона (-2,262, 2,262), поэтому можно отвергнуть нулевую гипотезу и признать, что наша корреляция не случайна. Это связано с тем, что p-значение весьма значимо: 0,000005976. Это намного ниже нашего порога в 0,05, так что мы имеем дело не с совпадением: корреляция обоснована. Вполне логично, что p-значение так мало, потому что расположение точек очень похоже на прямую. Крайне маловероятно, что они выстроились в ряд случайно.

Обучающая и тестовая выборки

К сожалению, практикующие специалисты по data science часто пренебрегают анализом, который я только что провел, когда вычислил коэффициент корреляции, статистическую значимость и коэффициент детерминации.

Основной метод, с помощью которого специалисты по машинному обучению борются с переобучением, заключается в том, чтобы разделять набор данных на обучающую и тестовую выборки. Обычно при этом 1/3 данных включается в тестовую выборку, а остальные 2/3 — в

обучающую, хотя бывают и другие пропорции. Обучающая выборка используется для того, чтобы подогнать линейную регрессию, а тестовая — чтобы оценить эффективность регрессии на данных, с которыми она раньше не сталкивалась. Этот прием обычно используется во всех видах машинного обучения с учителем, включая логистическую регрессию и нейронные сети.

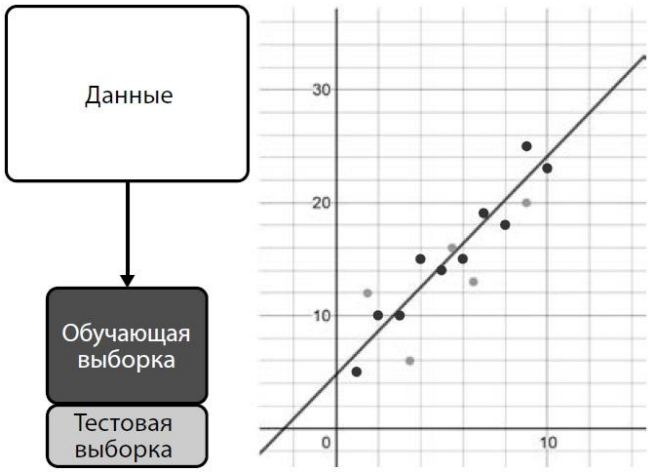


Рис. 5.18. Разделение данных на обучающую и тестовую выборки. Прямая регрессии подгоняется под обучающие данные (обозначены темным цветом) по методу наименьших квадратов, а затем проверяется на тестовых данных (обозначены светлым цветом), чтобы понять, насколько ошибочны предсказания на данных, которые не встречались ранее

Обучение — это подгонка регрессии. Обратите внимание, что подгонять регрессию — это то же самое, что обучать ее. Специалисты по машинному обучению предпочитают второй термин.

Глава 6. Логистическая регрессия и классификация

Логистическая функция — это S-образная кривая (также известная как *сигмоида*), которая в зависимости от значений входных переменных принимает значения между 0 и 1. Поскольку выходная переменная лежит в интервале от 0 до 1, с ее помощью можно представлять вероятность. Вот логистическая функция, которая выдает вероятность y для одного аргумента x:

$$(4)\ y = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x)}}$$

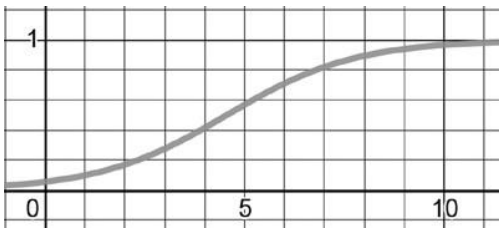


Рис. 6.6. График логистической функции

Подгонка логистической кривой

Чтобы подогнать логистическую функцию к заданной обучающей выборке, нам нужно найти коэффициенты β_0 и β_1. Метод наименьших квадратов здесь не подходит. Вместо этого нужен *метод максимального правдоподобия*. В отличие от линейной регрессии, не существует аналитического решения, с помощью которого вычисляются искомые коэффициенты.

Логит-функция

С начала XX века математиков интересовало, как отмасштабировать область значений линейной функции так, чтобы они лежали в интервале от 0 до 1, а значит, подходили для того, чтобы прогнозировать вероятность. В логистической регрессии для этого можно использовать функцию логарифма шансов, которая также называется логит-функцией или просто логитом.

$$(5)\ p = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x)}}$$

Линейная функция, которая стоит в показателе степени числа e, — это и есть логарифм шансов для интересующего нас события.

Возможно, вы удивитесь: «Подождите, я не вижу тут ни логарифма, ни шансов. Это просто линейная функция!» Немного терпения, и я покажу вам все математические подробности.

$$(6)\ \text{шансы} = \frac{p}{1-p}$$

$$(7)\ \text{логит} = ln\left(\frac{p}{1-p}\right)$$

Логарифмы и шансы интересным образом связаны друг с другом. Шансы на неудачу события находятся только в интервале от 0 до 1, а шансы на успех занимают всю числовую прямую от 1 до положительной бесконечности. Такая асимметрия выглядит неаккуратно. Однако логарифмирование масштабирует шансы так, что они становятся полностью линейными и симметричными. Если логит равен 0, это означает, что шансы на успех и на неудачу одинаковы. Логит, равный -1,05, линейно находится на том же расстоянии от 0, что и 1,05, поэтому сравнивать шансы гораздо проще.

Если p — это вероятность из логистической регрессии, а x — входная переменная, то верно равенство:

$$(8)\ ln\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x$$

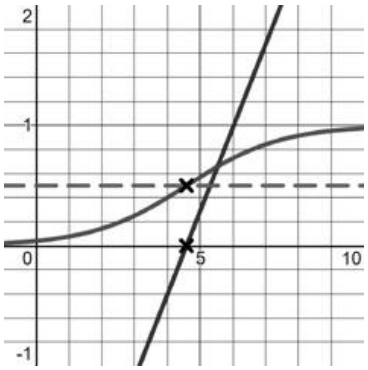


Рис. 6.10. Прямая логарифма шансов преобразуется в логистическую функцию, которая возвращает вероятность

Каждая логистическая регрессия на самом деле опирается на линейную функцию, и эта функция — логарифм шансов. Когда логит равен 0, то вероятность на логистической кривой равна 0,5.

Глава 8. Советы по дальнейшей карьере

Data science анализирует данные, чтобы получать результаты, которые можно непосредственно применить на практике. В действительности эта дисциплина объединяет в себе различные области, связанные с данными: статистику, анализ и визуализацию данных, машинное обучение, исследование операций, разработку программного обеспечения. Поэтому я говорю своим клиентам, что лучше всего считать, что data science — это разработка программного обеспечения на основе статистики, машинного обучения и оптимизации.

Эта глава важна, если вы хотите ориентироваться на рынке трудоустройства в области data science и эффективно применять полученные знания. Вряд ли вам будет приятно изучать статистические инструменты и машинное обучение только для того, чтобы обнаружить, что большинство вакансий побуждают вас переключиться на другую работу. Если так случится, рассматривайте эту ситуацию как возможность продолжить обучение и приобрести новые умения. Когда вы объедините свои фундаментальные математические знания с навыками программирования и разработки программного обеспечения, вы станете бесценным специалистом, потому что сможете преодолеть разрыв между IT и data science.

Обращайте внимание не на рекламную шумиху, а на практические решения и не замыкайтесь на чисто технических аспектах, чтобы не пострадать от «невидимой руки рынка». Старайтесь понимать мотивы руководства, а также людей в целом. Интересуйтесь, почему тот или иной метод или инструмент решает проблему, а не только как он работает с технической точки зрения.

Учитесь не ради того, чтобы учиться, а ради того, чтобы наращивать свои возможности и подбирать правильные инструменты, которые позволяют решать правильно поставленные задачи. Один из

самых эффективных способов обучения — взяться за проблему, которая вам интересна (а не зацикливаться на определенном инструменте). Потянув за эту ниточку, вы обнаружите еще одну интересную тему, а потом еще одну, и еще. Не упускайте из виду поставленную цель, продолжайте углубляться в нужные темы и вовремя отделывайтесь от ненужных. Такой подход оправдывается с лихвой и позволяет получить удивительно много знаний и опыта за короткое время.