

Логистическая регрессия для решения задач классификации

Вихляев Егор, ММТ-2

October 30, 2023

Abstract

1 Описание проблемы, для решения которой используется данная модель

Логистическая регрессия - это статистический метод, который обычно используется для решения задач классификации, а не задач регрессии, несмотря на свое название.

Вот несколько примеров проблем, для решения которых часто используется логистическая регрессия:

Бинарная классификация, где цель - разделить объекты на два класса. Например, определение, является ли письмо спамом или не спамом, определение, болен ли пациент определенной болезнью (да/нет), и многие другие подобные задачи.

Многоклассовая классификация, где объекты должны быть отнесены к одному из нескольких классов. Это делается с использованием множества бинарных классификаторов или методов, таких как "один против всех" или "многие против многих".

Прогнозирование вероятности события. Например, вероятность клиента купить продукт, вероятность ухода клиента из компании, вероятность прохождения студентом экзамена и т. д.

Медицинская диагностика, где определяется наличие или отсутствие болезни на основе медицинских показателей и тестов.

Финансовый анализ. В финансовой сфере логистическая регрессия может использоваться для определения вероятности дефолта заемщика, анализа кредитоспособности, антифрод-анализа и других задач.

Логистическая регрессия может применяться для анализа данных о клиентах, определения, склонны ли они к покупке определенных товаров или услуг, или к каким-либо другим действиям.

Логистическая регрессия хорошо подходит для задач, где требуется оценивать вероятность принадлежности объекта к классу, и она широко используется во многих областях, где необходима классификация и прогнозирование с учетом вероятностей.

2 Концептуальная постановка

На практике логистическая регрессия используется для решения задач классификации с линейно-разделяемыми классами. При этом, предполагается, что зависимая переменная принимает два значения и имеет биномиальное распределение.

3 Математическая постановка (постановка задачи восстановления логистической регрессии)

Пусть дана выборка (X, Y) , где $X = \begin{bmatrix} X_1 \\ \dots \\ X_m \end{bmatrix}$ – набор объектов, $Y = \begin{bmatrix} Y_1 \\ \dots \\ Y_m \end{bmatrix}$ – ответы на этих объектах. $x_i \in \mathbb{R}^n$. $y_i \in 0; 1$ – один из двух классов: либо 0, либо 1.

Нужно построить функцию $f : x_i \rightsquigarrow \tilde{y}_i \approx y_i$. Логистическая регрессия задает структуру данной функции.

$\tilde{y}_i = f(x_i; w, b) = \sigma(w^T x_i + b)$ – сигмоида от суммы скалярного произведения и смещения. $w \in \mathbb{R}^n$ – вещ. вектор, $b \in \mathbb{R}$ – вещ. число (смещение).

Логистическая регрессия для объекта x_i получает предсказание в два шага:

1. $z = (-w^T -) \begin{bmatrix} | \\ x_i \\ | \end{bmatrix} + b \in \mathbb{R}$ - получаем из вектора число.
2. Полученное число преобразуем в меру активации: $a = \sigma(z) = \frac{1}{1+e^{-z}}$ - логистическая функция или сигмоида.

Таким образом мы задали модель и способ вычисления предсказания. Поскольку в способе есть два неизвестных параметра w и b , то подберем их через функцию потерь:

$$L_i(\tilde{y}_i) = -y_i \log(\tilde{y}_i) - (1 - y_i) \log(1 - \tilde{y}_i).$$

На выборке $(X, Y) : L = \frac{1}{m} \sum_{i=1}^m L_i(\tilde{y}_i) \rightarrow \min(w, b) = \frac{1}{m} \sum_{i=1}^m L_i(\sigma(w^T x_i + b)) \rightarrow \min(w, b)$. Задача свелась к минимизации функции потерь. Минимизируем градиентным спуском.

$w_{j+1} := w_j - \alpha \frac{\partial L(w_j, b_j)}{\partial w}$, где w_j - параметр на текущем шаге, α - шаг градиентного спуска, $\frac{\partial L(w_j, b_j)}{\partial w}$ - градиент. Аналогично для параметра b .

Поскольку $L_i(w, b) = L_i(a(z(w, b)))$ - сложная функция, то ищем ее частную производную следующим образом:

$$\frac{\partial L(w_j, b_j)}{\partial w} = \frac{\partial L_i}{\partial a} \frac{\partial a}{\partial z} \frac{\partial z}{\partial w}.$$

Тогда:

$$\frac{\partial L_i}{\partial a} = -\frac{y_i}{a} + \frac{1 - y_i}{1 - a} = \frac{a - y_i}{a(1 - a)},$$

$$\frac{\partial a}{\partial z} = a(1 - a),$$

$$\frac{\partial z}{\partial w} = \frac{\partial (w^T x_i + b)}{\partial w} = \begin{bmatrix} | \\ x_i \\ | \end{bmatrix}.$$

Таким образом:

$$\frac{\partial L_i}{\partial w} = \frac{a - y_i}{a(1 - a)} * a(1 - a) * x_i = (a - y_i x_i),$$

при этом, аналогично,

$$\frac{\partial L_i}{\partial b} = a - y_i.$$