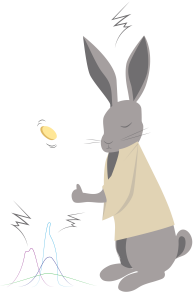
ПРИКЛАДНАЯ СТАТИСТИКА



Функции потерь

Блин блинский! Это потеря потерь!

Джейсон Стетхем

1. Какими бывают потери

Давайте представим себе машину. Она тормозит. Потому что пешеходный переход. Длина её тормозного пути зависит от разных факторов: скорости, гололёда, марки машины, шипастости шин и тп. Представим себе, что мы постоянно наблюдаем за одной и той же машиной на одной и той же дороге в одних и тех же условиях. В общем говоря, длина её тормозного пути y зависит только от скорости x с каким-то коэффициентом β, то есть

$$y = \beta x + \varepsilon.$$

В данном случае ε это шум, который накладывается на нашу взаимосвязь. В него входят различные случайные факторы, влияющие на тормозной путь (выскочившая белка, заевшая педаль и тп.). Если мы хорошо грамотно специфицировали модель, то математическое ожидание шума равно нулю.

У нас есть выборка. Мы немного понаблюдали за машиной и записали кучу измерений (x_i, y_i). Осталось только оценить коэффициент β. Возникает резонный вопрос: как это сделать?

Ответ прост. Решить насколько для нас страшно ошибиться в прогнозировании y и ввести функцию потерь. Обычно выбор конкретного вида функции зависит от поставленной задачи. Так

в эконометрике обычно выбирается квадратичная функция потерь. Оценка коэффициента находится путём минимизации квадрата ошибки, допущенной при прогнозировании

$$(y - \hat{y})^2 = (y - \beta x)^2 \to \min_{\beta}.$$

Давайте попробуем в явном виде проминимизировать такую функцию. Для каждого из наших наблюдений ошибка прогноза должна быть как можно меньше. То есть нужно минимизировать среднюю ошибку прогноза

$$\frac{1}{n} \cdot \sum_{i=1}^{n} (y_i - \beta x_i)^2 \to \min_{\beta}.$$

Берём производную, решаем уравнение, получаем ответ

$$\frac{2}{n} \cdot \left(\sum y_i x_i - \beta \sum x_i^2 = 0 \right) \Rightarrow \hat{\beta} = \frac{\sum x_i y_i}{\sum x_i^2}.$$

Взяв вторую производную, можно убедиться, что это действительно минимум. Сразу же после того, как была получена формула для оценивания бэтки,[1] возникает вполне естественный вопрос: откуда вообще взялась эта идея, минимизировать сумму квадратов отклонений?

Конечно, чем больше ошибка в прогнозе, тем сильнее нас карают за неё, но почему мы не взяли сумму модулей или четвёртых степеней? Чтобы ответить на этот вопрос, нужно ввести несколько вероятностных предположений.

Пусть ошибка в нашей регрессии зашумляет истинную взаимосвязь между переменными по нормальному распределению $\varepsilon \sim N(0, \sigma^2)$. Тогда мы можем выписать для нашей задачки правдоподобие

$$L = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{\varepsilon_i^2}{2\sigma^2}} = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y_i - \beta x_i)^2}{2\sigma^2}}$$

Прологарифмируем его

$$\ln L = -\frac{n}{2} \cdot \ln(2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - \beta x_i)^2.$$

Что мы видим? Для максимизации логарифма правдоподобия необходимо минимизировать сумму квадратов отклонений. Выходит, что метод наименьших квадратов на деле оказывается замаскированным методом максимального правдоподобия, его частным случаем. На практике

[1] Обычно в англоязычной литературе такая формула называется estimator, то есть оцениватель. Конкретная оценка называется estimate. Почему-то богатый русский язык не впитал это различие и стал называть оценкой и формулу и конкретное численной значение. Давайте исправлять это недоразумение и называть формулы оценивателями.

довольно часто функции потерь вытекают из каких-то функций правдоподобия.

Вторым важным наблюдением оказывается то, что выбор функции ошибки и распределения шума как-то взаимосвязаны между собой. Обратите внимание, что ошибки здесь имеют нулевое среднее и одинаковую дисперсию. Если вдруг мы увидим, что ошибки у нас имеют другую природу (ненулевое среднее или различные дисперсии), то с этим нужно что-то делать. Например, поискать другую функцию ошибки либо ввязаться в яростную борьбу с природой за предпосылки.

Эконометрика обычно проповедует путь борьбы. Дело в том, что оценки наименьших квадратов, при соблюдении предпосылок, обладают рядом няшных статистических свойств. Эти свойства открывают для нас целый мир, связанный с проверкой гипотез о различных взаимосвязях между переменными.

При этом главным профитом статистических процедур, проповедуемых в эконометрике, является величина эффекта. На выходе мы получаем величину $\hat{\beta}$, которую можно проинтерпретировать. Например, в нашем случае она будет означать, что при увеличении скорости на единицу, при прочих равных в среднем длина тормозного пути увеличивается на $\hat{\beta}$.

Как только мы немного видоизменим функцию потерь, например добавим для борьбы с переобучение регуляризатор, интерпретация сразу же будет утеряна. Дело в том, что регуляризация для улучшения прогнозных свойств модели вносит в неё искусственное смещение.

Интерпретация — это хорошо. Однако не стоит сковывать себя жесткими обязательствами. Мы свободны сами выбирать свою судьбу[2]. Никто не вынуждает нас останавливаться именно на такой функции потерь. Мы можем взять и использовать для решения задачи сумму модулей отклонений

$$|y - \hat{y}| = |y - \beta x| \to \min_{\beta}.$$

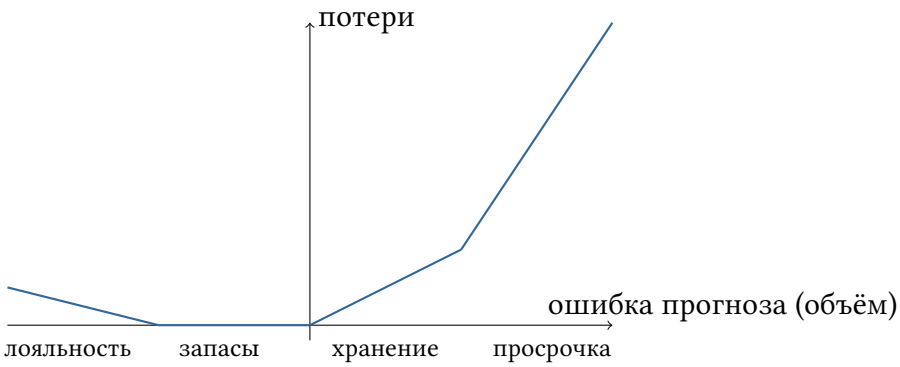
Конечно же мы потеряем няшные статистические свойства. Но тем не менее никто не мешает нам обратиться к великому и могучему бутстрапу и сбутстрапировать все доверительные интервалы и все критические значения для статистик, если это нам неожиданно потребуется.

Чаще всего условия выбора нам диктует задача, вставшая перед нами. Функцию потерь иногда приходится как следует выстрадать. Например, если речь идёт о числе товаров, которые мы должны хранить на складе, возникает необходимость использовать несимметричную кусочную функцию потерь.

Если мы завезли на склад слишком мало товара, потребителям не хватит его. Из-за того, что на товар будет наценка, а также из-за его нехватки, мы потеряем лояльность клиентов. Кривая потерь пойдёт под одним углом. Если нехватка будет небольшой, мы покроем её из запасов, потерь не будет. Если на складе будет избыток товара, мы потратим деньги на его хранение, кривая пойдёт под другим углом. Если избыток будет очень сильным, то возникнет просрочка.

[2]Если конечно вы верите в свободу воли.

Кривая пойдёт под третьим углом.



Функции потерь бывают разные, а свобода выбора — это очень страшно[3]. С выбором хочется не ошибиться и осознавать, какая функция потерь к каким последствиям (в плане прогнозов) может привести. Давайте попробуем это понять, а после будем реагировать на стимулы.

2. Про квадратичные потери потерь

Мы выяснили, что квадратичные потери соответствуют нормально распределённым ошибкам. Давайте теперь взглянем на них немного под другим углом. Пусть $L(y, \hat{y}) = (y - \hat{y})^2$ — наша функция потерь, наказание, которое мы несём за ошибку.

Ошибка на обучающей выборке $\frac{1}{n} \cdot \sum_{i=1}^{n} L(y_i, \hat{y}_i)$ — это просто эмпирическая оценка ожидаемых потерь $\mathbb{E}(L(y, \hat{y} \mid x)$. Этот факт позволяет по-новому взглянуть на старые функции потерь. Минимизируя $\mathbb{E}(L(y, \hat{y} \mid x)$, можно понять что именно мы получаем на выходе в качестве оценки:

$$\mathbb{E}((y - t)^2 \mid x) = \int (y - t)^2 \cdot f(y \mid x)\, dy \to \min_t .$$

Будем перебирать все возможные оценки t так, чтобы минимизировать математическое ожидание функции потерь. Найдём производную по t

$$\frac{\partial}{\partial t}\left(\int (y - t)^2 \cdot f(y \mid x)\, dy\right) = -2 \cdot \int (y - t) \cdot f(y \mid x)\, dy = 0$$
$$\int y \cdot f(y \mid x)\, dy - \int t \cdot f(y \mid x)\, dy = \mathbb{E}(y \mid x) - t \cdot 1 = 0 \Rightarrow t = \mathbb{E}(y \mid x)$$

Выходит, что в такой ситуации наилучшим прогнозом будет условное математическое ожидание. Именно из-за этого квадратичные потери оказываются чувствительны к выбросам. Подобный анализ позволяет иногда обнаружить, что функция потерь для решения задачи была выбрана не очень удачно.

[3] Вот так свобода и умирает под гром аплодисментов.

3. Про абсолютные потери потерь

Проделаем два точно таких же упражнения для MAE.

Упражнение 1

Пусть ошибки ε_i в задаче регрессии имеют распределение Лапласа с плотностью распределения

$$f_\varepsilon(t) = \frac{1}{2\sigma} e^{-\frac{|t|}{\sigma}}$$

Минимизации какой функции потерь в таком случае эквивалентен метод максимального правдоподобия?

Решение:

Выписываем правдоподобие

$$L = \frac{1}{(2\sigma)^n} \cdot e^{-\sum_{i=1}^{n} \frac{|y_i - \beta x_i|}{\sigma}} \to \max_{\beta,\sigma}$$

Прологарифмируем, получим

$$\ln L = -n \ln \sigma - \frac{1}{\sigma} \sum_{i=1}^{n} |y_i - \beta x_i| \to \max_{\beta,\sigma}.$$

Получается, что для максимизации правдоподобия и поиска β, нам нужно минимизировать абсолютные потери.

Упражнение 2

Пусть $L(y, \hat{y}) = |\hat{y} - y|$. Давайте выясним, что нам будет предсказывать такая модель в качестве оптимального прогноза.

Решение:

Выписываем задачу оптимизации через математическое ожидание функции потерь. Выборку y считаем случайной, x считаем фиксированным

$$\int |y - t| \cdot f(y \mid x)\, dy \to \min_t.$$

Найдём производную по t, при этом не будем забывать, что в нуле модуль не дифференцируется. К счастью, так как $\mathbb{P}(y = t \mid x) = 0$, одна точка никак не повлияет на наш интеграл.

$$\frac{\partial}{\partial t}\left(\int |y-t|\cdot f(y\mid x)\,dy\right)=\frac{\partial}{\partial t}\left(\int_{y\neq t} |y-t|\cdot f(y\mid x)\,dy\right)=$$
$$=\frac{\partial}{\partial t}\left(\int_{y>t}(y-t)\cdot f(y\mid x)\,dy-\int_{y<t}(y-t)\cdot f(y\mid x)\,dy\right)=$$
$$=\int_{y<t} f(y\mid x)\,dy-\int_{y>t} f(y\mid x)\,dy=0$$

Получается, что в данном случае для минимизации ожидаемых потерь, нужно, чтобы $\mathbb{P}(y < t \mid x) = \mathbb{P}(y > t \mid x)$. Ни для кого не секрет, что точка, в которой выполнится такое равенство, называется медианой. Получаем, что оптимальный прогноз $t = Med(y \mid x)$.

Именно из-за этого абсолютная ошибка нечувствительна к выбросам. На медиане они практически никак не сказываются, и прогноз не портится. Квадратичная ошибка к выбросам очень чувствительна. Одно большое значение довольно сильно искажает среднее.

4. Про логистические потери потерь

Пусть целевая переменная y принимает значения 0 и 1. Нам хочется по переменной x научиться прогнозировать y. Такая задача называется **классификацией**.

Задачу классификации можно попробовать решить методом максимального правдоподобия. Целевая переменная y принимает значение 1 с вероятностью p и значение 0 с вероятностью $1-p$. Если у нас завалялась выборка $y_1, \dots, y_n$, то по всем законам жанра можно выписать функцию правдоподобия:

$$L=\prod_{i=1}^{n} p^{y_i}\cdot(1-p)^{1-y_i}=p^{\sum y_i}\cdot(1-p)^{\sum(1-y_i)}$$

Прологарифмируем функцию правдоподобия и получим

$$\ln L=\sum y_i\cdot\ln p+\sum(1-y_i)\cdot\ln(1-p)=\sum_{i=1}^{n} y_i\ln p+(1-y_i)\ln(1-p).$$

Чтобы максимизировать функцию правдоподобия, нам нужно минимизировать следущую функцию потерь

$$L(y,\hat{y})=-y\cdot\ln(\hat{y})-(1-y)\cdot\ln(1-\hat{y}).$$

Эта функция потерь называет **логистической (logloss)**. Она часто используется в машинном обучении и эконометрике при решении задачи классификации. Вероятность p в данной

модели как-то должна зависеть от регрессора x. Функция, описывающая эту зависимость должна принимать значения на отрезке $[0; 1]$. В качестве такой функции берут какую-нибудь функцию распределения. Обычно это логистическое распределение (иногда функцию распределения логистической случайной величины называют **сигмоидой)**

$$P(y = 1 \mid x) = p = \frac{1}{1 + e^{-\beta x}}.$$

При использовании сигмоиды, оценки коэффициентов имеют интересную интерпретацию. Если мы найдём логарифм отношения шансов, то мы получим, что

$$\ln \frac{p}{1-p} = \beta x.$$

Выходит, что при изменении x на единицу, логарифм отношения шансов изменяется на β, то есть шансы на то, что $y = 1$ изменяются на $100 \cdot \beta\%$. При других функциях потерь хорошую интерпретацию для коэффициентов получить довольно сложно.

Когда мы конструировали логистические потери потерь, исходя из принципа правдоподобия, мы сразу же заложили в их природу то, что на выход в качестве прогноза будет идти вероятность $\mathbb{P}(y = 1)$. Тем не менее, давайте сделаем вид, что мы забыли это и проанализируем функцию потерь также, как мы делали это выше.

Мы хотели бы минимизировать условное математическое ожидание $\mathbb{E}(L(y, \hat{y}) \mid x)$. Случайной величиной в данном случае является переменная y, которая принимает два значения. Выписываем математическое ожидание

$$\sum_{k \in Y} (-y \ln t - (1 - y) \ln(1 - t)) \mathbb{P}(y = k \mid x) \to \min_t.$$

Обозначим для удобства $\mathbb{P}(y = 1 \mid x)$ как p. Тогда, учитывая что $Y = \{0, 1\}$, наша задача примет вид

$$-(1 - p) \cdot \ln(1 - t) - p \cdot \ln(t) \to \min_t.$$

Дело осталось за малым. Берём производную и находим экстремум.

$$\frac{\partial}{\partial t}(-(1 - p) \cdot \ln(1 - t) - p \cdot \ln(t)) = \frac{1 - p}{1 - t} - \frac{p}{t} = 0 \Rightarrow t = p = \mathbb{P}(y = 1 \mid x).$$

Выходит, что в логистической регрессии, минимизируя рассмотренную выше функцию потерь, мы получаем именно оценку вероятности.

5. И ещё одна функция потерь

Упражнение 1

Пусть переменная y_i — это лайки на странице Маши. Она получает их с какой-то интенсивностью λ, зависящей от числа постов за день x_i. То есть, $\lambda = \beta \cdot x_i$. Какую функцию потерь нужно минимизировать, чтобы получить оценку β, исходя из принципа максимизации правдоподобия?

Решение:

Это одна из версий Пуассоновской регрессии. Вероятность выпадения конкретного наблюдения составит

$$\mathbb{P}(y = y_i) = \frac{e^{-\beta x_i}(\beta x_i)^{y_i}}{y_i!}.$$

Выписываем функцию правдоподобия

$$L = \frac{e^{-\beta \sum x_i} \cdot (\beta x_1)^{y_1} \cdot \ldots \cdot (\beta x_n)^{y_n}}{y_1! \cdot \ldots \cdot y_n!} \to \max_{\beta}.$$

Прологарифмируем

$$\ln L = -\beta \sum x_i + \sum y_i \ln \beta x_i - \sum y_i! \to \max_{\beta}.$$

Откидываем все константные слагаемые и получаем функцию потерь

$$\beta \sum x_i - \sum y_i \ln \beta \to \min_{\beta}.$$

Обратите внимание, что в данном случае можно решить задачу влоб, тогда получится, что $\widehat{\beta} = \frac{\bar{x}}{\bar{y}}$. Выходит, что чувствительность интенсивности лайков к числу постов на стене равна тому, сколько постов приходится на один лайк.