Новости CrossValidated: выпуск 1

Андрей Костырка*

29 января 2016 г.

Аннотация

В этом разделе представлены компиляции ответов на наиболее интересные вопросы, задававшиеся на сайте StackExchange в разделе «Статистика» (stats). К каждому вопросу приводятся один или несколько ответов, получивших наибольшее количество пользовательских голосов. Мнение авторов ответов может не совпадать со мнением редакции.

Ключевые слова: статистика, вопросы, ответы, интернет.

1 Что такое over-fitting?

Вопрос. Где в реальной жизни чаще всего возникает проблема *переподгонки* (переобучения — не в смысле «обучения заново», а в смысле «чрезмерного обучения»)? Чем плоха чрезмерно точная подгонка модели под данные?

Исходный вопрос: http://stats.stackexchange.com/q/128616.

1.1 Подгонка под особенности шума

Зачастую набор данных является слишком простым, а модель — слишком «продвинутой», из-за чего оценивание даёт ложные либо нестабильные результаты оценивания. Дополнительные параметры сложных моделей иногда оцениваются по особенностям случайного шума, который совершенно не связан с самой структурой данных, но может образовывать статистические артефакты в единичной реализации.

*НИУ ВШЭ, Москва.

1.2 Модель плохо работает за пределами выборки

Несмотря на свою примитивность, линейные модели довольно неплохо дают общее представление об устройстве данных. Если же точки располагаются вдоль воображаемого нелинейного облака точек, то тогда специфические математические функции (экспонента, логарифм, полином k-й степени, синус и проч.) принимают значения, более близкие к значениям набора данных, однако за границами определённого диапазона их поведение теряет адекватность интерпретации.

Рассмотрим динамику населения США в XX веке (рис. 1). Линейная модель довольно хорошо описывает данные. Полином шестой степени более точно проходит через имеющиеся точки, однако даёт прогноз, согласно которому к 2050 году всё население США загадочным образом исчезнет. Подобное экстраполирование абсурдно, поэтому этот пример — классический случай переподгонки.

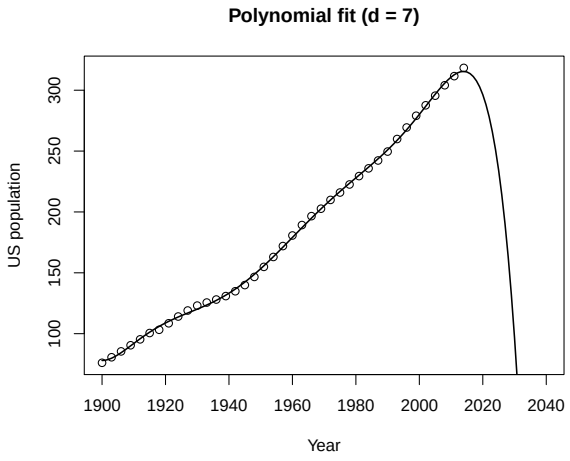

Рис. 1. Over-fitting динамики населения США по данным multpl.com

1.3 Система Птолемея и система Коперника

Птолемей полагал, что Земля находится в центре Вселенной, и вывел громоздкую систему вложенных сферических орбит, хорошо объяснявшую движение небесных тел. Однако реальные измерения систематически отличались от прогнозов, реализуемых в рамках системы Птолемея, поэтому астрономам-геоцентристам приходилось придумывать всё больше дополнительных сфер, пока, наконец, модель не стала настолько запутанной, что начала вызывать подозрения в истинности предпосылок, на которых основывалась.

Прим. ред. Система Птолемея лучше описывала видимые движения небесных тел, нежели система Коперника в её первоначальном виде. Систему Птолемея поэтому не стоит считать примером over-fitting'а, однако система Коперника — замечательный пример того, как за счёт небольшой потери предсказательной силы можно получить более элегантную модель.

1.4 Не бывает идеальной модели природы

Мы не знаем, как природа порождает данные, но лишь пытаемся её моделировать. Так как природа одна, то все её модели содержат ошибки того или иного уровня. Все модели ложны, но некоторые полезны[1]. Впрочем, также известно, что не все йогурты одинаково полезны.

В социальных науках дело осложняется большей степенью неопределенности и меньшей стабильностью зависимостей. Кроме того, большое внимание следует уделять моделированию процесса получения набора данных. Одно дело, если мы имеем данные по зарплате из бухгалтерской отчетности, и совсем другое — если мы опрашиваем людей по телефону.

1.5 Боязнь пропущенных переменных

Многие эконометристы полагают, что пропущенные переменные — это опаснейшая проблема, куда более острая, чем избыточные переменные. Чтобы выбрать из двух зол наименьшее, некоторые из этих эконометристов добавляют в модель степени регрессоров, а также всевозможные пересечения (кросс-произведения) и иррелевантные переменные. В самом общем случае добавление в уравнение множественной регрессии всех доступных в наборе переменных, которые могут потенциально обладать объясняющей силой, является переподгонкой: исследователь наблюдает не генеральную совокупность, а только выборку, поэтому он не может знать, какая из всех возможных спецификаций является верной.

Как водится, есть две новости: хорошая и плохая. Хорошая новость: включение лишних переменных не приводит к смещению оценок коэффициентов при значимых. Плохая новость: точность оценивания релевантных коэффициентов падает, ошибка регрессии растёт, доверительный интервал прогноза расширяется.

1.6 Мнение компилятора

Два критерия качества модели — goodness of fit (подгонка) и goodness of forecast (прогноз) — зачастую бывают недостижимы одновременно. Представьте себе отличную базу панельных данных с одним миллионом индивидов, каждый из которых наблюдается в течение пяти-десяти лет (такие базы,

[1] «All models are wrong, but some are useful», [Box, Draper, 1987].

например, имеются в распоряжении у французских статистических органов, а также у сотрудников Национальной школы статистики). Представим, что мы строим зарплатное уравнение в зависимости от стажа, пола, образования, возраста с квадратом и других стандартных факторов. Каждый индивид обладает своим индивидуальным эффектом, который легче всего измерить при помощи дамми-переменной для этого самого индивида. Предположим, у нас есть 8 000 000 наблюдений и 1 000 005 оцениваемых при помощи МНК коэффициентов (стаж, пол, возраст с квадратом, образование и миллион дамми). Оценённая модель будет обладать фантастически высоким показателем R^2 и отлично объяснять устройство данных; почти все коэффициенты при этом будут значимы! Однако представим, что в выборку попадает новый индивид — одна новая точка. Оценки индивидуального эффекта для него у нас попросту нет. Каким будет прогноз его заработной платы? Как в старом анекдоте: «Насчёт СССР мы не знаем, но на китайско-финской границе всё будет спокойно![2]»

Ещё более гипертрофированный пример: есть срез из 1 000 000 индивидов в один момент времени, и для каждого индивида в уравнение регрессии добавляется его дамми. Такая модель объяснит 100 % вариабельности переменной дохода, однако будет неспособна предсказать доход новых респондентов.

1.7 Вопрос читателям

Придумайте для какого-либо набора данных модель, которая обладает очень хорошей прогнозной силой (довольно точно предсказывает значения для точек как внутри диапазона значений — где модель оценивалась, — так и за его пределами — где необходимо угадать свойства объекта, не участвовавшего в обучении), однако скверной объясняющей способностью.

Прим. ред. Редакция заинтригована не меньше читателей!

2 Следует ли из статистической связи корреляция?

Вопрос. Многие студенты второго курса бывали биты за то, что утверждали, будто бы из корреляции следует каузальность, т. е. «они коррелируют, следовательно, одно является причиной другого». Любой зарубежный студент-отличник на устном экзамене повторяет мантру «correlation does not imply causation». Предположение о независимости случайных величин всегда сильнее предположения об их некоррелированности. Однако верно ли обратное? Обязательно ли из причинно-следственной связи следует корреляция?

Исходный вопрос: http://stats.stackexchange.com/q/26300.

[2] Анекдот 1975 года. Армянское радио спрашивают: — Что будет с Советским Союзом в 1984 году? Радио ответило: — Что будет с Союзом, мы не знаем, но на китайско-финской границе все будет спокойно...

Прим. ред. Здесь смешаны два разных вопроса. Первый — обязательно ли из статистической связи следует корреляция? Второй — обязательно ли статистическая взаимосвязь означает причинно-следственную? Автор оригинального текста отвечает на первый вопрос.

2.1 Линейная корреляция

Если понимать под корреляцией коэффициент корреляции Пирсона, рассчитываемый по формуле

$$\operatorname{Corr}(X,Y) = \frac{\operatorname{Cov}(X,Y)}{\sqrt{\operatorname{Var}(X)\operatorname{Var}(Y)}},$$

то причинно-следственная связь двух переменных не всегда приводит к возникновению линейной корреляции. Достаточно рассмотреть такой набор данных, как точки, лежащие на прямой $y = x^2$ в любом симметричном относительно начала координат диапазоне. В данном случае зависимость будет функциональной, однако коэффициент корреляции Пирсона будет равен нулю. Во многих справочниках присутствует изображение (рис. 2[3]), на котором показано, что у многих наборов данных две переменные явно связаны некоторой зависимостью, однако линейная корреляция равна нулю.

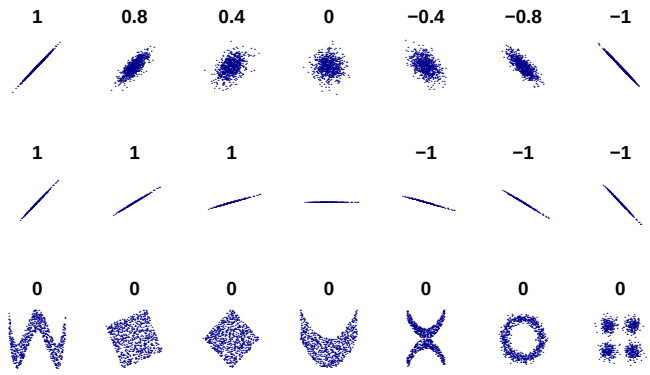

Рис. 2. Коэффициенты корреляции в различных наборах данных

Более подходящий термин для данного вопроса — это «взаимная информация». Справедливо утверждение о том, что причинно-следственная связь влечёт высокую *взаимную информацию*. Последняя имеет более сложное определение, чем корреляция, и измеряет уменьшение неопределённости относительно одной случайной величины при поступлении информации о другой случайной величине.

[3] Исходный примет взят с en.wikipedia.org/wiki/File:Correlation_examples2.svg.

2.2 Теоретический контрпример

Рассмотрим две случайные величины: $X \sim \mathcal{N}(0;1)$, $Y = X^2$. По определению $Y \sim \chi^2_1$. Трудно придумать более сильную причинно-следственную связь: X полностью определяет Y. Мы видим, что $\mathbb{E}(X) = 0$, $\mathbb{E}(Y) = 1$. Вычислим значение коэффициента ковариации.

$$\begin{aligned}\operatorname{Cov}(X,Y) &= \mathbb{E}\big((X-\mathbb{E}(X))(Y-\mathbb{E}(Y))\big) = \\ &= \mathbb{E}\big((X-0)(Y-1)\big) = \mathbb{E}(XY) - \mathbb{E}(X) = \mathbb{E}(X^3) - 0 = 0\end{aligned}$$

Мы воспользовались тем свойством, что нечётные центральные моменты стандартной нормальной случайной величины равны нулю. Следовательно, корреляция между ними также равна нулю.

Данное доказательство работает не только для стандартного нормального, но и вообще для любого симметричного относительно нуля распределения (равномерного от $-a$ до a, Лапласа, Стьюдента и проч.), у которого существуют хотя бы три центральных момента. Если каждое положительное значение величины X так же вероятно, как и противоположное ему, то при возведении в квадрат мы не можем сказать, связаны ли бóльшие значения квадрата случайной величины с положительными или отрицательными X.

2.3 Эмпирический контрпример

Если одно случайное событие является причиной другого случайного события, между ними обязана существовать некоторая взаимосвязь (односторонняя или двухсторонняя), которая может выражаться в нелинейной зависимости.

При проведении выборочных исследований домохозяйств иногда в качестве средства опроса используется телефон. При этом вероятность ответа на вопросы при телефонном интервью максимальна у людей, относящихся к среднему классу в терминах дохода, и значительно ниже у очень состоятельных или, наоборот, социально незащищённых граждан. На рис. 2 видно, что параболическая зависимость переменных (нижний ряд, центральное изображение, ветви параболы могут быть направлены вверх или вниз) влечёт нулевую корреляцию. Причинно-следственная связь вида «доход влияет на вероятность ответа на вопросы по телефону» заключается в том, что на верхнем конце распределения люди предпочитают не выдавать информации о себе (в том числе и потому, что к ним зачастую обращаются по телефону с просьбой «поделиться доходами»), а на нижнем конце распределения индивиды чаще являются должниками, имеющими непогашенный долг перед банком или знакомыми, и это является причиной их более осторожного поведения.

Прим. ред. Данные примеры показывают, что наличие статистической взаимосвязи может сопровождаться нулевой корреляцией. Тема причинно-следственных связей не раскрыта.

3 Почему вариабельность измеряют именно квадратами?

Вопрос. Почему в статистике для определения меры разброса случайной величины берутся *квадраты* отклонений от среднего, и почему стандартное отклонение считается как корень из математического ожидания квадрата? Разве ожидание модуля отклонений не покажет вариабельность данных?

$$\begin{array}{cc} \text{Почему вместо} & \text{не используют} \\ \sigma = \sqrt{\mathbb{E}\big((X-\mu_X)^2\big)} & \sigma = \mathbb{E}(|X-\mu_X|)? \end{array}$$

Исходный вопрос: http://stats.stackexchange.com/q/118.

3.1 Квадрат не единственная используемая функция

В некоторых моделях используется среднее абсолютное отклонение для определения меры разброса случайных величин. Так, при диагностике моделей временных рядов качество прогноза оценивается при помощи нескольких мер: среднее относительное отклонение, среднее абсолютное отклонение, среднее квадратичное отклонение.

Кроме того, стандартное отклонение, определяемое как корень из дисперсии, не является «стандартным» в статистической науке. Точно так же и главные компоненты в одноимённом методе не являются главным инструментом учёного: это всего лишь название.

3.2 Преимущества стандартного отклонения

Если стандартное отклонение призвано измерить разброс данных, то стоит сперва определить этот самый разброс. Функция квадратов отклонений обладает несколькими важными свойствами:

1. Функция квадрата непрерывно дифференцируема;
2. Она является достаточной статистикой для распределения Гаусса;
3. Она является разновидностью L^2-нормы, которая полезна при доказательствах сходимости;
4. Возведение в квадрат смещает вес в сторону бóльших отклонений — со всеми полезными и негативными последствиями.

Обе штрафные функции — квадрат и модуль — всегда возвращают неотрицательную величину, поэтому (за исключением вырожденного случая) сумма штрафов будет положительной.

Само по себе возведение в квадрат трудно интерпретируется, так как некоторые единицы измерения в квадрате (доллары, дни, станки) лишены физического смысла. Для возврата к оригинальным единицам считается квадратный корень из суммы.

Абсолютные отклонения (модули) назначают равные веса наблюдениям из всего диапазона значений, в то время как квадраты усиливают влияние крайних наблюдений. С алгебраической точки зрения работать намного удобнее с квадратами, в то время как модули не дают некоторых свойств (например, дисперсия равна разности ожидания квадрата и квадрата ожидания).

Удобство использования квадратов связано с теоремой Пифагора: $c^2 = a^2 + b^2$. Из неё следует, что дисперсии независимых случайных величин складываются, а стандартные отклонения — нет.

3.3 Недостатки абсолютного отклонения

Если функция модуля непрерывна всюду на $\mathbb{R}$, то её первая производная — нет (в нуле). Это усложняет аналитическое решение многих задач.

Если в линейной регрессии используется штрафная функция $L(e) = |e|$, то тогда полученная регрессия называется медианной, а в более общем случае ($(1-\alpha)|e|$ для $e < 0$ и $\alpha|e|$ для $e \geq 0$) — квантильной. Вычисление квантилей связано с задачами линейного программирования, которые могут становиться сложнее на порядок. При наличии n точек задача минимизации суммы квадратов решается за время $O(n)$, а суммы модулей — за $O(n \ln n)$, так как самый общий алгоритм подразумевает поиск решения.

3.4 Вопрос читателям

Чем бóльшие значения принимает штрафная функция при больших значениях аргумента, тем чувствительнее регрессия к большим выбросам, так как по сути происходит минимизация суммы с большой долей функции от максимальной компоненты. Рассмотрите три примера штрафной функции от остатков:

1. $L(e) = x \cdot \ln(|e| + 1)$;
2. $L(e) = x \cdot \ln^2(|e| + 1)$;
3. $L(e) = \sqrt{|e|}$.

Решите нормальные уравнения для задачи $\min \sum_i L(e_i)$ во всех трёх случаях для парной регрессии вида $y_i = \alpha + \beta x_i + \varepsilon_i$.

Сгенерируйте в любой эконометрической программной среде набор данных с известными свойствами и проведите серию экспериментов Монте-Карло, оценив в каждом случае уравнение регрессии методов наименьших штрафных функций, предложенных выше. Изучите распределение коэффициентов. Измерьте чувствительность коэффициентов к статистическим выбросам. Сравните эти оценки с оценками методов наименьших квадратов и наименьших модулей.

4 Как преобразовывать неотрицательные данные с нулями?

Если данные строго положительные, то в таком случае переменные иногда логарифмируют. Но что делать с неотрицательными данными, в которых присутствуют нули? Если рассматривать преобразование вида $\ln(x+c)$, то следует ли использовать $c = 1$, чтобы нули обратились в нули, либо оценивать $\hat{c}$, либо брать очень малое положительное значение? Есть ли другие преобразования?

Исходный вопрос: http://stats.stackexchange.com/q/1444.

4.1 Причины возникновения нулей

В первую очередь необходимо изучить природу данных и понять, почему некоторые наблюдения содержат нулевые значения. Каждую из нижеследующих причин необходимо рассматривать в отдельности:

1. Цензурирование данных (исследователь не располагает отрицательными наблюдениями, хотя они могли бы быть).
2. Пропущенные наблюдения (иногда при кодировании положительных переменных ноль используют как обозначение для пропуска в данных);
3. Естественный ноль (доход индивида может быть равен нулю, если он безработный);
4. Специфика чувствительности средства измерения переменной (инструмент не реагирует на количества, меньшие определённого порога). *Прим. ред.* На самом деле, это один из видов цензурирования.

Прим. ред. Очень многие путают цензурирование и усечение выборки. При цензурировании некоторые наблюдения в выборке заменяются на нули, в то время как настоящее значение переменной доподлинно неизвестно. Например, если в GoogleTrends попытаться узнать количество запросов по редкому ключевому слову, то GoogleTrends скажет, что их было 0, хотя на самом деле их было небольшое положительное количество. При усечении нулевые наблюдения не попадают в выборку. Например, если мы спросим посетителей автомастерской, сколько денег они потратили на ремонт машины за прошлый год, то мы не увидим тех людей, которые потратили ноль рублей.

Предлагаемые решения:

1. Использование моделей, учитывающих информацию об ограниченных значениях переменной (модель Хекмана, интервальная регрессия, модели времени жизни);
2. Выбрасывание из модели наблюдений, содержащих пропуски, если учтены все возможные последствия этого решения;
3. Числовое преобразование, упоминавшееся в вопросе;

4. Использование специальных LOD-моделей (Limit of Detection), непараметрических методов, а в первую очередь — изучение книги [Helsel, 2005].

Последний случай является самым тяжёлым, причём в эконометрике он практически не встречается (этот вопрос более актуален для специалистов, снимающих показания с реальных датчиков, обладающих порогом чувствительности).

Если переменная с нулями является объясняющей, то может быть полезно добавить дамми-переменную для наблюдений с нулями.

4.2 Числовое преобразование

В работе [Smithson, Verkuilen, 2006] предлагается использовать преобразование

$$x' = \frac{x(N-1)+s}{N}, \tag{4.1}$$

где N — число наблюдений, а s — волшебное число, о котором можно прочитать в исходной работе (но если лень, то можно использовать $s = 0{,}5$).

Кроме того, не следует забывать о преобразованиях Бокса—Кокса:

$$y^*(\lambda_1) = \begin{cases} \frac{y^{\lambda_1}-1}{\lambda_1}, & \lambda_1 \neq 0, \\ \ln y, & \lambda_1 = 0; \end{cases} \qquad y^*(\lambda_1,\lambda_2) = \begin{cases} \frac{(y+\lambda_2)^{\lambda_1}-1}{\lambda_1}, & \lambda_1 \neq 0, \\ \ln(y+\lambda_2), & \lambda_1 = 0. \end{cases}$$

Сами Бокс и Кокс в статье [Box, G. E. P., Cox, D. R., 1964] приводят алгоритм, позволяющий численно найти значения $\hat\lambda_1$ и $\hat\lambda_2$, максимизирующие значение функции правдоподобия, которая зависит не только от вектора зависимой переменной $\boldsymbol{y}$, но и от матрицы наблюдений $\boldsymbol{X}$. Существует пакет **geoR** для R, позволяющий оценить оба параметра, хотя при отсутствии возможности подогнать двухпараметрическое преобразование под данные можно воспользоваться значением $\lambda_2 = 0{,}5\min\{y_i\colon y_i \geq 0\}$ или рекомендацией «первый квартиль в квадрате делить на третий квартиль» [Stahel, 2013].

Преобразование Бокса—Кокса позволяет не только решить проблему нулей в данных, но и приблизить спецификацию к линейной.

Кроме того, существует альтернатива преобразованию Бокса—Кокса — гиперболический арксинус:

$$y^*(\theta) = \frac{\operatorname{arsh}\theta y}{\theta} = \frac{\ln(\theta y + \sqrt{\theta^2 y^2+1})}{\theta},$$

где $\theta > 0$. В статье [Burbidge, Magee, Robb, 1988] приводится метод оценивания параметра $\hat\theta$, а также предлагается (без формул) преобразование со сдвигом, похожим на сдвиг в преобразовании Бокса—Кокса, т. е. $y^*(\theta,\omega) = \frac{\operatorname{arsh}\theta(y+\omega)}{\theta}$.

4.3 Дешёвые и сердитые приёмы

Если нулей совсем немного и можно предположить, что они возникли в данных случайно, то можно удалить проблемные наблюдения и взять логарифмы. Также можно добавить небольшое $\hat{c}$, рекомендуемые значения для которого приведены в предыдущем пункте. Если результаты оценивания моделей почти не отличается, следовательно, модель довольно устойчива, поэтому следует выбирать тот набор данных, который позволяет сделать хорошо интерпретируемый вывод (пример: «в данных присутствовало 2 % безработных с нулевым доходом, однако их исключение не повлияло на значимость коэффициентов и не изменило двух первых значащих цифр оценок»).

Наконец, если требуется сделать монотонное преобразование $f(x)$ с $f' > 0$, $f'' < 0$, почему бы не рассмотреть квадратный корень? Если есть отрицательные значения, то можно использовать кубический корень.

4.4 Более сложные модели

Если данные непрерывные, то при наличии в них нулей следует обратить внимание на распределение значений: дискретный пик означает, что по той или иной причине малые значения были округлены до нуля. В любом случае, наиболее подходящие модели для данных, в которых присутствует большое количество нулей, должны учитывать вероятность того, что наблюдение будет нулём, а также условное распределение ненулевых значений (например, марковские модели для смешанных распределений, обобщённые линейные модели для дискретных данных, байесовские методы, симуляции Монте-Карло и проч.).

Список литературы

Box, G. E. P., *Cox, D. R.* An Analysis of Transformations // Journal of the Royal Statistical Society. Series B (Methodological). — 1964. — Т. 26, № 2. — С. 211—252.

Box G. E., *Draper N. R.* Empirical model-building and response surfaces. Т. 424. — Wiley New York, 1987.

Burbidge J. B., *Magee L.*, *Robb A. L.* Alternative Transformations to Handle Extreme Values of the Dependent Variable // Journal of the American Statistical Association. — 1988. — Т. 83, № 401. — С. 123—127.

Helsel D. R. Nondetects and data analysis: statistics for censored environmental data. — Wiley-Interscience, 2005. — (Statistics in practice).

Smithson M., *Verkuilen J.* A better lemon squeezer? Maxium-likelihood regression with beta-distributed dependent variables // Psychological Methods. — 2006. — Т. 11, № 1. — С. 54—71.

Stahel W. A. Statistische Datenanalyse: Eine Einführung für Naturwissenschaftler. — Springer-Verlag, 2013.