

Функции потерь

Терминология	2
Аннотация	3
1. Оценка максимального правдоподобия и ДКЛ	4
2. Оценка максимального правдоподобия для линейной регрессии	9
2.1 Линейная регрессия с нормально-распределённым шумом	9
2.2 Линейная регрессия с лапласовским шумом	10
2.3 Выводы	10
3. Байесовский вывод и правдоподобие	11
Literature	13

Терминология

- MAE - Mean Absolute Error, средняя абсолютная ошибка
- MAP оценка - оценка апостериорного максимума
- MLE - метод максимального правдоподобия. From Bayes' perspective, MLE is a special case of MAP that assumes uniform prior distribution of the parameters.

Maximizing the likelihood is asymptotically equivalent to minimizing KLD, thus finding Θ that defines the probability distribution that has a minimal distance (or divergence) to real probability from which data was generated.

- MSE - среднеквадратичная ошибка
- ДКЛ - дивергенция Кульбака-Лейблера

Аннотация

В этой статье мы займёмся выбором функций ошибки для задач регрессии и классификации. Сначала мы обоснуем минимизацию *среднеквадратичного отклонения* как функцию ошибки для линейной регрессии. Мы будем исходить из постановки задачи регрессии как функции с бесконечным количеством исходов, тем самым естественно предполагая, что функция ошибки в этом случае будет непрерывной (и даже кусочно-выпуклой). С другой стороны, задача классификации имеет дискретное количество исходов, и ее функция ошибки не имеет такой же естественной природы, как для регрессии. Мы попытаемся ввести метрику на пространстве распределений, приближающих значения классификации, и покажем, что хотя введенная величина может и не обладать всеми свойствами метрики, но тем не менее, она может служить для определения «расстояния» между распределениями, то есть быть успешно использована в качестве функции ошибки. Рассматриваемая нами величина называется *перекрёстной энтропией*, и именно она повсеместно используется в коммерческих библиотеках (Tensorflow, PyTorch) при построении моделей классификации.

1. Оценка максимального правдоподобия и ДКЛ

Существует большое число способов ввести метрику на пространстве распределений. Наиболее известными можно считать равномерную метрику (метрику Колмогорова):

$$\rho(F_x, F_y) = \sup\{|F(x) - F(y)| : x \in R^1\}$$

или, например, расстояние полной вариации:

$$\sigma(F_x, F_y) = \frac{1}{2} \int |d(F(x) - F(y))|$$

Одни метрики были заимствованы из функционального анализа, другие же, благодаря их особым свойствам, вводились по специальным случаям. К таким случаям можно отнести и пре-метрики, которые удовлетворяют лишь части аксиоматики метрик, но однако, часто используются для задания топологии пространства распределений, и в определенной степени играть роль расстояния на нем.

Такова пре-метрика, которая известна из теории информации: *дивергенция Кульбака-Лейблера (ДКЛ)*. Для дискретных распределений она определяется как:

$$D_{KL}(P||Q) = \sum_{x \in X} P(x) \log\left(\frac{P(x)}{Q(x)}\right) \quad (1)$$

Для непрерывных:

$$D_{KL}(P||Q) = \int_{-\infty}^{\infty} p(x) \log\left(\frac{p(x)}{q(x)}\right) dx$$

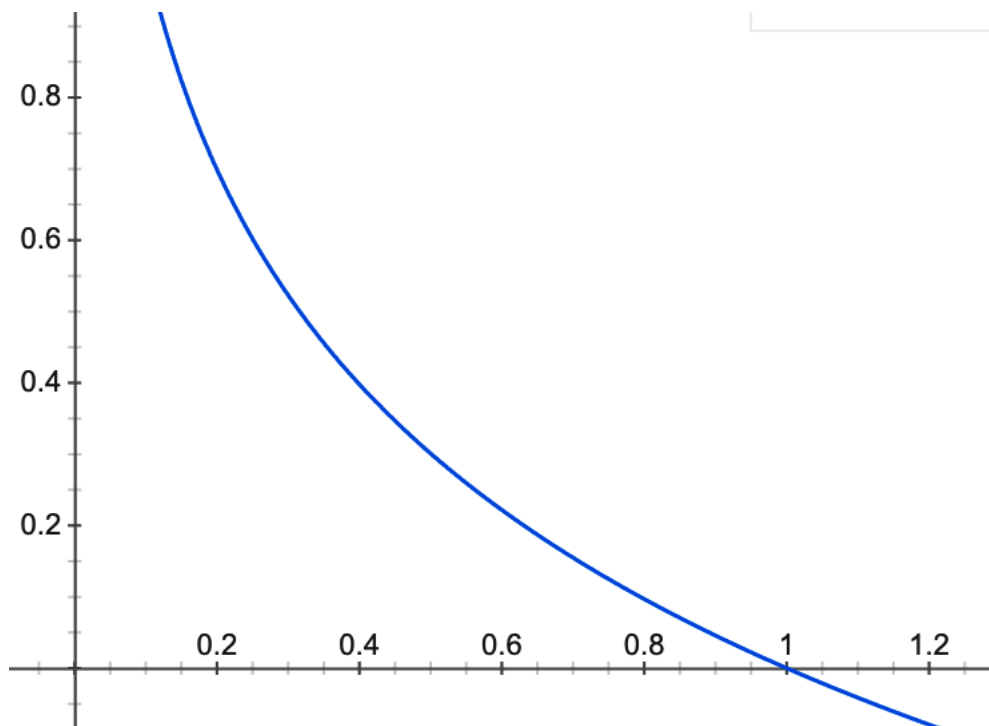
Эта дивергенция не является симметричной и не удовлетворяет неравенству треугольника:

$$D_{KL}(P||Q) \neq D_{KL}(Q||P)$$

Единственный факт, который роднит ДКЛ с метрикой, состоит в том что она не отрицательна и равна нулю только при $P = Q$ почти всюду.

Для того, чтобы объяснить смысл введённой величины отступим на шаг назад и попробуем формализовать интуитивное представление о том, что количество информации, которое несёт некое событие тем больше, чем это событие реже, т.е. чем меньше вероятность события, тем более оно информативно.

Это представление очень хорошо выражается функцией $I(x) = -\log(x)$, график которой приведён ниже:



По оси x здесь отложена вероятность события, по оси y - его «количество информации».

Можно заметить, что эта функция на отрезке $[0 \leq x \leq 1]$ прекрасно подходит к приведённому интуитивному выражению:

1. Она принимает 0 на значении 1 - максимально допустимом значении вероятности, т.е. информация, содержащаяся в событии, которое обязательно произойдёт (с вероятностью 1) - нулевая.
2. Чем меньше вероятность события, тем больше его собственная информация

$$\lim_{x \rightarrow +0} = \infty$$
3. $\forall x \in [0 \leq x \leq 1] : I(x) \geq 0$

Рассматриваемая величина была введена К. Шенноном в эпохальной работе [4], и получила название *собственной информации* события x :

$$I(x) = -\log p(x)$$

Она легко обобщается от одиночного события на всё (дискретное) распределение:

$$H(X) = - \sum_{i=1}^m p(x) \times \log p(x)$$

В этом случае она называется *энтропией* случайной величины.

Рассматривая энтропию как меру хаоса или неопределенности распределения, отметим теперь её особенности для известных распределений.

1. В общем случае, неравномерное распределение имеет меньшую энтропию чем равномерное

2. Равномерное распределение имеет наибольшую энтропию из всех возможных:

$$H(P) = - \sum_{i=1}^n \frac{1}{n} \log \frac{1}{n} = - \frac{n}{n} (-\log n) = \log n, \text{ где } n \geq 1 - \text{число испытаний.}$$

Очевидно, энтропия равномерного распределения неотрицательна.

3. Дискретное нормальное распределение имеет энтропию

$H(P) = \ln(\sigma \sqrt{2\pi e})$ независящую от матожидание распределения. (Вычисляется с помощью дискретного преобразования Абеля или интегрированием по частям для непрерывного случая) и увеличивающуюся только в зависимости от σ .

4. Распределение Лапласа (двойное экспоненциальное), которое часто используется в качестве предельного распределения в схемах суммирования случайного числа случайных величин, имеет энтропию:

$$H(X) = - \int_{-\infty}^{+\infty} \frac{2}{\lambda} e^{-\lambda|x-a|} \log \frac{2}{\lambda} e^{-\lambda|x-a|} dx = \log \frac{2}{\lambda}$$

также не зависящую от матожидания.

(Вычисляется теми же способами)

5. Наконец, энтропия биномиального распределения:

$$\begin{aligned} H(X) &= \sum_{m=0}^n C_n^m p^m q^{n-m} \log(C_n^m p^m q^{n-m})^{-1} = - \sum_{m=0}^n C_n^m p^m q^{n-m} [\log C_n^m + m \log p + (n-m) \log q] = \\ &= - \sum_{m=1}^n C_n^m p^m q^{n-m} \log C_n^m - n(p \log p + q \log q). \end{aligned}$$

Вообще говоря, информационная энтропия глубоко связана с энтропией физической. Природа представляется нам не терпящей порядка, т.е. любые проявления организованной структуры физического пространства могут рассматриваться как проявления временной аномалии. Равномерное распределение свойств с его максимальной энтропией и есть, собственно, суть второго начала термодинамики.

При сопоставлении двух распределений имеет смысл рассматривать *перекрестную энтропию*, которая определяется как:

$$H(P, Q) = - \sum_{x \in X} p(x) \log q(x)$$

Не возникает проблем с обобщением введенных величин и на непрерывные распределения. В этом случае рассматриваемая величина называется *дифференциальной энтропией* и выводится как первый член асимптотического разложения энтропии [5].

Здесь же нас интересует, прежде всего, дискретный случай, поэтому вернёмся к ДКЛ выпишем его дискретную форму поподробнее:

$$D_{KL}(P || Q) = \sum_{x \in X} P(x) \log\left(\frac{P(x)}{Q(x)}\right) = - \sum_{x \in X} p(x) \log q(x) + \sum_{x \in X} p(x) \log p(x) = H(P, Q) - H(P)$$

где $H(P, Q)$ - перекрестная энтропия между P и Q , а $H(P)$ - энтропия P .

Таким образом, в качестве важного промежуточного итога, мы имеем:

$$D_{KL}(P || Q) = H(P, Q) - H(P) \quad (2)$$

Дивергенция Кульбака-Лейблера применима также и к непрерывным распределениям. Например, найдем ДКЛ между двумя нормальными распределениями $p(x) = \mathbb{N}(x | \mu_1, \sigma_1)$ и $q(x) = \mathbb{N}(x | \mu_2, \sigma_2)$ (PRML, Bishop ex. 1.30, p.64)

$$\begin{aligned} D_{KL}(p || q) &= - \int p(x) \log q(x) dx + \int p(x) \log p(x) dx = \\ &= \frac{1}{2} \log(2\pi\sigma_2^2) + \frac{\sigma_1^2 + (\mu_1 - \mu_2)^2}{2\sigma_2^2} - \frac{1}{2} (1 + \log 2\pi\sigma_1^2) = \\ &= \log \frac{\sigma_2}{\sigma_1} + \frac{\sigma_1^2 + (\mu_1 - \mu_2)^2}{2\sigma_2^2} - \frac{1}{2} \end{aligned}$$

Последнее выражение дает 0 при $\mu_1 = \mu_2$ и $\sigma_1 = \sigma_2$.

Для дальнейших рассуждений напомним определение функции правдоподобия. Она представляет собой функцию от параметра распределения $f_x(x | \theta) : \Theta \rightarrow R$ определенную как:

$$L(\Theta) = \prod_{i=1}^{\infty} p(x_i | \theta) \quad (3)$$

Здесь $p(x_i)$ может быть выбрана или как функция вероятности, **или как функция плотности вероятности**.

Для заданного $p(x)$ (3) представляет собой условную вероятность

Её $\arg\max$ не изменится при логарифмировании, поэтому:

$$\Theta_{ML} = \arg \max_{\Theta} \prod_{i=1}^m p_{model}(x_i, \Theta) = \arg \max_{\Theta} \sum_{i=1}^m \log p_{model}(x_i, \Theta)$$

Argmax не изменится также при делении на m , поэтому:

$$\Theta_{ML} = \arg \max_{\Theta} \mathbb{E} \log p_{model}(x, \Theta) \quad (4)$$

Теперь, вспоминая (1), запишем:

$$D_{KL}(P_{data} || P_{model}) = \mathbb{E}_{data}[\log p_{data}(x) - \log p_{model}(x)]$$

но поскольку левая часть полученного выражения не зависит от P_{model} , то при минимизации дивергенции нам на самом деле остается минимизировать только

$$-\mathbb{E}[\log p_{model}(x)],$$

что совпадает с (4)

Таким образом, **максимизация правдоподобия эквивалентна минимизации дивергенции Кульбака-Лейблера.**

2. Оценка максимального правдоподобия для линейной регрессии

2.1 Линейная регрессия с нормально-распределённым шумом

Если рассматривать линейную регрессию как зависимость вида:

$$Y = w^T X + \epsilon \quad (5)$$

где ϵ - нормально распределенная случайная величина (шум) с матожиданием μ и дисперсией σ , т.е. $\epsilon \sim N(\mu, \sigma^2)$, независимая от X и Y , то значения Y тоже распределены нормально с плотностью вероятности, соответствующей многомерному нормальному распределению.

Функция правдоподобия такого распределения, в которой используется плотность вероятности, приобретает вид:

$$L(\Theta) = \prod_{i=1}^m p(y_i | x_i; w, \sigma) = \prod_{i=1}^m \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y_i - w_i x_i)^2}{2\sigma^2}},$$

где Θ - вектор (w, σ) .

После логарифмирования (т.к. $\log(\prod_{i=1}^m a_i b_i) = \sum_{i=1}^m \log a_i b_i$) это дает:

$$\begin{aligned} \ln \prod_{i=1}^m p(y_i | x_i; w, \sigma) &= \sum_{i=1}^m \ln p(y_i | x_i; w, \sigma) = \sum_{i=1}^m \ln \left[\frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2} \left(\frac{y_i - \hat{y}_i}{\sigma} \right)^2} \right] = \\ &= \sum_{i=1}^m \left[\ln(2\pi\sigma^2)^{-\frac{1}{2}} + \ln e^{-\frac{1}{2} \left(\frac{y_i - \hat{y}_i}{\sigma} \right)^2} \right] = \sum_{i=1}^m \left[-\frac{1}{2} \ln(2\pi) - \log \sigma - \frac{1}{2\sigma^2} (y_i - \hat{y}_i)^2 \right] = \\ &= -\frac{m}{2} \log(2\pi) - m \ln \sigma - \frac{1}{2\sigma^2} \sum_{i=1}^m (y_i - \hat{y}_i)^2 \end{aligned}$$

где \hat{y}_i - результат вычисления модели для элемента x_i , а m - число элементов выборки.

Но поскольку первые два члена правой части последнего выражения не зависят от параметров модели (σ - постоянная), то можно записать:

$$\Theta_{ML} = \arg \max_{\Theta} - \sum_{i=1}^m \frac{1}{2} |y_i - \hat{y}_i|^2 = \min \sum_{i=1}^m |y_i - \hat{y}_i|^2$$

Сравнивая это выражение с определением среднеквадратичной ошибки:

$$MSE = \frac{1}{m} \sum_{i=1}^m |\hat{y}_i - y_i|^2$$

можно легко заметить, что максимизация правдоподобия относительно искомого вектора Θ является, по сути, минимизацией среднеквадратичной ошибки для тех же параметров. Что, собственно, и показывает, что именно **среднеквадратичное отклонение является оптимальной функцией ошибки линейной регрессии**.

2.2 Линейная регрессия с лапласовским шумом

Если шум в линейной регрессии имеет распределение Лапласа:

$$p(\epsilon) = \frac{\alpha}{2} e^{-\alpha|\epsilon-\beta|}$$

с нулевым средним ($\beta = 0$), то логарифмическая оценка максимального правдоподобия дает:

$$Q_{ML} = \arg \min_q \frac{1}{m} \sum_{i=1}^m |a(x_i) - y_i|$$

т.е. среднюю абсолютную ошибку.

2.3 Выводы

Таким образом, линейная регрессия может быть определена без предположения о нормальном распределении шума. Её параметры могут быть рассчитаны согласно среднеквадратичному отклонению (MSE), однако этот метод будет оптимальным только в случае нормального распределения.

3. Байесовский вывод и правдоподобие

До сих пор мы обсуждали т. наз. частотный подход к оценке единственного значения параметра Θ . Наши дальнейшие суждения о неизвестных значениях данных строились на этой оценке.

В этой части мы изменим наш подход, рассматривая все возможные значения Θ для предсказания неизвестных данных. Мы будем полагать набор данных фиксированным, а параметр Θ - случайной величиной, распределённой согласно некоторому априорному распределению. Обычно это распределение выбирают близким к равномерному, т.е. с высокой энтропией, чтобы выразить высокую степень неопределенности Θ перед рассмотрением данных. Очевидно, что с получением новых данных, наша степень уверенности в правильности выбранного параметра пересматривается.

Как правило, мы начинаем с равномерного или нормального априорного распределения с высокой энтропией, изменяя наши суждения о параметре каждый раз в сторону уменьшения энтропии, т.е. заставляя искомый параметр Θ сужаться в область высокой вероятности.

[Занимаясь оценкой максимального правдоподобия мы целиком зависели от наблюдаемых данных.]

Рассмотрим то же выражение для линейной регрессии (5)

$$Y = w^T X + \epsilon,$$

но в этот раз будем считать вектор w случайной величиной, распределённой с некоторой априорной плотностью вероятности $p(w)$.

Тогда по определению условной вероятности:

$$P(X | Y)(w) \equiv \frac{\tilde{P}(X, Y)}{P(Y)} \quad (6),$$

где $\tilde{P}(X, Y)$ - совместная вероятность распределений X и Y .

По теореме умножения вероятностей совместная вероятность выражается через условную как

$$\tilde{P}(X, Y) = P(X)P(Y | X),$$

(если $P(X)$ - вероятность того, что пешеход пойдет на красный свет, а $P(Y | X)$ - вероятность того, что переходящий на красный свет пешеход, будет сбит, то их перемножение дает совместную вероятность - пример [отсюда](#)).

С другой стороны, поскольку переменные независимы, то

$$P(Y | X) = L_{Y,X}(w)$$

а значит

$$P(X | Y)(w) = \frac{P(X)L_{Y,X}(w)}{P(Y)} \quad (7 - \text{теорема Байеса})$$

где $L_{Y,X}(w)$ - функция правдоподобия распределения Y при наблюдаемых значениях случайной величины X , зависящая от параметра w . Это та же функция правдоподобия, которую мы определили в (3), однако здесь она используется как функция X при фиксированном Y .

[в другом контексте: частотный подход рассматривает w как фиксированный параметр, значение которого определяется оценкой, получаемой на различных данных выборки. Например, оценка максимального правдоподобия ищет значения w , которые доставляют максимум функции правдоподобия. В противоположность этому байесовский подход рассматривает единственную - наблюдаемую - выборку, неопределенность параметров которой выражается через распределение вероятности w .]

Знаменатель в (7) служит нормировочной константой, которая обеспечивает левую часть равенства необходимыми свойствами плотности вероятности, т. е.

$$P(Y) = \int L_{Z,Y}(w)P(X)dx = 0$$

Левая часть последнего равенства называется *апостериорной* вероятностью.

Вообще, для построения байесовского вывода требуются 3 вещи

1. Модель
2. Функцию плотности вероятности для априорного распределение
3. Сами данные - выборка
4. Имея эти данные, мы хотим получить апостериорное распределение

Произведение вероятностей в числителе формулы Байеса должно производиться для каждого элемента выборки, что практически неосуществимо для реальных наборов данных, содержащих до сотен тысяч элементов.

Однако для многих практических случаев знание сопряженных распределений позволяет заменить трудоемкие операции интегрирования простыми алгебраическими манипуляциями над параметрами распределений.

Если нельзя гарантировать условия сопряженности распределений, методы проб Монте-Карло и марковские сети могут давать приемлемые результаты.

С другой стороны, существуют и детерминированные схемы аппроксимации - variational Bayes, expectation propagation- которые последнее время используются все чаще для больших наборов данных.

Literature

1. Kullback S. (1959). Information theory and statistics. Dover Publications.
2. Bishop C. M. (2006). Pattern Recognition and Machine Learning. Springer.
3. Goodfellow I., Bengio Y., Courville A. (2016) Deep Learning. MIT Press.
4. Shannon C. E. A mathematical Theory of Communication.
5. Колмогоров А. Н. Теория информации и теория алгоритмов.

6.

7.