Главный закон подлости

«Все делают общие выводы из одного примера. По крайней мере, я делаю именно так.» Стивен Браст

Цифры обманчивы, особенно когда я сам ими занимаюсь; по этому поводу справедливо высказывание, приписываемое Дизраэли: «Существует три вида лжи: ложь, наглая ложь и статистика».

— Марк Твен

Статистику обвиняют в массе грехов: и во лжи и в возможностях манипуляций и, наконец, в непонятности. Но мне очень хочется реабилитировать эту науку, показав насколько сложна задача, для которой она предназначена. Посудите сами: теория вероятностей оперирует точными знаниями о случайных величинах в виде распределений или исчерпывающих комбинаторных подсчётов. Ещё раз подчёркну что располагать точным знанием о случайной величине возможно. Но что если это точное знание нам недоступно, а единственное чем мы располагаем – это наблюдения? У разработчика нового лекарства есть какое-то ограниченное число испытаний, у создателя системы управления транспортным потоком — лишь ряд измерений на реальной дороге, у социолога – результаты опросов, причём, он уверен в том, что на какие-то вопросы люди не дали искреннего или корректного ответа. Одно наблюдение не даёт ровным счётом ничего. Два – немного больше, три, четыре… сто… сколько нужно наблюдений чтобы получить какое-либо знание о случайной величине, в котором можно было бы быть уверенным с математической точностью? И что это будет за знание? Скороее всего, оно будет представленно в виде таблицы или гистограммы, дающей возможность оценить моменты распределения, быть может, глядя на гистограмму удастся угадать его точную форму. Но все эти результаты сами будут случайными величинами! Покуда мы не знаем распределения точно, все результаты наблюдений дают нам лишь вероятностное описание случайного процесса! Случайное описание случайного процесса – ещё бы здесь не запутаться или даже не подмухлевать!

Что же делает математическую статистику точной наукой? Её методы позволяют заключить наше незнание в чётко ограниченные рамки и дать вычислимую гарантию, того что в этих рамках наше знание точно. *Это язык, на котором можно рассуждать о неизвестных случайных величинах так, чтобы рассуждения имели смысл.* Такой подход очень полезен в философии, психологии или социологии, где очень легко пуститься в пространные рассуждения и дискуссии без всякой надежды на получение позитивного знания и, тем более, на доказательство. Грамотной статистической обработке данных посвящена масса литературы, ведь это абсолютно необходимый инструмент для медиков, социологов, экономистов, физиков, психологов… словом, для всех изучающих так называемый «реальный мир», отличающийся от идеального математического лишь степенью нашего незнания.

Теперь ещё раз взгляните на эпиграфы к этой главе и осознайте, что статистика, которую так пренебрежительно называют третьей степенью лжи – это единственное, чем располагает любой нематематик. Это ли не главный закон подлости мироздания! Все известные нам законы природы, от физических до экономических, строятся на математических моделях и их свойствах, но поверяются они статистическими методами в ходе измерений и наблюдений. В повседневности наш разум делает обобщения и подмечает закономерности, выделяет и распознаёт повторяющиеся образы, это, наверное, лучшее, что умеет человеческий мозг. Это именно то, чему в наши дни учат искусственный интеллект. Но разум экономит свои силы и склонен делать выводы по единичным наблюдениям, не сильно беспокоясь о точности или обоснованности этих выводов. И пока речь идёт об искусстве, характере домашних любимцев или обсуждении политики, об этом можно сильно не беспокоиться. Однако при строительстве самолёта, организации диспетческой службы аэропорта или тестировании нового лекарства, уже нельзя сослаться на то, что «мне так кажется», «интуиция подсказывает» и «в жизни всякое бывает». Тут приходится ограничивать свой разум рамками строгих математических методов.

Наша книжка не учебник и мы не будем детально изучать и доказывать статистические методы, но мне хотелось бы показать ход рассуждений и форму результатов, характерных для этой области знания.

Основными столпами матстатистики являются *закон больших чисел* и *центральная предельная теорема*. Первый, в вольной трактовке, говорит о том, что большое число наблюдений случайной величины отражает её распределение, так что наблюдаемые моменты: среднее, дисперсия и прочие характеристики, стремятся к точным значениям, соответствующим случайной величине. Иными словами, гистограмма наблюдаемых величин при бесконечном числе данных, стремится к истинному распределению.

Второй столп, опять же, в вольной трактовке, говорит, что наиболее вероятной формой распределения случайной величины является нормальное (гауссово) распределение. Точная формулировка звучит иначе: среднее значение большого числа идентично распределённых вещественных случайных величин, вне зависимости от их распределения, описывается нормальным распределением. Эту теорему обычно доказывают, применяя методы функционального анализа, но в контексте нашей книжки её проще понять вспомнив, что нормальное распределение имеет наибольшую энтропию при наименьшем числе ограничений. В этом смысле, нормальное распределение оптимально при описании неизвестной случайной величины, либо случайной величины, являющейся совкупностью многих других величин, распределение которых нам может быть неизвестно.

Эти два закона лежат в основе количественных оценок достоверности наших знаний, основанных на наблюдениях. В частности, в простейшем случае – для распределения Бернулли с параметром p, из них вытекает полезная в практике оценка достоверности предположения о значении параметра:

Испытывая случайную величину, подчинённую распределению Бернулли, можно быть уверенным в правильной оценке вероятности «успеха» p, если число «успешных» результатов находится в диапазоне n p \pm 2\sqrt{n p (1-p)}.

Это соответствует «правилу 2\sigma», которая даёт 95% вероятность того, что мы угадали параметр верно. Если заменить двойку перед корнем на тройку, то степень уверенности вырастет до 99.5%.

Может возникнуть вопрос: в каком смыле мы «угадываем» параметр? Неужели учёные угадывают или не угадывают свои знания? Здесь речь идёт о статитстическом подтверждении или опревержении предположения, которое можно сделать из каких-то общих оснований и математической модели. Такое предположение называется гипотезой и если число результатов, опровергающих эту гипотезу не выходит за пределы 2\sigma, то мы вынуждены признать, что у нам нет оснований для её опровержения.

Здесь можно подумать, что исследователи вынуждены совершать одну из классических логических ошибок, которая носит звучное латинское имя *Ad ignorantiam**(«к невежеству»).* Это аргументация истинности некоторого утверждения, основанная на отсутствии доказательства его ложности. Классический пример — слова, сказанные сенатором Джозефом Маккарти, когда его попросили предъявить факты для поддержки выдвинутого им обвинения, что некий человек является коммунистом: «У меня немного информации по этому вопросу, за исключением того общего заявления компетентных органов, что в его досье нет ничего, чтобы исключало его связи с коммунистами». Разница в том, что ...

В этой главе мы рассмотрим на ряд примеров использования статистического подхода.

В главе про монетку мы упомянули результат группы Перси Диакониса, говорящий о принипиальной, хоть и небольшой, нечестности процесса подбрасывания монетки. Напомню, что вероятность того, что монетка выпадет той же стороной, что была сверху при подбрасывании, оказалась равной 51%. Имеет ли смысл такое отклонение? Можно ли его заметить в экспериментах?

По мере накопления экспериментальных данных, стандартная ошибка среднего, отражающая погрешность, с которой может быть вычислена средняя величина, уменьшается пропорционально квадратному корню из числа испытаний: , здесь  — стандартное отклонение для исследуемого распределения. В нашем случае, для распределения Бернулли с вероятностью , которое равно . Чтобы уверенно выделить отклонение среднего в одну сотую, это отклонение должно превышать 2 стандартных отклонения. Таким образом, мы можем оценить число испытаний:

Постулат Персига  
Число разумных гипотез, объясняющих любое данное явление, бесконечно.

Прежде чем приступать к поиску причин совпадений непогоды и выходных, давайте выясним, сколько нужно провести наблюдений за погодой, чтобы быть уверенным в том, что эти совпадения неслучайны. Каждый дождливый день можно рассматривать как наблюдение случайной величины – дня недели, подчиняющегося распределению Бернулли с вероятностью 1/7. Примем в качестве гипотезы предположение, что все дни недели одинаковы с точки зрения погоды и дождь может пойти в любой из них равновероятно. Выходных у нас два, итого, получаем ожидаемую вероятность совпадения непогожего дня и выходного равной 2/7, эта величина будет параметром распределения Бернулли. Как часто идёт дождь? В разное время года по-разному, конечно, но в Петропавловске-Камчатском, в среднем, наблюдается девяносто дождливых или снежных дней в году. Так что поток дней с осадками имеет интенсивность около 90/365=1/4. Давайте посчитаем какое количество дождливых выходных мы должны зарегистировать, для того, чтобы быть уверенным в том, что существует некоторая закономерность. Напомню, что для распределения Бернулли число значимых отклонений от ожидаемой оценивается как . Вот что у нас получается:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Период наблюдений | Ожидаемое число наблюдений | Ожидаемое число положительных исходов | Значимое отклонение | Значимая доля непогожих в общем числе выходных дней |
| лето | 23 | 6 | 4 | 42% |
| год | 90 | 26 | 9 | 33% |
| 5 лет | 456 | 130 | 19 | 29% |

О чем говорят эти цифры? Если вам кажется, что который год подряд «лета не было», что злой рок преследует ваши выходные, насылая на них дождь, это можно проверить и подтвердить Однако в течение лета уличить злой рок можно лишь если больше двух пятых всех выходных окажутся дождливыми. Нулевая же гипотеза предполагает, что тоько четверть выходных должна совпасть с ненастной погодой. За пять лет наблюдений уже можно надеяться подметить тонкие отклонения выхадящие за пределы 5% и приступать к их объяснению.

Я воспользовался дневником погоды, который велся с 2014 по 2018 год и выяснил, что за пять лет случилось 459 ненастных дней из них 137 пришлись на выходные, это, действительно, больше ожидаемого числа, но всего на 7, а значимые отклонения начинаются с 19, так что это, как мы говорили в детстве: «не считается». Вот как выглядит гистограмма, показывающая распределение непогоды по дням недели.

Стоит ли после такого наблюдения искать причины совпадения или можно смело ссылаться на случайные флуктуации? Можно принять во внимание, что ненастные дни идут чередой и, может быть, мы неверно оценили вероятность попадания непогоды на выходные? Давайте проверим эту гипотезу. Если предположить, что дожди идут по два дня, то вероятность перекрыть выходные увеличивается до 3/7. При такой вероятности ожидаемое число совпадений для пяти лет должно составить 195+-21, то есть от 174 до 216 раз. Наблюдённая величина 137 не входит в этот диапазон и значит гипотезу об эффекте сдвоенных дней непогоды можно смело отвергать. И вообще, какие-то более тонкие эффекты рассматривать нет резона, поскольку наблюдения и, что самое главное, их количество, согласованно говорят в пользу самого простого объяснения.

Но недовольство у нас вызывает не пятилетняя и даже не годовая статистика, человеческая память не столь долгая. Обидно, когда дождь идёт на выходных три или четыре раза подряд! Как часто это может наблюдаться? Задачу можно сформулировать по-разному: либо «Какова вероятность того, что n дождливых дней подряд придутся на выходные?», либо «Какова вероятность того, что n выходных подряд окажутся дождливыми?» Оба варианта обидные, и достойны рассмотрения.

На первый вопрос можно отвечать таким образом: вероятность того, что дождливый день будет субботой или воскресеньем составляет 2/7. Для n таких дней, идущих подряд, мы должны получить вероятнгсть (2/7)^n. То есть для двух – 8%, для трёх – 2%, для четырёх – 0.5%. При этом мы неявно предполагаем, что между этими злосчастными выходными дождей не было. За пять лет наблюдений мы можем ожидать пять цепочек из двух и одну цепочку из трёх идущих подряд ненастных дней, пришедшихся на выходные. В моём дневнике таких цепочек было, соответственно четыре и одна – нечастое явление. Вероятность получения цепочек заданной длины в бернуллевских экспериментах описывается геометрическим распределением.

Со вторым вопросом дела обстоят так: разумно предположить, что непогожие дни образуют пуассоновский поток с интенсивностью ¼ дня. Мы прореживаем его с вероятностью 1-2/7, вычёркивая будние дни. Таким образом, получаем новый поток с интенсивностью 1/4\*2/7 = 1/14 дня, то есть, ненастные выходные случаются с интенсивностью раз в две недели. Однако, мы можем «сжать» поток данных, исключив из рассмотрения будние дни, как показано на рисунке В таком случае, мы снова получаем поток с интенсивностью ¼. Данные моего дневника подтверждают это, говоря, что поток ненастных выходных имеет интенсивность 137/(5\*2\*51) = 0,26. Интервалы между пуассоновскими событиями описываются экспоненциальным распределением, нас интересуют дискретные интервалы: 0,1,2,3 дня и т. д. Значит, мы вновь приходим к геометрическому распределению, но на сей раз, с сдругим параметром, равным 1/4. Два события, идущие подряд имеют интервал равный нулю, с вероятностью ¼ – 25%, для трёх событий мы ожидаем получить вероятнгсть (¼)^2 = 6%, для четырёх – (¼)^3 = 1.5% соответствующую двум нулевым интервалам следующим подряд. Для ответа на вопрос с какой вероятностью может быть испорчено два выходных подряд, мы можем посчитать значения:

Конечно, в наших рассуждениях мы увлекаемся упрощением. Пуассоновский процесс не имеет памяти, а погода на завтра, безусловно, зависит от сегодняшней погоды, но перерыв в неделю может часть этой памяти стереть. Так что экспериментальные данные говорят в пользую нашей простой модели и не требуют введения дополнительных усложнений.

Вообще-то, погоду нельзя описывать пуассоновским процессом – это динамическая система и состояние погоды не является независимым, оно прямо связано с предыдущими состояниями. Почему же наши наблюдения говорят в пользу простой стохастической модели? Дело в том, что мы отображаем закономерный процесс формирования осадков на множество дней недели, или, говоря на языке математики, на систему вычетов по модулю семь. Этот процесс способен порождать хаос из вполне упорядоченных рядов данных. Отсюда, к примеру, происходит видимая случайность в последовательности цифр десятичной записи большинства вещественных чисел.

Это тоже своеобразный закон подлости, которыя можно назвать проклятием Пифагора: Реальный мир нерауионален

Мы уже говорили о рациональных числах, тех, которые выражаются целочисленными дробями. Они имеют внутреннюю структуру, которая определяется двумя числами: числителем и знаменателем. Но при записи в десятичной форме мы наблюдаем скачки от регулярных представлений таких чисел, как 1/2, ¼, или 1/3 до периодических, но вполне беспорядочных, таких 1/17 или 1/113. Иррациональные числа не имеют конечной или периодической записи в десятичной форме и там в последовательности чисел чаще всего царит хаос. Но это не значит, что в этих числах нет порядка! Например, первое встретившееся математикам иррациональное число sqrt 2 в десятичной записи порождает хаотический набор цифр. Однако, с другой стороны, это число можно представить в виде бесконечной цепной дроби:

нетрудно показать, что эта цепочка, действительно равна корню из двух:

Цепные дроби с повторяюшимися коэффициентами записывают подобно периодическим десятичным дробям:

Знаменитое золотое сечение в этом смысле представляет собой самое просто устроенное иррациональное число:

Все рациональные числа представляются в виде конечных цепных дробей, часть иррациональных – в виде бесконечных, но периодических, их называют алгебраическими, те же, что не имеют конечной записи даже в такой форме – трансцендентными. Самое знаменитое из трансцендентных – число pi, оно порождает хаос как в десятичной записи, так и в виде цепной дроби.

А вот число Эйлера, оставаясь трансцендентным, в форме цепной дроби проявляет внутреннюю структуру, скрытую в десятичной записи:

Наверное, не один математик подозревал мир в коварстве, обнаруживая, что такое нужное, такое фундаментальное число pi, сопровождающее циклические явления и формы, всплывающее всюду, где мы используем изотропность нашего мира (то есть, инвариантность относительно поворотов), появляющееся в нормировке гауссового распределения, в во многих других задачах, имеет столь неуловимо сложную хаотическую структуру. Конечно, его можно представить в виде сумм вполне изящных числовых рядов но эти ряды напрямую не говорят о природе этого числа и они не универсальны. Я верю, что математикам будущего откроется какое-нибудь новое представление чисел, столь же универсальное, как цепные дроби, которое позволит выявить строгий порядок, скрытых природой в числе pi.