METL

TP 1

Tientso Ning

Softmax

1. Define softmax.

Softmax: $\begin{cases} \mathbb{R}^n \to [0,1]^n \\ x \mapsto \frac{e^x}{\sum_{j=1}^n e^{x_j}} \end{cases}$

With $x = (x_1, \ldots, x_n), n \in \mathbb{N}$ $\sum_{k=1}^n \text{softmax}(x) = 1$

$\text{softmax}(x)_i = \frac{e^{x_i}}{\sum_{i=1}^n x^j}$

2. Show softmax is invariant to constants.

Let $x \in \mathbb{R}^n, c \in \mathbb{R}^n, n \in \mathbb{N}$

n = dim(x), where c is a constant vector.

$\text{softmax}(x + c)i = \frac{e^{(x_i + c_i)}}{\sum_{i=1}^n e^{(x_i + c_j)}}$

$= \frac{e^{x_i} e^{c_i}}{\sum_{i=1}^n e^{x_j + c_j}}$

$= \frac{e^{c_i} e^{x_i}}{e^{c_j} \sum_{j=1}^n e^{x_j}}$

Since c is a constant.

$= \frac{e^{x_i}}{\sum_{j=1}^n e^{x_i}}$

$= \text{softmax}(x)_i, \forall i \in [1, n]$

$\forall x \in \mathbb{R}^n, \forall c \in \mathbb{R}^n, \text{softmax}(x + c) = \text{softmax}(x)$

Sigmoid

1. Define sigmoid and gradient.

sigmoid: $\begin{cases} \mathbb{R}^n \to [0,1]^n \\ x \mapsto \frac{1}{1+e^{-x}} \end{cases}$ with $x \in \mathbb{R}, n \in \mathbb{N}$ $\sigma(x) = \frac{1}{1+e^{-x_i}}$

$$\begin{aligned}\frac{d}{dx}\sigma(x) &= \frac{d}{dx}\left(\frac{1}{1+e^{-x}}\right)\\ &= \frac{d}{dx}\left(1+e^{-x}\right)^{-1}\\ &= \frac{d}{dx}\left(1+e^{-x}\right)^{-1}\\ &= -\left(1+e^{-x}\right)^{-2}\left(-e^{-x}\right)\\ &= \frac{e^{-x}}{\left(1+e^{-x}\right)^2}\\ &= \frac{1}{1+e^{-x}}\cdot\frac{e^{-x}}{1+e^{-x}}\end{aligned}$$

$= \frac{1}{1+e^{-x}} \cdot \frac{(1+e^{-x})-1}{1+e^{-x}} = \frac{1}{1+e^{-x}} \cdot \left(1 - \frac{1}{1+e^{-x}}\right) = \sigma(x) \cdot (1-\sigma(x))$

2. Derive gradient with respect to CE loss.

$CE = -\sum_i^n y_i \log(\hat{y}_i)$

$\frac{\partial CE}{\partial z} = -\sum_k y_k \log(\hat{y}_k)$

$= -\sum_k y_k \cdot \frac{\partial \log(\hat{y}_k)}{\partial \hat{y}_k} \times \frac{\partial \hat{y}_k}{\partial z}$

$= -\sum_k y_k \frac{1}{\hat{y}_k} \times \frac{\partial \hat{y}_k}{\partial z}$

Since we can have i = k and i != k,

$= -y_i(1-\hat{y}_i) - \sum_{k\neq i} y_k \frac{1}{\hat{y}_k}(-\hat{y}_k \cdot \hat{y}_i)$

$= -y_i(1-\hat{y}_i) + \sum_{k\neq 1} y_k \cdot \hat{y}_k$

$= -y_i + y_i\hat{y}_i + \sum_{k\neq 1} y_k \cdot \hat{y}_k$

$= \hat{y}_i\left(y_i + \sum_{k\neq 1} y_k\right) - y_i$

Due to the one-hot encoding, we will have the values inside the parenthesis equal to 1, making the final form:

$= \hat{y}_i - y_i$

3. Derive the gradients.

$z_1 = xw_1 + b_1 \;\; z_2 = hw_2 + b_2$

$\frac{\partial CE}{\partial z_2} = \hat{y} - y$, from the previous derivations.

$\frac{\partial CE}{\partial h} = (\hat{y}-y)\frac{\partial z_2}{\partial h} = (\hat{y}-y)w_2^\top$

where $\frac{\partial z_i}{\partial x_i} = w_i^\top$.

Note that the jacobian $\frac{\partial z_2}{\partial h} = \frac{\partial}{\partial h}\sum W_2 h = \sum W_2 \frac{\partial h}{\partial h} = W_2$ and that since we are taking the row order, we use W_2^T

$\frac{\partial CE}{\partial z_1} = (\hat{y} - y) w_2^T \frac{\partial h}{\partial z_1} = (\hat{y} - y) w_2^T \odot \sigma'(z_1)$ since the layer is the application of the sigmoid, in element-wise fashion.

$\frac{\partial CE}{\partial x} = (\hat{y} - y) w_2^\top \odot \sigma'(z_1) \frac{\partial z_1}{\partial x}$

$= (\hat{y} - y) w_2^\top \odot \sigma^2(z_1) w_1^\top$

4. The number of parameters is `(Dx + 1)H + (H+1)Dy` Since Dx is the dimension of x, and the layer `xW1 + b` requires the bias to be of the same shape H, and HDx is the dimension of xW1. The second part is from the requirement that it matches the output dimensions, since its a one layer NN. Dy corresponds to the output dimension. The bias has to be the same shape Dy, and the layer produces output dimension HDy.