Логистическая регрессия

это статистическая модель, используемая для прогнозирования вероятности возникновения некоторого события путём подгонки данных к логистической кривой.

Логистическое уравнение, изначально появилось при изучении изменений численности населения.

- скорость размножения популяции пропорциональна её текущей численности, при прочих равных условиях
- скорость размножения популяции пропорциональна количеству доступных ресурсов, при прочих равных условиях. Таким образом, второй член уравнения отражает конкуренцию за ресурсы, которая ограничивает рост популяции.

ЛР применяется для прогнозирования вероятности возникновения некоторого события по значениям множества признаков.
Для этого вводится так называемая *зависимая переменная* , принимающая лишь одно из двух значений — как правило, это числа 0 и 1, и множество *признаков* — вещественных , на основе значений которых требуется вычислить вероятность принятия того или иного значения зависимой переменной. Как и в случае линейной регрессии, для простоты записи вводится фиктивный признак x_0=1

Делается предположение о том, что вероятность наступления события $y = 1$ равна:

$$\mathbb{P}\{y = 1 \mid x\} = f(z),$$

где $z = \theta^T x = \theta_0 + \theta_1 x_1 + \ldots + \theta_n x_n$, x и θ — векторы-столбцы значений независимых переменных $1, x_1, \ldots, x_n$ и параметров (коэффициентов регрессии) — вещественных чисел $\theta_0, \ldots, \theta_n$, соответственно, а $f(z)$ — так называемая *логистическая функция* (иногда также называемая сигмоидом или логит-функцией):

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

Так как y принимает лишь значения 0 и 1, то вероятность принять значение 0 равна:

$$\mathbb{P}\{y = 0 \mid x\} = 1 - f(z) = 1 - f(\theta^T x).$$

Для краткости функцию распределения y при заданном x можно записать в таком виде:

$$\mathbb{P}\{y \mid x\} = f(\theta^T x)^y (1 - f(\theta^T x))^{1-y}, \quad y \in \{0, 1\}.$$

Фактически, это есть распределение Бернулли с параметром, равным $f(\theta^T x)$.

Эта модель часто применяется для решения задач классификации — объект можно отнести к классу , если предсказанная моделью вероятность , и к классу в противном случае. Получающиеся при этом правила классификации являются линейными классификаторами.

Softmax — это обобщение логистической функции для многомерного случая. Функция преобразует вектор размерности в вектор той же размерности, где каждая координата полученного вектора представлена вещественным числом в интервале [0,1] и сумма координат равна 1.

На логистическую регрессию очень похожа пробит-регрессия, отличающаяся от неё лишь другим выбором функции . Softmax-регрессия обобщает логистическую регрессию на случай многоклассовой классификации, то есть когда зависимая переменная принимает более двух значений. Все эти модели в свою очередь являются представителями широкого класса статистических моделей — обобщённых линейных моделей.