

# Машинное обучение, ФКН ВШЭ

## Домашнее задание №7

**Задача 1.** На лекции и семинаре был рассмотрен метод обратного распространения ошибки в общем случае. Рассмотрим частный случай с полносвязным слоем, у которого  $d_{in}$  входных и  $d_{out}$  выходных нейронов. В качестве функции активации используем сигмоидальную:  $g(t) = \frac{1}{1+\exp(-t)}$ .

В полносвязном слое  $i$ -й выходной нейрон для слоя  $l$  можно выразить следующим образом:

$$x_i^l = g\left(\sum_{j=1}^{d_{in}} w_{ij}^l x_j^{l-1} + b_i^l\right)$$

Требуется вычислить производную функции потерь  $L(z, y)$  по весу  $w_{ij}^l$  полносвязного слоя  $l$ :

$$\frac{\partial L}{\partial w_{ij}^l}$$

Выражение может включать в себя величины, посчитанные во время прямого прохода по нейронной сети, и величины, полученные со следующих (по порядку прямого прохода) слоёв во время обратного прохода. Функция потерь дифференцируема по выходам сети  $z$ .

**Задача 2.** Рассмотрим вместо полносвязного слоя из задачи 1 свёрточный слой. Пусть на вход поступает изображение размера  $H \times W$ , свёрточный слой имеет размер  $k_1 \times k_2$ . Тогда применение свёрточного слоя можно выразить следующим образом:

$$x_{ij}^l = g\left(\sum_{m=0}^{k_1-1} \sum_{n=0}^{k_2-1} w_{mn}^l x_{i+m, j+n}^{l-1}\right)$$

Требуется вычислить производную функции потерь  $L(z, y)$  по весу  $w_{mn}^l$  свёрточного слоя  $l$ :

$$\frac{\partial L}{\partial w_{mn}^l}$$

Выражение может включать в себя величины, посчитанные во время прямого прохода по нейронной сети, и величины, полученные со следующих (по порядку прямого прохода) слоёв во время обратного прохода. Функция потерь дифференцируема по выходам сети  $z$ .

**Задача 3.** С ростом количества слоёв нейронные сети могут выделять всё более сложные структуры в исходном пространстве признаков. Однако обучение глубоких

нейронных сетей с помощью градиентных методов оптимизации вызывает некоторые сложности. Одной из таких проблем является проблема затухающих градиентов, когда градиенты для весов первых слоёв оказываются близкими к нулю, из-за чего первые слои обучаются медленнее последних. Подумайте и ответьте, почему так происходит. Предлагается рассматривать сигмоиду и гиперболический тангенс в качестве функции активации во всей сети.

**Задача 4.** Рассмотрим нелинейное преобразование ReLU:

$$g(x) = \begin{cases} x, & x \geq 0 \\ 0, & x < 0 \end{cases}$$

Несмотря на свою простоту, оно позволяет сети выучивать сложную структуру и при этом легко вычисляется. Предлагается убедиться в нелинейности, подобрав коэффициенты сети для классификации с 2 входами и 1 выходом так, чтобы она выдавала ответы такие же, как функция XOR (нули на наборах из двух нулей и двух единиц, единицу на наборах с одной единицей). Также покажите, что для такой зависимости нельзя построить линейный классификатор, не допускающий ни одной ошибки.

Сеть должна иметь один скрытый слой из двух нейронов с активацией ReLU и одним выходным нейроном без нелинейного преобразования. Класс выдаваемый сетью определяется с помощью некоторого порога (например, 0.5, что эквивалентно при использовании сигмоидальной функции выбору класса с максимальной вероятностью). Свободных членов в сети нет. Достаточно найти любое подходящее решение.