2021

Лекция 5. Логистическая регрессия

МЕТОДЫ МАШИННОГО ОБУЧЕНИЯ

ПАПУЛИН С.Ю. (papulin.study@yandex.ru)

Дополнительные материалы доступны по ссылке: https://github.com/mlmethods

Содержание

1. Метод максимального правдоподобия

Методы максимального правдоподобия (Maximum Likelihood Estimation – MLE) используются для оценки параметров (так же как метод оценки апостериорного максимума, про который в одной из следующих лекций).

Метод максимального правдоподобия – метод оценки, который позволяет по имеющейся выборке оценить параметры вероятностного распределения генеральной совокупности. Другими словами, у нас есть выборка и задача MLE оценить параметры вероятностного распределения данных, из которых была взята эта выборка.

Функция правдоподобия $L(\boldsymbol{x}|\theta)$ для каждого значения θ определяет меру правдоподобия получения наблюдения $\boldsymbol{x}$

$$\theta_{MLE} = \underset{\theta}{\operatorname{argmax}} L(\boldsymbol{x}|\theta) = \underset{\theta}{\operatorname{argmax}} p(\boldsymbol{x}|\theta) = \underset{\theta}{\operatorname{argmax}} \prod_{i=1}^{n} p(x_i|\theta)$$

Или с использованием логарифма:

$$\theta_{MLE} = \underset{\theta}{\operatorname{argmax}} \log \prod_{i=1}^{n} p(x_i|\theta) = \underset{\theta}{\operatorname{argmax}} \sum_{i=1}^{n} \log p(x_i|\theta)$$

Замечание

В данном случаем для функции правдоподобия используется обозначение $L(x|\theta)$. Существуют и другие варианты, например, $L(\theta|x)$, $L(x;\theta)$ или $L_X(\theta)$

Замечание

Для дискетной случайной величины используется функция вероятности (pmf), а для непрерывной – функция плотности вероятности (pdf)

1.1. Оценка параметра дискретного распределения Бернулли

Частный случай

Дана выборка 1110. Найти оценку вероятности p появления 1?

Функция правдоподобия в этом случае будет иметь вид:

$$L(x_1, x_2, x_3, x_4|p) = p \cdot p \cdot p \cdot (1-p) = p^3 \cdot (1-p)$$

Чтобы найти

$$\hat{p} = \underset{p}{\operatorname{argmax}} L(x_1, x_2, x_3, x_4|p)$$

Возьмем производную и приравняем полученное выражение к нулю:

$$\frac{d}{dp}L(x_1, x_2, x_3, x_4|p) = 3p^2 - 4p^3 = 0$$

В итоге оценка вероятности будет равна

$$\hat{p} = 3/4$$

Общий случай

Дана выборка $x_1, x_2, \dots, x_n$. Полагаем, что значение каждого наблюдения подчиняется закону Бернулли, т. е. $x_1, x_2, \dots, x_n \sim \mathcal{B}(x|p)$, где $x \in \{0,1\}$, p – переменная, соответствующая вероятности появления 1.

Распределение Бернулли имеет вид:

$$p(x|p) = \begin{cases} p, \text{если } x = 1 \\ q = 1 - p, \text{если } x = 0 \end{cases}$$

Или можно записать, как

$$p(x|p) = p^x(1-p)^{(1-x)}$$

Необходимо найти оценку параметра p по имеющейся выборке?

Решение:

Наша задача найти:

$$\hat{p} = \underset{p}{\operatorname{argmax}} L(x_1, x_2, \dots, x_n|p)$$

Или эту же задачу можно представить как

$$\hat{p} = \underset{p}{\operatorname{argmax}}[\log L(x_1, x_2, \dots, x_n|p)]$$

Функция правдоподобия имеет вид:

$$L(x_1, x_2, \dots, x_n|p) = \prod_{i=1}^{n} p(x_i|p)$$

Возьмём логарифм от функции правдоподобия:

$$\log L(x_1, x_2, \dots, x_n|p) = \log \prod_{i=1}^{n} p(x_i|p) = \sum_{i=1}^{n} \log p(x_i|p) = \sum_{i=1}^{n} \log\left[p^{x_i}(1-p)^{(1-x_i)}\right]$$
$$= \sum_{i=1}^{n} [x_i \log p + (1 - x_i)\log(1-p)]$$

Получаем

$$\log L(x_1, x_2, \dots, x_n|p) = \sum_{i=1}^{n} [x_i \log p + (1 - x_i)\log(1-p)]$$

Чтобы найти максимум, возьмем производную по p и приравняем полученное выражение к нулю:

$$\frac{d}{dp}\sum_{i=1}^{n}[x_i \log p + (1-x_i)\log(1-p)] = \sum_{i=1}^{n}\left[\frac{x_i}{p} + \frac{1-x_i}{1-p}\right] = 0$$

Таким образом,

$$\hat{p} = \frac{\sum_{i=1}^{n} x_i}{n}$$

1.2. Оценка параметров нормального закона распределения

Дана выборка $x_1, x_2, \dots, x_n$. Полагаем, что значение каждого наблюдения подчиняется нормальному закону распределения, т. е. $x_1, x_2, \dots, x_n \sim \mathcal{N}(x|\mu, \sigma^2)$

Необходимо найти оценки параметров μ, σ^2?

Решение:

Нормальное распределение данных с одним признаком:

$$p(x_i|\mu,\sigma^2) = \mathcal{N}(x_i|\mu,\sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2\sigma^2}(x_i-\mu)^2}$$

Функция правдоподобия:

$$L(x_1, x_2, \dots, x_n|\mu,\sigma^2) = \prod_{i=1}^{n} p(x_i|\mu,\sigma^2) = \prod_{i=1}^{n} \mathcal{N}(x_i|\mu,\sigma^2) = \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^n e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2}$$

Логарифм функции правдоподобия:

$$\log L(x_1, x_2, \dots, x_n|\mu,\sigma^2) = -\frac{n}{2}\log 2\pi - \frac{n}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2$$

Тогда неизвестные параметры μ и σ^2 можно оценить следующим образом:

$$\hat{\theta}_{MLE} = (\hat{\mu}_{MLE}, \hat{\sigma}^2_{MLE})$$
$$= \underset{\mu,\sigma^2}{\operatorname{argmax}} L(x_1, x_2, \dots, x_n|\mu,\sigma^2) = \underset{\mu,\sigma^2}{\operatorname{argmax}} \left[-\frac{n}{2}\log 2\pi - \frac{n}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2\right]$$

Приравняем частные производные по параметрам к нулю:

$$\frac{\partial L(x_1, x_2, \dots, x_n|\mu,\sigma^2)}{\partial \mu} = 0$$

$$\frac{\partial L(x_1, x_2, \dots, x_n|\mu,\sigma^2)}{\partial \sigma^2} = 0$$

В итоге получим, что

$$\hat{\mu}_{MLE} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

$$\hat{\sigma}^2_{MLE} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \hat{\mu}_{MLE})^2$$

1.3. Оценка параметров линейной регрессии

Дана выборка $\{(x_i, y_i)\}_{1..n}$

Линейная регрессия в общем виде представляется как:

$$y_i = \boldsymbol{\theta}^T \boldsymbol{x}_i + \varepsilon_i$$

Полагаем, что

$$\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$$

$$\sigma^2 = \mathrm{var}[\varepsilon_i]$$

Найти оценки параметров θ и σ^2?

Функция плотности распределения для y_i:

$$p(y_i|X) = \mathcal{N}(y_i|\boldsymbol{\theta}^T\boldsymbol{x}_i, \sigma^2) = \frac{1}{\sqrt{2\pi}\sigma}\exp\left[-\frac{1}{2\sigma^2}(y_i - \boldsymbol{\theta}^T\boldsymbol{x}_i)^2\right]$$

Функция правдоподобия для всей выборки будет иметь вид:

$$L(\mathbf{y}|X;\boldsymbol{\theta},\sigma^2) = \prod_{i=1}^{n} p(y_i|X;\theta,\sigma^2) = \left[\frac{1}{\sqrt{2\pi}\sigma}\right]^n \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \boldsymbol{\theta}^T\boldsymbol{x}_i)^2\right]$$
$$= [2\pi\sigma^2]^{-n/2}\exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \boldsymbol{\theta}^T\boldsymbol{x}_i)^2\right]$$

Возьмем логарифм от функции правдоподобия

$$\log L(\mathbf{y}|X;\boldsymbol{\theta},\sigma^2) = -\frac{n}{2}\log 2\pi\sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \boldsymbol{\theta}^T\boldsymbol{x}_i)^2 =$$
$$= -\frac{n}{2}\log 2\pi - \frac{n}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \boldsymbol{\theta}^T\boldsymbol{x}_i)^2$$

Наша задача сводится к тому, чтобы найти такие параметры θ, которые бы максимизировали функцию правдоподобия (или её логарифм):

$$\widehat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta}}{\mathrm{argmax}}[\log L(\mathbf{y}|X;\boldsymbol{\theta},\sigma^2)]$$

Чтобы найти оценки неизвестных параметров $\boldsymbol{\theta}$ и σ^2, найдем частные производные и приравняем их к нулю:

$$\nabla \log L(\mathbf{y}|X;\boldsymbol{\theta},\sigma^2) = \mathbf{0}$$

В частности, частные производные по $\boldsymbol{\theta}$:

$$\nabla_\theta \log L(\mathbf{y}|X;\boldsymbol{\theta},\sigma^2) = \frac{\partial}{\partial\boldsymbol{\theta}}\log L(\mathbf{y}|X;\boldsymbol{\theta},\sigma^2) = \mathbf{0}$$

И частная производная по σ^2:

$$\frac{\partial}{\partial \sigma^2} \log L(\mathbf{y}|X; \boldsymbol{\theta}, \sigma^2) = 0$$

Если решим систему уравнений для частных производные по $\boldsymbol{\theta}$, то найдем значения оценок параметров $\widehat{\boldsymbol{\theta}}$. Далее эти значения надо подставить для оценки σ^2. В итоге получим, что

$$\widehat{\boldsymbol{\theta}} = (X^T X)^{-1} X^T \mathbf{y}$$

и

$$\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^{n} (y_i - \widehat{\boldsymbol{\theta}}^T \boldsymbol{x}_i)^2$$

2. Классификация

Цель классификации – получить входной вектор $\boldsymbol{x}$ и соотнести его с одним из K дискретных классов c_k, где $k = 1, \dots, K$. Как правило, классы представляются как несовместные события, то есть каждому входному вектору присваивается один и только один класс.

Для вероятностных моделей с двумя классами наиболее удобным представлением целевого значения является переменная y со значениями 0 или 1, $y \in \{0,1\}$. При этом $y = 0$ представляет класс c_0, а $y = 1$ класс c_1. Иногда для представления классов используются значения -1 и 1.

Примеры:

- Медицинская диагностика
- Кредитный скоринг
- Распознавание образов

Существует три подхода в классификации:

- дискриминантная функция (discriminant function)

  Напрямую назначает входному вектору определенный класс.

- дискриминативная модель (discriminative model)

  Представляет вероятность $p(c_k|x)$ как параметрическую модели и оценивает параметры на обучающем множестве. После получения вероятностей для всех классов принимает решение о принадлежности некоторому классу. Пример: логистическая регрессия, метод опорных векторов

- порождающая модель (generative model)

  Использует условные по классу вероятности $p(x|c_k)$ и априорные вероятности $p(c_k)$ для классов, и затем вычисляется апостериорная вероятность $p(c_k|x)$ по теореме Байеса:

$$p(c_k|x) = \frac{p(x|c_k)p(c_k)}{p(x)}$$

Пример: наивный байесовский классификатор, смешанная гауссовская модель (Gaussian Mixture Model), скрытая марковская модель

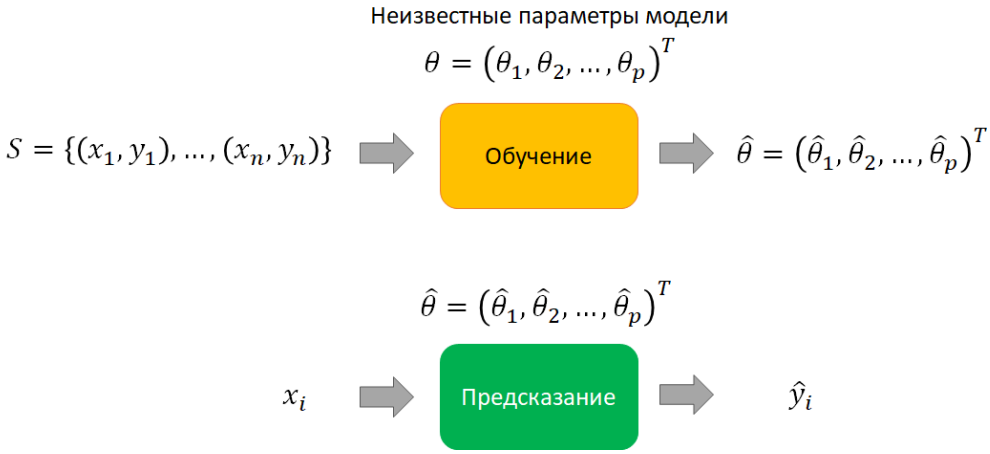
Порождающая модель строит вероятностную модель для каждого класса и определяет, насколько входной вектор (наблюдаемое значение) соответствует каждой из моделей, и на основе этого дает заключение о принадлежности к какому-либо классу. Дискриминативная модель обучается для распознавания/различения классов. Такая модель может отличить один класс от другого, но не имеет представления о моделях самих классов.

3. Логистическая регрессия

3.1. Общие сведения

Логистическая регрессия является методом классификации и относится к классу методов обучения с учителем. Таким образом, для обучения используется множество с известными целевыми значениями (метками) y (например, 0 или 1) для каждого наблюдения/экземпляра.

Кроме того, логистическая регрессия относится к параметрическим методам, то есть необходимо найти значения параметров, при которых модель будет иметь наименьшую ошибку классификации.



Задача заключается в оценке параметров θ таким образом, чтобы $\hat{y}$ для каждого наблюдения соответствовала бы действительному значению y.

Для этого необходимо:

1) Задать метрику определения близости $\hat{y}$ и y. Выражается в виде расстояния и называется функция потерь (**loss/cost function**)
2) Решить задачу оптимизации, т. е. найти способ определения параметров, для которых функция потерь будет минимальна

Параметр θ_i определяет, насколько важным является входной признак для принятия решения по отнесению наблюдения к конкретному классу. Например, если параметр положительный, то он ассоциируется с классом 1, если отрицательный, то с классом 0.

3.2. Логистическая функция для задачи классификации

Линейная регрессия для классификации

Функцию линейной регрессии можно записать следующим образом

$$z = \sum_{i=0}^{p} \theta_i x_i = \boldsymbol{\theta}^T \boldsymbol{x}.$$

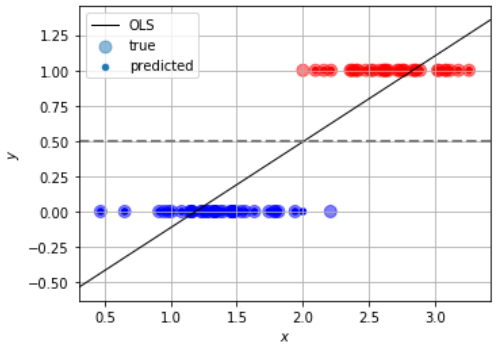
Замечание

Здесь и далее полагаем, что $x_0 = 1$

Чтобы построить классификатор на её основе, необходимо задать критерий отнесения значения z к тому или иному классу. Для этому можно использовать простой порог, например:

$$\hat{y} = \begin{cases} 1, \text{если } z > 0.5 \\ 0 \text{ иначе} \end{cases}$$

Пример применения метода наименьших квадратов (МНК) для задачи классификации:



Проблемы классификации с линейной регрессией:

- С одним признаком

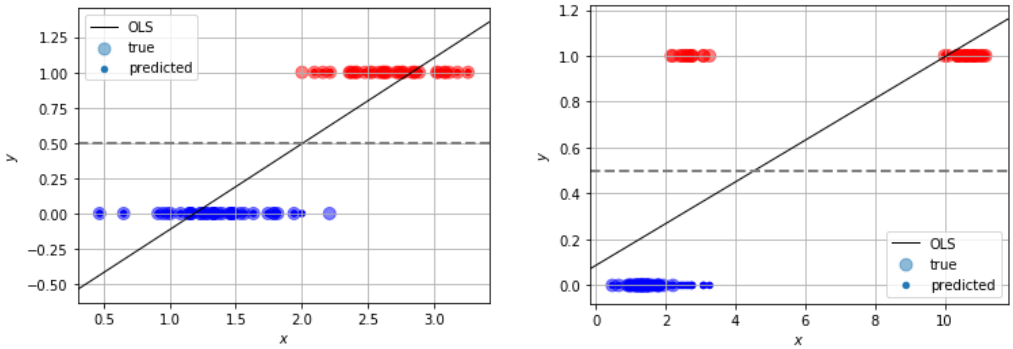


Рисунок x – Классификакция посредством линейно регрессии для одномерного пространства признаков

- С двумя признаками

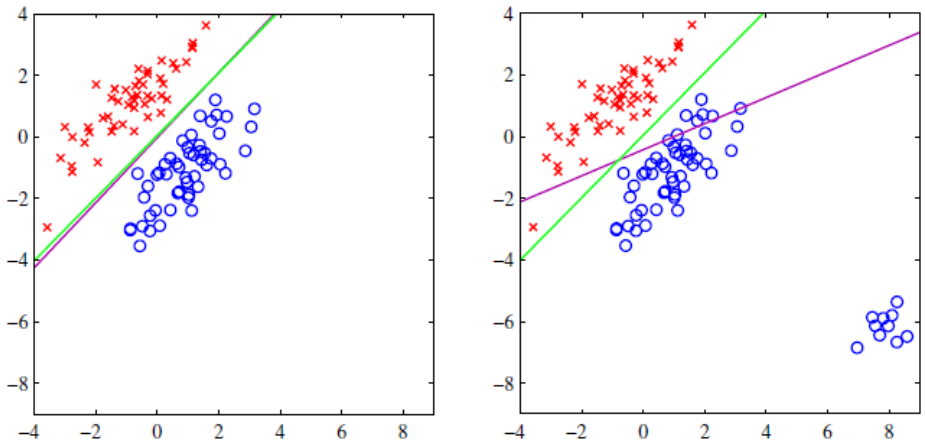


Рисунок x – Классификакция посредством линейно регрессии для двумерного пространства признаков

Логистическая функция *(сигмоида) (Logistic/sigmoid function)*

Как было показано ранее, классификатор на основе линейной регрессии чувствителен к выбросам. Что избежать данной проблемы, рассмотрим сигмоидальную функцию вида

$$\sigma(z) = \frac{1}{1 + e^{-z}}$$

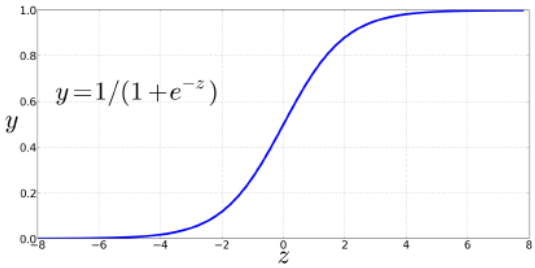


Рисунок x – Сигмоидальная функция

Если вместо z подставить функцию линейной регрессии с одним признаком, то получим следующую сигмойду:

$$\sigma(x) = \frac{1}{1 + e^{-(\theta_0 + \theta_1 x)}},$$

где параметры θ_0 и θ_1 контролируют смещение по x и крутизну кривой, соответственно.

В общем виде для p признаков сигмоидальная функция примет вид

$$\sigma(\boldsymbol{x}) = \frac{1}{1 + e^{-\boldsymbol{\theta}^T \boldsymbol{x}}},$$

так как

$$z = \sum_{i=0}^{p} \theta_i x_i = \boldsymbol{\theta}^T \boldsymbol{x},$$

Преимущества использования сигмоидальной функции:

- Возвращает действительное значение
- Область значения от 0 до 1, что соответствует значениям вероятностей
- Выбросы (outlier) будут стремиться к 0 или 1
- Дифференцируема

Представим логистическую функцию как вероятность принадлежности к классу 1:

$$p(y = 1|\boldsymbol{x}) = h_\theta(\boldsymbol{x}) = \frac{1}{1 + e^{-\boldsymbol{\theta}^T \boldsymbol{x}}}$$

Тогда принадлежность к классу 0 можно записать как

$$p(y = 0|\boldsymbol{x}) = 1 - h_\theta(\boldsymbol{x}) = \frac{e^{-\boldsymbol{\theta}^T \boldsymbol{x}}}{1 + e^{-\boldsymbol{\theta}^T \boldsymbol{x}}}$$

Сумма вероятностей принадлежности классу 1 или классу 0:

$$p(y = 0|\boldsymbol{x}) + p(y = 1|\boldsymbol{x}) = 1$$

Функция принятия решения

$$\hat{y} = \begin{cases} 1, \text{если } p(y = 1|\boldsymbol{x}) > 0.5 \\ 0 \text{ иначе} \end{cases}$$

Каким образом найти параметры $\boldsymbol{\theta}$? Метод максимального правдоподобия поможет нам разобраться как решается эта задача. Далее рассмотрим какой будет функция правдоподобия в данном случае и как найти неизвестные параметры $\boldsymbol{\theta}$.

3.3. Функция потерь логистической регрессии

Если рассматривать, что элементы выборки $\boldsymbol{x}_1, \boldsymbol{x}_2, \dots, \boldsymbol{x}_n$ независимы, то функцию правдоподобия для классификации (дискриминативной) можно записать как

$$L(\mathbf{y}|X, \boldsymbol{\theta}) = \prod_{i=1}^{n} p(y_i|\boldsymbol{x}_i, \boldsymbol{\theta}).$$

Задача заключается в поиске параметров $\boldsymbol{\theta}$, при которых функция правдоподобия примет максимальное значение:

$$\widehat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta}}{\operatorname{argmax}}\, L(\mathbf{y}|X, \boldsymbol{\theta})$$

Другими словами, можно сказать, что необходимо найти такие параметры $\boldsymbol{\theta}$, при которых мы с наибольшей вероятностью получим исходную выборку целевых значений $y_1, y_2, \dots, y_n$.

Замечание

Ранее отмечалось, что оценку неизвестного параметра в методе максимального правдоподобия можно записать следующим образом

$$\theta_{MLE} = \underset{\theta}{\operatorname{argmax}} L(\boldsymbol{x}|\theta) = \underset{\theta}{\operatorname{argmax}} p(\boldsymbol{x}|\theta) = \underset{\theta}{\operatorname{argmax}} \prod_{i=1}^{n} p(x_i|\theta)$$

Вероятность появления некоторого y_i при заданных $\boldsymbol{x}_i$ и $\boldsymbol{\theta}$, что можно записать в следующем виде

$$p(y_i|\boldsymbol{x}_i, \boldsymbol{\theta}) = p(y_i = 1|\boldsymbol{x}_i, \boldsymbol{\theta})^{y_i} p(y_i = 0|\boldsymbol{x}_i, \boldsymbol{\theta})^{1-y_i}$$

Ранее приведенное выражение (распределение Бернулли) можно пояснить следующим образом:

- если действительное значение $y_i = 1$:

$$p(y_i|\boldsymbol{x}_i, \boldsymbol{\theta}) = p(y_i = 1|\boldsymbol{x}_i, \boldsymbol{\theta})^1 \underbrace{p(y_i = 0|\boldsymbol{x}_i, \boldsymbol{\theta})^0}_{1}$$

- если действительное значение $y_i = 0$:

$$p(y_i|\boldsymbol{x}_i, \boldsymbol{\theta}) = \underbrace{p(y_i = 1|\boldsymbol{x}_i, \boldsymbol{\theta})^0}_{1} p(y_i = 0|\boldsymbol{x}_i, \boldsymbol{\theta})^1$$

Замечание

Распределение Бернулли имеет вид:

$$p(x|p) = \begin{cases} p, \text{если } x = 1 \\ q = 1 - p, \text{если } x = 0 \end{cases}$$

Или можно записать, как

$$p(x|p) = p^x (1-p)^{(1-x)}$$

Если учесть, что вероятность появления $y_i = 1$ при заданном $\boldsymbol{x}_i$ и $\boldsymbol{\theta}$ можно представить как сигмоидальную функцию, получим

$$p(y_i = 1|\boldsymbol{x}_i, \boldsymbol{\theta}) = h_{\theta,i}$$

и

$$p(y_i = 0|\boldsymbol{x}_i, \boldsymbol{\theta}) = 1 - h_{\theta,i}.$$

Получаем следующую общую запись вероятности

$$p(y_i|\boldsymbol{x}_i, \boldsymbol{\theta}) = h_{\theta,i}^{y_i} \left(1 - h_{\theta,i}\right)^{1-y_i}.$$

Приведенное выражения обладает рядом важных свойств:

- Максимизация вероятности также максимизирует логарифм вероятности

$$\log p(y_i|\boldsymbol{x}_i, \boldsymbol{\theta}) = \log \left(h_{\theta,i}^{y_i} \left(1 - h_{\theta,i}\right)^{1-y_i} \right) = y_i \log h_{\theta,i} + (1 - y_i) \log\left(1 - h_{\theta,i}\right)$$

Замечание

Для распределения Бернулли

$$\log\left[p^{x_i}(1-p)^{(1-x_i)}\right] = x_i \log p + (1-x_i)\log(1-p)$$

- Данная функция является вогнутой
- Меняем знак, чтобы получить выпуклую функцию
- Локальный минимум выпуклой функции является глобальным минимумом

Полученное выражение называется перекрестная энтропия, которую далее будем использовать как функцию потерь (cross-entropy loss function):

$$l_{CE,i}(\boldsymbol{\theta}) = -\log p(y_i|\boldsymbol{x_i}, \theta) = -y_i \log h_{\theta,i} - (1-y_i)\log\left(1-h_{\theta,i}\right)$$

или

$$l_{CE,i}(\boldsymbol{\theta}) = -y_i \log \frac{1}{1+e^{-\boldsymbol{\theta}^T \boldsymbol{x}_i}} - (1-y_i)\log\left(1-\frac{1}{1+e^{-\boldsymbol{\theta}^T \boldsymbol{x}_i}}\right)$$

Мы хотим, чтобы потери были небольшими, если оценка h_θ близка к действительному значению y, и наоборот большими, если значения значительно отличаются.

Идеальный классификатор присвоит вероятность $p(y_i|\boldsymbol{x}_i, \boldsymbol{\theta}) = 1$ корректному исходу ($y = 1$ или $y = 0$) и вероятность $p(y_i|\boldsymbol{x}_i, \boldsymbol{\theta}) = 0$ для некорректного. Другими словами, если действительное значение класса равно $y = 1$, то значение вероятности отнесения классификатором наблюдения x к классу 1 должно быть равно 1, а к классу 0, соответственно, 0. Аналогично для $y = 0$, значение вероятности отнесения классификатором наблюдения $\boldsymbol{x}$ к классу 0 должно быть равно 1, а к классу 1, соответственно, 0. Для такого классификатора функция потерь будет иметь нулевое значение.

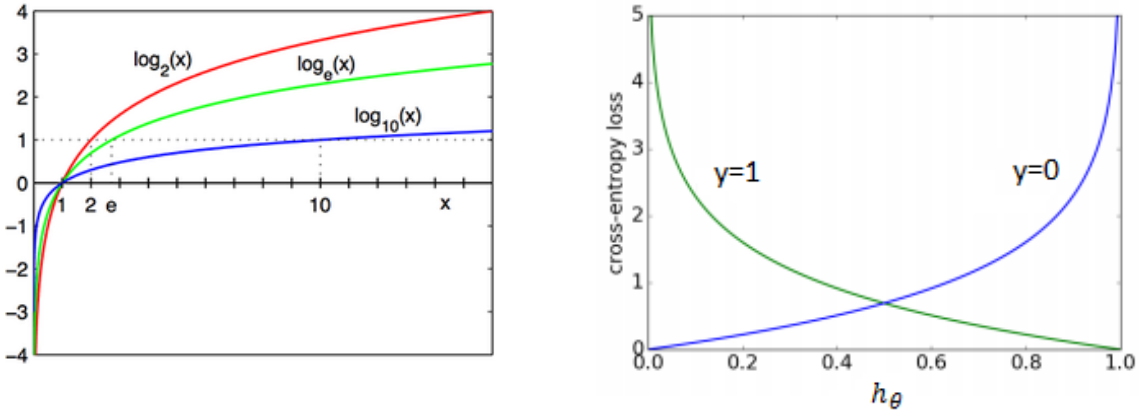


Рисунок – Функция ошибки для одного наблюдения

Обучение логистической регрессии происходит на множестве наблюдений. Для выборки из n наблюдений функцию правдоподобия можно записать следующим образом

$$L(\mathbf{y}|X, \boldsymbol{\theta}) = \prod_{i=1}^{n} h_{\theta,i}^{y_i}\left(1-h_{\theta,i}\right)^{1-y_i}.$$

Если взять логарифм от функции правдоподобия, то получим:

$$\log L(\mathbf{y}|X, \boldsymbol{\theta}) = \log \prod_{i=1}^{n} p(y_i|\boldsymbol{x}_i, \boldsymbol{\theta}) = \sum_{i=1}^{n} \log p(y_i|\boldsymbol{x}_i, \boldsymbol{\theta}) = \sum_{i=1}^{n} \left[y_i \log h_{\theta,i} + (1-y_i)\log\left(1-h_{\theta,i}\right)\right]$$

Для оценки неизвестных параметров $\boldsymbol{\theta}$ необходимо решить следующую задачу

$$\widehat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta}}{\operatorname{argmax}} \log L(\mathbf{y}|X, \boldsymbol{\theta}) = \underset{\boldsymbol{\theta}}{\operatorname{argmax}} \sum_{i=1}^{n} \left[y_i \log h_{\theta,i} + (1 - y_i) \log\left(1 - h_{\theta,i}\right) \right]$$

Замечание

Логарифм функции максимального правдоподобия для распределения Бернулли

$$\log L(x_1, x_2, \dots, x_n|p) = \sum_{i=1}^{n} [x_i \log p + (1 - x_i) \log(1 - p)]$$

Оценка неизвестного параметра

$$\hat{p} = \underset{p}{\operatorname{argmax}} [\log L(x_1, x_2, \dots, x_n|p)]$$

Однако в качестве функции потерь для задачи классификации посредством логистической регрессии, как правило, используют среднюю перекрестную энтропию, которая имеет следующий вид

$$L_{CE}(\boldsymbol{\theta}) = \frac{1}{n} \sum_{i=1}^{n} l_{CE,i}(\boldsymbol{\theta}) = -\frac{1}{n} \sum_{i=1}^{n} \left[y_i \log h_{\theta,i} + (1 - y_i) \log\left(1 - h_{\theta,i}\right) \right]$$

По сути, мы меняем задачу максимизации на задачу минимизации.

Замечание

В контексте классификации расстояние Кульбака-Лейблера (Kullback–Leibler divergence) определяет ассиметричную меру того, как одно случайное распределение P (наблюдаемые данные) отличается от другого Q (предсказания)

$$D_{KL}(P \parallel Q) = H(P, Q) - H(P) = -\sum_{i} p(x) \log q(x) + const,$$

где $H(P, Q)$ – кросс-энтропия между двумя распределениями вероятностей случайных величие P и Q

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

и $H(P)$ – энтропия (мера неопределенности)

$$H(P) = -\sum_{i} p(x) \log p(x) = const$$

В случае с логистической регрессией при одном наблюдении (x, y)

$$H(P, Q) = -y \log h_\theta - (1 - y) \log(1 - h_\theta)$$

Таким образом, для оценки параметров логистической регрессии необходимо решить следующую задачу:

$$\widehat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta}}{\operatorname{argmin}} L_{CE}(\boldsymbol{\theta}) = \underset{\boldsymbol{\theta}}{\operatorname{argmin}} \frac{1}{n} \sum_{i=1}^{n} l_{CE,i}(\boldsymbol{\theta})$$

Почему не использовать MSE в качестве функции потерь:

$$L_{MSE}(\boldsymbol{\theta}) = \frac{1}{n} \sum_{i=1}^{n} (y_i - h_{\theta,i})^2 \, ?$$

Данная функция не является выпуклой (без доказательства)

Замечание

В контексте классификации расстояние Кульбака-Лейблера (Kullback–Leibler divergence) определяет ассиметричную меру того, как одно случайное распределение P (наблюдаемые данные) отличается от другого Q (предсказания)

$$D_{KL}(P \parallel Q) = H(P, Q) - H(P) = -\sum_{i} p(x) \log q(x) + const,$$

где $H(P, Q)$ – кросс-энтропия между двумя распределениями вероятностей случайных величин P и Q

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

и $H(P)$ – энтропия (мера неопределенности)

$$H(P) = -\sum_{i} p(x) \log p(x) = const$$

В случае с логистической регрессией при одном наблюдении (x, y)

$$H(P, Q) = -y \log h_\theta - (1 - y) \log(1 - h_\theta)$$

В итоге для задачи классификации получаем

$$\widehat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta}}{\operatorname{argmin}} D_{KL}(P \parallel Q) = \underset{\boldsymbol{\theta}}{\operatorname{argmin}} H(P, Q)$$

3.4. Обучение (оценка параметров)

Обучение логистической регрессии производится посредством итеративных методов с использованием частных производных различных порядков. В данном случае нет аналитического решения (закрытой формы) подобно методу наименьших квадратов в линейной регрессии.

Как было отмечено ранее для обучения нам необходимо решить следующую задачу оптимизации:

$$\widehat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta}}{\operatorname{argmin}} L_{CE}(\boldsymbol{\theta}) = \underset{\boldsymbol{\theta}}{\operatorname{argmin}} \frac{1}{n} \sum_{i=1}^{n} l_{CE,i}(\boldsymbol{\theta})$$

Рассмотрим два варианта: градиентный спуск и стохастический градиентный спуск.

3.4.1. Градиентный спуск

Выборка:

$$S = \{(x_1, y_1), \dots, (x_n, y_n)\}$$

Замечание: чтобы использовать более общую запись, принимаем $x_{i0} = 1$

Итерация:

$$\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} - \alpha \nabla L_{CE}(\boldsymbol{\theta})$$

Функция потерь:

$$L_{CE}(\boldsymbol{\theta}) = \frac{1}{n} \sum_{i=1}^{n} l_{CE,i}(\boldsymbol{\theta}) = -\frac{1}{n} \sum_{i=1}^{n} y_i \log h_{\theta,i} + (1 - y_i) \log\left(1 - h_{\theta,i}\right)$$

Частная производная по θ_j:

$$\frac{\partial L_{CE}(\boldsymbol{\theta})}{\partial \theta_j} = \sum_{i=1}^{n} \left(h_{\theta,i} - y_i\right) x_{ij} = \sum_{i=1}^{n} \left(\frac{1}{1 + e^{-\theta^T x_i}} - y_i\right) x_{ij}$$

Вывод

1)

$$\frac{d}{dx} \log x = \frac{1}{x}$$

и

$$\frac{d}{dx} \sigma(x) = \frac{d}{dx} \frac{1}{1 + e^x} = \sigma(x)\left(1 - \sigma(x)\right)$$

2)

$$\begin{aligned}
\frac{\partial l_{CE}(\boldsymbol{\theta})}{\partial \theta_j} &= -\frac{\partial}{\partial \theta_j} \left(y \log h_\theta + (1 - y) \log(1 - h_\theta)\right) = -\frac{y}{h_\theta} \frac{\partial}{\partial \theta_j} h_\theta - \frac{1 - y}{1 - h_\theta} \frac{\partial}{\partial \theta_j} (1 - h_\theta) \\
&= -\left(\frac{y}{h_\theta} - \frac{1 - y}{1 - h_\theta}\right) \frac{\partial}{\partial \theta_j} h_\theta = -\left(\frac{y - h_\theta}{h_\theta (1 - h_\theta)}\right) \frac{\partial}{\partial \theta_j} h_\theta \\
&= -\left(\frac{y - h_\theta}{h_\theta (1 - h_\theta)}\right) h_\theta (1 - h_\theta) \frac{\partial}{\partial \theta_j} \theta^T x = -\left(\frac{y - h_\theta}{h_\theta (1 - h_\theta)}\right) h_\theta (1 - h_\theta) x_j \\
&= -(y - h_\theta) x_j = (h_\theta - y) x_j
\end{aligned}$$

3)

$$\frac{\partial L_{CE}(\boldsymbol{\theta})}{\partial \theta_j} = \sum_{i=1}^{n} \left(h_{\theta,i} - y_i\right) x_{ij}$$

Градиент $L_{CE}(\theta)$:

$$\nabla L_{CE}(\boldsymbol{\theta}) = \begin{bmatrix} \frac{\partial L_{CE}}{\partial \theta_0} \\ \vdots \\ \frac{\partial L_{CE}}{\partial \theta_p} \end{bmatrix}$$

3.4.2. Стохастический градиентный спуск

Одно наблюдение:

$$(x, y),$$

где

$$x = \left(x_1, x_2, \dots, x_p\right)^T$$

Итерация:

$$\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} - \alpha \nabla l_{CE,i}(\boldsymbol{\theta})$$

Функция потерь:

$$l_{CE}(\boldsymbol{\theta}) = -y \log h_\theta - (1 - y) \log(1 - h_\theta)$$

Частная производная по θ_j:

$$\frac{\partial L_{CE}(\boldsymbol{\theta})}{\partial \theta_j} = (h_\theta - y) x_j = \left(\frac{1}{1 + e^{-\boldsymbol{\theta}^T \boldsymbol{x}}} - y\right) x_j$$

Градиент $l_{CE}(\theta)$:

$$\nabla l_{CE}(\boldsymbol{\theta}) = \begin{bmatrix} \frac{\partial l_{CE}}{\partial \theta_0} \\ \vdots \\ \frac{\partial l_{CE}}{\partial \theta_p} \end{bmatrix}$$

Алгоритм. Стохастический градиентный спуск

1	load(S)	Загрузка обучающего множества
2	initialize($\boldsymbol{\theta}$)	Начальное значение
3	while stopCriteria is False:	Количество эпох, проверка сходимости
4	shuffle(S)	Тасовка
5	for $(\boldsymbol{x}, y)$ in S:	Цикл по всем элементам S
6	$\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} - \alpha \nabla l_{CE}(\boldsymbol{\theta}, \boldsymbol{x}, y)$	Обновление значения
7	$\widehat{\boldsymbol{\theta}} \leftarrow \boldsymbol{\theta}$	
8	return $\widehat{\boldsymbol{\theta}}$	

3.5. Предсказание

После того, как были получены оценки параметров $\widehat{\boldsymbol{\theta}}$ на обучающем множестве, модель готова предсказывать значения классов для новых данных. Для оценки качества обученной модели, как правило, используется тестовое множество. Для нового наблюдения x_* предсказание вычисляется следующим образом:

$$\hat{y} = \begin{cases} 1, \text{если} \dfrac{1}{1 + e^{-\hat{\boldsymbol{\theta}}^T \boldsymbol{x}_*}} > 0.5 \\ 0 \text{ иначе} \end{cases}$$

Особенность логистической регрессии является то, что данных метод дает вероятностное представление о принадлежности к классам:

$$\hat{p}\left(y = 1 \middle| \boldsymbol{x}_*, \hat{\boldsymbol{\theta}}\right) = \frac{1}{1 + e^{-\hat{\boldsymbol{\theta}}^T \boldsymbol{x}_*}}$$

В некоторых случаях используют вероятности вместо предсказания класса $\hat{y}$.

3.6. Взвешенная логистическая регрессия

Кросс-энтропия для взвешенной регрессии будет иметь вид

$$l_{CE}^{w} = -w_1 \, y \log h_\theta - w_0 (1 - y) \log(1 - h_\theta),$$

где w_0 и w_1 – веса классов 0 и 1, соответственно.

Далее всё аналогично тому, как было рассмотрено ранее.

Если веса классов неизвестны, то в общем виде их можно определить следующим образом:

$$w_i = \frac{n}{K \cdot n_i},$$

где n – количество наблюдений; K – количество классов; n_i – количество наблюдений i-го класса.

Взвешенная логистическая регрессия может быть полезной, например, при использовании несбалансированной по классам выборке. В этом случае наблюдений одного класса значительно больше, чем другого, и мы хотим придать больше веса классу с меньшим количеством наблюдений.

4. Многоклассовая классификация

4.1. Полиномиальная логистическая регрессия (multinomial logistic regression/softmax regression)

В многоклассовой классификации количество классов больше двух. Пусть k обозначает класс c_k из множества C. Всего классов K, то есть $|C| = K$ и $1 \leq k \leq K$.

Вероятность принадлежности классу k при заданном входном векторе x:

$$p(y = k|\boldsymbol{x}) = \text{softmax}(z_k) = \frac{e^{z_k}}{\sum_{j=1}^{K} e^{z_j}},$$

где

$$z_k = {\boldsymbol{\theta}_k}^T \boldsymbol{x}.$$

Для всех классов:

$$\text{softmax}(\boldsymbol{z}) = \left(\frac{e^{z_1}}{\sum_{j=1}^{K} e^{z_j}} \quad \frac{e^{z_2}}{\sum_{j=1}^{K} e^{z_j}} \quad \cdots \quad \frac{e^{z_K}}{\sum_{j=1}^{K} e^{z_j}} \right)^T,$$

где

$$\boldsymbol{z} = (z_1 \quad z_2 \quad \cdots \quad z_K)^T.$$

При этом должны выполняться условия:

$$\sum_{k=1}^{K} \text{softmax}(z_k) = 1$$

и

$$0 \leq \text{softmax}(z_i) \leq 1.$$

Обучение

Общая задача:

$$\widehat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta}}{\text{argmin}}\, L_{CE}(\boldsymbol{\theta}) = \underset{\boldsymbol{\theta}}{\text{argmin}} \frac{1}{n} \sum_{i=1}^{n} l_{CE,i}(\boldsymbol{\theta})$$

Кросс-энтропия как функция потерь в многоклассовой логистической регрессии для одного наблюдения:

$$l_{CE}(\boldsymbol{\theta}) = -\sum_{k=1}^{K} 1\{y = k\} \log p(y = k|\boldsymbol{x}) = -\sum_{k=1}^{K} 1\{y = k\} \log \frac{e^{\boldsymbol{\theta}_k{}^T \boldsymbol{x}}}{\sum_{j=1}^{K} e^{\boldsymbol{\theta}_j{}^T \boldsymbol{x}}},$$

где

$$1\{y = k\} = \begin{cases} 1, \text{если } y = k \\ 0 \text{ иначе} \end{cases}$$

Частная производная по параметрам k-ой модели $\boldsymbol{\theta}_k$:

$$\nabla_{\theta_k} l_{CE}(\boldsymbol{\theta}) = \frac{\partial l_{CE}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}_k} = \begin{bmatrix} \frac{\partial l_{CE}(\boldsymbol{\theta})}{\partial \theta_{k,0}} \\ \vdots \\ \frac{\partial l_{CE}(\boldsymbol{\theta})}{\partial \theta_{k,p}} \end{bmatrix} = -[1\{y = k\} - p(y = k|x)]x_k = -\left[1\{y = k\} - \frac{e^{\boldsymbol{\theta}_k{}^T \boldsymbol{x}}}{\sum_{j=1}^{K} e^{\boldsymbol{\theta}_j{}^T \boldsymbol{x}}} \right] x_k,$$

где

$$\boldsymbol{\theta}_k = \begin{bmatrix} \theta_{k,0} \\ \vdots \\ \theta_{k,p} \end{bmatrix}.$$

Градиент $\nabla l_{CE}(\boldsymbol{\theta})$:

$$\nabla l_{CE}(\boldsymbol{\theta}) = \begin{bmatrix} \frac{\partial l_{CE}}{\partial \theta_{1,0}} & \cdots & \frac{\partial l_{CE}}{\partial \theta_{K,0}} \\ \vdots & \ddots & \vdots \\ \frac{\partial l_{CE}}{\partial \theta_{1,p}} & \cdots & \frac{\partial l_{CE}}{\partial \theta_{K,p}} \end{bmatrix}.$$

Предсказание:

$$\hat{y} = \underset{1 \le k \le K}{\operatorname{argmax}}\, p(y = k|\boldsymbol{x}_*) = \underset{1 \le k \le K}{\operatorname{argmax}}\, \operatorname{softmax}\left(\widehat{\boldsymbol{\theta}}_k{}^T \boldsymbol{x}\right) = \underset{1 \le k \le K}{\operatorname{argmax}}\, \frac{e^{\widehat{\boldsymbol{\theta}}_k{}^T \boldsymbol{x}_*}}{\sum_{j=1}^{K} e^{\widehat{\boldsymbol{\theta}}_j{}^T \boldsymbol{x}_*}}$$

4.2. One-vs-rest

Для каждого класса строится бинарный классификатор h_k, где $1 \le k \le K$. Целевые значения обучающего набора данных для k-го классификатора представляется как набор, состоящий из двух классов: 1 – элементы k-го класса, 0 – все остальные. В качестве выходного значения каждый бинарный классификатор должен выдавать вероятностную оценку принадлежности тому или иному классу.

В итоге предсказание для некоторого наблюдения $\boldsymbol{x}$ будет иметь вид:

$$\hat{y} = \underset{1 \le k \le K}{\operatorname{argmax}}\, h_k(\boldsymbol{x})$$

Проблемы:

- Несбалансированная выборка при бинарной классификации

4.3. One-vs-one

Для каждой пары классов строится отдельный бинарный классификатор f_{ij}, где $1 \le i, j \le K$ и $i \ne j$. Всего получается

$$\frac{K(K-1)}{2}$$

классификаторов. Для каждого из них используется только те элементы обучающего множества, которые соответствуют данному бинарному классификатору. При этом выходным значением должно быть предсказание конкретного класса, а не вероятности.

В качестве итогового предсказания будет класс, предсказанный большинством бинарных классификаторов.

Проблемы:

- Может иметь одинаковое количество голосов для нескольких классов

Список литературы

- Christopher M. Bishop. 2006. Pattern Recognition and Machine Learning (Information Science and Statistics). Springer-Verlag, Berlin, Heidelberg. URL: https://www.microsoft.com/en-us/research/uploads/prod/2006/01/Bishop-Pattern-Recognition-and-Machine-Learning-2006.pdf
- Chapter 5. Logistic Regression // Speech and Language Processing. Daniel Jurafsky & James H. Martin URL: https://web.stanford.edu/~jurafsky/slp3/
- Chapter 4. Classification // An Introduction to Statistical Learning by Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshir. URL: http://faculty.marshall.usc.edu/gareth-james/ISL/