Глава 1

Кто-то теряет, а кто-то находит

Блин блинский! Это потеря потерь!

Джейсон Стетхем

В течение книги мы несколько раз говорили о том, что в байесовской статистике, на выходе, мы получаем гораздо больше, чем просто точечную оценку. Мы получаем целое распределение. Также мы сказали, что если от нас требуется указать точечную оценку, мы должны спросить: «А как нас накажут за ошибку?» и выбрать точечную оценку исходя из этого. Вспомните одно из чудес первой главы. Для принцессы было важно дадут ей молодильное яблоко или превратят в тыкву. В зависимости от стимула она давала разный ответ.

В этой главе мы поговорим о стимулах. Мы проанализируем классические функции потерь и увидим, что они предлагают нам в качестве прогнозов. Мы узнаем про энтропию, дивергенцию Кульбака-Лейблера, а также увидим, что правдоподобие обожает играть в шпионские игры и постоянно маскируется. Надеюсь, что нам удастся разоблачить его.

1.1 Про то 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 суммарную ошибку прогноза

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

Берём производную, решаем уравнение, получаем ответ

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

Взяв вторую производную, можно убедиться, что это действительно минимум. Сразу же после того, как была получена формула для оценивания бэтки[1]

[1]Обычно в англоязычной литературе такая формула называется estimator, то есть оцениватель. Конкретная оценка называется estimate. Почему-то богатый русский язык не впитал

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 функцию максимального правдоподобия

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

Прологорифмируем полученное добро

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

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 довольно часто функции потерь вытекают из каких-то функций правдоподобия. Убедиться в этом вы можете, попытавшись решить упражнение 1.

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.

это различие и стал называть оценкой и формулу и конкретное численной значение. Давайте исправлять это недоразумение и называть формулы оценивателями.

При этом главным профитом статитических процедур, проведуемых в эконометрике, является величина эффекта. На выходе мы получаем величину $\widehat{\beta}$, 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 $\widehat{\beta}$.

Как только мы немного видоизменим функуцию 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. Не будем забегать вперёд. Разговор о регуляризации и о двух великих вопросах анализа данных ожидает нас в следущей главе. Постарайтесь удержать мысль о взаимосвязи между вероятностными распределениями и функциями потерь у себя в голове до следущей главы. Там она получит должное развитие. Здесь же мы собрались немного по иному поводу.

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

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

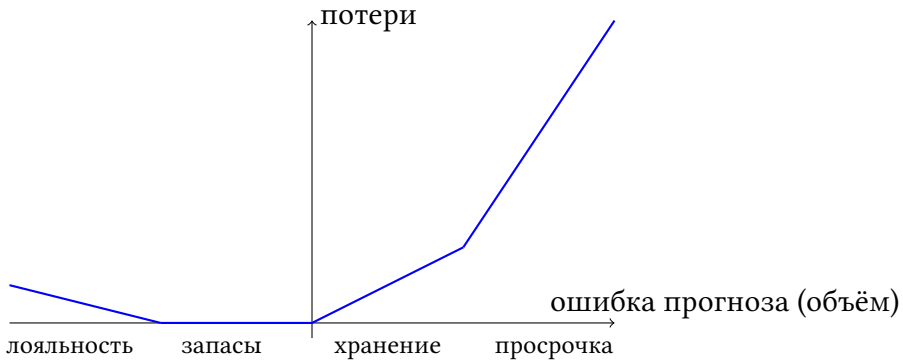
Конечно же мы потеряем няшные статистические свойства. Но тем не менее никто не мешает нам обратиться к великому и могучему бустрапу и сбутсрапировать все доверительные интервалы и все критические значения для статистик, если это нам неожиданно потребуется.

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

[2]Если читатель внимательно читал вторую главу, он помнит, что это неточно.

пойдёт под другим углом. Если избыток будет очень сильным, то возникнет просрочка. Кривая пойдёт под третьим углом.



Одним словом говоря, функции потерь бывают разные. А свобода выбора — это очень страшно[3]. Хочется с ним 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 как Спящая Красавица.

Для разнообразия пока что не будем говорить о y и $\hat{y}$ (вернёмся к ним через страницу). В байесовском мире у нас есть β с каким-то апостериорным распределением. И мы хотели бы выбрать из него точечную оценку $\hat{\beta}$.

Пусть $L(\beta, \hat{\beta})$ — наша функция потерь, наказание, которое мы несём за ошибку. Тогда наш выбор оценки заключается в минимизации апостериорных ожидаемых потерь, то есть

$$E(L(\beta, t) \mid y) = \int L(\beta, t) \cdot f(\beta \mid y)\, d\beta \to \min_t .$$

Будем перебирать все возможные оценки t так, чтобы минимизировать математическое ожидание функции потерь. Обратим внимание на то, что случайной величиной в данном случае является β. Значение t мы пытаемся выбрать самостоятельно. Давайте попробуем подставить в этот интеграл конкретные функции потерь и посмотрим что у нас из этого получится.

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

1.2 Про квадратичные потери потерь

Пусть $L(\beta, \hat{\beta}) = (\hat{\beta} - \beta)^2$, тогда

$$\int (\beta - t)^2 \cdot f(\beta \mid y)\, d\beta \to \min_t$$

Найдём производную по t

$$\frac{\partial}{\partial t}\left(\int (\beta - t)^2 \cdot f(\beta \mid y)\, d\beta\right) = -2 \cdot \int (\beta - t) \cdot f(\beta \mid y)\, d\beta = 0$$

$$\int \beta \cdot f(\beta \mid y)\, d\beta - \int t \cdot f(\beta \mid y)\, d\beta = E(\beta \mid y) - t \cdot 1 = 0 \Rightarrow t = E(\beta \mid y)$$

Из этих довольно простых рассуждений мы получаем несколько интересных выводов. Во-первых, при квадратичных потерях, оптимальной байесовской точечной оценкой будет апостериорное среднее.

«Во-вторых» связано с машинным обучением и эконометрикой. Предположим, нам захотелось методом наименьших квадратов обучить модель и получить исходя из неё прогноз $\hat{y}$. За каждый плохой прогноз мы понесём наказание $(y_i - \hat{y}_i)^2$. Выходит, что в такой ситуации наилучшим прогнозом будет условное математическое ожидание.

Ошибка на обучающей выборке $\frac{1}{n} \cdot \sum_{i=1}^{n} L(y_i, \hat{y}_i)$ — это просто эмпирическая оценка ожидаемых потерь $E(L(y, \hat{y} \mid x)$. Этот факт позволяет по-новому взглянуть на старые функции потерь. Минимизируя $E(L(y, \hat{y} \mid x)$, можно понять что именно мы получаем на выходе в качестве оценки. В случае квадратичной функции потерь, это ни что иное как $E(y \mid x)$. Убедиться в этом можно, проделав ровно то же самое, что мы проделали выше, но заменить β на y, а $\hat{\beta}$ на $\hat{y}$

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

Подобный анализ позволяет иногда обнаружить, что функция потерь для решения задачи была выбрана не очень удачно. Например, в упражнении 3 целевая переменная y принимает значения 0 и 1. Нам хочется оценить вероятность того, что y примет значение 1. Хотелось бы, чтобы модель выдавала нам

число, лежащее на отрезке от нуля до единицы. Решив упражнение, можно убедиться, что использовать для этих целей абсолютное отклонение, не очень удачная идея, так как оно будет выдавать в качестве ответа либо 0 либо 1, но никак не вероятность.

1.3 Про абсолютные потери потерь

Пусть $L(\beta, \hat{\beta}) = |\hat{\beta} - \beta|$, тогда

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

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

$$\begin{aligned}
\frac{\partial}{\partial t}\left(\int |\beta - t| \cdot f(\beta \mid y)\, d\beta\right) &= \frac{\partial}{\partial t}\left(\int_{\beta \neq t} |\beta - t| \cdot f(\beta \mid y)\, d\beta\right) = \\
&= \frac{\partial}{\partial t}\left(\int_{\beta > t} (\beta - t) \cdot f(\beta \mid y)\, d\beta - \int_{\beta < t} (\beta - t) \cdot f(\beta \mid y)\, d\beta\right) = \\
&= \int_{\beta < t} f(\beta \mid y)\, d\beta - \int_{\beta > t} f(\beta \mid y)\, d\beta = 0
\end{aligned}$$

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

Снова делаем два вывода. Во-первых, в случае такой функции потерь апостериорная медиана является оптимальной байесовской точечной оценкой. Во-вторых, в случае, если мы используем такую функцию потерь для обучения модели по признаку x и ответу y, то алгоритм в качестве оценки будет выдавать нам $Med(y \mid x)$.

Надеемся, что для читателя теперь стало понятно, почему 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.

1.4 Про жадные потери потерь

Я запутался

1.5 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**. На самом деле мы уже решали такую задачу в упражнениях для прошлой главы, когда говорили о поломке шатла.

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

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

Выходит, что 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.

Мы немного отвлеклись от генеральной линии повествования. Давайте вернёмся к ней. 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 $P(y = 1)$. 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 $E(L(y, \hat{y}) \mid x)$. Случайной величиной в данном случае является переменная y, которая принимает два значения. Выписываем математическое ожидание

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

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

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

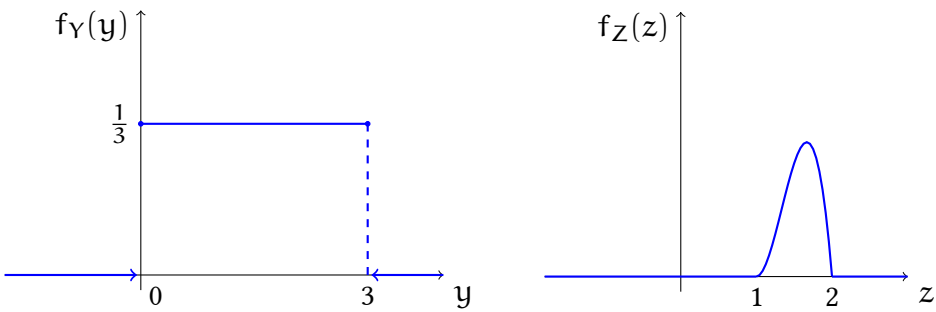
Дело осталось за малым. Берём производную и находим экстремум.

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

Выходит, что в логистической регрессии, минимизируя рассмотренную выше функцию потерь, мы получаем именно оценку вероятности. Думаю, что концепция, в принципе, ясна. Не будем топтаться на месте и перейдём к следующему сюжету.

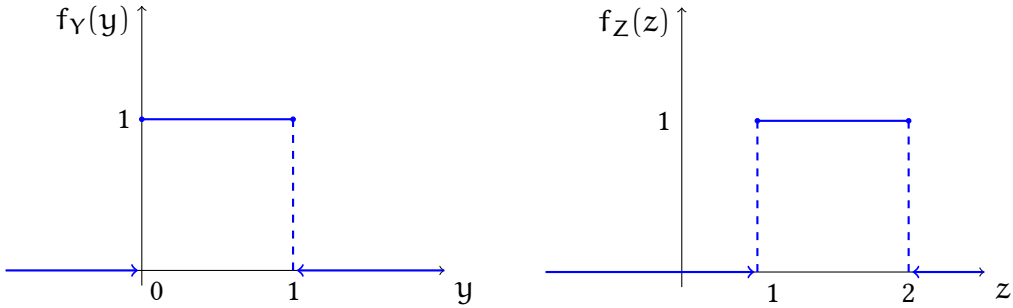
1.6 Про энтропию

Приятно было бы начать этот раздел с картинки. Давайте посмотрим на распределение двух случайных величин. Какая из них предсказуемее: левая или правая?



Случайная величина Z (правая) сконцентрирована на довольно узком отрезке. Вероятность того, что она выпадет за его пределами крайне мала. Случайная величина Y (левая) сконцентрирована на широком отрезке. Она равновероятно может выскочить из любой его части. Логика подсказывает, что она непредсказуемее. Если мы решим спрогнозировать эти две случайные величины, правый случай нам будет обуздать легче. Ошибка, которую мы будем допускать, окажется меньше чисто из-за пикообразной природы этой случайной величины.

Давайте теперь посмотрим на ещё две картинки. Какая непредсказуемее: левая или правая?



Случайные величины Y и Z отличаются друг от друга только отрезком. Одна распределена от 0 до 1, вторая от 1 до 2. Их форма одинакова. Они принимают разные значения, но одинаково непредсказуемо. Если мы попробуем спрогнозировать их, сложность этой затеи будет одинакова.

Для того, чтобы улавливать такие вещи придумали специальную метрику. Она называется энтропией. **Энтропия** — это мера непредсказуемости случайной величины Y, это то количество информации, которое я получаю, наблюдая случайную величину Y. Она никак не опирается на те значения, которые принимает случайная величина и для дискретного случая определяется как

$$H(Y) = E(-\ln P(Y)).$$

Для непрерывного случая энтропия определяется как

$$H(Y) = E(-\ln f_Y(y)).$$

Попробуем немного пожить с энтропией. После того, как мы освоимся, можно будет поглубже обсудить её смысловую составляющую. Посчитаем энтропию для случайной величины:

Y	1	17	26
$P(Y = k)$	$\frac{1}{3}$	$\frac{1}{3}$	$\frac{1}{3}$

Энтропия никак не смотрит на то, какие именно значения принимает случайная величина. Её интересует только то, как вероятность размазана по этим значениям:

$$H(Y) = -\frac{1}{3} \cdot \ln\left(\frac{1}{3}\right) - \frac{1}{3} \cdot \ln\left(\frac{1}{3}\right) - \frac{1}{3} \cdot \ln\left(\frac{1}{3}\right) = \ln 3.$$

Для случайной величины, принимающей 4 значения с вероятностями $\frac{1}{4}$ энтропия будет равна $\ln 4$, а в общем случае для $Y \sim U[0; a]$ энтропия составит

$$H(Y) = \int_0^a \frac{1}{a} \cdot \left(-\ln\left(\frac{1}{a}\right)\right) dt = \ln a.$$

Чем больше значений принимает равномерная случайная величина, тем она непредсказуемее. Кстати говоря, для вырожденного распределения

Y	42
$P(Y = k)$	1

энтропия окажется нулевой. Вырожденная случайная величина очень даже определена. Попробуем более сложную ситуацию, найдём энтропию для нормально распределённой случайной величины $Y \sim N(0, \sigma^2)$:

$$H(Y) = E(-\ln(f_Y(y))) = \int_{-\infty}^{+\infty} f_Y(y) \cdot \ln f_Y(y)\, dy = \\ = E\left(\frac{1}{2}\ln(2\pi\sigma^2) + \frac{Y^2}{2 \cdot \sigma^2}\right) = \frac{1}{2}\ln(2\pi\sigma^2) + \frac{1}{2}.$$

Если попробовать подставить в формулу разные значения σ, то можно получить следущую примерную табличку:

σ	1	10	100
$H(Y)$	$\ln 4.13$	$\ln 41.3$	$\ln 413$

Выходит, что случайные величины $X \sim U[0; 4]$ и $Y \sim N(0, 1)$ в плане непредсказуемости очень похожи. Небольшой флэшбэк. Когда мы впервые

говорили о априорных распределениях, мы сказали, что часто статистики говорят, что вообще ничего не знают о параметрах и выражают своё незнание либо нормальным распределением с большой дисперсией либо равномерным на очень большом отрезке. Энтропия в какой-то степени является обоснованием того, почему эти два подхода замоделировать своё априорное незнание эквивалентны.

Энтропия обладает несколькими интересными математическими свойствами:

1. Она всегда неотрицательна $H(Y) \geqslant 0$.
2. Для конечного числа исходов m, равномерное распределение над этими исходами будет давать максимальную энтропию.
3. Если X и Y независимы, то $H(X \cdot Y) = H(X) + H(Y)$.
4. Если X и Y имеют одинаковые распределения, но принимают разные значения, то $H(X) = H(Y)$. Энтропия это функция над распределениями, а не значениями.

На самом деле формула для энтропии была получена как раз исходя из этих свойств[4]. Освоились? Давайте теперь поговорим про смысл.

Итак, энтропия отражает то, насколько случайная величина непредсказуема. Другой характеристикой, описывающей вариабильность случайной величины является дисперсия. Отличие энтропии от дисперсии в том, что ей плевать на значения, которые принимает случайная величина. Если найти энтропию, для равномерного, нормального и экспоненциального распределений, которые мы использовали для моделирования наших априорных ожиданий чаще всего, то мы получим, что:

$$
\begin{aligned}
N(0, \sigma^2): &\qquad H(Y) = \frac{1}{2} \cdot \ln(2\pi e) + \ln \sigma \\
U[a; b]: &\qquad H(Y) = \frac{1}{2} \cdot \ln(12) + \ln \sigma, \quad \sigma^2 = \frac{1}{12}(b-a)^2 \\
Exp(\alpha): &\qquad H(Y) = 1 + \ln \sigma, \quad \sigma^2 = \alpha^{-2}
\end{aligned}
$$

[4]Помните мы недавно обсуждали как формула для простейшего потока события тоже была получена из свойств этого потока? Можно показать, что эта формула единственна. Тут такая же ситуация. Честно говоря, это восхищает.

Энтропия для всех трёх распределений выражается через дисперсию и может быть записана, как логарифм стандартного отклонения плюс некоторая константа. Этот факт заставляет начать сомневаться в целесообразности введения нового понятия.

На самом деле с целесообразностью всё в порядке. Для мультимодальных распределений обнаруженная нами закономерность нарушается. Можно довольно легко подобрать случайные величины X и Y, для которых $\mathrm{Var}(X) > \mathrm{Var}(Y)$, но при этом $H(X) < H(Y)$. Всё бы хорошо, но это порождает новый вопрос. Какая из случайных величин, X или Y непредсказуемее? Какую метрику для этого использовать? В прочем, ответ вас не удивит. Выбор метрики зависит от задачи.

Ещё раз, ещё раз. Вся мощь энтропии заключена в том, что она описывает неопределённость, заложенную в случайную величину абстрагируясь от её значений и порядка этих значений. Дисперсия, в свою очередь, очень пристально смотрит на значения случайной величины и их порядок.

Давайте представим себе парочку ситуаций.

Пример когда плевать на значения

Пусть теперь на наш остров надвигается шторм. Хочется понять будет ли шторм разрушительным. Пусть X это скорость ветра в километрах в час. Давайте сравним между собой два распределения:

1. $X_1 \sim N(100, 10^2)$.
2. $X_2 \sim$ смесь $N(50, 3^2)$ и $N(200, 3^2)$ с весами 0.5.

Как раз в данном случае одно из распределение бимодальное. Мы можем увидеть, что $\mathrm{Var}(X_1) < \mathrm{Var}(X_2)$, но при этом $H(X_1) > H(X_2)$, так как случайная величина X_2 более плотно сосредоточена на двух конкретных значениях. С практической точки зрения было бы правильным говорить, что случайная величина X_2 несёт в себе большую неопределённость, так как мы не знаем будет ли обычный ветер или разрушительный шторм. В случае X_1 мы знаем, что ветер будет сильным, но неразрушительным. В данном случае имеет смысл руководствоваться дисперсией, так как для нас важны значения, которые принимает случайная величина.

Задачка про робота

Ещё раз, ещё раз. При использовании в формуле энтропии двоичного логарифма, мы можем интерпретировать её как среднее количество бит информации, которое мы тратим на кодирование. На практике нам обычно хочется полегче считать энтропию. Поэтому мы отойдём от двоичного основания назад к натуральному, так как с натуральными логарифмами работать намного приятнее.

Энтропия довольно часто используется в машинном обучении. Например, с помощью неё обучают деревья. Кроме того, на ней базируется понятие спутанности (perplexity), которое определяется как

$$\mathrm{Perplexity}(Y) = e^{H(X)}$$

Перплексия (спутанность) довольно часто используется в моделях, связанных с обучением без учителя, но с этим мы встретимся позже. Сейчас наша дорога ведёт нас к очередному новому понятию, **дивергенции Кульбака-Лейблера**.

1.7 Про дивергенцию

Дивергенция Кульбака-Лейблера пришла в теорию вероятностей из теории информации. По сути KL-дивергенция — это мера разницы между двумя вероятностными распределениями P и Q. Обычно считают, что P — это истиное распределение, а Q — его приближение. Пусть распределение P сложное и страшное и мы хотим заменить его на простое и хорошее.

картинка с простым распределением и со страшным

В таком случае дивергенция служит оценкой качества приближения и отражает то, какое количество информации мы потеряли, заменив распределение P на распределение Q. Обычно KL-дивергенцию обозначают как $KL(P||Q)$.

Для дискретного случая дивергенцию можно найти как

$$KL(P||Q) = \sum P(x) \cdot \ln \frac{P(x)}{Q(x)}.$$

Для непрерывного случая формула аналогична, но сумма заменяется на интеграл, а вероятности на плотности

$$KL(P||Q) = \int f_P(x) \cdot \ln \frac{f_P(x)}{f_Q(x)}.$$

Выше мы сказали, что KL-дивергенция позволяет измерять расстояния между распределениями. Однако уже по формуле видно, что дивергенция Кульбака-Лейблера расстоянием не является. Из курса матана долгопомнящий читатель может вспомнить, что для того, чтобы функция $\rho(x, y)$ была расстоянием, необходимо выполнение трёх свойств:

1. Неотрицательность: $\rho(x, y) \geqslant 0$. Расстояние от одного объекта до другого всегда положительно. Если расстояние равно нулю, то объекты находятся в одном месте.

2. Симметричность: $\rho(x, y) = \rho(y, x)$. От первого объекта до второго ехать столько же сколько обратно.

3. Неравенство треугольника: $\rho(x, y) + \rho(y, z) \geqslant \rho(x, z)$. Если мы ехали из одной точки в другую и по пути заехали куда-то ещё, то ехать придётся подольше, если наша остановка была нам не по пути.

С первым свойством всё хорошо. Со вторым и третьим начинаются проблемы. Дивергенция несимметрична, $KL(P||Q) \neq KL(Q||P)$. Зачем так сложно? Почему бы просто не взять и не использовать обычные давно известные метрики. Мало того, что они нам знакомы, так ещё и симметричны. Например, почему бы не взять $\rho(P, Q) = \max_t |f_P(x) - f_Q(x)|$?

Для ответа на этот вопрос снова посмотрим на парочку картинок.

картинки

На левой картинке расположены разные распределения. Однако в терминах $\rho(P, Q)$ эти две плотности будут похожи. Метрика забьёт на вероятностные различия и сконцентрируется на функциональных особенностях плотностей. На второй картинке наоборот для двух похожих распределений $\rho(P, Q)$ выдаст сильную разницу. Одним словом говоря, дивергенция хорошо зарекомендовала себя на практике даже несмотря на то, что она не является расстоянием в привычном для на смысле. Скорее даже наоборот, ассиметрия помогает нам, так как обычно мы всегда хотим заменить забубенистое распределение, которое мы видим в данных чем-то более простым, а это односторонняя ситуация.

Поглядим на формулу для поиска энтропии чуть пристальнее и попробуем над ней немного поколдовать

$$KL(P||Q) = \sum P(x) \cdot \ln \frac{P(x)}{Q(x)} = \\ = \sum P(x) \ln P(x) - \sum P(x) \ln Q(x) = H(P) + H(P, Q).$$

Получается, что KL-дивергенция в явном виде выражается через энтропию распределения P. Второе слагаемое называется **перекрёстной или кросс энтропией**. Перекрёстная энтропия интерпретируется как риск использования распределения Q, если данные пришли из P.

Выходит, что оценка качества приближения распределения P распределением Q складывается из двух составляющих: энтропии P и риска использования Q вместо P. Смысл мы обсудили, теперь давайте попробуем пощупать KL-дивиргенцию ручками.

Пусть у нас есть случайная величина X, имеющая распределение P. Это распределение кажется нам слишком сложным и мы хотим заменить его на Q.

X	1	10	100
P	$\frac{1}{2}$	$\frac{1}{4}$	$\frac{1}{4}$
Q	$\frac{1}{3}$	$\frac{1}{3}$	$\frac{1}{3}$

Давайте посчитаем дивергенцию между этими двумя распределениями.

$$KL(P||Q) = \left[\frac{1}{2}\ln(3) + \frac{1}{4}\ln(3) + \frac{1}{4}\ln(3)\right] - \\ - \left[\frac{1}{2}\ln(2) + \frac{1}{4}\ln(4) + \frac{1}{4}\ln(4)\right] = 0.059.$$

Теперь сделаем то же самое для непрерывной случайной величины. Найдём дивергенцию между $N(0, 1)$ и $N(0, 4)$.

$$KL(N(0,1)||N(0,4)) = \int_{-\infty}^{+\infty} f_P(t) \cdot (-\ln(f_Q(t))\,dt - \int_{-\infty}^{+\infty} f_P(t) \cdot (-\ln(f_P(t))\,dt = \int_{-\infty}^{+\infty} f_P(t) \cdot \left(-\ln \frac{f_Q(t)}{f_P(t)}\right)\,dt.$$

Найдём второй множитель

$$-\ln \frac{f_Q(t)}{f_P(t)} = -\ln\left(\frac{1}{2} \cdot e^{\frac{3}{8}t^2}\right) = \ln 2 - \frac{3t^2}{8}.$$

Теперь добьём дивергенцию

$$\int_{-\infty}^{+\infty} f_P(t)(\ln 2 - \frac{3t^2}{8})\,dt = \ln 2 \cdot 1 - \frac{3}{8}E(X^2) = \ln 2 - \frac{3}{8}.$$

Если заменить натуральный логарифм на двоичный, можно снова уйти в биты. Но обычно для удобства используется именно натуральный логарифм.

Пример про классификацию

В байесовских методах с помощью KL-дивергенции можно оценивать информационный выигрыш при переходе от априорного распределения к апостериорному. В таком контексте формула примет вид

$$KL(\beta \mid y_{n+1}, y||\beta \mid y) = \int f(\beta \mid y_{n+1}, y) \cdot \ln \frac{f(\beta \mid y_{n+1}, y)}{f(\beta \mid y)}.$$

Полученная в ходе вычислений цифра будет отражать то, какое количество дополнительной информации касательно параметра β мы получили, пронаблюдав дополнительный y_{n+1}. Другой путь осознать насколько полезным оказалось новое наблюдение — посмотреть на разность апостериорной и априорной энтропий

$$\Delta H = H(\beta \mid y_{n+1}, y) - H(\beta \mid y).$$

Никто не гарантирует, что эта разность получится положительной. В случае дивергенции Кульбака-Лейблера мы можем быть уверены, что прирост информации окажется либо нулевым либо положительным.

Здесь мы можем увидеть разницу между информацией и неопределённостью во всей красе. Энтропия измеряет неопределённость. При поступлении нового наблюдения она может как возрасти так и уменьшиться. KL-дивергенция отражает то, сколько дополнительной информации мы извлекли из нового наблюдения. Прирост информации есть всегда. Даже когда она увеличивает неопределённость.

Конечно же, если говорить в терминах математических ожиданий, то ожидаемый прирост информации всегда равен ожидаемому уменьшению неопределённости. Конечно же, при условии, что модель была верно специфицирована.

По аналогии мы можем применять метрики из теории информации для распределения y_{new}, которое мы используем при построении прогнозов.

снова монетка и подбрасывания, наблюдение за информацией

1.8 Про поимку шпиона

Пришло время срывать маски. Если на Земле действуют прогрессоры значительно более развитых рас, то в чём будут, скорее всего, состоять их действия? Если правдоподобие хочет, чтобы мы его не заметили, то что оно сделает?

Вариант первый: скажет, что оно это функция потерь. Случаи, когда происходит подобная метаморфоза, мы рассмотрели выше. Сейчас мы попытаемся словить пару более тонких ситуаций. Чтобы сделать это, немного поколдуем с оценками максимального правдоподобия, которые мы получали максимизируя функцию правдоподобия:

$$\hat{\beta} = \arg\max_{\beta} L(y \mid \beta) = \arg\max_{\beta} \prod f(y \mid \beta).$$

Обычно мы пользовались логарифмическим правдоподобием. Давайте домножим его на -1, и будем искать вместо максимума минимум. Также давайте домножим его на $\frac{1}{n}$. Эта константа ни на что не повлияет, так как мы сможем избавиться от неё после взятия производной, но интерпретации даст много

$$\hat{\beta} = \arg\max_{\beta} \sum \ln f(y_i \mid \beta) = \arg\min_{\beta} \left(-\frac{1}{n} \sum \ln f(y_i \mid \beta) \right).$$

Колдуем дальше. Если бы мы могли заглянуть в хрустальный шар и увидеть там истиное значение параметра β_0, тогда бы по закону больших чисел среднее сходилось бы к математическому ожиданию

$$-\frac{1}{n} \sum \ln f(y_i \mid \beta_0) \to E(-\ln f(y \mid \beta_0)) = H(Y)$$

и задача максимизации правдоподобия была бы эквивалентна задаче минимизации энтропии. К сожалению, у нас нет хрустального шара. Тем не менее, мы обладаем математикой, в которой издревле завалялся классический трюк. Добавим нужное нам слагаемое. Не будем забывать, что чтобы купить что-нибудь нужное, надо продать что-нибудь ненужное. Вычтем из функции то же самое слагаемое.

$$\begin{aligned} \frac{1}{n} \sum -\ln f(y_i \mid \beta) = \\ = \frac{1}{n} \sum (-\ln f(y_i \mid \beta) + \ln f(y_i \mid \beta_0) - \ln f(y_i \mid \beta_0)) = \\ = \frac{1}{n} \sum \left[\log \frac{f(t \mid \beta_0)}{f(y \mid \beta)} - \ln f(y \mid \beta_0) \right]. \end{aligned}$$

Снова используем закон больших чисел и получим, что

$$\frac{1}{n} \sum \left[\log \frac{f(t \mid \beta_0)}{f(y \mid \beta)} - \ln f(y \mid \beta_0) \right] \to KL(f(y \mid \beta_0 \| y \mid \beta)) + H(Y).$$

На второе слагаемое мы не можем оказывать никакого влияния своими манипуляциями. Вся наша статистическая работа идёт с β, которое находится в первом слагаемом. Выходит, что когда мы хотим получить β максимально близкое к β_0, мы на самом деле минимизируем дивергенцию.

Итак, в текущей главе мы посмотрели на несколько различных понятий и методов, которые оказались довольно сильно переплетены между собой[5]. Мы поговорили о том, к чему приводит использование различных функций потерь в контексте прогнозов, а также ввели несколько новых понятий, перекочевавших в теорию вероятностей из теории информации. Оказалось, что эти понятия тесно связаны с такой классической штукой, как правдоподобие.

В следующей главе мы разовьём идеи, связанные с потерями и поговорим про регуляризацию. Немного позже мы вернёмся к дивергенции и обсудим такую важную штуку, как вариационные приближения. Будет интересно. Но давайте не будем забегать вперёд и для начала прорешаем упражнения к этой главе. В них надо будет распутать пару клубков.

мем со скуби-ду и типо вот кто ты на самом деле

1.9 Ещё задачи

Упражнение 1. Давайте попробуем посмотреть, к каким функциям потерь могут привести нас разные распределения.

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?

2. 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?

3. Ещё пунктов!

[5] Потяни за нить, за ней потянется клубок.

1. 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.

2. Подобная модель называется пуассоновской регрессией. Вероятность выпадения конкретного наблюдения составит

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

Выписываем функцию правдоподобия

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

Прологарифмируем

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

Откидываем все константные слагаемые и получаем функцию потерь

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

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. Звучит логично.

Упражнение 2. Исследователь Дмитрий захотел выиграть грант на свои байесовские исследования. По условиям конкурса он должен предложить богатой государственной комиссии какие-нибудь интересные байесовские точечные оценки и объяснить из каких-таких предположений они появились.

Дмитрий собирается попробовать следующие функции. К чему это приведёт и получит ли бравый исследователь грант?

1. $L(\beta, \beta_F) = (\beta - \hat{\beta})^3$
2. $L(\beta, \beta_F) = (\beta - \hat{\beta})^2 + \lambda \cdot \hat{\beta}^2$
3. $L(\beta, \beta_F) = (\phi(\beta) - \phi(\hat{\beta})) - \phi'(\hat{\beta}) \cdot (\beta - \hat{\beta})$, где ϕ — любая бесконечно дифференцируемая функция.
4. $L(\beta, \hat{\beta}) = \begin{cases} \alpha \cdot (\hat{\beta} - \beta), \text{ если } \hat{\beta} > \beta \\ (1 - \alpha) \cdot (\beta - \hat{\beta}), \text{ если } \hat{\beta} \leqslant \beta. \end{cases}$
5. $L(\beta, \hat{\beta}) = \frac{|\beta - \hat{\beta}|}{\beta}$. Как можно найти оптимум для такой функции потерь?

1. Пусть $L(\beta, \hat{\beta}) = (\hat{\beta} - \beta)^3$, тогда

$$\int (\beta - t)^3 \cdot f(\beta \mid y)\, d\beta \to \min_t$$

Найдём производную по t

$$\frac{\partial}{\partial t}\left(\int (\beta - t)^3 \cdot f(\beta \mid y)\, d\beta\right) = -3 \cdot \int (\beta - t)^2 \cdot f(\beta \mid y)\, d\beta = 0$$

$$\int (\beta^2 - 2\beta t + t^2) f(\beta \mid y)\, d\beta = t^2 - 2t\, E(\beta \mid y) + E(\beta^2 \mid y) = 0$$

$$D = 4\, E^2(\beta \mid y) - 4\, E(\beta^2 \mid y) = -4 \cdot Var(\beta \mid y) < 0$$

Действительного оптимального прогноза не существует. За такие исследования Дмитрий грант явно не получит...

2. Пусть $L(\beta, \widehat{\beta}) = (\beta - \widehat{\beta})^2 + \lambda\beta_\Gamma^2$, тогда

$$\int ((\beta - t)^2 + \lambda t^2) \cdot f(\beta \mid y)\, d\beta \to \min_t$$

Найдём производную по t

$$\frac{\partial}{\partial t}\left(\int ((\beta - t)^2 + \lambda t^2) \cdot f(\beta \mid y)\, d\beta\right) =$$

$$= 2 \cdot \int (-2(\beta - t) + 2\lambda t) \cdot f(\beta \mid y)\, d\beta = 0$$

$$\int \beta \cdot f(\beta \mid y)\, d\beta - \int (\lambda + 1) \cdot t \cdot f(\beta \mid y)\, d\beta = E(\beta \mid y) - (\lambda + 1) \cdot t \cdot 1 = 0$$

$$\Rightarrow \quad t = \frac{E(\beta \mid y)}{1 + \lambda}$$

Хммм... Получилось что-то дельное. Чем большее значение принимает λ, тем сильнее наша точечная байесовская оценка стягивается к нулю. Уже лучше, но на грант всё ещё не тянет. Запомните этот пример. В следующей главе мы будем обсуждать регуляризацию, и эта формула снова всплывёт.

3. Пусть $L(\beta, \widehat{\beta}) = (\phi(\beta) - \phi(\widehat{\beta})) - \phi'(\widehat{\beta}) \cdot (\beta - \widehat{\beta})$, где ϕ — любая бесконечно дифференцируемая функция, тогда

$$\int ((\phi(\beta) - \phi(t)) - \phi'(t) \cdot (\beta - t)) \cdot f(\beta \mid y)\, d\beta \to \min_t$$

Найдём производную по t

$$\frac{\partial}{\partial t}\left(\int (\phi(\beta) - \phi(t) - \phi'(t) \cdot (\beta - t)) \cdot f(\beta \mid y)\, d\beta\right) =$$

$$= \int (-\phi'(t) - \phi''(t) \cdot (\beta - t) + \phi'(t)) \cdot f(\beta \mid y)\, d\beta = 0$$

$$\phi''(t) \cdot (E(\beta \mid t) - t) = 0 \quad \Rightarrow \quad t = E(\beta \mid y)$$

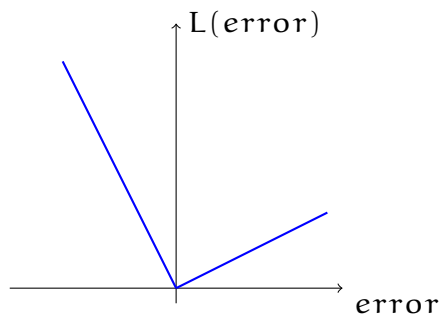
Получаем интересный вывод. Математическое ожидание доставляет минимум более широкому классу функций потерь, а не только квадратичной функции. Приведённая в этой задаче функция называется дивиргенцией (функцией потерь) Брегмана.

В случае $\phi(x) = x^2$, получаем уже знакомую нам квадратичную функцию потерь:

$$\phi(\beta) - \phi(t) - \phi'(t) \cdot (\beta - t) = \beta^2 - t^2 - 2t \cdot (\beta - t) = (\beta - t)^2$$

В случае $\phi(x) = p \cdot \ln p$ можно получить дивергенцию Кульбака-Лейбнера, а в случае $\phi(x) = -\ln(p)$ дивергенцию Итакура-Саито. Но это уже совсем другая история. Если до Дмитрия это никто не заметил, наклёвывается грант.

4. Мы имеем дело с квантильными потерями потерь. Они почти как MAE, но только положительная и отрицательная ошибки, имеют разные цены. Например, в такой ситуации недопрогноз более критичен:



Давайте выясним, что нам даёт в качестве прогноза, квантильная функция потерь.

$$\int_{\beta > t} \alpha \cdot (\beta - t) f(\beta \mid y)\, d\beta + \int_{\beta \leqslant t} (1 - \alpha) \cdot (t - \beta) f(\beta \mid y)\, d\beta \to \min_t$$

Берём производную, приравниваем к нулю

$$-\int_{\beta > t} \alpha f(\beta \mid y)\, d\beta + \int_{\beta \leqslant t} (1-\alpha) f(\beta \mid y)\, d\beta.$$

Получается, что

$$\alpha P(\beta > t \mid y) = (1-\alpha) P(\beta \leqslant y \mid y) \Rightarrow P(\beta < t \mid y) = \alpha.$$

На выходе получаем, что оптимальный прогноз находится как условный квантиль уровня α для нашего апостериорного распределения.

5. С этим пунктом есть проблемы.

Эта функция потерь называетя MAPE. Она довольно часто применяется, если нам принципиально, на сколько процентов мы ошиблись, а не абсолютное значение. Если мы предсказали 1, а в реальности получилось 10 — это не то же самое, что мы предсказали 1001, а получилось 1010. В первом случае ошибка в процентном плане катастрофичнее, и MAPE улавливает это. Давайте проанализируем эту функцию потерь.

Если по аналогии с MAE попробовать изучить эту функцию потерь, мы получим, что прогноз возникает из странного равенства

$$\int_{\beta \leqslant t} \frac{1}{\beta} f(\beta \mid y)\, d\beta = \int_{\beta > t} \frac{1}{\beta} f(\beta \mid y\, d\beta)$$

И что? Можно ил это интерпретировать как exp(Med(ln b)) и придумать грамотное процентное описание? ХЕЛП!

Возникает новый вопрос. MAPE не дифференцируемая, более того при нулевом правильном ответе мы уходим в бесконечность. Как правильно оптимизировать такую функцию?

На практике можно применить следущую хитрость. Давайте попробуем преобразовать наше β функцией $f(\beta)$ так, чтобы свести задачу к оптимизации абсолютной функции потерь.

Мы хотим, чтобы $\frac{|\beta-\hat{\beta}|}{\beta} \approx |f(\beta) - f(\hat{\beta})|$.

Если разложить f в ряд Тэйлора до первого члена, то можно получить, что

$$\frac{|\beta - \hat{\beta}|}{\beta} \approx |f(\beta) - f(\hat{\beta})| \approx |f'(\beta)||\beta - \hat{\beta}|.$$

Чтобы получить после взятия производной MAPE, нам нужно, чтобы $f'(\beta) = \frac{1}{\beta}$. То есть, если мы хотим оптимизировать MAPE, нам нужно прологарифмировать таргеты, обучиться на логарифмах, оптимизируя MAE, а после восстановить предсказания с помощью экспонент.

Упражнение 3. Перед Винни-Пухом стоит задача классифицировать пчёл на правильных и неправильных. В его распоряжении есть выборка (y_i, x_i). Переменная y_i принимает значение 1, если пчела правильная и значение 0, если пчела неправильная. Переменная x_i — это густота мёда пчелы.

Выборка собрана, исследовательский энтузиазм зашкаливает. Есть только одна беда. Непонятно какую именно функцию потерь лучше использовать. Однако есть варианты:

1. $L(y, a(x)) = (y - a(x))^2$
2. $L(y, a(x)) = |y - a(x)|$

Винни очень бы хотелось на выходе обязательно получить оценку вероятности принадлежности пчелы к определённому классу. Какую из функций лучше использовать исследователю?

1. Пусть $L(y, a(x)) = (y - a(x))^2$. Тогда получаем задачу

$$\sum_{y \in Y} (y - t)^2 P(y \mid x) \to \min_t$$

Снова обозначим $p = P(y = 1 \mid x)$, а после возьмём производную

$$p \cdot (1-t)^2 + (1-p)(-t^2) \to \min_t$$
$$-2p(1-t) - 2(1-p) \cdot (-t) = 0 \quad \Rightarrow \quad t = p.$$

Получается, что функция потерь удовлетворяет условиям Пуха.

2. Пусть $L(y, a(x)) = |y - a(x)|$. Тогда получаем задачу

$$\sum_{y \in Y} |y - t| P(y \mid x) \to \min_t$$
$$p \cdot |1 - t| + (1-p) \cdot |-t| \to \min_t$$

Рассмотрим ситуации, в которых $t < 0$, $0 < t < 1$, $t > 1$.

Дорешать

В конечном итоге видим, что величина t принимает значения либо 1 либо 0 и никак не тянет на оценку вероятности. Такая функция не удовлетворяет условию Пуха.

Доделать про Ивана и Царя!

Упражнение 4. Царь предлагает Ивану не просто голову с плеч или половину царства, но ещё и промежуточный вариант, ничего (0). Функция как бы склеивается из кусочков. Сюда же задача про царя-извращенца, который платит за то, что мы не угадали

Упражнение 5. Случайная величина Y принимает два значения: 0 с вероятностью p и 1 с вероятностью $1 - p$. Постройте график зависимости энтропии от p. При каком p энтропия будет максимальной? Проинтерпретируйте это. Является ли функция монотонной? Выпуклой?

Упражнение 6. ещё упражнение про энтропию - ? добавить везде пунт с поиском перплексии

Упражнение 7. Найдите дивергенцию Кульбака-Лейбнера, если она определена,

1. из биномиального $\mathrm{Bin}(\frac{1}{3}, 2)$ в равновероятное на $0, 1, 2$;
2. из равновероятного на $0, 1, 2$ в биномиального $\mathrm{Bin}(\frac{1}{3}, 2)$;
3. из $\mathrm{N}(0, 1)$ в $\mathrm{N}(0, \sigma^2)$;
4. из $\mathrm{N}(0, \sigma^2)$ в $\mathrm{N}(0, 1)$;
5. из $\mathrm{N}(0, 1)$ в $\mathrm{Exp}(1)$;
6. из $\mathrm{Exp}(1)$ в $\mathrm{N}(0, 1)$.

Упражнение 8.

Добавить упражнений на взаимосвязь всего этого в один клубок и добавить байесовщинки :3

Поискать какие-нибудь задачи дожития с экспоненциальным распределением