

Теоретическая часть

1. Посчитайте производную функции $\tanh(z)$ и выразите ее через саму функцию $\tanh(z)$, считая что z — скаляр. Преобразуйте ответ, так, чтобы при вычислении $\tanh(z)$ и ее производной была только одна операция экспоненцирования.

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

$$\begin{aligned} (\tanh(z))' &= \left(\frac{\sinh(x)}{\cosh(x)} \right)' = \frac{(\cosh(x) * \sinh'(x) - \sinh(x) * \cosh'(x))}{\cosh^2(x)} \\ &= \frac{(\cosh(x) * \cosh(x) - \sinh(x) * \sinh(x))}{\cosh^2(x)} = 1 - \tanh^2(x) \end{aligned}$$

2. Запишите оценочную функцию кросс-энтропия $L(W^1, \dots, W^L, x_{\{1\}}, \dots, x_{\{N\}})$ для нейронной сети с одним скрытым слоем ($L=2$), а затем в общем виде для нейронной сети с $L-1$ скрытыми слоями, для случая мультиклассовой классификации (считаем, что есть K классов). В качестве активации для скрытого слоя используется $\tanh(z)$, для выходного слоя — $\text{softmax}(z)$.

$$L = - \sum_d t_d * (z_d^L - \log \sum_c e^{z_c^L})$$

3. Скрытый слой — матрица $M \times N$ — $M * N$ параметров. Выходной слой — матрица $N \times K$ — $N * K$ параметров. Всего у такой двухслойной сети — $N * (K + M)$ параметров
4. Выведите формулу для δ^L — градиента оценочной функции по предактивациям в последнем слое z^L .

$$\begin{aligned} \delta_k^L &= \frac{\partial L}{\partial z_k^L} = - \sum_d t_d * (1_{d=k} - \frac{1}{(\sum_c e^{z_c^L})} * e^{z_k^L}) \\ &= - \sum_d t_d * (1_{d=k} - a_k^L) = \sum_d t_d * a_k^L - \sum_d t_d * 1_{d=k} = a_k^L \sum_d t_d - t_k \\ &= a_k^L - t_k, \text{ где } a^L = \frac{e^{z^L}}{\sum_c e^{z_c^L}}, t_k - \text{истинная метка } k - \text{го выхода} \end{aligned}$$

5. Выведите формулу для подсчета δ^l — градиента оценочной функции по z^l — через δ^{l+1}

$$\begin{aligned} \delta_j^l &= \frac{\partial L}{\partial a_j^l} * \frac{\partial a_j^l}{\partial z_j^l} \\ \frac{\partial L}{\partial a_j^l} &= \sum_k \frac{\partial L}{\partial z_k^{l+1}} * \frac{\partial z_k^{l+1}}{\partial a_j^l} = \sum_k \delta_k^{l+1} * W_{kj}^{l+1} \\ \frac{\partial a_j^l}{\partial z_j^l} &= f'(z_j^l) \\ \delta_j^l &= f'(z_j^l) * \sum_k \delta_k^{l+1} * W_{kj}^{l+1} \end{aligned}$$

где $a_j^l = f(z_j^l)$, f — функция активации

6. Выведите формулу для $\nabla_{W^l} L$ — градиента оценочной функции по весам W^l , используя δ^l .

$$\nabla_{w^l} L = \frac{\partial L}{\partial W_{ji}^l} = \frac{\partial L}{\partial a_j^l} * \frac{\partial a_j^l}{\partial z_j^l} * \frac{\partial z_j^l}{\partial W_{ji}^l} = \delta_j^l * \frac{\partial z_j^l}{\partial W_{ji}^l} = \delta_j^l * a_j^{l-1}$$