

Линейная и логистическая регрессия

Ниже рассматриваются две простейшие статистические модели: регрессионная модель — линейная регрессия и модель классификации — логистическая регрессия.

Регрессия — математическое выражение, отражающее зависимость эндогенной переменной y от независимых переменных x .

Классификация — систематизированное распределение объектов на определенные классы на основании сходства и различия в признаках, описывающих объекты.

Линейная регрессия

Пусть $D = \{(x_i, y_i) : X_i \in \mathbf{R}^n, y^i \in \mathbf{R}\}$ некоторый набор данных. Необходимо восстановить отношение между зависимой переменной y^i и независимой переменной $x_i^T = [x_1^i, \dots, x_n^i]$.

В модели линейной регрессии отношение между X и y предполагается линейным, тогда *гипотеза*, отражающая линейную зависимость X и y может быть записана следующим образом:

$$\hat{y} = h_\theta(x) = \theta_0 + \sum_{j=1}^n x_j \theta_j,$$

где $\theta = [\theta_0, \theta_1, \dots, \theta_n]^T$ — неизвестные параметры (или веса) модели. Параметр θ_0 отражает предположение о том, что данные изначально не центрированы, а сдвинуты на какую-то константу.

Если к признакам всех объектов добавить фиктивный признак равный единице, гипотезу можно переписать компактнее:

$$h_\theta(x) = \sum_{j=0}^n x_j \theta_j$$

или в векторном виде:

$$h_\theta(X) = X\theta,$$

где

$$X = \begin{bmatrix} 1 & x_1^1 & x_2^1 & \dots & x_n^1 \\ 1 & x_1^2 & x_2^2 & \dots & x_n^2 \\ \dots & \dots & \dots & \dots & \dots \\ 1 & x_1^m & x_2^m & x_3^m & x_n^m \end{bmatrix}, \quad \theta = \begin{bmatrix} \theta_0 \\ \theta_1 \\ \dots \\ \theta_n \end{bmatrix}$$

Теперь, чтобы регрессионная модель обучалась, необходимо определить функцию, измеряющую, насколько хорошо параметры θ восстанавливают имеющуюся зависимость. Для этого введем среднеквадратичную функцию ошибки:

$$J(\theta) = \frac{1}{2m} \sum_{i=1}^m (\hat{y}^i - y^i)^2 = \frac{1}{2m} \sum_{i=1}^m (h_\theta(x^i) - y^i)^2$$

Приняв $\vec{y} = [y^1, y^2, \dots, y^m]^T$, запишем функцию потерь в матричной форме:

$$J(\theta) = \frac{1}{2m}(X\theta - \vec{y})^T(X\theta - \vec{y})$$

Функцию ошибок, очевидно, необходимо минимизировать. Исходя из этого получаем следующую задачу оптимизации:

$$\frac{1}{2m} \sum_{i=1}^m (h_{\theta}(x^i) - y^i)^2 \rightarrow \min_{\theta}$$

или

$$\frac{1}{2m} \|X\theta - \vec{y}\|^2 \rightarrow \min_{\theta}$$

Для решения оптимизационной задачи воспользуемся методом градиентного спуска. Будем итеративно обновлять веса модели, сдвигаясь в направлении антиградиента функции потерь с некоторым шагом:

$$\theta_j := \theta_j - \alpha \frac{\partial}{\partial \theta_j} J(\theta) = \theta_j - \alpha \frac{1}{m} \sum_{i=1}^m (h_{\theta}(x^i) - y^i) x_j^i, \quad j \in \{0, 1, \dots, n\}$$

В матричной форме

$$\theta := -\alpha \nabla J(\theta) = \theta - \frac{\alpha}{m} X^T (X\theta - \vec{y})$$

Для предотвращения проблемы переобучения используется регуляризация. Введем "штраф" на большие веса модели. Теперь функция потерь будет иметь вид:

$$J(\theta) = \frac{1}{2m} \sum_{i=1}^m (\hat{y}^i - y^i)^2 = \frac{1}{2m} \sum_{i=1}^m (h_{\theta}(x^i) - y^i)^2 + \lambda \sum_{j=1}^n \theta_j^2$$

Необходимо заметить, что параметр θ_0 в последней сумме не участвует, то есть не штрафует. Этот факт лишает возможности компактной записи функции потерь в матричном виде с сохранением введенных обозначений.

Выражения для обновления весов методом градиентного спуска с введенным регуляризатором примет вид:

$$\begin{aligned} \theta_0 &= \theta_0 - \alpha \frac{1}{m} \sum_{i=1}^m (h_{\theta}(x^i) - y^i) x_0^i \\ \theta_j &:= \theta_j - \alpha \left[\left(\frac{1}{m} \sum_{i=1}^m (h_{\theta}(x^i) - y^i) x_j^i \right) + \frac{\lambda}{m} \theta_j \right], \quad j \in \{1, 2, \dots, n\} \end{aligned}$$

Логистическая регрессия

Рассмотрим задачу классификации. Данные теперь имеют вид $D = \{(x_i, y_i) : x_i \in \mathbf{R}^n, y_i \in \{0, 1\}\}$. Будем предсказывать вероятность принадлежности объекта x^i к классу y^i . То есть $h_\theta(x) = P(y = 1|x, \theta) = 1 - P(y = 0|x, \theta)$ Это, очевидно, накладывает условия на гипотезу, а именно:

$$0 \leq h_\theta(x) \leq 1$$

Воспользуемся для удовлетворения данного условия сигмодой:

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

$$z = \theta^T x$$

$$h_\theta(x) = g(\theta^T x)$$

После этого необходимо установить решающую границу. Сделаем это следующим образом:

$$h_\theta(x) \geq 0.5 \rightarrow y = 1$$

$$h_\theta < 0.5 \rightarrow y = 0$$

Зададим для нашей гипотезы функцию потерь так, чтобы она имела единственный минимум:

$$J(\theta) = \frac{1}{m} \sum_{i=1}^m \text{Cost}(h_\theta(x^i), y^i)$$

$$\begin{cases} \text{Cost}(h_\theta(x), y) = -\log(h_\theta(x)), & y = 1 \\ \text{Cost}(h_\theta(x), y) = -\log(1 - h_\theta(x)), & y = 0 \end{cases}$$

Можем упростить запись функции $\text{Cost}(\cdot)$:

$$\text{Cost}(h_\theta(x), y) = -y \log(h_\theta(x)) - (1 - y) \log(1 - h_\theta(x))$$

Окончательно функция потерь имеет вид:

$$J(\theta) = -\frac{1}{m} \sum_{i=1}^m [y^i \log(h_\theta(x^i)) + (1 - y^i) \log(1 - h_\theta(x^i))]$$

В матричной форме:

$$J(\theta) = \frac{1}{m} (-y^T \log(g(X\theta)) - (1 - y)^T \log(1 - g(X\theta)))$$

Осталось лишь записать выражения для обновления весов модели:

$$\theta_j := \theta_j - \frac{\alpha}{m} \sum_{i=1}^m (h_\theta(x^i) - y^i) x_j^i, \quad j \in \{0, 1, \dots, n\}$$

Или в матричном виде:

$$\theta := \theta - \frac{\alpha}{m} X^T (g(X\theta) - \vec{y})$$

Аналогично тому, как мы вводили регуляризацию для линейной регрессии, сделаем это для логистической регрессии. Начнем с функции потерь:

$$J(\theta) = -\frac{1}{m} \sum_{i=1}^m [y^i \log(h_\theta(x^i)) + (1 - y^i) \log(1 - h_\theta(x^i))] + \frac{\lambda}{2m} \sum_{j=1}^n \theta_j^2$$

Выражения для градиентного спуска:

$$\theta_0 = \theta_0 - \frac{\alpha}{m} \sum_{i=1}^m (h_\theta(x^i) - y^i) x_0^i$$

$$\theta_j = \theta_j - \alpha \left[\left(\frac{1}{m} \sum_{i=1}^m (h_\theta(x^i) - y^i) x_j^i \right) + \frac{\lambda}{m} \theta_j \right], \quad j \in \{1, 2, \dots, n\}$$

Выше была описана $l2$ (*ridge*) регуляризация, если же необходимо "обнулить" некоторые веса в модели, можно воспользоваться $l1$ (*lasso*) регуляризацией:

$$J(\theta) = -\frac{1}{m} \sum_{i=1}^m [y^i \log(h_\theta(x^i)) + (1 - y^i) \log(1 - h_\theta(x^i))] + \frac{\lambda}{m} \sum_{j=1}^n |\theta_j|$$

Веса в этом случае будут обновляться по следующему правилу:

$$\theta_0 = \theta_0 - \frac{\alpha}{m} \sum_{i=1}^m (h_\theta(x^i) - y^i) x_0^i$$

$$\theta_j = \theta_j - \alpha \left[\left(\frac{1}{m} \sum_{i=1}^m (h_\theta(x^i) - y^i) x_j^i \right) + \frac{\lambda}{m} \text{sign}(\theta_j) \right], \quad j \in \{1, 2, \dots, n\}$$