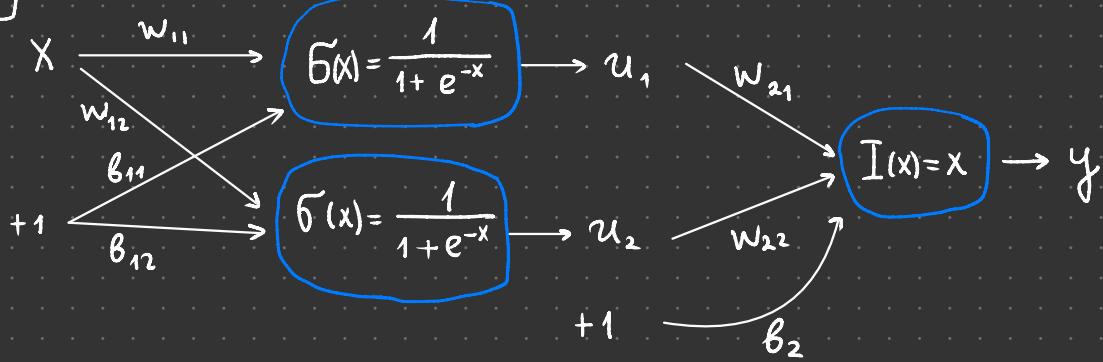


1



2 Решение задачи:

$$MSE = L(X, y) = \frac{1}{n} \sum_{i=1}^n (\hat{y}_i - y_i)^2$$

- $\frac{\partial L}{\partial \hat{y}_i} = \frac{1}{n} \cdot 2(\hat{y}_i - y_i)$

- $\frac{\partial \hat{y}_i}{\partial w_{2h}} = \frac{\partial}{\partial w_{2h}} \left(\sum_{h=1}^2 w_{2h} u_h(x_i) + b_2 \right) = u_h(x_i)$

- $\frac{\partial \hat{y}_i}{\partial b_2} = \frac{\partial}{\partial b_2} \left(\sum_{h=1}^2 w_{2h} u_h(x_i) + b_2 \right) = 1$

- $\frac{\partial \hat{y}_i}{\partial u_h(x_i)} = \frac{\partial}{\partial u_h} \left(\sum_{h=1}^2 w_{2h} u_h(x_i) + b_2 \right) = w_{2h}$

- $\frac{\partial MSE}{\partial w_{2h}} = \sum_{i=1}^n \frac{\partial MSE}{\partial \hat{y}_i} \cdot \frac{\partial \hat{y}_i}{\partial w_{2h}}$

- $\frac{\partial MSE}{\partial b_2} = \sum_{i=1}^n \frac{\partial MSE}{\partial \hat{y}_i} \cdot \frac{\partial \hat{y}_i}{\partial b_2}$

- $\frac{\partial MSE}{\partial u_h} = \sum_{i=1}^n \frac{\partial MSE}{\partial \hat{y}_i} \cdot \frac{\partial \hat{y}_i}{\partial u_h}$

$$\bullet \frac{\partial \delta(x)}{\partial x} = \frac{+e^{-x}}{(1+e^{-x})^2} \stackrel{\pm 1}{=} \frac{1}{1+e^{-x}} - \frac{1}{(1+e^{-x})^2} = \sigma'(1-\sigma)$$

$$\bullet \frac{\partial u_h(x_i)}{\partial w_{ih}} = \frac{\partial}{\partial w_{ih}} \left(\delta(w_{ih}x_i + b_{ih}) \right) = x_i \cdot \sigma'(w_{ih}x_i + b_{ih})$$

$$\bullet \frac{\partial u_h(x_i)}{\partial b_{ih}} = \frac{\partial}{\partial b_{ih}} \left(\delta(w_{ih}x_i + b_{ih}) \right) = \sigma'(w_{ih}x_i + b_{ih})$$

$$\bullet \frac{\partial \text{MSE}}{\partial w_{ih}} = \sum_{i=1}^n \frac{\partial \text{MSE}}{\partial \hat{y}_i} \cdot \frac{\partial \hat{y}_i}{\partial u_h(x_i)} \cdot \frac{\partial u_h(x_i)}{\partial w_{ih}}$$

$$\bullet \frac{\partial \text{MSE}}{\partial b_{ih}} = \sum_{i=1}^n \frac{\partial \text{MSE}}{\partial \hat{y}_i} \cdot \frac{\partial \hat{y}_i}{\partial u_h(x_i)} \cdot \frac{\partial u_h(x_i)}{\partial b_{ih}}$$

3 Если выборка очень большая, то обычный градиентный спуск может быть слишком медленным и будет требовать много ресурсов.

Рассмотрим другие методы обновления параметров:

■ Стохастический градиентный спуск

⊕ быстрее сходится

⊖ менее стабильный

■ Mini-batch градиентный спуск

⊕ содержит в себе преимущества SGD и стабильность полного градиентного спуска



Вывод: 1) Наша нейронная сеть имеет 2

слоя:

первый — принимает $x \in \mathbb{R}$ и возвращает w_1, w_2
с функцией активации $\sigma(x) = \frac{1}{1+e^{-x}}$

второй — возвращает $y \in \mathbb{R}$ с тождественной функцией акт.

Таким образом, всего общихших параметров — 7

2) Мы посчитали градиенты, необходимые для
градиентного спуска

$$\frac{\partial \text{MSE}}{\partial w_{1h}} \quad \text{и} \quad \frac{\partial \text{MSE}}{\partial b_{1h}}$$

3) При большой выборке стоит обращать внимание
на другие способы обновления параметров